Predicting switching behavior on health insurer, by acoustic features from call center speech

Sam van Kesteren
Tilburg University
ANR: 682695

Thesis submitted in fulfillment of the requirements for the degree of master of science in communication and information sciences, master track data science business and governance research serves as the thesis for the master Data Science Business & Governance, at the School of Humanities and Digital Sciences of Tilburg University

Author: Sam van Kesteren,

Supervisor Tilburg University: dr. M. Postma, External supervisor:
M. van Os,

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands January 2019

Abstract

Recent research developments in voice acoustics have motivated this thesis, i.e., the successful prediction of future behavior by humanly judged voice utterances. The approach in this thesis deviates from that of earlier studies, by applying machine learning models to a data set that is constructed from real-world speech recordings. The aim is to predict consumers' switching behavior regarding health insurance by analyzing the acoustics of the consumers' voices. A data set of 3,887 conversations is built from a comparison website's call center recordings. Prosodic and spectral feature are extracted from thin slices of the speech recordings and enriched with fiver non-acoustic features about the call and the customer to complete a data set of 114 features. Five experiments are executed with a different combination of features each to determine the predictive capacity of voice acoustics as indicators of future switching behavior. Results show that non-acoustic features by themselves manage to outperform a random guessing baseline. The addition of features from the extracted pitch and intensity improves the performance of the non-acoustic features, highlighting the potential of voice acoustics in the prediction of future behavior. Models trained on a data set that included extracted MFCC features or consisted of only acoustic features did not prove successful performance. Implications for future research are discussed to build upon these preliminary findings.

Table of Contents

1. Introduction

This thesis aims to predict consumers' switching behavior regarding health insurance by analyzing the acoustics of the consumers' voices. Since the introduction of free provider choice in the Netherlands in 2006, consumers struggle to switch health insurance effectively. Well-informed switching decisions are difficult in the current market's complexity as the search for information is too complicated and confusing (Lako, Rosenau, & Daw, 2011; Van Beest, Lako, & Sent, 2012). Assistance comes from comparison websites, increasing the chance of consumers switching as they are assisted in their choice of health policy (Han & Urmie, 2017). Their service can be improved and personalized by analyzing data from customer interactions and applying machine learning models to measure consumers' attitudes and behaviors (Rust & Huang, 2014).

Besides support via their websites, comparison websites provide counselling over the phone through their call centers. Data that flows from these conversations is huge and potentially rich in information. Therefore, insights in the customer-specific decision-making process might be derived from this data, since voice to voice communication offers more emotional cues than written communication (e.g. Goldberg, & Grandey, 2007; Rueff-Lopes et al., 2015). The analysis of recorded call center speech with modern big data techniques is a challenging task, but predicting switching behavior from this data can open up the next step in personalization and customer service for call centers.

1.1 The case at Independer

This thesis project is carried out in cooperation with Independer. Independer is a Dutch comparison website for insurance and other financial products. The website is built to be self-explanatory for customers. However, Independer's call center is available for support and it is possible to close a contract over the phone. Call center conversations, numbering 173,416, were recorded from September 23, 2016, up until January 3, 2017. These recordings are the data that is used for this thesis. In 2016, a total of 173,432 customers closed a contract for new health insurance via Independer, either online or via the phone. In the Netherlands, a total of 1.17 million people switch health insurance (Nederlandse Zorgautoriteit, 2016), giving Independer a market share of 14.8% of all customers who switch. In this thesis, Independer will be further referred to as 'the company'.

1.2 Problem statement

This thesis focusses on call center interactions between customers and agents from a comparison website. Speech data from recorded calls is analyzed, both acoustic and non-acoustic parameters are extracted to predict switching behavior of health insurer.

The problem statement is: *To what extent can acoustic features from call center speech successfully predict switching behavior of health insurer?*

To support the structure of this research, five experiments are executed that each contributes by addressing the formulated research objectives:

Research objective 1: *To identify the extent to which non-acoustic features from call center speech can successfully predict switching behavior of health insurer.*

Research objective 2: *To establish the effect of adding pitch and intensity from call center speech to models using non-acoustic features on the performance of predicting switching behavior.*

Research objective 3: *To establish the effect of adding MFCCs from call center speech to models using non-acoustic features, pitch and intensity on the performance of predicting switching behavior.*

Research objective 4: *To identify the extent to which pitch and intensity from call center speech can successfully predict switching behavior of health insurer.*

Research objective 5: *To establish the effect of adding MFCCs from call center speech to models using pitch and intensity on the performance of predicting switching behavior.*

1.3 Scientific relevance

In the past, predictive models based on acoustic features of speech have been used to classify emotions by using databases with acted speech fragments (e.g. El Ayadi, Kamel, & Karray, 2011). The approach in this thesis deviates from that of earlier studies, by using natural speech data instead of acted speech. Although acted speech differs in articulatory movements from natural speech (Erickson, Menezes, & Fujino, 2004), the demonstrated predictive ability opens the potential for using acoustic features from natural speech. Classifying emotions is not a straightforward task, but predicting long-term decisions from momently speech acoustics might be even more challenging. However, predicting future behavior from speech has been done before. Human judges managed to predict future voting behavior of citizens by analyzing nonverbal characteristics of telephone speech utterances (Rogers, ten Brinke, & Carnay, 2016). The likelihood of voting was successfully predicted by the level of confidence in the voice from people making self-predictions on their voting behavior. The use of machine learning techniques from previous studies in voice acoustics combined with the potential of predicting future behavior from voice shows the relation to multiple research fields and the opportunity of them combined.

1.4 Practical relevance

The insights from a model that relates acoustic speech features to future behavior have practical implications for customer service companies. Because this thesis makes use of automated analysis where no human judges are needed, the development of automated human-robot interactions, like chatbots, will benefit from the insights. Chatbots are emerging in many online service delivering businesses and will be an important subject of innovation (Pereira et al., 2016). Besides, predictions on future behavior allow robots to anticipate and therefore better assist customers in, for example, choosing the right health insurance. Even when bots are not used yet, the prediction of switching customers has benefits for service or insurance companies as they can adjust their assistance to the expected behavior of the customer.

2. Related work

This section describes the scientific context of this thesis and discusses the related research fields. Previous research about behavior predictions from voice, the production of speech within the human body, influences between speakers in a conversation, and consumers switching behavior regarding health insurances are discussed.

2.1 Predicting behavior from speech

Linking voice characteristics from call center speech to predicting future behaviour is a premature research topic. Voice activity, interruption, and hesitation were found to be valuable features for identifying successful calls in a call center (Atassi & Smékal, 2014). That means calls that resulted in a sale. Furthermore, the rate of speaking from customers is successfully used to label problematic calls (Pandharipande & Kopparapu, 2012). These features reveal promising outcomes in call center research but do not yet extricate the predictive potential from prosodic features.

Conversations contain much more data in the form of acoustic features from the caller's voice, than call structures or dialogue features can reveal. By using, for example, the frequency, the intensity and the energy of the caller's voice, researchers managed to detect emotional states of the caller (Devillers, Vaudable, & Chastagnol, 2010; Vaudable & Devillers, 2012). These studies are important first steps in the field of real-life speech research in call center settings. An important finding that results from these research efforts is the successful application of prosodic features for automated predictions of emotional states. These preliminary works substantiate the use of acoustic features, but additional research should be looked upon to predict behavior from voice.

Although the assessment of voice features for predicting behavior currently makes use of human judgements, the potential in the automated use of voice features is visible in this field of research. Judgements on thin slices of voice clips managed to successfully predict the effectiveness of salespeople, by assessing their personality, trustworthiness, motivations and affect (Ambady, Krabbenhoft, & Hogan, 2006). The availability of semantic content had no impact on the quality of the predictions, indicating the value of speech characteristics. A closer link to predicting behavior is found in the work of Rogers, ten Brinke and Carney (2016): human observers managed to successfully predict people's future voting behavior. Thin sliced speech fragments with self-stated predictions on voting were used by the observers. People's voices were successfully judged and used to differentiate between people that would follow through on their predictions or not. Uncertainty and deception are the psychological traits causing the vocal cues that were used to separate between voters. Uncertainty is detected in a

voice within milliseconds by a human observer due to the quiet voice and the rise of the pitch at the end of a sentence (Jiang & Pell, 2015). Deception comes from a combination of cognitive load and arousal, which may result in a higher pitch in the voice as an effect of tensions and nervousness (Zuckerman, DePaulo, & Rosenthal, 1981). If these humanly observed voice features resulted in successful predictions of future behavior, the potential of automated retrieval and modelling of voice features is promising. The challenge remains in extracting the right features in a reliable way, building data sets from real-life speech data, and successfully applying models to disclose the potential of acoustic features (Eyben et al., 2016).

2.2 Speech production

To build a data set from acoustic features of speech, it is essential to scrutinize the way speech is produced in the human body. This illustrates the connections of the relevant muscular movements and will provide insight into the parameters that can be extracted for automated analysis. The exposition focusses solely on phonetics: the acoustic, perceptual, and production

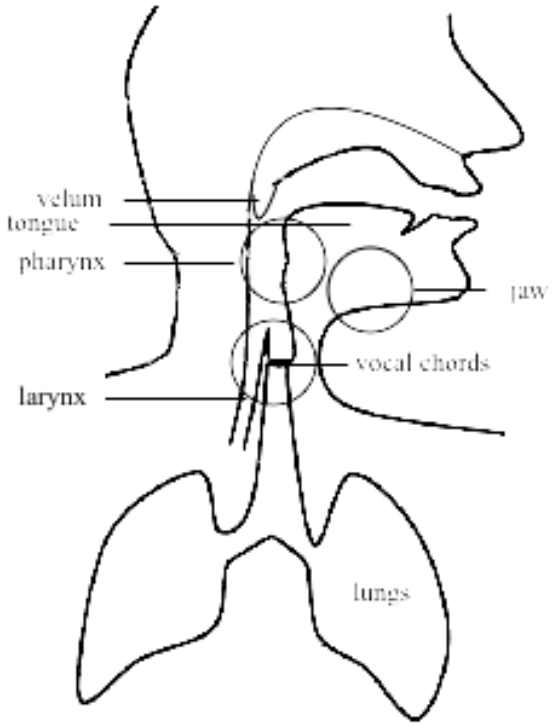

Figure 1. Speech organs in the human body, (adapted from Goswami et al., 2013)

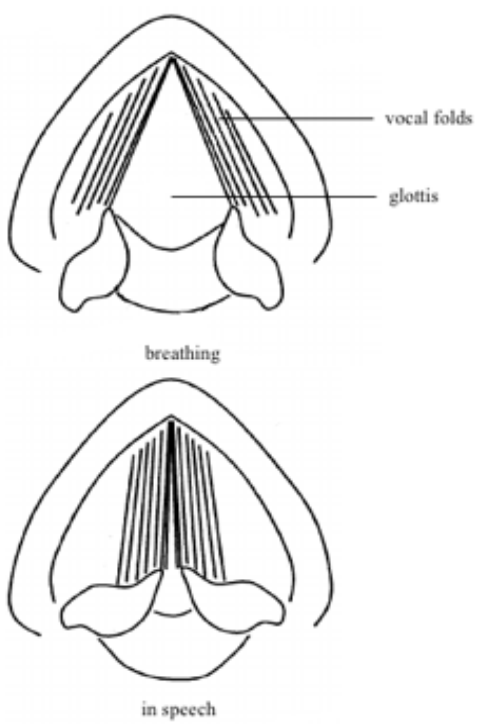

Figure 2. Vocal cords in the human body, (adapted from Story, 2002)

aspects of spoken language, in contrast to linguistics, which deal with written language (Schuller et al., 2013).

The production of speech in the human body involves multiple speech organs, as illustrated in figures 1 and 2. The signal starts from the air in the lungs that is forced through the vocal folds (or cords) and the space between them (glottis) into the pharynx. The generated

airflow is controlled by the tension and length of the vocal folds, which modulate the size of the glottis forming different frequencies of the speech signal. From that voice source, the signal enters the pharynx where it is affected by articulatory muscles as the velum, the tongue, the jaw, the lips, and movements of the tongue. The interplay between the voice source activities and the articulatory controls shape the total waveform of the speech signal (Lee et al., 2014). This is an automatic process where changes in any muscle's tension will impact the dynamics of the vocal tract (Cummins et al., 2015). The result is a complex interplay of speech muscles, making the extraction of valuable parameters challenging.

The distinctive components of the voice signal can be represented by features. These features can be extracted from the pure speech signal by representing the distinctive elements of the voice in an abstract manner (Kinnunen & Li, 2010). In total, many different features can be extracted, but three categories are distinguished: prosodic, source, and spectral features. Prosodic features are characterized by variations in the stress, rhythm, speech rate, and intonation of the voice. Source features give information about the source of the speech signal, with jitter and shimmer as important measures. Spectral features represent the speech spectrum, which is the distribution of different frequencies in the voice (Lee et al., 2014; Cummins et al., 2015). These features are examples of standard features that are used for automated emotion detection, better known as low-level descriptors (LLD's). This thesis will focus on three specific features.

2.2.1 Pitch

The first extracted feature is the pitch. This parameter is determined by the rate of vibration from the vocal folds and the opening of the glottis (Praksah & Gaikwad, 2015), they work as a filter for the speech signal passing through it and have its reflection on formant frequencies. Pitch relates to the fundamental frequency or formant 0 (f0). Hertz is the unit of measurement and relates to the number of speech wave cycles per second.

One of the major challenges when extracting pitch is to use a robust estimation algorithm since pitch extraction is prone to errors (Schuller et al., 2011). For this thesis, the extraction technique from the COVAREP toolbox (Degottex et al., 2014) is used, which has proven itself as a robust measure. Besides errors in the estimation of pitch values, personal differences like gender play a role in pitch differences as well. Table 1 displays the differences between genders and low-pitched and high-pitched voices.

Table 1

Differences in pitch value

Gender	Low-pitched voice (Hz)	High-pitched voice (Hz)
Male	98-125	152-178
Female	115-151	189-225

Note. The data in Table 1 is derived from "Effect of the tone of voice and the perception of the face in the formation of impressions on media speakers", by M. T. Soto Sanfiel, 2008, Comunicación y sociedad, 10, p. 143.

As part of prosodic speech, pitch is a feature that is strongly influenced by emotions and represents important cues for affective speech (Cowie et al., 2001; Lee et al., 2014). Among others, researchers found support for pitch as a predictor of happiness (Kim, Lee, & Narayanan, 2010), stress and anger (Eyben et al., 2016), arousal (Yildirim et al., 2004), and for boredom (Goudbeek & Scherer, 2010). As shown above, pitch is a widely-used feature in emotion classification and is found to be a robust parameter. At last, using pitch features is found to be most suited for binary emotion classification (Busso, Lee, & Narayanan, 2009), strengthening the support for the use of pitch in this thesis.

2.2.2 Intensity

The second extracted feature is intensity (or loudness). Like pitch, this feature is part of the prosodic features and originates from the source signal (Koolagudi & Rao, 2012). The source signal is formed when air from the lungs is pressed through the vocal folds in the larynx (Lee et al., 2014). This creates signal energy sound pressure (measured in Pascal (Pa)), which is normally described as an amount of pressure per unit of time. The variation of pressure differences in the atmosphere is enormous, but when speaking of sound pressure, levels between 10^{-5} Pa and 10^{2} Pa are relevant (Zwicker & Fastl, 2013). The sound pressure level (SPL) is introduced to coop with the wide range of pressure differences:

$$SPL = 20 \log\left(\frac{p}{p_0}\right) dB$$

$p_0 = 20\mu$ Pa as the reference level of pressure.

Just as the SPL, the intensity (I) of sound is a logarithm of a ratio power, though with an energy reference level instead of pressure. The equitation of the sound intensity level (SIL) is therefore similar to that from the sound pressure level:

$$SIL = 10 \log\left(\frac{I}{I_0}\right) dB$$

$I_0 = 10^{-12}$ W/ m^2 as the reference level of energy.

Like the sound intensity level, when the intensity increases, the perceived sound increases logarithmically by the observation of the human ear (Schuller et al., 2011). However, measures of intensity or loudness strongly depends on frequencies: 100dB at 100Hz results in a different perceived intensity than 100dB at 500Hz. Therefore, to make the measurement of intensity closer to the human hearing, the intensity from the speech signal is weighted in different frequency ranges (Weninger et al., 2013). The intensity in dB is then calculated by taking the sum of all frequency ranges and apply this to the logarithmic scale. In this thesis, the methods from the COVAREP initiative are used to extract the intensity form the voice signal (Degottex et al., 2014).

The intensity of the voice has proven to be a valuable indicator of different emotional states. Prosodic features in general, with intensity as an important part of them, are found to be relevant as a predictor of arousal in humans (Scherer, 2003; Goudbeek & Scherer, 2010). Next to classifying emotional states, intensity has also proved valuable in classifying distinct emotions such as high intensity in human voices, that is related to joy, anger, stress, and fear and low intensity, that is associated with disgust and sadness (Ramakrishnan, 2012; Ververidis & Kotropoulos, 2006; Scherer, 2003). Altogether, intensity proved to be an important feature in classifying speech instances.

2.2.3 MFCC

The third and last feature that is extracted is the Mel-Frequency Cepstrum Coefficient (MFCC). MFCCs are part of spectral shape parameters and are one of the most used features in speech recognition research from the last decades (Eyben et al., 2016; Cutajar et al., 2013). They represent energy in different frequency bands, like the cochlear human auditory system's operation. Compared to pitch and intensity, MFCCs are rather complex and a diversity of algorithms is used in the extraction process (Zheng, Zhang, & Song, 2001). This process consists of a sequence speech signal transformations. This paragraph provides a step-by-step description of these techniques.

MFCCs are extracted from the frequency domain, which is converted from the initial time-based speech signal. Short time windows, typically 25 ms or 30 ms, are shifted with 10 ms steps to cover the full audio fragment (Young et al., 2002; Cutajar et al., 2013). Within each window, the spectrogram of the signal is transformed into a spectrum by making use of a Fourier transform. This transformation represents the sum of multiple sine and cosine functions and reveals where energy is located in the speech signal (Harris, 1978). The peaks in the spectrum display the dominant frequency areas in the signal, the fundamental frequency and

the formants are visible as peaks of the spectrum (Cummins et al., 2015). The energy-frequency representation of the spectrum enables further extraction of the MFCC.

In the human auditory perception, frequencies are non-linearly dispersed over the spectrum causing the need for applying filter banks to the spectrum to obtain a linear frequency distribution (figure 3). The human perception of sound is linear up and until 1000 Hz, but

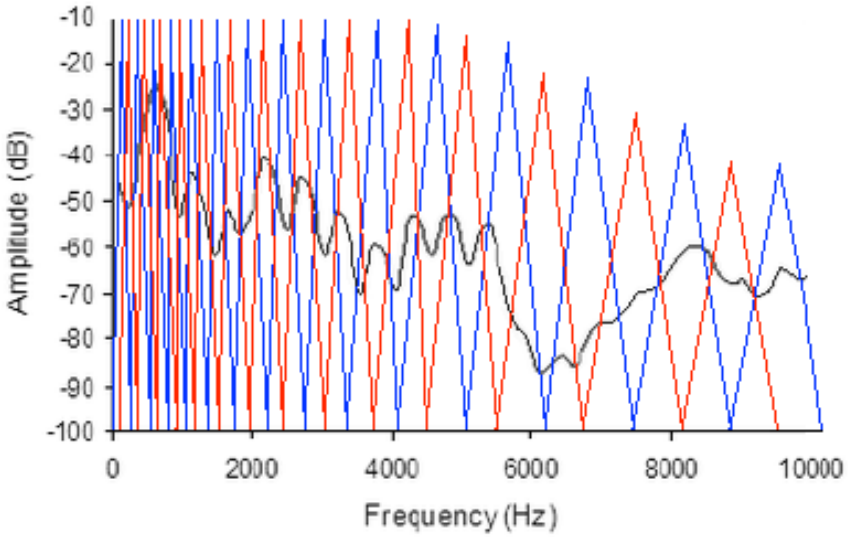

Figure 3. Spectrum of speech signal and Mel scale filter bank.

increases logarithmic from there. To mimic this distribution of frequencies, the spectrum is filtered using the Mel scale. This scale is constructed to deal with the subjective distance between frequencies at different levels (Wicks, 1998). The conversion to a linear Mel frequency scale is done with the formula suggested by Fant (1973):

$$F_{mel} = \frac{1000}{\log(2)} * \left[1 + \frac{F_{Hz}}{1000}\right]$$

The obtained spectrum with linear frequency distribution has Mel frequencies (F_{mel}) as unit of measurement and supports the further calculation of MFCC's.

Since MFCCs represent energy from frequency bands, the energy is calculated from each Mel-filter bank. To retrieve this energy, first, the Mel spectrum is transformed with an inverse Fourier transform. The output from this transform is called the cepstrum (Bogert, 1963). From this cepstrum the log energy output is calculated with this formula (Cutajar et al., 2013; Ganchev, Fakotakis, & Kokkinakis, 2005):

$$X_i = log_{10}\left(\sum_{k=0}^{N-1} |X(k)| \cdot H_i(k)\right)$$

for $i = 1, 2, \ldots, M$

Individual filter banks are represented in the formula by $H_i(k)$ and the number of coefficients is given by M. Accordingly, the real algorithmic energy is represented from the filter band's spectrum. A cosine transformation is used to obtain the cepstrum coefficients and therefore complete the calculation of the MFCCs. The formula that is used for this transformation is (Davis & Mermelstein, 1980; Zheng, Zhang, & Song, 2001; Mporas et al., 2007):

$$MFCC_j = \sum_{i=1}^{M} X_i \, cos\left(j \cdot \left(i - \frac{1}{2}\right)\frac{\pi}{M}\right)$$

for $j = 1, 2, \ldots, J - 1$ that is the number of MFCC's that are needed.

In the full procedure of extracting the MFCC from a speech signal, many detailed changes can be made by, for example, varying the number of coefficients that are used and the inclusion of the 0^{th} coefficient (Zheng, Zhang, & Song, 2001). The 0^{th} coefficient is referred to as the spectrum's average value or, said differently, the collection of energy from all frequency bands (Picone, 1993). Due to questions about the reliability of the measure, this 0^{th} coefficient is excluded in most researches. That practice is followed in this thesis, congruent to the approach from the COVAREP initiative (Degottex et al., 2014). Besides the exclusion of the 0^{th} coefficient, the number of total MFCCs used has an impact on the performance of a model. Lower coefficients were found to carry more phonetic speech information and the higher coefficients more non-speech information, as found in research that discriminates speech signals from music signals using MFCCs (Mubarak, Ambikairajah, & Epps, 2005). The lower frequency MFCCs are linked to the distribution of spectral energy (Eyben et al., 2016), making them suitable to classify emotions in speech (e.g. Neiberg, Elenius, & Laskowski, 2006). MFCCs from the higher frequency bands are more related to small energy dispersions in the speech signal, which makes them more reliable to identify semantic content (Eyben et al., 2016). For that reason, a balanced number of MFCC coefficients should be used that fits the scope of the research.

Overall, MFCCs are found to be robust against noise in the signal but tend to be affected by the textual content (Schuller et al., 2011). Poor performance is often also related to this textual dependency (Ververidis & Kotropoulos, 2006). LFPCs are proposed because they include pitch values. However, pitch is already included as a feature in the feature set of this thesis, taking away the disadvantages from using MFCCs. Therefore, the use of MFCCs seemed to be justified in the research context of this thesis.

2.3 Interspeaker influence in call center conversations

Even when acoustic features can be extracted from the voice of a customer and it is understood how these features are produced within the human body, it is important to understand the theory regarding human to human interaction. The interspeaker influence in telephone calls has implications for the measurements of the customer's voice and should be considered when natural speech data is analyzed. General theories of human interaction are applicable, as well as more specific findings from studies executed in call centers.

People have the tendency to adjust their speech and vocal patterns towards others in a conversation. Communication accommodation theory shows how people modify their speech when interacting with others. Social differences are the reason of speech adjustment in this theory, divergence is used to emphasize the contrast and convergence is used to minimize the inequality between people (Giles, Coupland, & Coupland, 1991). Synchrony and similarity in a conversation build rapport and affiliation (Bernieri, Reznick, & Rosenthal, 1988) and since call centers focus on the assistance of customers, minimizing the social difference between expert and customer is beneficial. Therefore, convergence is expected to be more dominantly present in the calls than divergence.

Both the expert and the customer adjust their speech to be more like the other. This accommodation of the voice is also called vocal mimicry (Sun, Truong, Pantic, & Nijholt, 2011). Verbally, mimicry can result in copying words or expressions of the other in a conversation (e.g. Gonzales, Hancock, & Pennebaker, 2010). Non-verbal mimicry is about the imitation of the way of talking, the effects have a long history in the literature. For example, people were found to mimic speech rate (e.g. Webb, 1972), the rhythm of speech (Capella & Planalp, 1981), and speech accents (Giles, 1973). Mimicry can be beneficial in a customer service call center setting for the outcome of the final decision because peoples' general social orientation is affected by it, making them more inclined to help others (Van Baaren et al., 2004). For that reason, the adjustments in speech that come from interspeaker influences can have a positive impact on telephone conversations in which experts give advice to consumers on switching health policies.

The effect of speech adjustment in dyadic interaction goes beyond imitation of the vocal signal. Vocal mimicry is part of an overarching theory, called emotional contagion. This theory describes the synchronization of vocal features (mimicry) as the first phase and emotional conversion that succeeds as the second phase, called feedback (Hatflied, Caciopo, & Rapson, 1993). After mimicking vocal cues, an emotional state is experienced that is compatible with those vocal cues. Therefore, the emotional states of an individual are affected by the vocal cues

of another person in a conversation (Hatfield et al., 1995). Emotional contagion is one of the most influential processes in interpersonal communication and is found to happen fast and unconscious when two people interact in a call center. Furthermore, vocal mimicry was found to be stronger for negative emotions than for positive emotions (Rueff-Lopes et al., 2015). Altogether, vocal signals are automatically copied by both the expert and the customer and the congruent emotions are experienced by the other person in the conversation.

The management of emotions in a call center interaction has implications for the outcomes of that call. In general, emotions are found to be a valuable predictor for the emotions of the customer in service encounters (Pugh, 2001; Berger & Grandey, 2006). Furthermore, the emotional competence of an employee in customer service has a positive effect on the loyalty and satisfaction of the customer (Delcourt et al., 2013). To this extent, positive emotions expressed by call center experts can be expected to have positive effects on the outcome of a call. This is partially proven by the emotional contagion theory. Considering that the emotional expressions from a call center expert, via speech, spontaneously have an impact on the emotions of the customer. Beyond this automatic process, adopting the right strategy to deal with customer emotions will decrease negative and increase positive emotions from the customer in a call center (Little et al., 2013). Both conscious and unconscious processes affect the conversation in a call center, yet positive emotions can be expected to have the most beneficial influences on the outcome of the call.

2.4 Consumers' switching behavior

When predictions are made on switching health insurer, it is valuable to understand the differences in the probability of switching within the customer population. Not every customer in the health insurance market is as likely to switch like any other. Since the introduction of the regulated competition in the Dutch health insurance market in 2006, 37 % of all consumers switched insurer at least once (Romp & Merckx, 2017). Nevertheless, differences are found in demographics between switching and non-switching customers. Young, healthy people are more likely to switch health policies than elderly people with poorer health (Boonen, Laske-Aldershof, & Schut, 2016; De Jong, Van den Brink-Muinen, & Groenewegen, 2008; Duijmelinck, & van de Ven, 2016). Results about gender differences are more ambiguous, as women were found to be less likely to switch health insurer (Hendriks et al., 2010) and, at the same time, they are more inclined to search for information about the different policies (Rademakers et al., 2014). These findings show a bit of the disunity in switching behavior of Dutch consumers.

To predict switching behavior of customers, differences in demographics do not appear to provide sufficient information. The distinctions found in groups of consumers switching policy are set in a different perspective by the study of Hendriks et al. (2010). The rate of consumers switching health policies is consistent over different groups at 31%, but only when people intend to switch. Consumers differ in their intention to switch, but not in the conversion to actual switching behavior. For that reason, it is important to look at possible bumps in the road towards formulating the intention to switch health insurance when looking to predict switching behavior.

Defining the construct of having the intention to switch is not a clear-cut task. Recent literature finds no support for critical reflection on price and service quality by customers, even though an active role of the consumer is an important aspect of the new health insurance system (De Jong, Van den Brink-Muinen, & Groenewegen, 2008). It is found that consumers rather hand over the switching to a group purchasing organization, because of lower transaction costs for the individual (Lako et al., 2011). Furthermore, the number of choice options, the available information, and the possibility to switch via the internet are other reasons not to switch health insurer (Duijmelinck, Mosca, & van de Ven, 2015). However, when individuals search for health plan information, their likelihood of switching increases as well as their sensitivity to price (Boonen et al., 2016). Therefore, the searching behavior from customers seems to be closely linked to the intention to switch. In this thesis, consumers that actively search for information and compare different policies are considered to have the intention to switch. Hence, completely satisfied consumers have no interest in searching for information on alternative policies.

Defining switching behavior is more straightforward. In 2017, customers had a choice between 24 different health insurers (Romp & Merckx, 2017). For the definitions of switching between these insurers, this thesis will use the definition of the Dutch Healthcare Authority. They consider a consumer to be switching when he or she changes health insurance provider, only changing to a different package from the same insurance company is not considered “switching” (Nederlandse Zorgautoriteit, 2016).

Understanding the conversion from the intention to switch to actual switching behavior is important. Theory of planned behavior describes perceived behavioral control as an important aspect of the intention towards a certain action (Ajzen, 2002). Lesser perceived behavioral control over switching behavior can be experienced by consumers when they face barriers in their intention to switch, like the time and effort needed to find alternative policies and the uncertainty about the quality of service from the new insurance (Hendriks et al., 2010).

Research from the US' health insurance market supports the finding that perceived behavioral control is a valuable predictor for the way consumers process information in their search of new health insurance, as well as assistance in choosing the right policy (Han & Urmie, 2017). Assistance comes from knowledgeable experts that provide counselling via websites or over the phone. When consumers perceive that the benefits outweigh the costs of changing health policy, actual switching occurs (Duijmelinck, Mosca, & van de Ven, 2015). Therefore, perceived behavior control and assistance from experts are important conditions for converting the switching intentions into actual switching behavior.

3. Methods

In this section, multiple data sources are described that are used to construct a data set. The steps towards building the final data set for the experiments are defined and the mathematical models are introduced that are trained for the classification task of this thesis.

3.1 The data

The data that is received from the company comes in a variety of forms and in different files. Most importantly, 173,416 call center calls are recorded and provided by the company. Furthermore, internal planning data and additional characteristics of the call and the customer is provided as meta data.

3.1.1 Speech data

The company has recorded calls from their call center in the period from September 23, 2016, till January 3, 2017. The distribution of different all the calls in this period is visualized in figure 5.

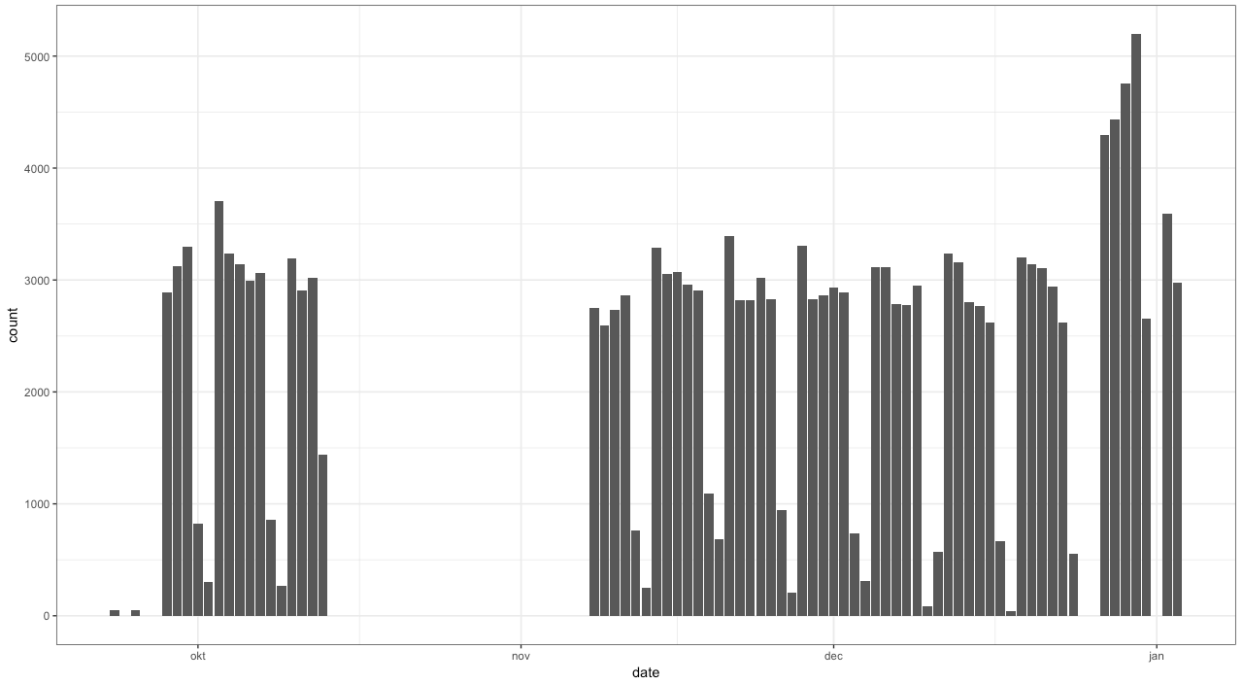

Figure 5, distribution of call center calls from 23-09-2016 till 03-10-2017.

All the recordings are stored in folders per day, which contain folders of hours per day (for example 09:00 – 10:00). Each individual call comes as a .wav audio file, with a unique filename, an example given: “094543086ZxO&^1018_+33.012.wav”. This filename consists of some meaningful sub-parts. The numbers “094543086” from the example refer to the time of the day that the call was recorded, accurate in tens of microseconds (10^{-5th} second). Furthermore, the “O” refers to an outbound call, “1018” refers to the workstation from where the call was recorded and “33.012” refers to the length of the conversation in seconds (accurate till a thousandth of a second).

3.1.2 Non-acoustic data

The meta data on the characteristics of each call and the planning of all workstations is provided in two separate files. The calls' characteristics consist of a file with 173,416 rows and 47 columns. From this data, useful features are extracted after deleting 18 columns with redundant features. Columns are redundant when they contained duplicate information from other rows, contained no values at all, or had inexplicable information like "757ac46ede1df0c25". Additionally, columns were deleted that contained sensitive private information from the customer with no value for the analysis (for example the name of the customer). The exact features extracted and there meaning are described in section 3.3.

The planning of the workstations consists of 158,341 rows and 5 columns. This data covers the name of the agent, the workstation, the initials of the agent, the product group that the agent was assigned to, and the date plus time. The planning covers information from October 1, 2016 untill December 31, 2016. For these days, every workstation is connected to an agent's name, a product group, and a time of the day. This information allowed to assign a product group to each call center conversation by making use of the date and workstation number. Unfortunately, workstation numbers do not correspond to one product group per day. It is possible that someone has a call from a different product group in between. An explanation for this might be the fact that when the telephone's exchange capacity is full on one product group but has free agents on another, a call can be directed to that free agent from a different product group. Forasmuch as the time in the planning, this does unfortunately not correspond to the times of the call center calls. For that reason, calls could not be connected to a product group by their time on the day. The solution to this is found in assigning the most occurring product group per workstation per day to all calls from that day and workstation. This approach allows a few calls to be assigned to the wrong product group, though this is only true for a very small number of calls because of the reason mentioned above.

3.2 Constructing the data set

For the analysis of the calls, not all 173,416 calls will be used. Multiple selection criteria contribute to the construction of a subset for the experiments. In the next subsections, all criteria are specified and illustrated.

3.2.1 Product group

As stated above, the full data set of 173,416 calls comes from a range of different product groups, these need to be limited down to calls about health insurance. Because the planning is limited to data from October up to and including December, the data set is reduced to this range in dates. The total number of calls that remains is 157,438 (90.8%). Furthermore, the planning revealed a total of 2,403 incorrect agent names (for example: "%13668%"). These names and all corresponding calls were omitted from the full data set, resulting in a data set of 130,913 calls (75.5%). In this set, the present product groups are "Auto", "Bancair Leven", "BZR", and "Pakket en TP".

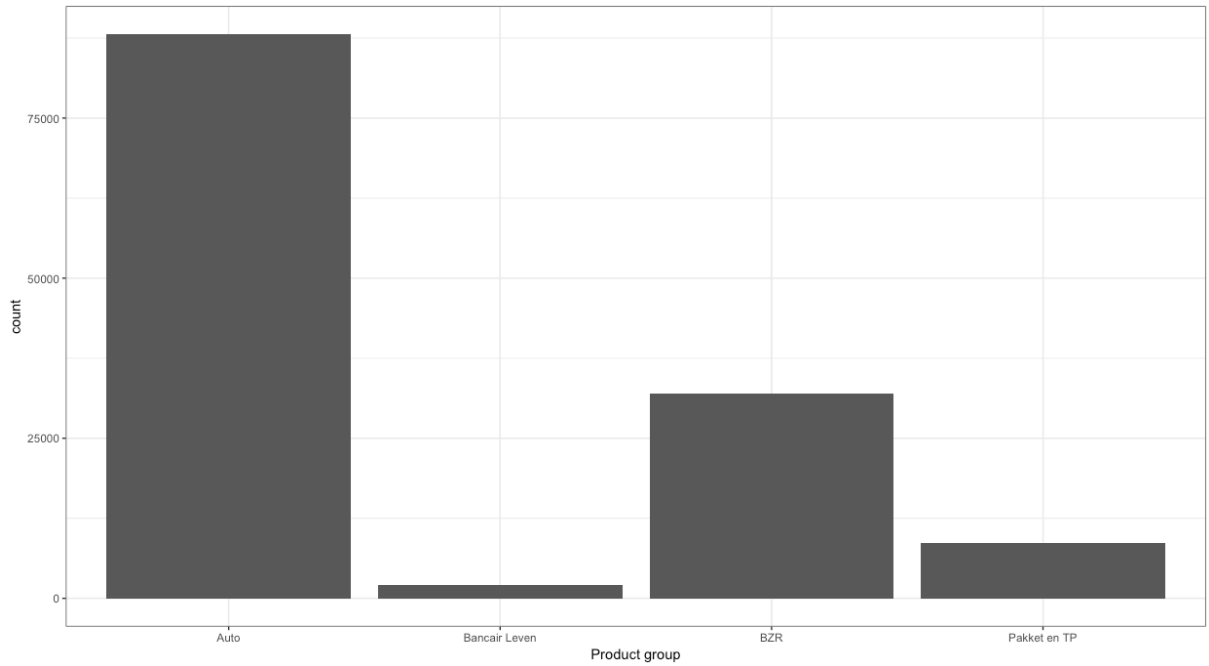

Figure 6, distribution of the product groups.

Since the health insurance calls need to be extracted, only the "BZR" calls (about health insurance, fire insurance and travel insurance) are selected from the 130,913 calls, leaving 32,032 conversations (18.5%). From these calls, another selection is made. Calls are selected from the health insurance season. Health insurers in the Netherlands are obliged by law to publish their premium six weeks before the end of the year (art 17 clause 7, Health Insurance Act 2005). In 2016 this meant before November 19. Thereby, as of November 19, 2016, all premiums from Dutch health insurers are published, allowing customers to make a fully informed comparison. Switching health insurance provider is only possible until December 31. The period that allows a customer to effectively compare and change health insurers is from the November 19, up to and including the December 31 (also called the health insurance

season). Calls from that period are selected for further analysis in this thesis. This leaves 23,155 conversations in the data set (13.4%).

3.2.2 Duration of the call

The duration of the recorded calls varies from 0 to 3,592 seconds, with different occurrences for every duration. The distribution of the telephone calls' duration is visualized by figure 7.

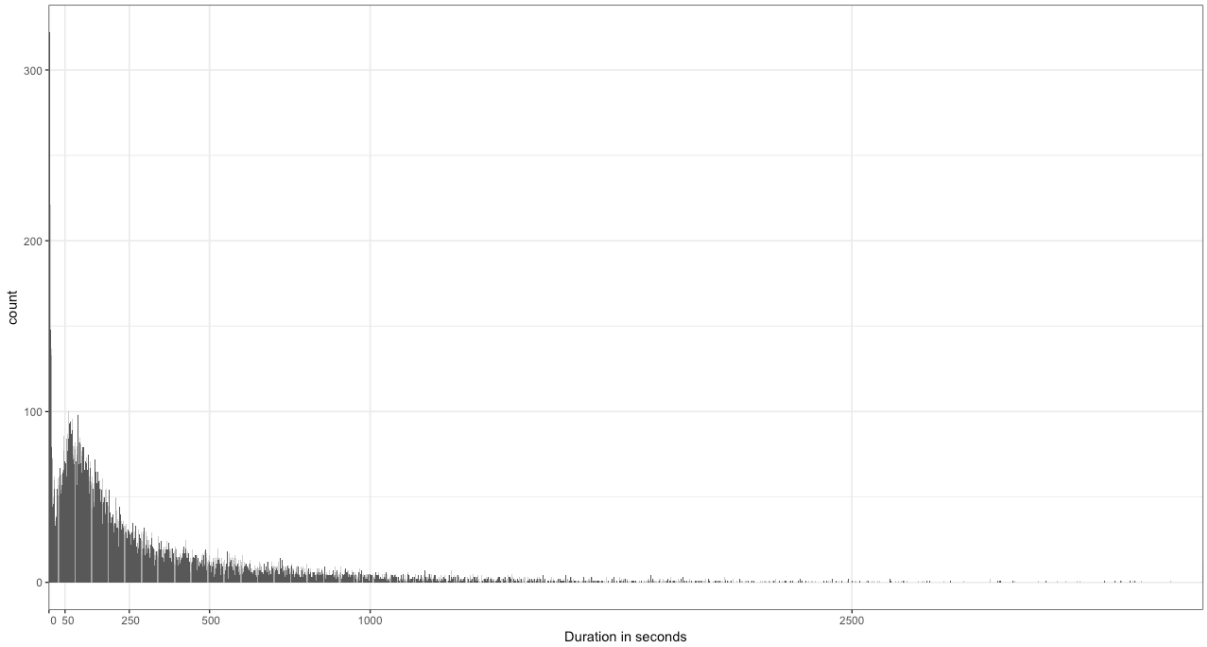

Figure 7, distribution of the calls' duration in seconds.

Listening to a sample of the conversations revealed that short calls appeared to be not valuable for the analyses. Zero, one, or two seconds is logically not enough to build an informative conversation between the customer and the agent. However, a conversation from around 30 seconds can already contain valuable information for the analysis. Therefore, the right threshold had to be found to conclude which calls should be excluded from the analyses. Based on expert knowledge (M. van Os, personal communication, February 14, 2017) and insights derived from data exploration, the threshold is set to a minimum duration of 30 seconds. This assumption is tested by taking a random sample of 50 calls with a duration of 20 to 30 seconds from the full data set. Results from the sample show no reasons to change the threshold to a shorter duration in seconds, as shown by table 3.

Table 3

Results from calls with a duration of 20 to 30 seconds

Sample	N	Percentage	Mean	Standard deviation	Min	Max
Total calls	50	100%	24.8	4.3	6	30
Uninformative conversation	35	70%	24.6	4.7	6	30
Informative conversation	15	30%	25.3	3.1	21	30

All the calls with a duration of 29 seconds or less were excluded from the analysis, resulting in a total set of 20,824 calls (12%) after removing 2,331 short conversations.

3.2.3 Direction of the call

The 20,824 calls consist out of inbound and outbound calls, with a division of 14,894 inbound (71.5%) and 5,916 outbound calls (28.4%). For 14 calls, the direction was not known and therefore these calls were omitted from the data set. Characteristics of inbound and outbound calls can be different, as found during data exploration. This is also confirmed by expert knowledge from the company (M. van Os, personal communication, February 14, 2017). Inbound calls always start with a customer calling the company, whereas in outbound calls the agent calls a customer and might even never talk to the customer if the phone is not answered. This assumption is tested by taking a random sample of 50 outbound calls from the 5,916 total of outbound calls. This sample showed that outbound calls are not useful for the analysis, because the number of real conversations between an agent and a customer is too low. Frequent problems in outbound calls appeared to be agents calling insurance companies, agents leaving voicemails to customers, and internal calls of agents calling colleagues. All these defective conversations in outbound calls resulted in the exclusion of all outbound calls, leaving 14,894 inbound calls (8.6%) for the analysis.

Table 4

Results from the outbound call sample

Sample	N	Percentage
Total calls	50	100%
Informative calls	23	46%
Uninformative calls	27	54%
Call to other company	10	20%
Voicemail from customer	7	14%
Internal call	5	10%

3.2.4 Data about the customer

From the 14,894 calls, a further selection is made based on the availability of more data from the caller. This data was found in the company's customer database. The telephone number from which the call was made connects the call to the customer's profile from which the data could be retrieved. Not all phone numbers were recognized in the company's database. Only when a caller has been a customer before, is a current customer, or has a contract with a future start date, a profile in the company's database could be found.

Since this thesis focusses on switching behavior, first time insured should be excluded from the analysis. That means that every 18-year-old (at that age people are legally bound to have their own health insurance) is removed from the data set, leaving 3,887 (2.2% of the total recorded) calls for the final analysis. The birthdate from the caller was used to calculate the age. This is done by dividing the number of weeks from the caller's birthdate by 52.25 and rounding it to the lowest whole number.

3.3 Features from meta data

The meta data that is provided by the company is used for more purposes than constructing the final data set. The non-audio data that was provided from the customer database also adds features to the data set that can be used for the classification task. From the customer's profile, valuable data was retrieved: birthdate and gender.

Five non-acoustic features in total are derived from the meta data: the duration of the call, the hour of the day the call was made, the customer's gender, the customer's age, the number of days till the deadline of the health insurance season, and the target feature of this research. That is the binary feature indicating if a customer has switched health insurance or not.

3.4 Acoustic features

To analyze the speech from the selected calls, acoustic features are extracted. The first step in this procedure is selecting the right short time window for the analysis, called a thin slice. Afterwards, features are extracted and high statistical functions calculated from them.

3.4.1 Thin slices

A short section from each conversation is taken for analysis: a thin slice. Support for taking thin slices is found in widely used speech databases for speech and emotion recognition. These databases are built out of audio fragments of a few seconds (Schuller, Steidl, & Batliner, 2009; Schuller et al., 2010) or only one sentence (Burkhardt et al., 2005; Bänziger, Mortillaro, & Scherer, 2012; Grimm, Kroschel, & Narayanan, 2008). Whereas these databases hold short recordings in their full length, other research uses the same approach as this thesis. Segments

were used with a fixed length, meaning that a longer signal was cut down to a certain time limit (Povolny et al., 2016). More general, thin slices were found to have an important predictive value in social psychology (Ambady, Bernieri, & Richeson, 2000).

Before any thin slice was taken, the right time window in the selected calls should be found. A hundred calls were selected randomly from the 3,887 recordings. These calls were manually listened to with the purpose of establishing the common structure of an inbound call and the time that each distinctive part of the call lasted on average. Six distinctive segments were found in the structure of the calls: the introduction of the agent, the introduction of the customer, main question from the customer, exchange of information and answering the question, conclusion, and the ending.

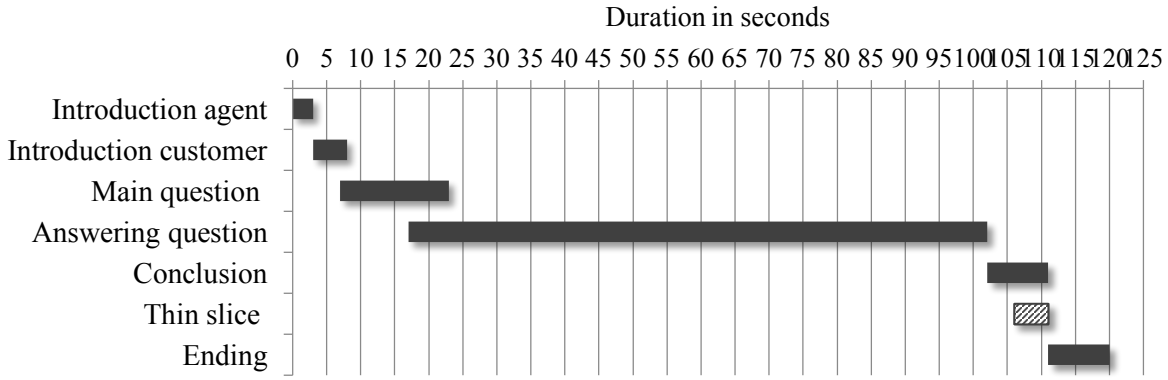

Figure 8, An example structure of a call with a duration of 120 seconds.

The focus of this thesis is on the conclusion of the conversation. This section was found to be the crux of the whole call since the agent gives the final solution on the questions from the customer. On average, the conclusion took 9 seconds (standard deviation = 6.24 seconds), as well as the ending of the call (standard deviation = 7.00 seconds). For the thin slice, the chosen time window is calculated by starting at the end of the conversation. Accordingly, the start of the thin slice is -14 seconds and the end is -9 seconds (total length of 5 seconds), as it aims at selecting the end of the conclusion from the call. The full structure and selection of the thin slice is graphically displayed in figure 8.

Another random sample of 100 recording was taken to check the proportion of speech from the agent and the customer in the conclusion (results of the sample can be found in table A1 in Appendix I). Results revealed that in the conclusion, on average, both the agent and the customer were present. Furthermore, some calls did not have a clear conclusion, for example, because the customer was connected to another agent.

3.4.2 Acoustic feature extraction

Acoustic features were extracted by making use of the COVAREP toolbox (Degottex et al., 2014). The COVAREP project aims at more reproducible research in the field of voice analysis by making state-of-the-art algorithms for feature extraction freely available. Their extraction methods are chosen for pitch and intensity in this research. The work of Pandit (2015) was used for extracting the MFCC. The basic structure of Pandit's MFCC extraction is very much related to the approach from the COVAREP project. Octave (an open source Matlab tool) is used for the feature extraction since the extraction algorithms from COVAREP and Pandit are provided in this coding language.

3.4.3 High-level statistical functions

High-level statistical functions (HSFs) are calculated from the low-level descriptors, being the basic extracted prosodic features (pitch and intensity) and the spectral features (MFCC's). The raw values that are derived from the feature extraction hold a short-time temporal structure, for example, a pitch value is derived every 100th of a second. The functions or HSFs that are calculated have a long-time or supra-segmental time structure because they hold values for the entire utterance (Anagnostopoulos et al., 2015). Predictive value rather lies within utterance or frame wise variations than in static short-term LLDs (Anagnostopoulos et al., 2015; Mirsamadi, Barsoum, & Zhang, 2017). Therefore, the final data set consists of many HSFs, that describe the temporal variations and contours of the LLD's at the thin slices level.

3.4.3.1 Pitch HSFs

The extraction of the raw pitch values itself is done every 100th of a second for the full 5-second-length of the thin slice, resulting in 500 pitch values. Global and local statistics are calculated from the raw pitch values because the most explanatory features from pitch values are found to be the continuous variations (Busso, Lee, & Narayanan, 2009). Global calculated values are: mean, minimum, maximum, range, delta, standard deviation, skewness, and kurtosis (Anagnostopoulos et al., 2015; Busso, Lee, & Narayanan, 2009). Local statistics are calculated per frame; each frame holds 10 raw pitch values. The mean per frame is taken and the differences between each two-consecutive means calculated. A list of all these differences is taken and from those values, the mean, minimum, maximum, range and standard deviation are calculated (as used by Anagnostopoulos, & Iliou, 2010). Besides, the mean interquartile range is calculated by detracting the 25th percentile from the 75th percentile of all the frame means. In a similar procedure, the interquartile ranges from the minima and maxima are calculated from lists of all local minima and local maxima.

3.4.3.2 Intensity HSFs

Intensity is retrieved from the raw signal with an output value for every 1000th of a second. The HSFs that are taken from these values for every thin slice are: minimum, maximum, delta, range, standard deviation, and the signal to noise ratio (Mirsamadi, Barsoum, & Zhang, 2017; Anagnostopoulos et al., 2015). This measure is the proportion of true signal in the thin slice compared to noise. Sound with an intensity lower than 35 dB is handled as noise and everything above 35 dB is taken as part of the true signal in this calculation. Each of the statistical functions is calculated over the global signal, giving a total of six HSFs that are calculated from the intensity of the thin slice.

3.4.3.3 MFCC HSFs

From the recordings, 12 MFCCs are retrieved and HSFs are calculated per coefficient. As done for the pitch and intensity measures, the MFCCs' HSFs are calculated from the raw values of the coefficients, also known as the energy measures. Each thin slice had a 500 by 13 output matrix consisting of 500 vectors with the 0th till the 12th coefficient for every frame of the MFCC extraction. As described before, the 0th coefficient (or power) is excluded in this research and therefore a 500 by 12 output matrix remains for every thin slice. The HSFs are calculated per coefficient and over the 500 frames for every file. In total, seven HSFs are calculated: mean, minimum, maximum, median, delta, range, and standard deviation. This results in a total of 84 features that are added to the total data set by the MFCC calculations.

Comparing the deltas between the vectors of MFCCs is an often-used approach, but probably most suitable for detailed prediction tasks like voice recognition. For a rather robust task like categorizing audio fragments, summarizing the sequence of MFCC vectors and calculating HSFs over them can be more applicable. A related approach is used to label real-life call center audio as negative or positive in emotions in previous research, indicating the usability of this method (Vaudable & Devillers, 2012).

3.5 Algorithms

Different classifying algorithms are trained to distinct patterns in the data. This subsection describes the chosen models and their proven value in applications from other researches with similar data sources. Multiple factors are considered for the choice of the classifiers. First, the tolerance of high dimensionality affects the performance of classifiers. This problem is often worked around by applying feature selection, but it also relates to classification performances. Some algorithms proved to work well on real-life speech data, but did still suffer from an increasing number of features, like the k-nearest neighbor algorithm for example. Support vector machines are supposed as the answer to this problem (Schuller et al.,

2010). Thereafter, an algorithm's capability of being applied to a smaller data set should also be considered. Naïve Bayes is often chosen when researchers deal with small data sets. At last, the skewness of classes in the data can also affect the classification performance of an algorithm. Support vector machines might benefit from this for example (Eyben et al., 2016).

When all the characteristics of the data are evaluated, different types of classifiers can be chosen. Linear classifiers base their decision on the linear product from the feature vectors, while non-linear classifiers use a weighted combination of the values from the feature vectors. These non-linear classifiers come with a risk of overfitting, due to many dimensions of freedom (Koolagudi & Rao, 2012). The chosen model should fully depend on the nature of the data. In real-world settings, like the speech data from this thesis, the nature of the data is not known and therefore a broad approach where multiple algorithms are tested is used. This method supports the preliminary character of this thesis in the existing body of literature.

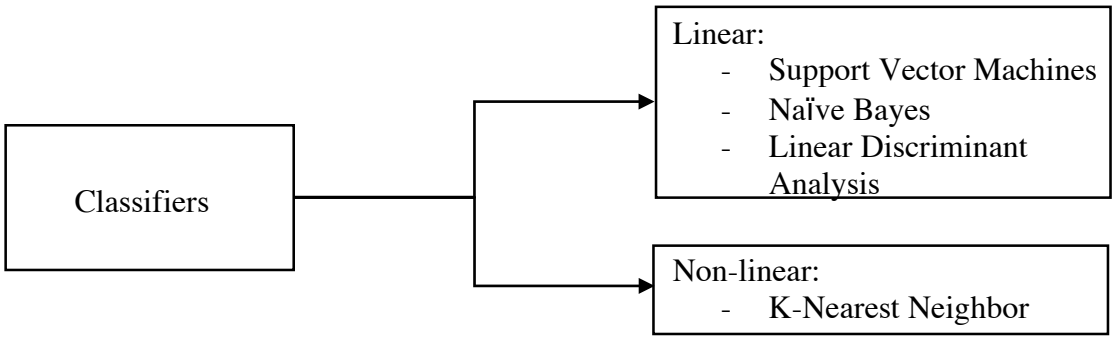

Figure 4. Types of classifiers

3.5.1 Naïve Bayes

The naïve Bayes algorithm is a popular algorithm in data mining, especially when it comes to text classification. This classifier assumes that all the independent variables are independent of each other, an assumption that is often not true in real-life situations. However, the naïve Bayes model has proven high performance in complex real-world situations (Rish, 2001). On top of that, it performs well with a small number of training instances and implementation is rather simple. The logic within the algorithm makes use of the conditional probability and assigns the most likely class to an unseen instance based on its feature vector. This classifier is often used as a baseline to benchmark the performances of other algorithms against.

3.5.2 K-nearest neighbor

The k-nearest neighbor algorithm (k-NN) places all instances in a multi-dimensional space and assigns unclassified points to the same classification as the nearest sample point that is already classified. It is assumed that classified observations that are close together belong to

the same class since the distribution of all points is independent and based on the feature values. Therefore, it is reasonable to expect that data points that are close in the multidimensional space have similar feature values increasing the probability to belong to the same class (Cover & Hart, 1967). The distance between the data points is often measured in Euclidian distance, but other measures are also possible. Tuning of the k-NN algorithm is done by adjusting the value for *k,* meaning that a different number of neighbors is considered when an unclassified data point is predicted a class label. Nowadays, like naïve Bayes, the k-NN classifier is often used as a benchmark or baseline model to compare the performance of other models against.

3.5.3 Linear discriminant analysis

Linear Discriminant Analysis (LDA) is a widely-used algorithm for both classification and dimensionality reduction. The model creates new axes that explain the variance of the data points, maximizing the variation between classes and minimizing variation within classes. It operates like principal component analysis (PCA), but, in contrast to PCA, LDA maximizes the distance between two classes of data points whereas PCA finds new axes that explain most of the variance. With LDA, categories in the data are separated by plotting it on a two-dimensional scale. In speech recognition research, LDA has been used for classification (Balakrishnama & Ganapathiraju, 1998).

3.5.4 Support vector machines

The most widely chosen classifier in acoustic speech research are support vector machines (SVM) (Eyben et al., 2016). These can be both a linear and a non-linear classifier, depending on the kernel that is used to separate the different classes. It is assumed that there is a hyperplane between the different classes that separates the data points. The data points that are on the edge of the hyperplane are called the support vectors, these help to determine the shape of the hyperplane. The algorithm looks to optimize the perpendicular distance from the support vectors to the center of the hyperplane. As a linear classifier, the support vector algorithm looks for the largest possible distance between the two classes (Burges, 1998). Non-linear applications of support vector machines make use of roughly the same principle, only they map the data into an infinite dimensional space and look for more complex hyperplane structures. This thesis only applies linear SVM models.

SVMs can be tuned by using different values for the complexity measure cost or *C*. This value can be varied to optimize the performance of the classifier. The value for *C* determines the penalty that is assigned to errors, being data points that are not separated correctly by the position and size of the hyperplanes (Burges, 1998). Furthermore, it is important when applying an SVM to a data set to normalize the data (Eyben et al., 2016).

3.6 Explorative data analysis

In total, 114 features are used for the prediction task in this thesis. Table 5 displays an overview of all features that are used in this research. Appendix II holds a more extensive table that shows the name and type of all features. Mean values and standard deviations of all the continuous variables are disclosed as well. Besides, Appendix III holds distribution plots of all the features that were obtained from the meta data.

Table 5

Features in the data set

Category	Number of features	Features
Acoustic	108	18 Pitch features, 6 Intensity features, and 84 Mel-Frequency Cepstral Coefficient features
Non-acoustic	5	Duration, Hour of the day, Gender, Age, Days till deadline
Target	1	Switched
Total	114	

For the 108 acoustic features and their high-level statistical functions, a similar table is displayed in Appendix II, showing the minimum, median, mean, and maximum values per feature. The most important variation between switching and non-switching customers is discussed.

Table 6

Difference between switching and non-switching customers, for the non-acoustic continuous features

Feature name	Min	Median	Mean	Max
durationInSeconds_switched	30	309.5	507.60	3592
durationInSeconds_not_switched	30	175.0	294.50	3592
age_switched	19	36.0	39.64	91
age_not_switched	19	36.0	39.26	96
daysTillDeadline_switched	0	3.0	10.24	42
daysTillDeadline_not_switched	0	11.0	14.57	42

The first non-acoustic feature to be discussed is the duration of the call, which is measured in seconds. Most of the calls have a duration between one and three minutes, but the average is shifted to more than 6 minutes (380.6 seconds) due to high outliers increasing to a maximum of almost one hour (3592 seconds). An important finding that can be derived from

table 6 is that customers that switched from health insurance have on average longer calls with the company than customers that do not switch. Figure 9 shows that calls over 8 minutes (480 seconds) are more likely to come from switching customers, the outliers are also from that same group of customers which skews their average call duration.

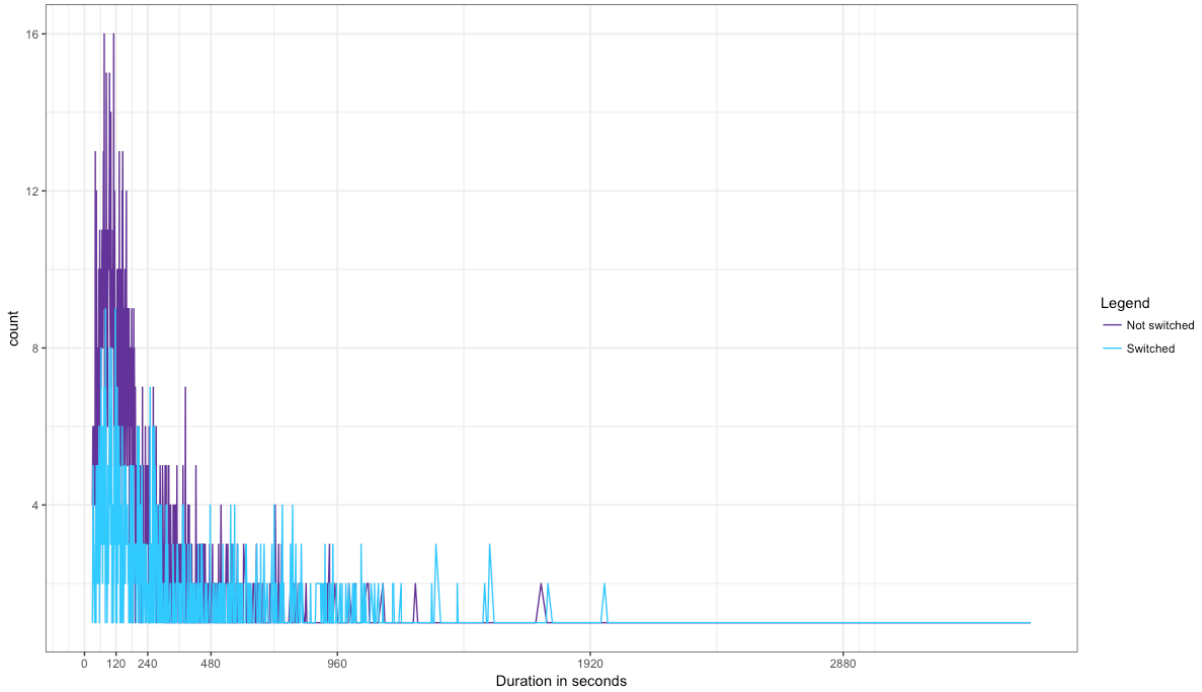

Figure 9, call duration in seconds for both classes of the target feature.

The second non-acoustic feature is the age of the customer, which shows no big diversity between both groups of customers. There are some specific ages where a higher number of

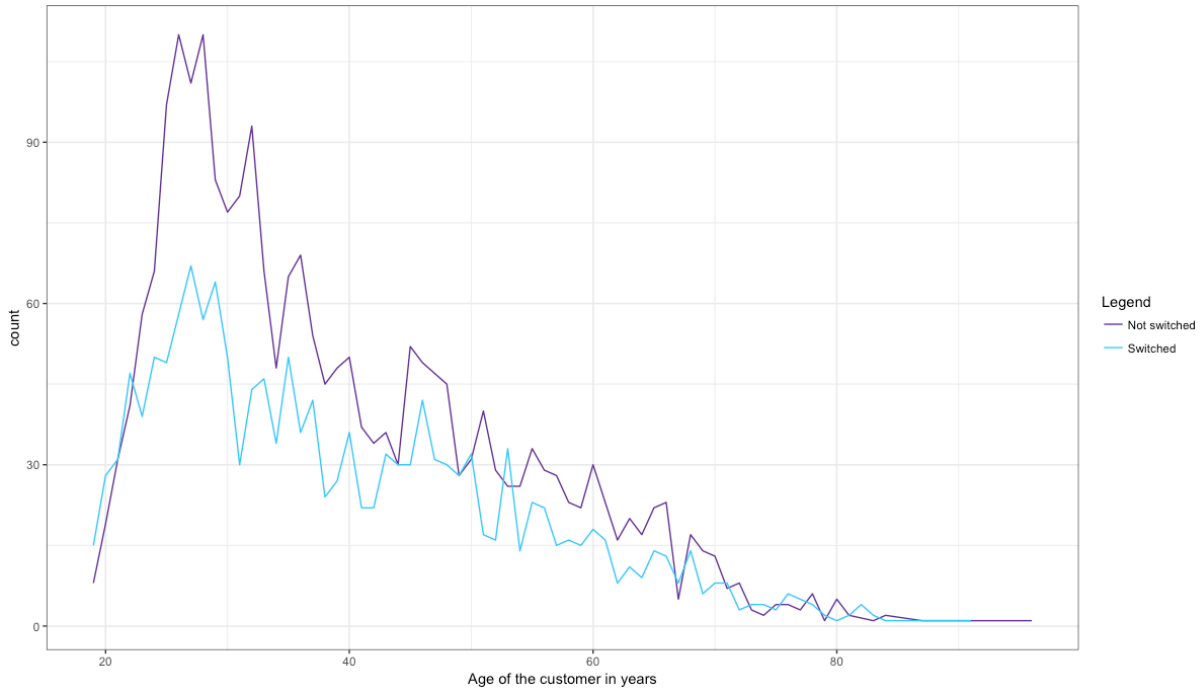

Figure 10, the age of the customer for both classes of the target feature.

customers switched insurer than not (like age 53), but these can be considered as random coincidences. The general trend shows a big peak between 26 and 28, with a gradually decreasing graph to a maximum age of 96. Table 7 shows similar results, with a slightly higher mean age for switching customers with an even lower maximum. However, this is still a small difference.

The number of days till the deadline of the health insurance season is the third feature in this explorative data analysis that is considered. The statistical differences between the two groups of customers (switching and non-switching) reveal that, on average, calls from customers that switched take place closer to the deadline of December 31. Figure 11 exhibits the line graphs for both classes of customers and the combined total line, revealing that most of the calls take place in the last five days before the deadline. From December 29, onward, more calls from switching customers are recorded than vice versa.

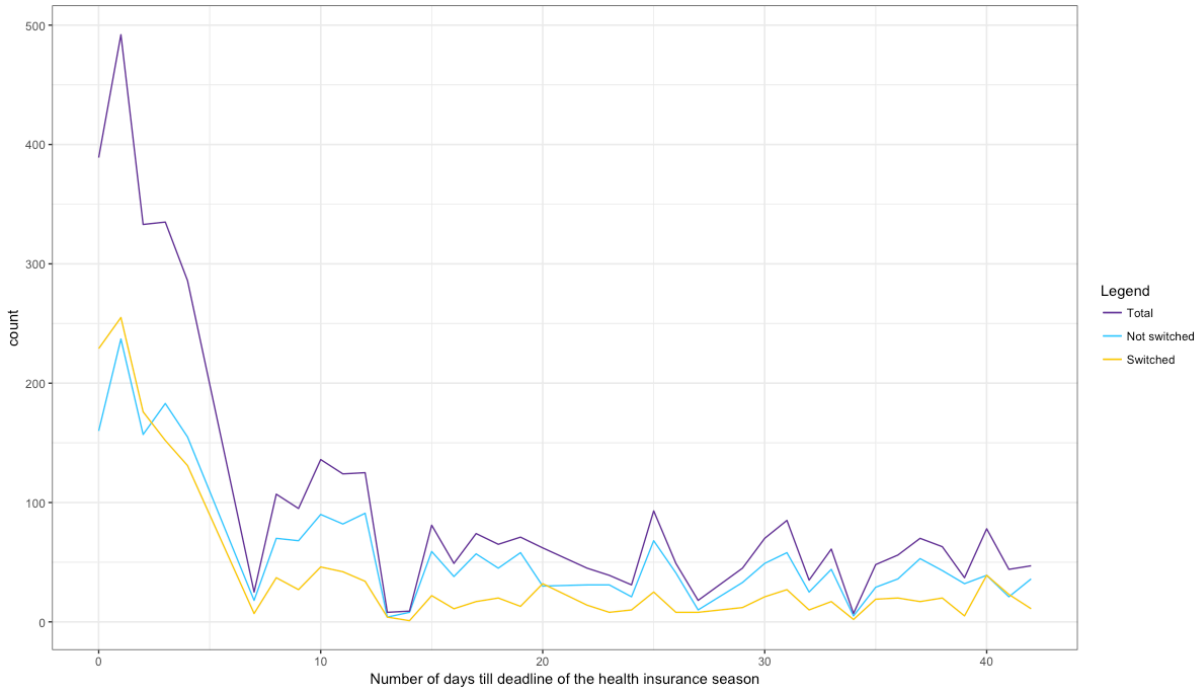

Figure 11, number of days till the health insurance season's deadline per group of customers and total.

Another pattern seems to emerge from the "Total" line in figure 11 as well, it looks like a seven-day week rhythm is recognized. The dates of each call are converted to get insight into the number of calls per day of the week. Figure 12 reveals that, although most calls are made

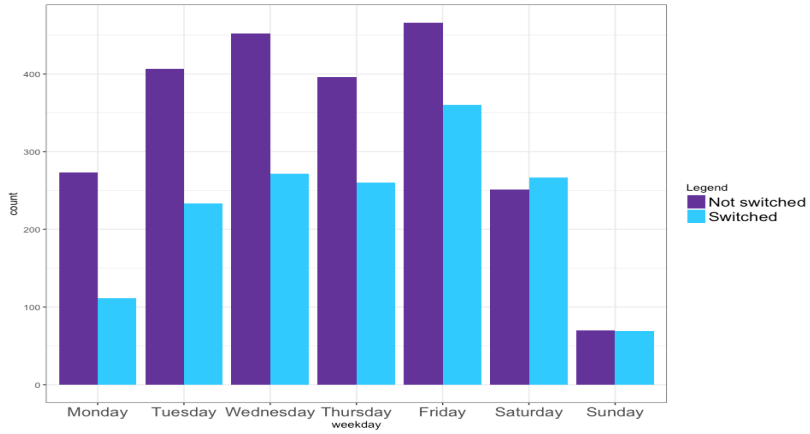

Figure 12, the day of the week and the number of calls for both classes of the target feature

on working days, a bigger share from all the weekend calls comes from customers that switched health insurer. Besides weekend days and the last days before the deadline, also specific hours of the day are more likely to hold calls from customers that switched. Figure 13 shows that between the hours 11:00 and 17:00 relatively many calls are recorded from customers who switched, compared to the hours in the morning and evening (08:00 – 11:00 and 17:00 – 21:00). In the last five days before the deadline on December the 31st, the company extends the call center's opening hours. In these late-night calls, even more customers switched health insurance than not.

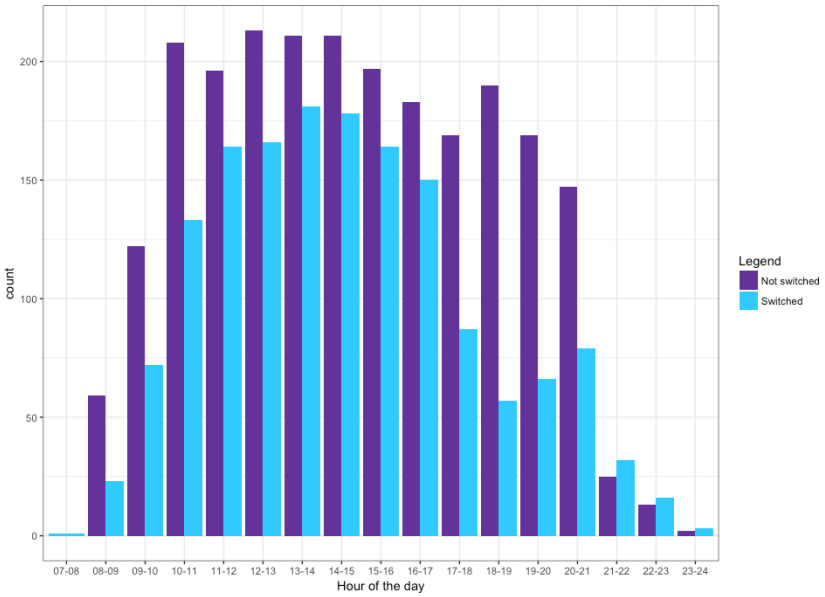

Figure 13, the number of calls per hour of the day for both classes of the target feature.

Another feature about the customer that is considered is the gender. The different switching behavior per gender is visible in figure 14, revealing that although most of the callers are men, a bigger share of all the women switched.

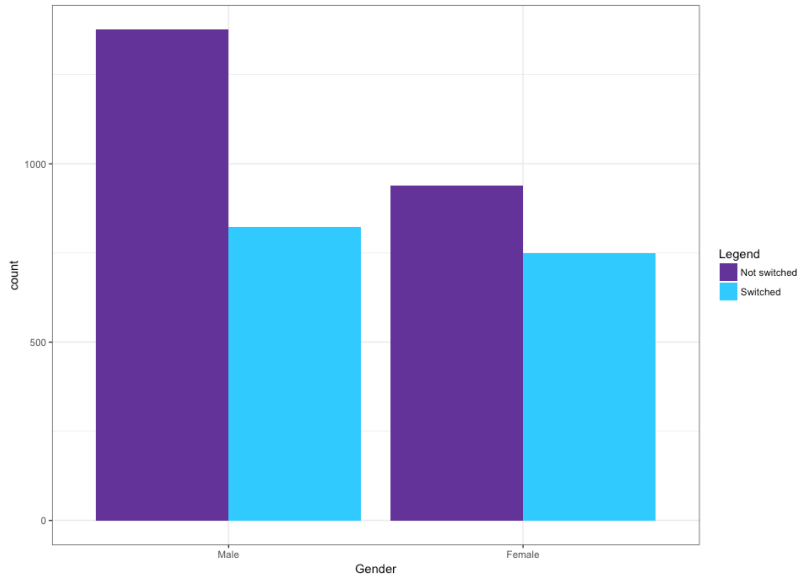

Figure 14, the gender of the customer, split by the number of switched and non-switched callers.

Finally, the target feature shows a split between 40.4% of the customers switching from health insurer and 59.6% that do not switch. The distribution of the target variable is relatively balanced, which is important for the performance of some of the algorithms that are known to suffer from class imbalance (for example: support vector machines).

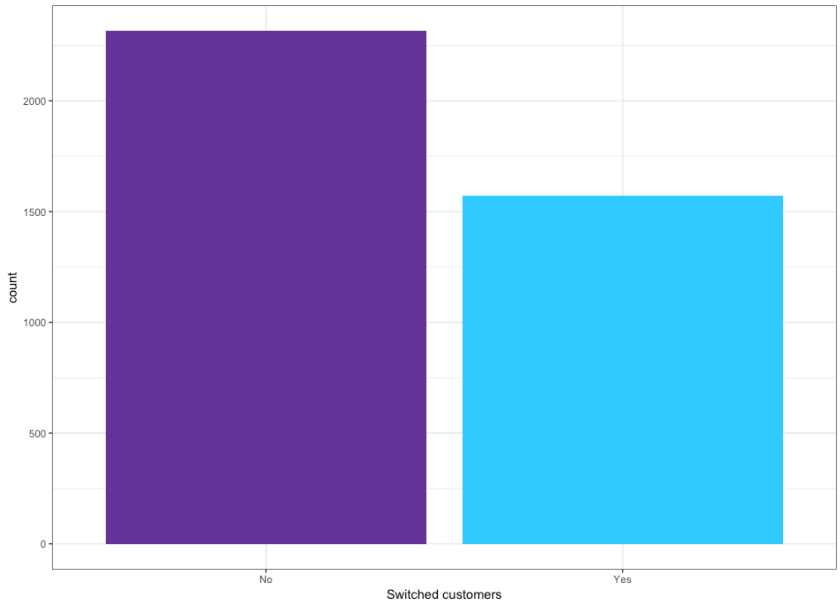

Figure 15, split between the number of switched and not switched customers

The explorative data analysis also provides a better understanding of the acoustic features. The descriptive statistics from the acoustic features are disclosed in Appendix II. They indicate a high average diversity of acoustic values within each thin slice. Examples from that are the relatively high standard deviation for the mean pitch values (30.3) and the high standard deviation on the delta of the intensity (60.8). The high mean range of intensity (18.98) suggests a variation of intensity within the thin slice. The statistics from MFCC values also indicate diversity within the thin slices, because the mean range is rather high for each coefficient (from 3.6 to 12.44). As well, the standard deviation of the deltas for each coefficient is also high, which might point out that calls differ a lot from each other. About the MFCC values can be noted that the mean ranges and the standard deviations of the ranges decrease when the coefficients increase. This suggests that the biggest differences within and between the calls are expressed in the lower MFCC coefficients. As indicated in subsection 3.4.3.3, for robust classification tasks, lower MFCCs might have more predictive value than higher MFCCs.

After assessing the statistics from the acoustic features, the data set is split into two sets. One contains all customers that did switch health insurance and the other data set contains all the customers that did not switch. As the general statistics from the acoustic features already highlighted, there seems to be strong variety within and between the thin slices from the calls. The split of the data into the two classes of the output feature can reveal if the variance in the data is explained by separating switching and non-switching customers or that the diversity of the feature values is more widespread through the data set. Appendix IV holds double boxplots for all the 108 acoustic features, one boxplot for switching customers and one boxplot for non-switching customers. This allows for a direct comparison between the two classes. On top of that, appendix V displays the difference in statistical values for both classes of the target feature. These boxplots and statistical numbers together help to better understand the distinctive capacity of the data to discriminate between acoustic speech characteristics from both types of customers. The findings are discussed per group of acoustic features: pitch features, intensity features and MFCC features.

When looking at the pitch features' boxplots, the first learning is that the mean values for switching and non-switching customers are very close together. The size of the whiskers is rather small and appears to be equal for both plots per feature, indicating low variability. However, there is more difference between the lowest and the highest value for F0 in the calls, which is visible in the longer whiskers form the delta_F0 feature. The boxplots show almost identical plots per pitch feature, the only alteration is in the outliers. These findings are confirmed by table AV.1 in Appendix V.

The boxplots for the intensity features present similar patterns as the pitch features do. The means appear to be identical for both groups of customers, more variability is found in the minima, deltas and ranges of the intensity. Besides, the maximum values for the intensity show very strong outliers to their lowest values. It can be stated there is no clear distinctiveness in the intensity features, based on this data exploration.

Lastly, the boxplots and statistical numbers for the MFCC features are assessed. As for the pitch and intensity features, the mean values are almost identical between the two plots per feature. The numerical values in the table from appendix V reveals that there is some difference, though very small. The MFCC features display rather small variability in their boxplots, except for outliers in their deltas.

4. Experimental procedure

The classification task, which is at the core of this thesis, is performed in five distinct experiments. This section describes how data is split into train and test sets, the experimental procedures within each experiment, and the evaluation criteria that are used to measure the effectiveness of each model.

4.1 Data splitting

The full data set of 3,887 instances is split into an 80% train set (3,110 instances) and a 20% test set (777 instances). While splitting the data randomly in two sets, the target variable's class distribution is kept constant for both data sets (59.6% not switching). Validation scores from the applied models are derived by using 5-Fold cross-validation. In this process, train and validation data is separated five times in an 80% - 20% separation, so that all data points are used for validation exactly once. The error estimation is averaged over all the 5 folds to get the effectiveness score of the model. That output is used to tune applied algorithms and choose the best performing one before applying it to test data.

4.2 Experiments

This thesis consists of five experiments, each experiment using a different feature set. The size of a data set is important because the combination of different features holds predictive value and chances of including the right features increases (Eyben, Batliner, & Schuller, 2010). On the contrary, selecting the right features can be important in speech research. This would best be determined by findings in literature from previous studies within similar contexts. However, since this research is applied in a new setting, using real-life data, there is no best feature set that can be derived from previous research. Even though standard sets of parameters are proposed (e.g. Eyben et al., 2016), they are not inevitably applicable to all other related research settings and the perfect set has not been found yet (Anagnostopoulos, Iliou, & Giannoukos, 2015). For those reasons, no definite feature selection is performed in this research.

Five feature sets are built up in sequential steps, increasing the share of acoustic features in every step. In the first experiment, only features from meta data are used to train a predictive model. As subsection 2.4 highlighted, customer characteristics might already hold predictive power and therefore features from meta data are taken as a first feature set. Pitch and intensity features are added to the feature set in the second experiment. Thereafter the MFCC features are added in the third experiment and algorithms are trained on the full data set. In experiment four and five, the features from meta data are excluded to research the predictive power of models trained on acoustic features only, in line with the approach of more traditional studies

on emotion recognition from acted speech databases (e.g. El Ayadi, Kamel, & Karray, 2011). Experiment four consists of prosodic features only as pitch and intensity are, compared to MFCCs, relatively easy to extract and analyze and are therefore added before MFCC features. Besides this rather practical argument, prosodic features have a strong proven scientific record of classifying speech utterances (Scherer, 2003; Goudbeek & Scherer, 2010). The fifth and final experiment consists of a feature set from all acoustic features and therefore MFCC features are added to test the discriminative potential of acoustic features only. An overview of the features used is provided in table 7.

Table 7

Five experiments and the feature sets used

Experiment	Features used
Experiment 1	5 non-acoustic features
Experiment 2	5 non-acoustic features, 16 pitch features, and 6 intensity features
Experiment 3	5 non-acoustic features, 16 pitch features, 6 intensity features, and 84 MFCC features
Experiment 4	16 pitch features and 6 intensity features
Experiment 5	16 pitch features, 6 intensity features, and 84 MFCC features

Within each experiment, the same methodological setup is executed. Data pre-processing is an important first step in this experimental setup. The large differences in feature ranges can be unfavorable for the performance of predictive models because it results in an unbalanced vector space. Normalizing the features' range is the solution that is used (Guyon & Elisseeff, 2003). Two often used methods are applied in each experiment: z-scores and min-max normalization. Z-scores are a widely recognized method in statistics whereas min-max normalization linearly transforms data by using the minimum and maximum values from a feature to range all values to new values between zero and one (Al Shalabi, Shaaban, & Kasasbeh, 2006).

While normalization provides solutions to the unbalanced vector space, the relatively big number of features might also lead to poor performance. To reduce the dimensional complexity, dimensionality reduction is performed by applying principal component analysis (PCA). This technique looks to summarize the variance of all features in newly created components. Application of PCA is done before cross-validation and after normalization to prevent for overfitting.

The validation scores in each experiment are collected in five distinct steps to find the optimal pre-processing and normalization. The steps are:

1. Raw features without pre-processing
2. Features are normalized using z-scores
3. Features are normalized using min-max normalization
4. Features are normalized using z-scores and PCA is applied
5. Features are normalized using min-max normalization and PCA is applied

Within each experiment, four algorithms are applied to the data set and validation scores reported. The four models used are the naïve Bayes, k-NN, LDA, and linear SVM (as described in section 3.5). All experiments are executed using the caret package in R, the packages that are used for the algorithms are (k-NN is included in the caret package):

- MASS (for LDA)
- naivebayes (for Naïve Bayes)
- kernlab (for SVM)

Algorithms are trained using the standard tuneLength functionality in caret which is kept constant at a value of ten. For k-NN, that only has "k" as tuning parameter, this means that the following values for k are used: 5, 7, 9, 11, 13, 15, 17, 19, 21, 23. The naïve Bayes algorithm has three tuning parameters: "useKernel" which allows for a kernel density estimate instead of Gaussian density estimate for continuous features and is set to true or false, "adjust" which allows to adjust the width of the kernel density that can be set to 0 or 1, and "fl" which allows the use of the Laplance smoother by retrieving 0 or 1. LDA has only "dimen" as a tuning parameter when the MASS package is used which is held at one, meaning that only one linear combination of predictors or discriminant function is created. The SVM classifier is tuned by changing the values for "C" or the cost. The values that are used for the cost are: 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128. These tuning options are kept constant over all experiments and methods of pre-processing.

4.3 Evaluation criteria

Choosing the right evaluation criteria is determined by the structure of the data set, the field of research, and the prediction task at hand. To effectively assess the algorithms' scores, a confusion matrix is considered. All predicted labels, both correct and incorrect, are displayed in the confusion matrix in table 8. True positive labels refer to instances that were correctly classified as positive examples and true negative labels are correctly classified as negative examples. False positive instances are labelled by the classifier as positive but are negative examples. The same logic applies vice versa to the false negative examples.

Table 8

Confusion matrix

	Predicted Yes	Predicted No
Actual Yes	True positive (TP)	False negative (FN)
Actual No	False positive (FP)	True negative (TN)

The performance of a classifier can be represented by various evaluation criteria, each optimizing for different capabilities of the algorithm. For example, the ability to correctly identify classes, reduce failure or the overall ability to discriminate between classes. Currently, classifiers are mainly optimized by making use of accuracy scores. This method makes no difference between correct classification into one of the classes, any correct score is improving the accuracy (Sokolova, Japkowicz & Szpakowicz, 2006). It provides an easily interpretable performance indicator that shows the ability to correctly classify unseen instances. Purpose of the evaluation is to compare algorithms and to research the relevance to a specific field of research (Sokolova et al., 2006). However, the ability to distinguish between classes can be relevant, especially when there is a big class imbalance, costs of misclassification are not known, or when collecting data is difficult and labor intensive. Criteria that distinct between correctly classifying in one of the classes are sensitivity and specificity. The relationship between the correctly labelled positive and correctly labelled negative instances is visualized by the Receiver Operating Characteristic (ROC). The curve consists of many points, each representing a classifier's performance with a specific class distribution. The total area under the curve is taken as the general measurement for an algorithm's performance. Its value holds the predicted probability that any randomly selected positive instance has a higher chance of being labelled as positive than negative. The ROC curve was found to hold more discriminative power than accuracy on real-world data (Huang & Ling, 2005).

An important goal of real-world machine learning applications is increasing profits and decreasing costs. The classification that is optimized using accuracy assumes equal costs for false classified instances, but in real world situations, this is often not the case. On top of that, class distributions are often different between real-world situations and existing benchmark data sets. Comparing accuracy scores from test environments might, therefore, say little about performance on real-world tasks (Provost, Fawcett, & Kohavi, 1998). ROC curves, on the contrary, describe the best performing model independent of misclassification costs and class distributions, making it very suitable for building models for real-world applications.

An example scenario is considered by Huang and Ling (2005), where real-world models are applied to a direct marketing campaign. When models are optimized for accuracy, the profit is determined by the percentage of correctly classified buyers and all predicted buyers should be targeted. However, in many real-world examples, only a top percentage of the predicted buyers is approached. Mean profit is found to be higher for models optimized by ROC than for accuracy in those examples. For that reason, optimizing models using the ROC curve could be more valuable in real-world examples than using accuracy.

Within this thesis, models' performances will be evaluated both on accuracy as on ROC scores. Because the most important purpose of model evaluations is to compare algorithms and to research the relevance to a specific field of research, accuracy is the first score that is looked at (Sokolova et al., 2006). However, ROC can be a strong indicator of performance in practical applications, as shown by the case of Huang and Ling (2005), and is therefore also considered.

5. Results

In this section, training results from cross-validation are displayed and discussed per experiment and thereafter results on the holdout test set are discussed. Before any experiment results, the baseline scores are reported and expressed both as for accuracy and ROC. The baseline is retrieved by using the Zero Rule (ZeroR) classifier, this algorithm simply predicts all instances to be of the majority class. As there are 1852 negative instances in the training set and 1258 positive, the accuracy can be calculated by:

$$\frac{1852}{1852 + 1258} = 0.595$$

The ROC for the ZeroR classifier is of course 0.5 because the sensitivity equals 0 and the specificity equals 1. Hence, the area under the curve (and therefore the ROC with this distribution) is 0.5. The scores from this classifier are used as a baseline to benchmark other algorithms' performances against. A complete table of all training scores from cross-validation is included in Appendix VI. For completeness and transparency, all training scores from cross-validation are discussed per experiment and the results on the holdout test set are discussed thereafter.

5.1 Experiment 1 cross-validation results

In the first experiment, validation scores are collected using the 5 non-acoustic features and in the five experiment steps as explained in section 4.2. Table 9 displays the training results of the first experiment reported in both accuracy and ROC scores.

Table 9

Experiment 1, training results on cross-validation

Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
	accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
none	0.621	0.683	0.645	0.651	0.649	0.673	0.651	0.667
z-scores	0.624	0.683	0.647	0.648	0.657	0.672	0.651	0.672
min-max	0.610	0.685	0.627	0.625	0.651	0.675	0.653	0.671
z-scores + PCA	0.588	0.681	0.649	0.646	0.655	0.674	0.651	0.668
min-max + PCA	0.599	0.659	0.626	0.623	0.657	0.673	0.651	0.671

When training the classifiers, all models outperform the ZeroR baseline on accuracy and ROC. There is a difference though, between reported classifier's performance in accuracy and

in ROC. Performance of the models expressed in accuracy scores has a lower relative performance compared to ROC, especially for naïve Bayes, LDA and SVM. This indicates that these algorithms are better capable of ranking positive instances higher than negative instances, but have more difficulty selecting a threshold to classify instances to a specific class. Naïve Bayes applied on z-scores and PCA seems to be most affected by this, with accuracy performance at 0.588 and ROC at 0.681. The best overall ROC score (ROC = 0.685) is achieved by naïve Bayes with min-max normalization, kernel density estimate was used. The highest accuracy score (accuracy = 0.657) is achieved by the LDA classifier on both min-max normalization with PCA and z-scores, scoring 6.2 percentage points above the baseline. Therefore, LDA trained on z-scores data from this experiment will be the model applied to test data. Little variance between the scores for the different pre-processing methods is found per classifier. PCA was not expected to improve the performance of this small feature set since the number of components is not reduced.

5.2 Experiment 2 cross-validation results

In the second experiment, validation scores are collected using 5 non-acoustic features, 18 pitch features, and 6 intensity features. The same five experimental steps are applied as in the first experiment. Table 10 displays the training results of cross-validation in the second experiment, reported in both accuracy and ROC scores.

Table 10

Experiment 2, training results on cross-validation

Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
	accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
none	0.614	0.632	0.632	0.617	0.658	0.671	0.650	0.662
z-scores	0.608	0.647	0.625	0.602	0.652	0.667	0.650	0.669
min-max	0.610	0.640	0.615	0.602	0.649	0.671	0.648	0.665
z-scores + PCA	0.606	0.639	0.627	0.600	0.654	0.671	0.652	0.668
min-max + PCA	0.609	0.636	0.610	0.603	0.653	0.671	0.649	0.664

All cross-validation results from the classifiers outperform the ZeroR baseline on accuracy and ROC. As in the first experiment the algorithms' scores expressed in ROC display a higher relative performance than accuracy scores. This indicates that the models are better in making the split between the classes than assigning the right class to an unseen instance.

Dimensionality reduction with PCA resulted in a reduction from 27 continuous features to 18 components, only improving the performance of LDA and SVM. Best training performances on accuracy are reported when applying LDA and SVM, with the highest peak by LDA using none pre-processed data (accuracy = 0.658). A similar for ROC scores, where LDA and SVM outperform the rest. LDA is even just a bit higher on all data sets (ROC = 0.671), except for the one with z-scores. In this experiment, there is no general best choice in pre-processing. The overall best performing model is LDA applied on the none pre-processed data, which might indicate that the algorithms do not suffer from the unbalanced vector space. However, this model might overfit by using the differences in vector spaces improperly to its advantage. Nevertheless, based on the highest cross-validation scores in ROC and accuracy the LDA model trained on none pre-processed data will be applied to the holdout test set.

5.3 Experiment 3 cross-validation results

In experiment number three, cross-validation scores are collected using the full feature set of 113 acoustic and non-acoustic features. The same five experimental steps are applied as in the previous experiments. Table 11 displays the training results of the third experiment, reported in both accuracy and ROC scores.

Table 11

Experiment 3, training results on cross-validation

Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
	accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
none	0.602	0.632	0.629	0.608	0.642	0.648	0.641	0.655
z-scores	0.616	0.631	0.597	0.571	0.634	0.656	0.647	0.650
min-max	0.611	0.640	0.608	0.607	0.634	0.655	0.641	0.658
z-scores + PCA	0.612	0.635	0.601	0.568	0.650	0.655	0.655	0.665
min-max + PCA	0.615	0.631	0.619	0.619	0.654	0.658	0.649	0.660

Cross-validation scores from all the classifiers outperform the ZeroR baseline on accuracy and ROC and as in the first two experiments, the algorithm's scores expressed in ROC display a higher relative performance than accuracy scores. LDA and SVM models clearly outperform the other two classifiers in this experiment both on accuracy and ROC. Small improvement of the scores from LDA and SVM models can be seen when PCA is applied. PCA reduces the dimensionality from 111 continuous features to 60 components. Normalization of

the data is beneficial for LDA and SVM scores, but the difference between min-max and z-scores is very small. Naïve Bayes and k-NN might suffer from the high dimensionality, although this is only known from k-NN. On the contrary, cross-validation performances do not increase after applying PCA. Naïve Bayes is not expected to suffer from the high dimensionality due to the assumed independence between the features. The best validation scores from the third experiment are achieved with SVM on the data set with set z-scores and PCA applied (accuracy = 0.655 and ROC = 0.665), the cost was tuned to 0.5 in the best performing model. This model will, therefore, be applied to the test set.

5.4 Experiment 4 cross-validation results

In the fourth experiment, cross-validation scores are collected using 18 pitch features and 6 intensity features. The same five experimental steps are applied as in the previous experiments. Table 12 displays the training results from cross-validation of the fourth experiment, reported in both accuracy and ROC scores.

Table 12

Experiment 4, training results on cross-validation

Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
	accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
none	0.569	0.512	0.559	0.499	0.595	0.479	0.595	0.515
z-scores	0.559	0.512	0.555	0.510	0.594	0.485	0.595	0.523
min-max	0.557	0.510	0.557	0.514	0.595	0.502	0.595	0.514
z-scores + PCA	0.588	0.498	0.563	0.509	0.595	0.493	0.595	0.514
min-max + PCA	0.583	0.497	0.550	0.501	0.596	0.489	0.595	0.516

In contrast to the first three experiments, the cross-validation scores from models applied in the fourth experiment do not outperform the ZeroR baseline when it comes to accuracy and only exceeds the 0.5 baseline for ROC by a small margin in few instances. LDA and SVM models match baseline performances on accuracy and only SVM scores above the 0.5 baseline with ROC. LDA and SVM probably have their decision boundary on the edge of the feature space and therefore assigning same classes to all instances, which matches the ZeroR performance. It is not expected that any model from the fourth experiment will outperform the baseline when applied to the test data. Therefore, no trained models from the fourth experiment will be applied to the test data set.

5.5 Experiment 5 cross-validation results

In the fifth experiment, cross-validation scores are collected using 18 pitch features, 6 intensity features, and 84 MFCC features. The same five experimental steps are applied as in the previous experiments. Table 13 displays the training results from cross-validation of the fifth experiment, reported in both accuracy and ROC scores.

Table 13

Experiment 5, training results on cross-validation

Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
	accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
none	0.547	0.552	0.558	0.491	0.568	0.518	0.595	0.526
z-scores	0.547	0.544	0.575	0.516	0.570	0.535	0.595	0.519
min-max	0.543	0.547	0.564	0.531	0.564	0.524	0.595	0.532
z-scores + PCA	0.565	0.540	0.562	0.533	0.581	0.532	0.595	0.540
min-max + PCA	0.562	0.518	0.571	0.536	0.570	0.534	0.595	0.535

The cross-validation performance of the models trained in the fifth experiment does not outperform the ZeroR baseline in accuracy scores. Only the SVM classifier match the accuracy scores of the baseline, probably because all instances are assigned to the same class and no real separating hyperplane is found. ROC validation scores in the fifth experiment turn out to be just over the 0.5 baseline in some cases, but that can be due to some lucky shots. It is not expected that any model from the fifth experiment will outperform the baseline when applied to the test data. Therefore, no trained models from the fifth experiment will be applied to the test data set.

When all experiments are compared, LDA applied on the data set from the second experiment without pre-processing has displayed the highest performance on accuracy (0.658). The best ROC score is the achieved by Naïve Bayes (0.685), applied on min-max normalized data of the first experiment. LDA has a generally strong performance throughout the experiments, whereas k-NN and Naïve Bayes perform best on a small data set. SVM relatively has the best cross-validation performance with a high dimensional data set, as in experiment 3. SVM benefits from dimensionality reduction with PCA though, as does LDA. The effect of PCA is less eminent in experiment 2. The cross-validation scores do not reveal important performance differences between normalization with z-scores or min-max. Two of the best

performing models make use of z-score normalization, but the difference with min-max scores are very small. From the first three experiments, the best model per experiment is applied to the holdout test data.

5.6 Test set results

The best scoring model from each experiment is applied to the hold out test set if the performance is at least above the ZeroR baseline (59.5% accuracy and 0.5 ROC). This means that from each of the first three experiments, one model is applied on the test set with the relevant features. From experiment 1, the LDA classifier is taken and it is applied to the normalized test set with z-scores. The LDA classifier is also used from the second experiment, but then applied to the test without pre-processing. From experiment 3, the SVM classifier is applied to the normalized test set. The same dimensionality reduction using PCA is applied on this test set as in experiment 3. Test set results in both accuracy and ROC are displayed in table 14.

Table 14

Test set results

Experiment	Pre-processing	model	accuracy	ROC
1	z-scores	LDA	0.646	0.683
2	none	LDA	0.662	0.679
3	z-scores + PCA	SVM	0.634	0.652

The highest accuracy score is achieved by the Linear Discriminant Analysis applied to the test set that contains the features from the second experiment, without any form of pre-processing (accuracy = 0.662). The performance of this model outperforms the ZeroR baseline by 6.6 percentage points, which is 1.6 percentage points better than the next best performing model on accuracy scores (LDA on meta features in z-scores). Performance of the SVM classifier on the full feature set after PCA and z-score normalization is also above the ZeroR baseline, but only by 3.8 percentage points. The models' scores expressed in ROC display a slightly different picture, namely the highest score is achieved by the smallest feature set (experiment 1) and the lowest score is achieved by the full feature set from experiment 3. All ROC scores outperform the baseline of 0.5, with the highest score achieved by LDA from experiment 1 (ROC = 0.683).

6. Discussion and conclusion

In this section, an answer to the problem statement is given. Thereafter, the research objectives are addressed in relation to the experimental results, limitations of models and data are explained, the contribution of this thesis to the existing body of literature is discussed, and recommendations for future research are given.

6.1 Answer to the problem statement

The aim of this thesis was to predict consumers' switching behavior regarding health insurance by analyzing the acoustics of the consumer's voice. The formulated problem statement of this thesis is: *To what extent can acoustic features from call center speech successfully predict switching behavior of health insurer?*

Voice recordings from call center conversations have been used in this thesis to execute five experiments, each looking to predict consumers' switching behavior on health insurer with different feature sets. It is demonstrated that models with a mix of acoustic and non-acoustic features could outperform a random guessing baseline. Pitch and intensity features improved the performance of a basic model consisting of only five non-acoustic features, resulting in the best performing model (6.6% better than a random guess). This highlights the potential of acoustic features as predictors for future behavior. However, the addition of MFCC features and the performance of models with only acoustic features indicate that the findings of this thesis are too prelaminar to state that acoustic features from call center speech can successfully predict switching behavior of health insurer.

The ability to predict future behavior by vocal cues from call center speech has proven to be a challenging task. Promising results from human judges in predicting future behavior from acoustics in speech (Rogers et al., 2016) are partially reproduced with machine learning technique in this thesis. Results of the five experiments have demonstrated that models trained on acoustic features only are not able to predict future switching behavior. However, the predictive ability of voice acoustics is underlined by the fact that it managed to improve the performance of a model trained on non-acoustic features. With that, the approach from previous research in automated emotion recognition on acted speech databases (e.g. El Ayadi et al., 2011) is linked to new research applications, opening interesting opportunities for future research.

6.2 Discussion of research objectives

Research objective 1: *To identify the extent to which non-acoustic features from call center speech can successfully predict switching behavior of health insurer.*

Experiment 1 is executed to address this research objective and based on the experiment's results it is stated that non-acoustic features from call center speech can successfully predict switching behavior of health insurer, but only by a few percents. LDA with z-score normalization is most successful on this specific feature set. LDA trained on the call's duration, the hour of the day, the customer's gender, the customer's age, and the number of days till the health season's deadline managed to outperform a random guessing baseline by 5%. Z-score normalization was used on this small data set, which deals with the unbalanced vector space. The performance of the model expressed in ROC is relatively higher to the accuracy, which indicates that the model does a better job in ranking all instances than choosing the right cutoff point for class separation. Because the random guessing baseline is only outperformed by a few percents, this result is interpreted as a preliminary outcome.

Even though the implementation of LDA is not too complex, from a practical standpoint it is advised to see this result as a begin of more experiments with non-acoustic feature sets for future behavior predictions of customers. The results from experiment 1 can be explained in relation to previous research as it found support in the predictive value of personal information as a feature in studying switching behavior on health insurers. The age of an individual is an important indicator for one's tendency to switch as found by Boonen, Laske-Aldershof, & Schut, 2016 and De Jong, Van den Brink-Muinen, & Groenewegen, 2008. Besides, gender is an important predictor of switching behavior as well, as men and women are found to be different in their likelihood to switch from health insurer (Lako, Rosenau, & Daw, 2011; Hendriks et al., 2010). Some studies even found gender to be the most predictive factor for switching behavior on health insurance (Rademakers et al., 2014). Notwithstanding the demonstrated performance, the construction of this small data set and its application to predicting future switching behavior is a preliminary finding.

Research objective 2: *To establish the effect of adding pitch and intensity from call center speech to models using non-acoustic features on the performance of predicting switching behavior.*

Experiment 2 is executed to address this research objective and based on the experiment's results it is stated that the addition of pitch and intensity features from call center speech results in an improvement of predicting switching behavior with non-acoustic features. Adding pitch and intensity results in a lower ability to split both customer groups (indicated by

lower ROC), but improves the ability to assign the right label to a new customer (indicated by higher accuracy). Linear discriminant analysis is most powerful, applied on a data set without normalization or dimensionality reduction. As this model performs only 6.6% above the random guessing baseline, the findings should be interpreted as preliminary. The potential of acoustic features in predicting future behavior is proven but might need more research to endorse and expand.

The results from experiment 2 can be explained in relation to previous research. Converting the intention to switch from health insurer into actual switching behavior is affected by the perceived behavioral control of a customer (Ajzen, 2002). Stressful customers would not be expected to experience complete control if the stress is not caused by any external factor. Pitch and intensity are proven indicators for stress in a customer's voice (Ververidis & Kotropoulos, 2006; Eyben et al., 2016). Although these direct relations might need affirmation in future research.

Research objective 3: *To establish the effect of adding MFCCs from call center speech to models using non-acoustic features, pitch and intensity on the performance of predicting switching behavior.*

Experiment 3 is executed to address this research objective and based on the experiment's results it is stated that the addition of MFCC features from call center speech, as extracted in this thesis, results in decreased performance of predicting switching behavior with non-acoustic features, pitch features and intensity features. Adding MFCC features to the second experiment's mixed model decreased the predictive performance. This data set of 113 independent features still outperforms a random guess by 3.8%, but the lower performance than the first two experiments indicates that MFCC features, as used in this research, deteriorate the ability to predict a customer's switching behavior of health insurers. The support vector machine performs best on this high dimensional data set and benefits from z-score normalization and dimensionality reduction with PCA.

The results from experiment 3 can be related to previous research. The chosen method of calculating HSFs from MFCCs is in line with the work of Vaudable and Devillers (2012) as they also applied it on real-life call center data. However, they make use of an intermediate step of predicting negative, positive and neutral labels per part of the conversation and classify afterwards based on these labelled subparts. With their approach, they do not manage to automatically detect positive instances which again is found difficult with these features in this thesis. Furthermore, MFCCs are described as textual depend before (Schuller et al., 2011), which might have had an impact on the performance of MFCCs in this thesis as well. Since

textual analysis is not part of the experiments, it is not possible to judge the potential interference. More general, the addition of MFCC features adds a lot of dimensionality to the total data set which might not be beneficial in real-life situations (Eyben et al., 2016).

Research objective 4: *To identify the extent to which pitch and intensity from call center speech can successfully predict switching behavior of health insurer.*

Experiment 4 is executed to address this research objective and based on the experiment's results it is stated that pitch and intensity, as extracted in this thesis, do not predict switching behavior of health insurer better than a random guessing experiment. Cross-validation scores from models trained on high statistical functions from the pitch and intensity of call center speech never exceeded the baseline and therefore no models from the fourth experiment are applied to the test set.

Research objective 5: *To establish the effect of adding MFCCs from call center speech to models using pitch and intensity on the performance of predicting switching behavior.*

Experiment 5 is executed to address this research objective and based on the experiment's results it is stated that MFCCs, as extracted in this thesis, do not add to the predictive power of models trained on pitch and intensity features to predict switching behavior of health insurer. Cross-validation scores from models trained on high statistical functions from MFCCs of call center speech do not outperform a random guessing baseline and therefore no models from the fifth experiment are applied to the test set.

6.3 Research limitations

This thesis has presented promising findings, but they should be interpreted in relation to limitations that belong to this research. First, the data from the recordings are recorded on one track, which makes it not possible to split between the voice of the customer and the company's agent. This could add noise to the data and affecting the performance of the extracted features. Although, interspeaker influence in call center conversations, described in subsection 2.3, explains that vocal signals are copied between persons and convergence of speech is likely to occur. This will limit the effect of noise that is caused by the way of recording because the two voices are expected to exhibit similar patterns. Other sources of noise in the signal might be caused by external factors from the surroundings of the customer.

Furthermore, in this thesis, the assumption is made that the most predictive power lies within the conclusion of the conversation. On top of that, the conclusion is automatically extracted from every recording. Computational power restricts the amount of information that can be processed and because acoustic features are extracted per 10 or 1 millisecond(s), it does not allow analysis of the full conversation's length. The extraction of thousands of data points

from every conversation would also add enormously to the dimensionality of the data set. Thin slices are therefore taken. Nevertheless, a very structured approach is chosen by taking random samples from the data set to manually test the assumptions. This restricts the potential loss of predictive value from the conversations.

Limitations regarding the extraction of features is tried to keep minimal by making use of the COVAREP toolbox and Pandit's widely used method for MFCC extraction (2015). That also adds to the comparability of the findings in the field of research. However, the computation of the high statistical functions is still very different between studies. Especially when it comes to MFCCs. The approach of using deltas and delta-delta features often taken in contrast to the approach of this thesis. Still, the support for Vaudable and Devillers' alternative method (2012) is not completely contradicted as the delta-delta features are used to capture the uniqueness of the voice for example in speech recognition (e.g. Cutajar, et al., 2013).

For the experimental setup in this thesis, a basic approach is chosen to support the progressive and explorative nature of this research. As all five experiments are separate studies, it is likely that more tailored pre-processing and model tuning will benefit the performance of the models. As argued for example by Ramakrishan (2012), MFCCs would benefit from cepstral mean normalization instead of z-scores or min/max normalization. Still, the experimental setup from this thesis already achieves promising results and suits the preliminary character.

6.4 Recommendations and directions for future research

This thesis serves as a step towards better understanding of customer behavior and is exploratory in the field of processing acoustics from real-life speech recordings. Directions for future research are multiple. The first direction of future research should aim at reducing noise in the data. The potential of acoustic features is shown in this thesis, but improved methods of speech recording and the ability to isolate the customer's voice are expected to improve the performance of acoustic models. On top of that, noise reduction algorithms could be applied to the same data set to improve the models' performance.

The second direction of future research concerns the experimental setup, the application of pre-processing and the tuning of algorithms. More methods of pre-processing can be tested, for example, grid search techniques to achieve more fine-tuned algorithms. New algorithms can be tested on the data as well, like neural networks and hidden Markov models. As well as feature selection methods to achieve more understanding of the performance of the current feature set.

From a practical standpoint, it is recommended to expand on the predictions with non-acoustic feature sets as well. If 5 non-acoustic features already outperform a random guessing

baseline, there is even more potential in larger feature sets. The practical advantage of this approach lies within the relative ease of implementation and feature extraction and can, therefore, be the low hanging fruit for companies that are looking to improve their call center analytics.

References

Ajzen, I. (2002). Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. Journal of applied social psychology, 32(4), 665-683.

Al Shalabi, L., Shaaban, Z., & Kasasbeh, B. (2006). Data mining: A preprocessing engine. Journal of Computer Science, 2(9), 735-739.

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. Advances in experimental social psychology, 32, 201-271.

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin-slice judgments to evaluate sales effectiveness. Journal of Consumer Psychology, 16(1), 4-13.

Anagnostopoulos, C. N., & Iliou, T. (2010). Towards emotion recognition from speech: definition, problems and the materials of research. Semantics in Adaptive and Personalized Services, 127-143.

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review, 43(2), 155-177.

Atassi, H., & Smékal, Z. (2014, November). Automatic identification of successful phone calls in call centers based on dialogue analysis. In Cognitive Infocommunications (CogInfoCom), 2014 5th IEEE Conference on (pp. 425-429). IEEE.

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological science, 16(10), 814-819.

Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18, 1-8.

Barger, P. B., & Grandey, A. A. (2006). Service with a smile and encounter satisfaction: Emotional contagion and appraisal mechanisms. Academy of management journal, 49(6), 1229-1238.

Bänziger, T., Mortillaro, M., & Scherer, K. R. (2012). Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion, 12(5), 1161.

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. Journal of personality and social psychology, 54(2), 243.

Bogert, B. P. (1963). The quefrency alanysis of time series for echoes: Cepstrum pseudo-autocovariance, cross-cepstrum, and saphe cracking. Time series analysis, 209-243.

Boonen, L. H., Laske-Aldershof, T., & Schut, F. T. (2016). Switching health insurers: the role of price, quality and consumer information search. The European Journal of Health Economics, 17, 339.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L. (2005). Statistical analysis of a telephone call center: A queueing-science perspective. Journal of the American statistical association, 100(469), 36-50.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of german emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

Busso, C., Lee, S., & Narayanan, S. (2009). Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on audio, speech, and language processing, 17(4), 582-596.

Cappella, J. N., & Planalp, S. (1981). Talk and silence sequences in informal conversations III: Interspeaker influence. Human Communication Research, 7(2), 117-132.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech Communication, 71, 10-49.

Cutajar, M., Gatt, E., Grech, I., Casha, O., & Micallef, J. (2013). Comparative study of automatic speech recognition techniques. IET Signal Processing, 7(1), 25-46.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, 28(4), 357-366.

De Jong, J. D., Van den Brink-Muinen, A., & Groenewegen, P. P. (2008). The Dutch health insurance reform: switching between insurers, a comparison between the general population and the chronically ill and disabled. BMC Health Services Research, 8(1), 58.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014, May). COVAREP—A collaborative voice analysis repository for speech technologies. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on (pp. 960-964). IEEE.

Delcourt, C., Gremler, D. D., Van Riel, A. C., & Van Birgelen, M. (2013). Effects of perceived employee emotional competence on customer satisfaction and loyalty: The mediating role of rapport. Journal of Service Management, 24(1), 5-24.

Devillers, L., Vaudable, C., & Chastagnol, C. (2010). Real-life emotion-related states detection in call centers: a cross-corpora study. In Eleventh Annual Conference of the International Speech Communication Association.

Duijmelinck, D. M., Mosca, I., & van de Ven, W. P. (2015). Switching benefits and costs in competitive health insurance markets: A conceptual framework and empirical evidence from the Netherlands. Health policy, 119(5), 664-671.

Duijmelinck, D. M., & van de Ven, W. P. (2016). Switching rates in health insurance markets decrease with age: empirical evidence and policy implications from the Netherlands. Health Economics, Policy and Law, 11(2), 141-159.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572-587.

Erickson, D., Menezes, C., & Fujino, A. (2004, October). Some articulatory measurements of real sadness. In INTERSPEECH.

Eyben, F., Batliner, A., & Schuller, B. (2010, April). Towards a standard set of acoustic features for the processing of emotion in speech. In Proceedings of Meetings on Acoustics 159ASA (Vol. 9, No. 1, p. 060006). ASA.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.

Fant, G. Speech Sounds and Features. The MIT Press, Cambridge, MA, USA, 1973.

Ganchev, T., Fakotakis, N., & Kokkinakis, G. (2005, October). Comparative evaluation of various MFCC implementations on the speaker verification task. In Proceedings of the SPECOM (Vol. 1, pp. 191-194).

Giles, H. (1973). Accent mobility: A model and some data. Anthropological linguistics, 87-105.

Giles, H., Coupland, N., & Coupland, I. U. S. T. I. N. E. (1991). 1. Accommodation theory: Communication, context, and. Contexts of accommodation: Developments in applied sociolinguistics, 1.

Goldberg, L. S., & Grandey, A. A. (2007). Display rules versus display autonomy: emotion regulation, emotional exhaustion, and task performance in a call center simulation. Journal of occupational health psychology, 12(3), 301.

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. Communication Research, 37(1), 3-19.

Goswami, S., Dutta, D., Deka, P., Sarma, D., & Bardoloi, B. (2013). ZCR Based Identification of Voiced, Unvoiced and Silent Parts of Speech Signal in Presence of Background Noise. In *International Conference on Computation and Communication Advancement (IC3A)* (p. 135).

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. The Journal of the Acoustical Society of America, 128(3), 1322-1336.

Grimm, M., Kroschel, K., & Narayanan, S. (2008, June). The Vera am Mittag German audio-visual emotional speech database. In Multimedia and Expo, 2008 IEEE International Conference on (pp. 865-868). IEEE.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

Han, J., & Urmie, J. (2017). Medicare Part D Beneficiaries' Plan Switching Decisions and Information Processing. Medical Care Research and Review, 1077558717692883.

Harris, F. J. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE, 66(1), 51-83.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. Current directions in psychological science, 2(3), 96-100.

Hatfield, E., Hsee, C. K., Costello, J.,Weisman, M. S., & Denney, C. (1995). The impact of vocal feedback on emotional experience and expression. Journal of Social Behavior and Personality, 10, 293–312.

Health insurance act 2005. (2005, June 16th). Accessed at July 2nd, 2017, from http://maxius.nl/zorgverzekeringswet/artikel17/lid7

Hendriks, M., De Jong, J. D., Den Brink-Muinen, V., & Groenewegen, P. P. (2010). The intention to switch health insurer and actual switching behaviour: are there differences between groups of people?. Health Expectations, 13(2), 195-207.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310.

Jiang, X., & Pell, M. D. (2015). On how the brain decodes vocal cues about speaker confidence. cortex, 66, 9-34.

Kim, J., Lee, S., & Narayanan, S. S. (2010). A study of interplay between articulatory movement and prosodic characteristics in emotional speech production. In Eleventh Annual Conference of the International Speech Communication Association.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International journal of speech technology, 15(2), 99-117.

Lako, C. J., Rosenau, P., & Daw, C. (2011). Switching health insurance plans: results from a health survey. Health Care Analysis, 19(4), 312-328.

Lee, C.C., Kim, J., Metallinou, A., Busso, C., Lee, S., & Narayanan, S.S. (2014). Speech in affective computing. In R. Calvo and S. D'Mello and J. Gratch and A. Kappas (Ed.), The Oxford Handbook of Affective Computing (pp. 170-183). New York, NY, USA: Oxford University press.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Little, L. M., Kluemper, D., Nelson, D. L., & Ward, A. (2013). More than happy to help? Customer-focused emotion management strategies. Personnel Psychology, 66(1), 261-286.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on(pp. 2227-2231). IEEE.

Mishne, G., Carmel, D., Hoory, R., Roytman, A., & Soffer, A. (2005, October). Automatic analysis of call-center conversations. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 453-459). ACM.

Mporas, I., Ganchev, T., Siafarikas, M., & Fakotakis, N. (2007). Comparison of speech features on the speech recognition task. Journal of Computer Science, 3(8), 608-616.

Mubarak, O. M., Ambikairajah, E., & Epps, J. (2005, August). Analysis of an mfcc-based audio indexing system for efficient coding of multimedia sources. In Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on (Vol. 2, pp. 619-622). IEEE.

Nederlandse Zorgautoriteit. (2016). Markscan zorgverzekeringsmarkt 2016. Received from https://www.nza.nl/1048076/1048181/Marktscan_Zorgverzekeringsmarkt_2016.pdf

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. In Ninth International Conference on Spoken Language Processing.

Pandharipande, M. A., & Kopparapu, S. K. (2012, May). A novel approach to identify problematic call center conversations. In Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on (pp. 1-5). IEEE.

Pandit, P. (2015). Speaker Diarization of Broadcast News Audio. (Dual Degree Dissertation, Department of Electrical Engineering, Indian Institute of Technology Bombay, India). Retrived from https://www.slideshare.net/ParthePandit/parthepandit10d070009ddpthesis-53767213

Pereira, M. J., Coheur, L., Fialho, P., & Ribeiro, R. (2016). Chatbots' Greetings to Human-Computer Communication. arXiv preprint arXiv:1609.06479.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. Proceedings of the IEEE, 81(9), 1215-1247.

Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., ... & Lamel, L. (2016, October). Multimodal Emotion Recognition for AVEC 2016 Challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (pp. 75-82). ACM.

Praksah, C., & Gaikwad, V. B. (2015). Analysis Of Emotion Recognition System Through Speech Signal Using KNN, GMM & SVM Classifier. IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), 10(2), 55-67.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998, July). The case against accuracy estimation for comparing induction algorithms. In ICML (Vol. 98, pp. 445-453).

Pugh, S. D. (2001). Service with a smile: Emotional contagion in the service encounter. Academy of management journal, 44(5), 1018-1027.

Rademakers, J., Nijman, J., Brabers, A. E., de Jong, J. D., & Hendriks, M. (2014). The relative effect of health literacy and patient activation on provider choice in the Netherlands. Health Policy, 114(2), 200-206.

Ramakrishnan, S. (2012). Recognition of emotion from speech: A review. In Speech Enhancement, Modeling and Recognition-Algorithms and Applications. InTech.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). New York: IBM.

Rogers, T., ten Brinke, L., & Carney, D. R. (2016). Unacquainted callers can predict which citizens will vote over and above citizens' stated self-predictions. Proceedings of the National Academy of Sciences, 113(23), 6449-6453.

Romp, M. & Merckx, P. (2017). Verzekerden in beeld (Jaargang 22, april 2017). Retrieved from: chrome-extension://oemmndcbldboiebfnladdacbdfmadadm/https://www.vektis.nl /images/Verzekerden_in_beeld_2017.pdf

Rueff-Lopes, R., Navarro, J., Caetano, A., & Silva, A. J. (2015). A Markov Chain Analysis of Emotional Exchange in Voice-to-Voice Communication: Testing for the Mimicry Hypothesis of Emotional Contagion. Human Communication Research, 41(3), 412-434.

Rust, R. T., & Huang, M. H. (2014). The service revolution and the transformation of marketing science. Marketing Science, 33(2), 206-221.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1), 227-256.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9), 1062-1087.

Schuller, B., Steidl, S., & Batliner, A. (2009). The interspeech 2009 emotion challenge. In Tenth Annual Conference of the International Speech Communirepcation Association.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In Eleventh Annual Conference of the International Speech Communication Association.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. Computer Speech & Language, 27(1), 4-39.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

Soto Sanfiel, M. T. (2008). Efecto del tono de voz y de la percepción del rostro en la formación de impresiones sobre los hablantes mediáticos. Comunicación y sociedad, (10), 129-161.

Story, B. H. (2002). An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, *23*(4), 195-206.

Sun, X., Truong, K. P., Pantic, M., & Nijholt, A. (2011, October). Towards visual and vocal mimicry recognition in human-human interactions. In Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on (pp. 367-373). IEEE.

Van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004). Mimicry and prosocial behavior. Psychological science, 15(1), 71-74.

Van Beest, F., Lako, C., & Sent, E. M. (2012). Health insurance and switching behavior: Evidence from the Netherlands. Health, 4(10), 811.

Vaudable, C., & Devillers, L. (2012, March). Negative emotions detection as an indicator of dialogs quality in call centers. In Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on (pp. 5109-5112). IEEE.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), 1162-1181.

Webb, J. T. (1972). Interview synchrony: An investigation of two speech rate measures in an automated standardized interview. Studies in dyadic communication New York.

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. Frontiers in psychology, 4.

Wicks, M.A.: 'The mel frequency scale and coefficients'. 1998. Available from: http://kom.aau.dk/group/04gr742/pdf/MFCC_worksheet.pdf

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Lee, S., ... & Busso, C. (2004, October). An acoustic study of emotions expressed in speech. In INTERSPEECH.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ... & Valtchev, V. (2002). The HTK book. Cambridge university engineering department, 3, 175.

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. Journal of Computer Science and Technology, 16(6), 582-589.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. Advances in experimental social psychology, 14, 1-59.

Zwicker, E., & Fastl, H. (2013). Psychoacoustics: Facts and models (Vol. 22). Springer Science & Business Media.

Appendix I: Results of explorative call analysis.

Table AI.1

Proportion of speech in the conclusion

Sample	N	Mean % agent	Mean % customer	Standard deviation
Calls in sample	100	58.5	41.5	32.5
Conclusion	82	62.1	37.2	31.0
No conclusion	18	41.2	58.8	34.1

Appendix II: Feature descriptions

Table AII.1

Descriptive statistics from discrete features

hourDay	count	%	gender	count	%	switched	count	%
07-08	1	0.0	male	2199	56.6	yes	1572	40,4
08-09	82	2.1	female	1688	43.4	no	2315	59,6
09-10	194	5.0						
10-11	341	8.8						
11-12	360	9.3						
12-13	379	9.8						
13-14	392	10.1						
14-15	389	10.0						
15-16	361	9.3						
16-17	333	8.6						
17-18	256	6.6						
18-19	247	6.4						
19-20	235	6.0						
20-21	226	5.8						
21-22	57	1.5						
22-23	29	0.7						
23-24	5	0.1						

Table AII.2

Names and descriptive statistics from the numeric continuous features

Feature name	Mean	Std. dev.	Min	Max
durationInSeconds	380.64	418.27	30	3,592
age	39.41	14.15	19	96
daysTillDeadline	12.81	13.17	0	42
framewise_mean	0.19	2.17	-6.91	7.41
framewise_stdev	73.97	17.23	23.24	138.64
framewise_min	-179.28	48.6	-359.65	-44.55
framewise_max	182.01	47.26	48.15	365.65
framewise_range	361.3	82.91	103.65	698.95
mean_slopes	-0.01	1.47	-5.74	6.95
global_IQR_mean	98.3	32.29	9.5	255.75
global_IQR_min	70.34	29.04	8	255.5
global_IQR_max	175.02	54.25	13.5	352.5
delta_F0	8.62	138.5	-405.5	390.5
maximum_F0	496.19	6.95	422.5	500
minimum_F0	84.19	19.37	50	176.5
mean_F0	250.77	30.03	148.18	378.34
stdev_F0	94.94	17.05	39.41	152.17
range_F0	412	20.77	289.5	450
median_F0	230.83	43.76	105.75	440
skewness_F0	0.7	0.47	-1.13	4.44
kurtosis_F0	0.08	1.2	-1.56	22.14
SNR	0.97	0.11	0	1
delta_db	49.49	18.98	-18.15	98.27
maximum_db	77.27	6.26	28.72	91.45
minimum_db	16.47	17.44	-12.33	59.18
stdev_db	8.03	3.12	1	25.41
range_db	60.8	17.12	22.35	102.39
mean_coefficient1	-0.51	1.09	-7.4	4.07
min_coefficient1	-7.18	1.5	-11.87	-1.2
max_coefficient1	5.27	1.28	-3.56	10.45

median_coefficient1	-0.58	1.25	-7.5	4.71
delta_coefficient1	-0.19	3.7	-12.78	11.59
stdev_coefficient1	2.42	0.48	0.4	6.17
range_coefficient1	12.44	1.9	2.41	19.49
mean_coefficient2	-1.29	0.67	-3.81	2.05
min_coefficient2	-6.16	0.94	-9.54	-0.69
max_coefficient2	3.75	1.1	-0.35	8.2
median_coefficient2	-1.16	0.77	-3.91	2.33
delta_coefficient2	-0.1	2.83	-10	9.48
stdev_coefficient2	1.93	0.32	0.36	2.97
range_coefficient2	9.92	1.45	2.4	15.28
mean_coefficient3	-1.19	0.47	-3.42	1.19
min_coefficient3	-5.16	0.93	-8.87	-1.06
max_coefficient3	2.09	0.77	-1.14	5.33
median_coefficient3	-1.07	0.5	-3.43	0.94
delta_coefficient3	0.09	1.88	-7.46	6.65
stdev_coefficient3	1.28	0.22	0.35	2.22
range_coefficient3	7.25	1.15	2.16	11.88
mean_coefficient4	-1.43	0.47	-3.29	0.5
min_coefficient4	-5.12	0.72	-7.7	-0.94
max_coefficient4	1.89	0.71	-0.22	5.63
median_coefficient4	-1.3	0.54	-3.41	0.33
delta_coefficient4	0.02	2.01	-6.66	8.77
stdev_coefficient4	1.35	0.2	0.36	2.2
range_coefficient4	7.01	0.98	2.1	10.85
mean_coefficient5	-0.71	0.35	-2.13	1
min_coefficient5	-3.36	0.58	-5.89	-0.58
max_coefficient5	1.73	0.52	0.19	4.28
median_coefficient5	-0.67	0.38	-2.17	0.93
delta_coefficient5	0.03	1.35	-4.6	5.21
stdev_coefficient5	0.92	0.13	0.34	1.58
range_coefficient5	5.09	0.7	2	8.05
mean_coefficient6	-0.47	0.3	-1.4	0.75

min_coefficient6	-3.1	0.55	-5.51	-0.55
max_coefficient6	1.92	0.48	0.55	4.35
median_coefficient6	-0.41	0.33	-1.45	0.94
delta_coefficient6	0.03	1.36	-5.42	4.82
stdev_coefficient6	0.9	0.12	0.34	1.53
range_coefficient6	5.02	0.69	1.85	8.26
mean_coefficient7	-0.19	0.26	-1.15	0.91
min_coefficient7	-2.37	0.45	-4.84	-0.58
max_coefficient7	1.91	0.44	0.66	4.2
median_coefficient7	-0.17	0.27	-1.2	0.82
delta_coefficient7	-0.04	1.09	-4.5	4.6
stdev_coefficient7	0.75	0.1	0.34	1.22
range_coefficient7	4.28	0.59	2.03	7.48
mean_coefficient8	-0.5	0.29	-2.08	1.14
min_coefficient8	-2.55	0.43	-4.68	-0.64
max_coefficient8	1.53	0.41	0.46	3.37
median_coefficient8	-0.48	0.31	-2.4	0.84
delta_coefficient8	0.03	1.13	-4.29	3.97
stdev_coefficient8	0.75	0.11	0.3	1.32
range_coefficient8	4.08	0.56	1.9	6.49
mean_coefficient9	-0.19	0.25	-1.11	0.89
min_coefficient9	-2.08	0.39	-4.13	-0.52
max_coefficient9	1.84	0.56	0.55	4.39
median_coefficient9	-0.18	0.25	-1.28	0.76
delta_coefficient9	0.05	0.99	-4.03	3.63
stdev_coefficient9	0.68	0.1	0.32	1.16
range_coefficient9	3.93	0.62	1.64	6.77
mean_coefficient10	-0.23	0.21	-1.08	0.58
min_coefficient10	-1.95	0.34	-3.74	-0.54
max_coefficient10	1.66	0.52	0.32	3.73
median_coefficient10	-0.23	0.21	-1.13	0.53
delta_coefficient10	0.01	0.9	-3.48	3.31
stdev_coefficient10	0.61	0.09	0.28	1.15

range_coefficient10	3.61	0.61	1.56	6
mean_coefficient11	-0.11	0.23	-0.83	0.78
min_coefficient11	-1.8	0.37	-4.14	-0.56
max_coefficient11	1.79	0.57	0.46	3.85
median_coefficient11	-0.12	0.23	-0.87	0.8
delta_coefficient11	0.06	0.91	-4	4.27
stdev_coefficient11	0.61	0.11	0.3	1.14
range_coefficient11	3.6	0.69	1.67	7.67
mean_coefficient12	-0.07	0.24	-1.09	1.05
min_coefficient12	-1.75	0.44	-4.26	-0.54
max_coefficient12	1.78	0.58	0.38	3.78
median_coefficient12	-0.09	0.23	-1.1	1.14
delta_coefficient12	0.05	0.89	-3.92	3.4
stdev_coefficient12	0.59	0.12	0.29	1.22
range_coefficient12	3.53	0.78	1.62	6.63

Appendix III: Distribution of features.

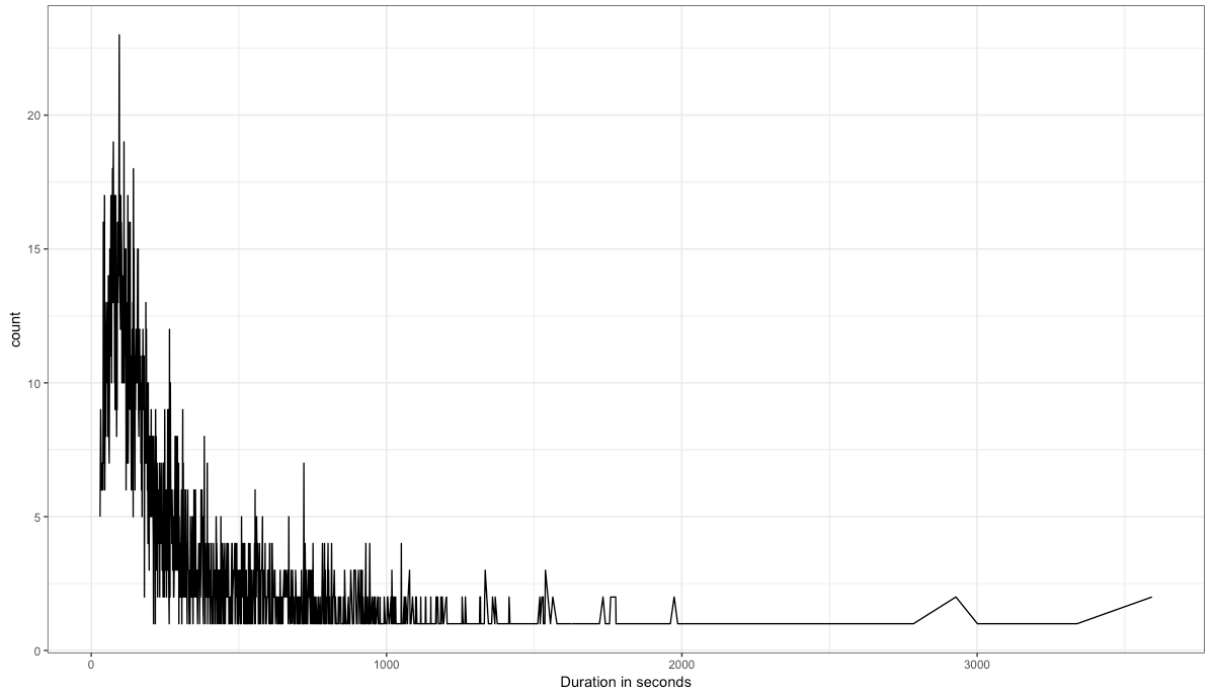

Figure AIII.1, distribution of duration in seconds.

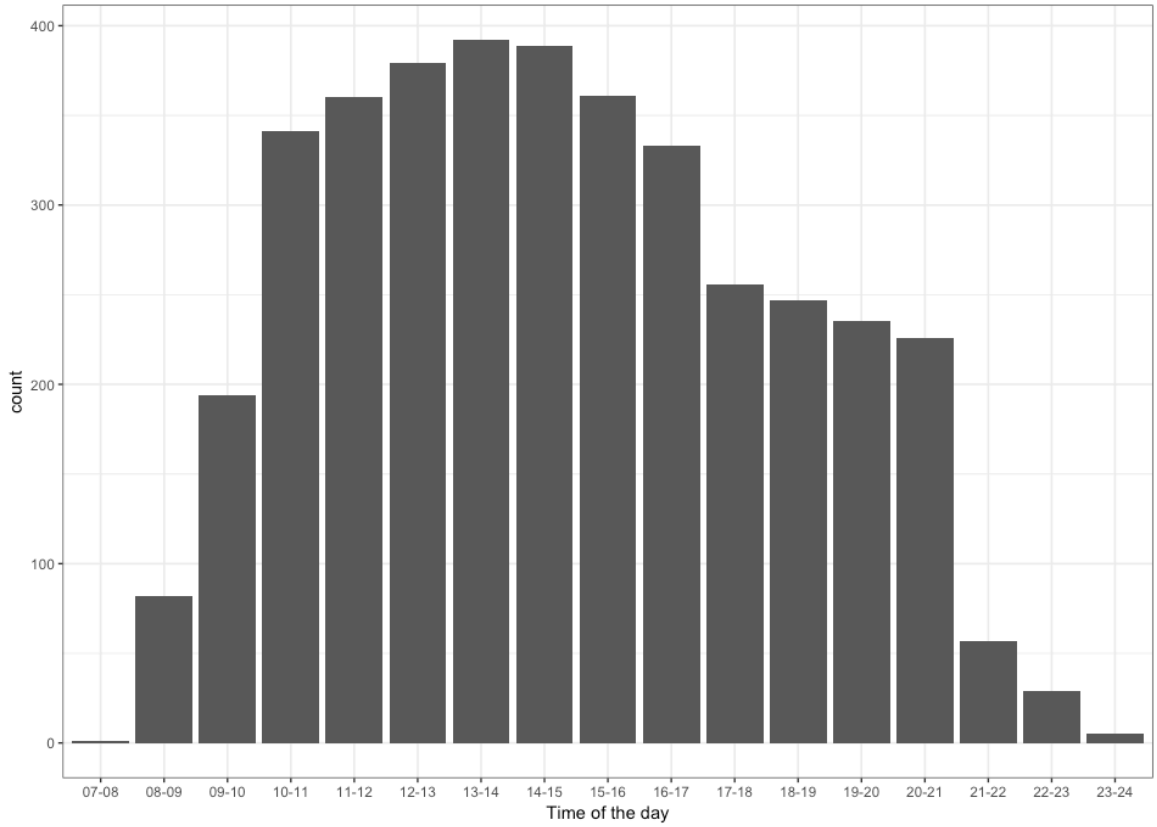

Figure AIII.2, distribution of the time of the day.

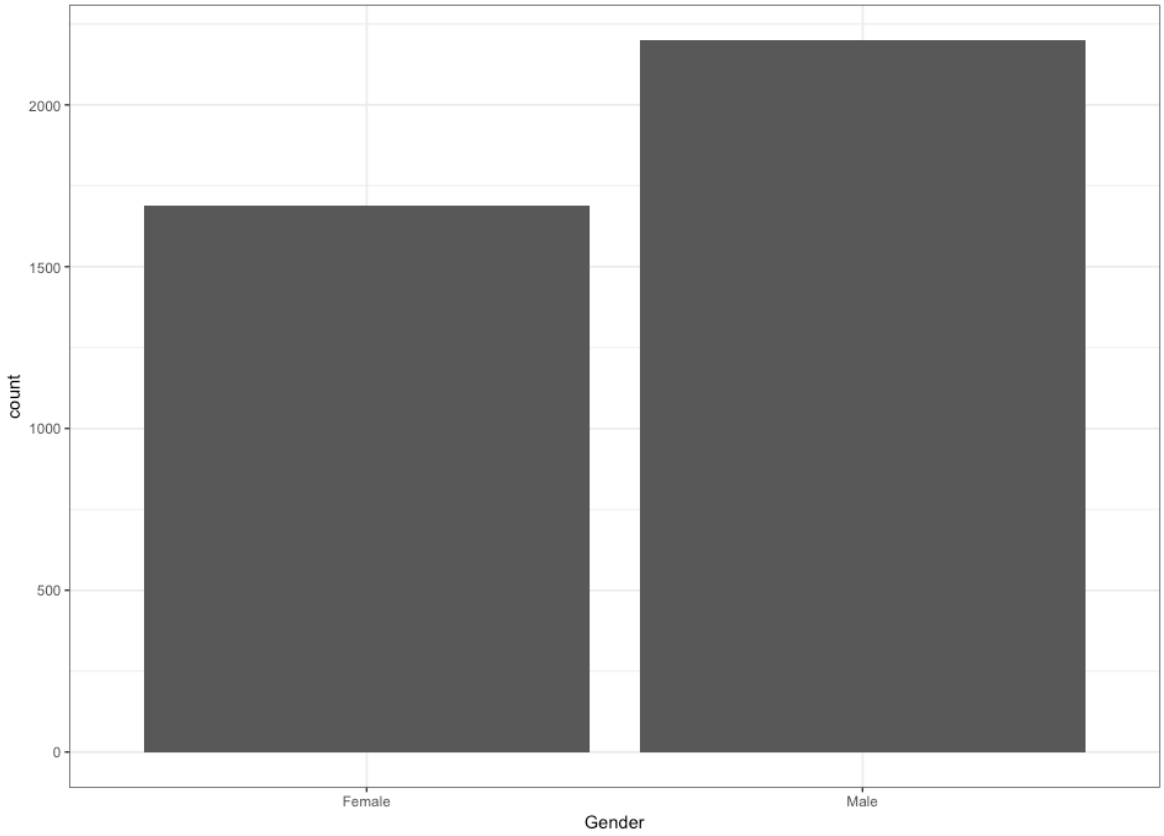

Figure AIII.3, distribution of gender.

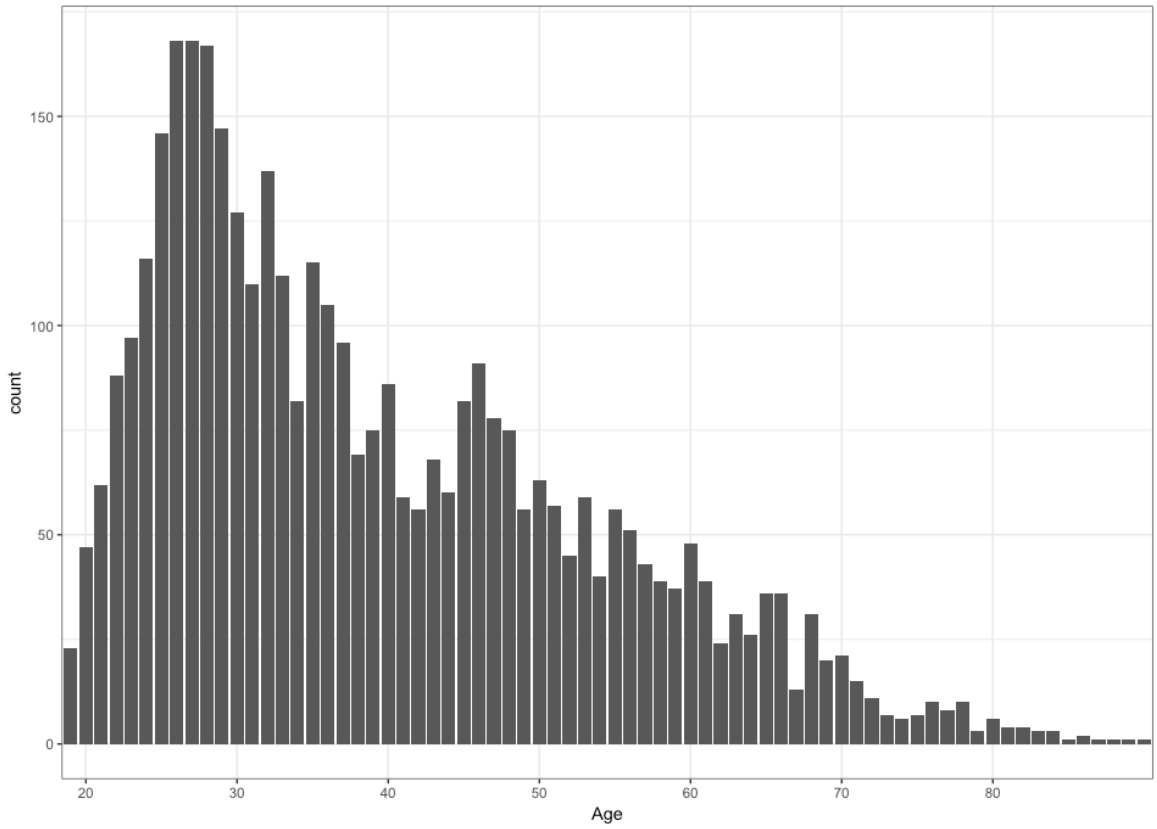

Figure AIII.4, distribution of age.

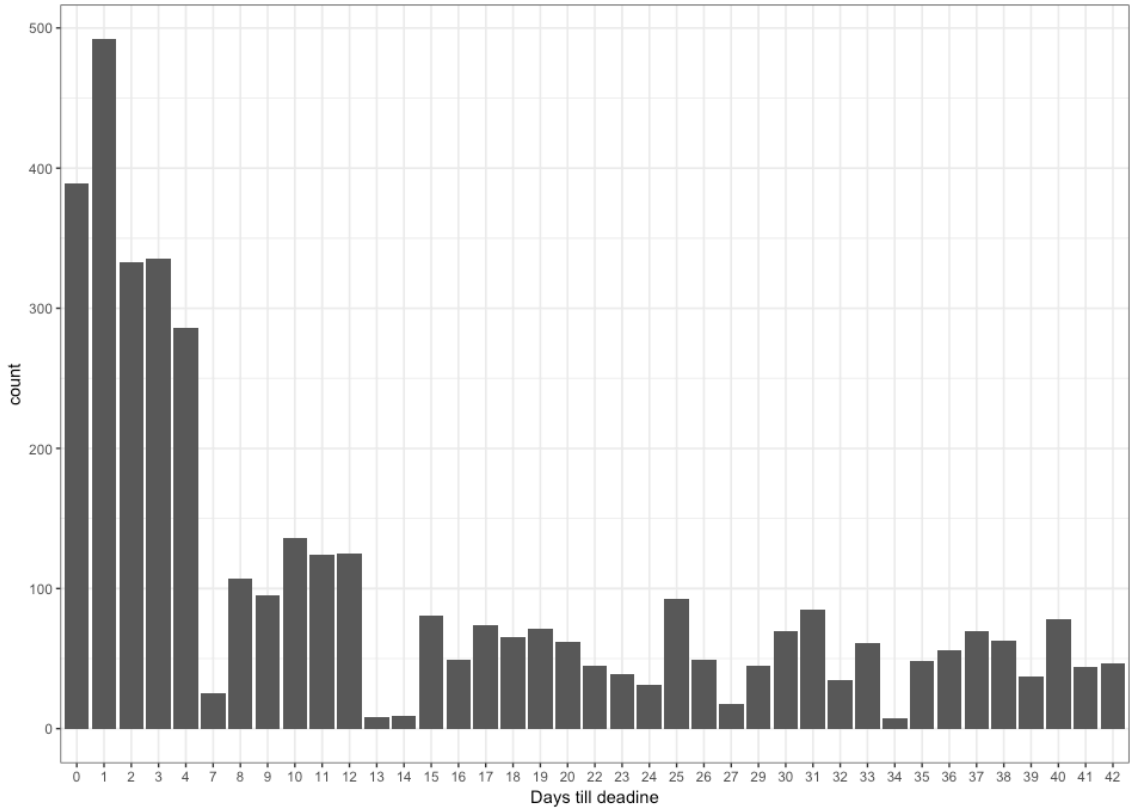

Figure AIII.5, distribution of number of days till the deadline of the health insurance season.

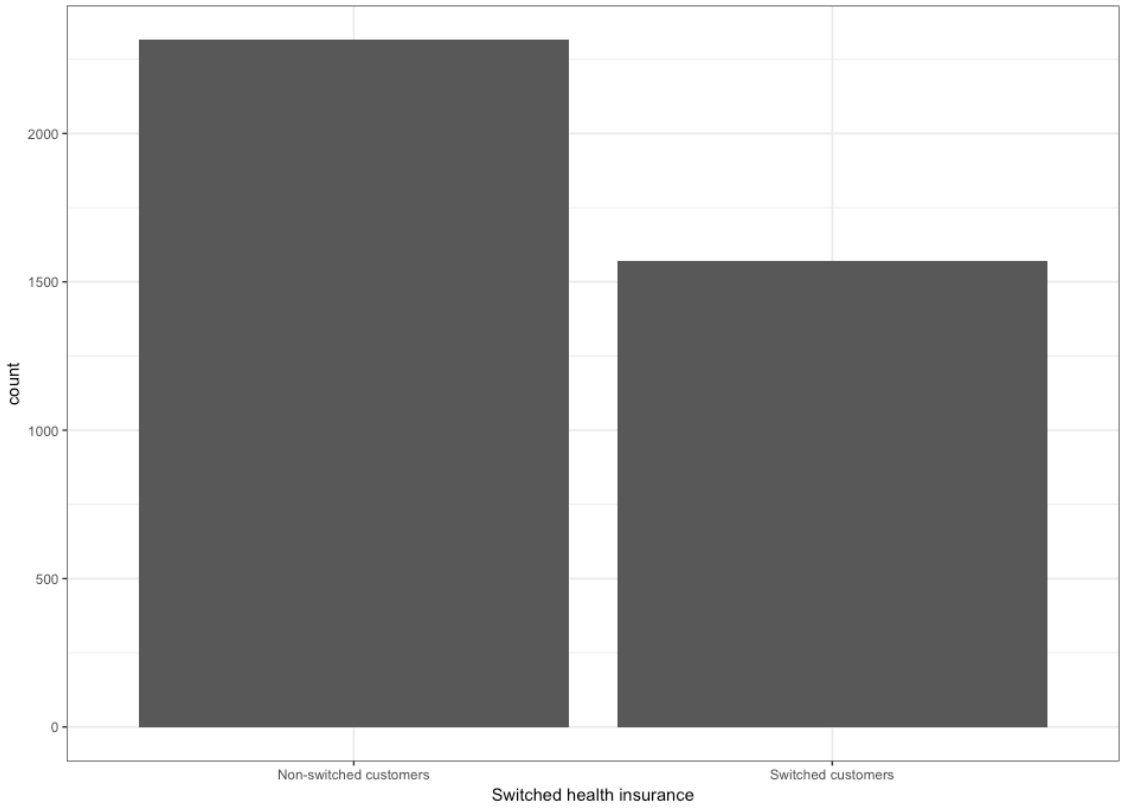

Figure A7, distribution of switched versus non-switched customers.

Appendix IV: Switching and non-switching customers, visualizations per acoustic feature.

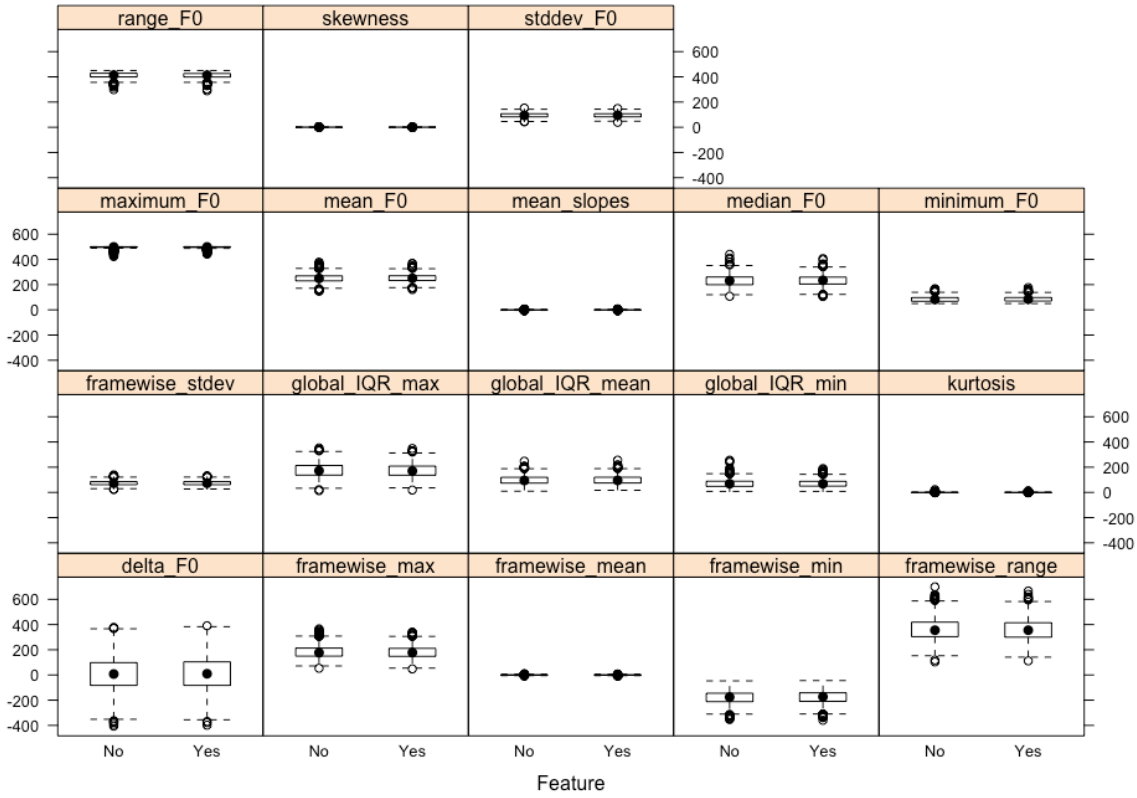

Figure AIV.1 box plots for all the pitch related features.

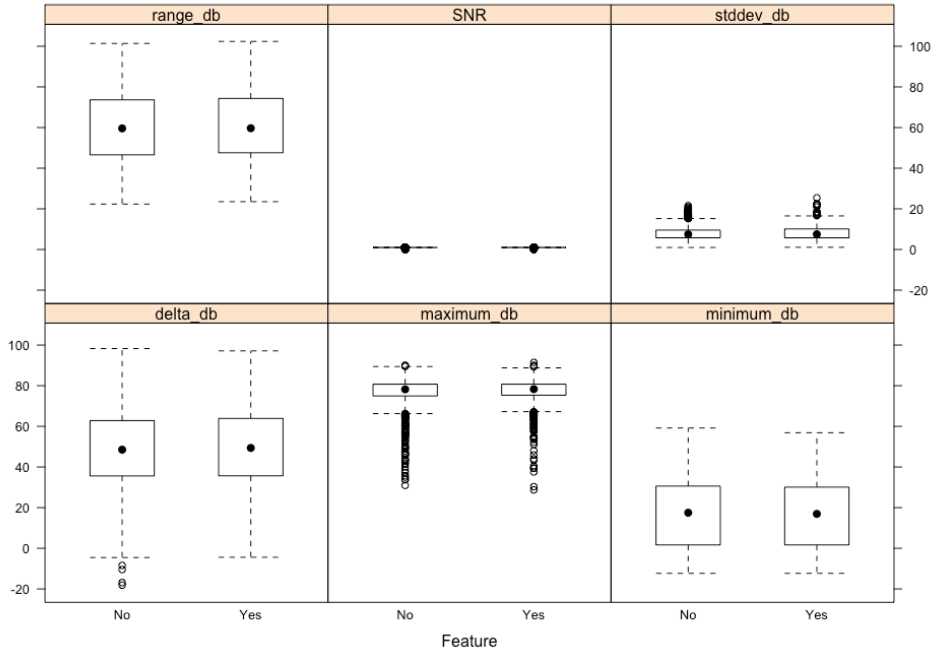

Figure AIV.2 box plots for all the intensity related features, for both classes of the target variable.

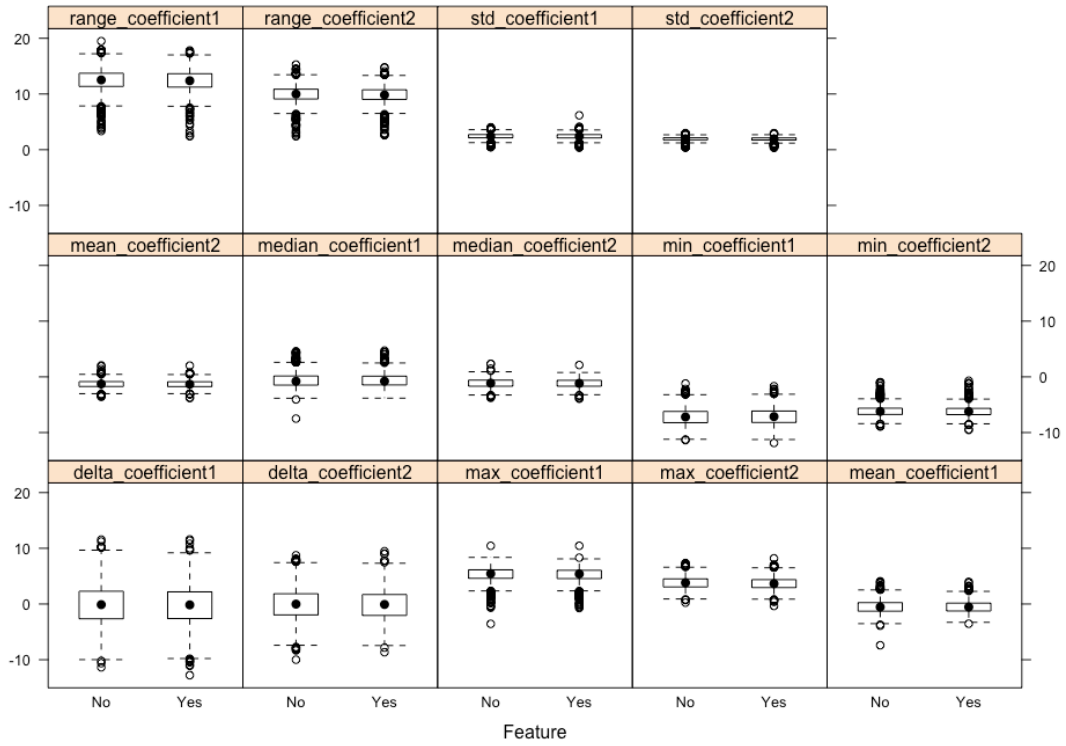

Figure AIV.3 box plots for coefficients 1 and 2 from the MFCC features, for both classes of the target variable.

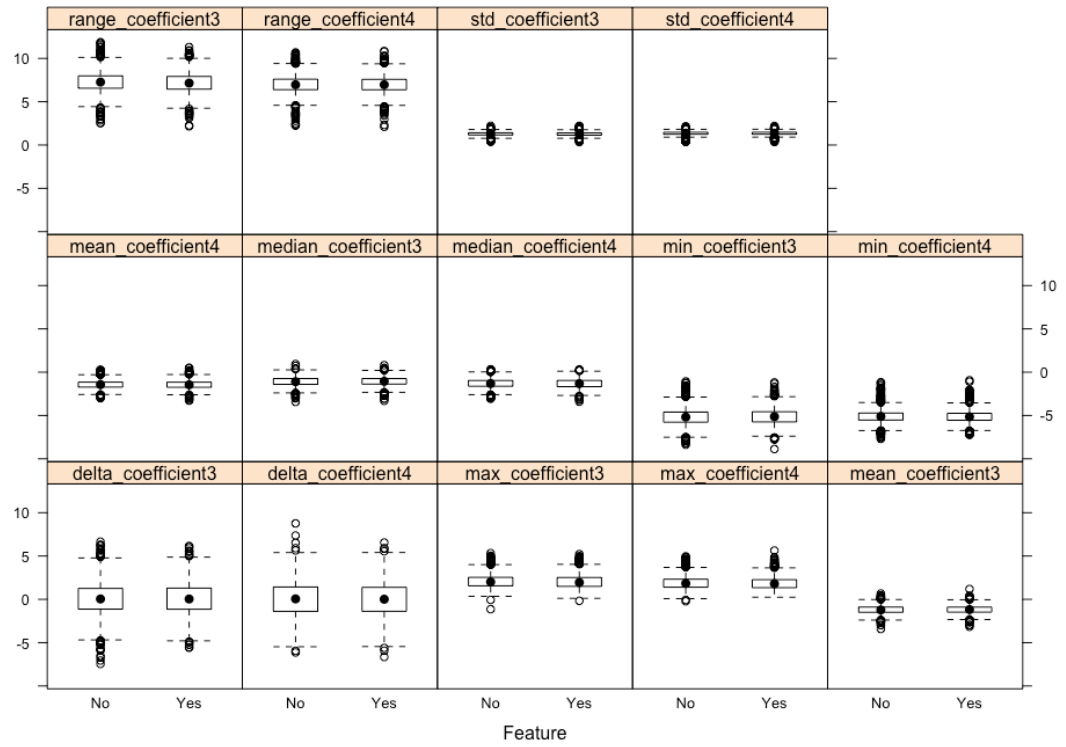

Figure AIV.4 box plots for coefficients 3 and 4 from the MFCC features, for both classes of the target variable.

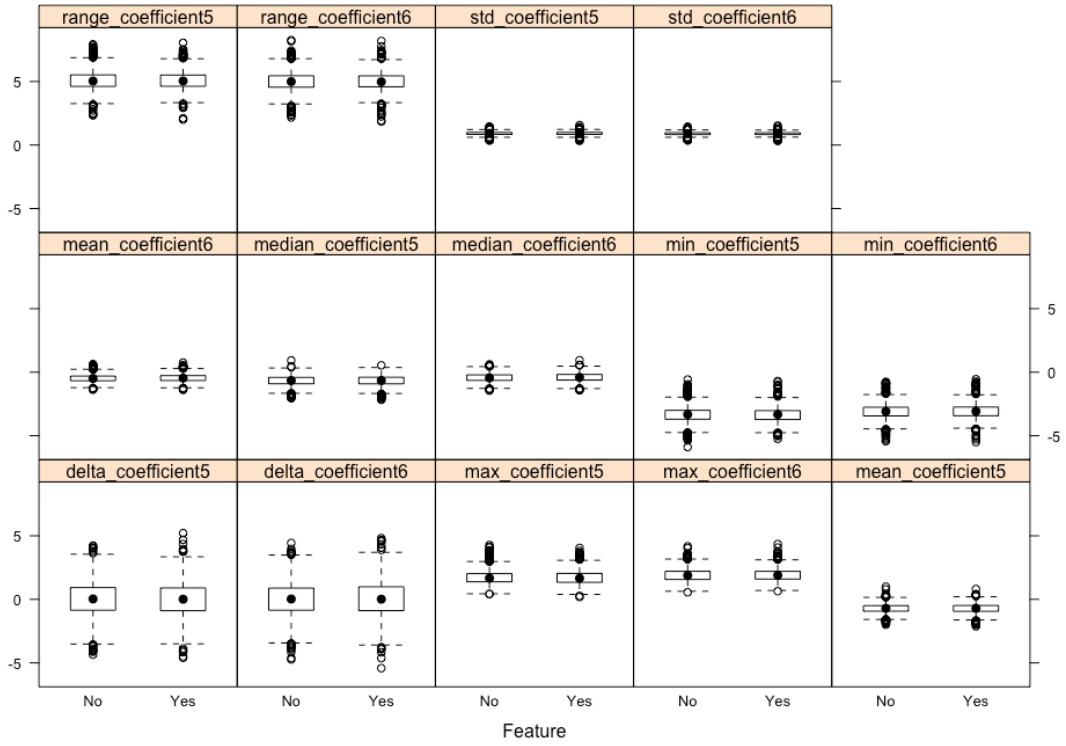

Figure AIV.5 box plots for coefficients 5 and 6 from the MFCC features, for both classes of the target variable.

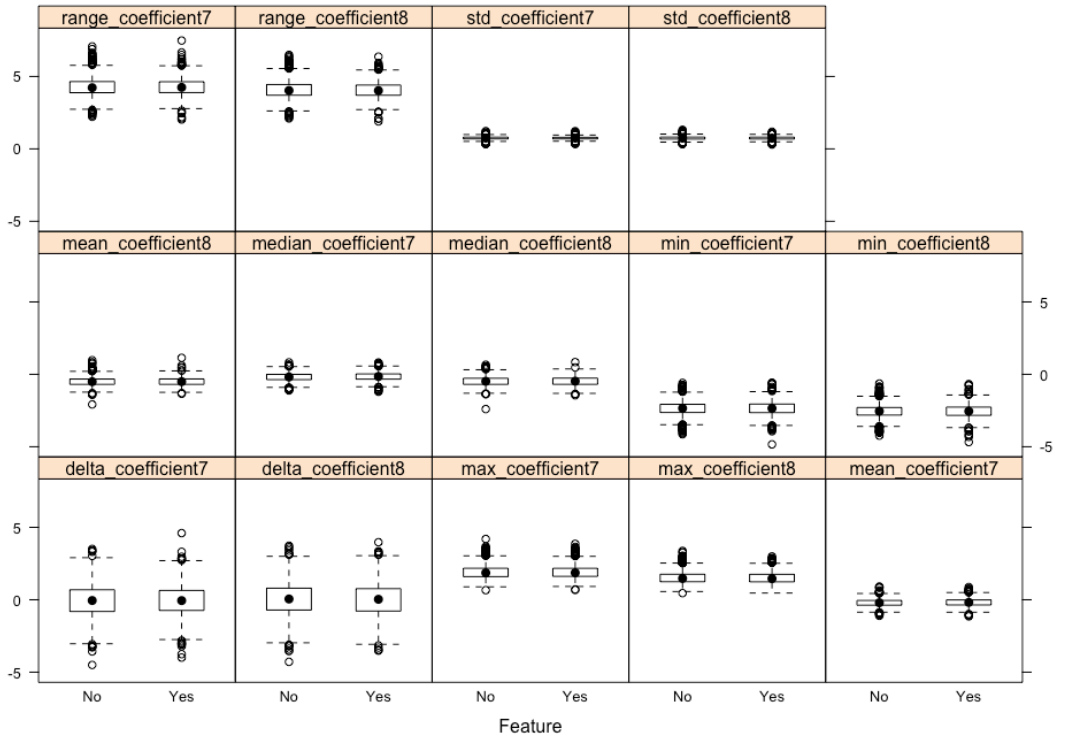

Figure AIV.6 box plots for coefficients 7 and 8 from the MFCC features, for both classes of the target variable.

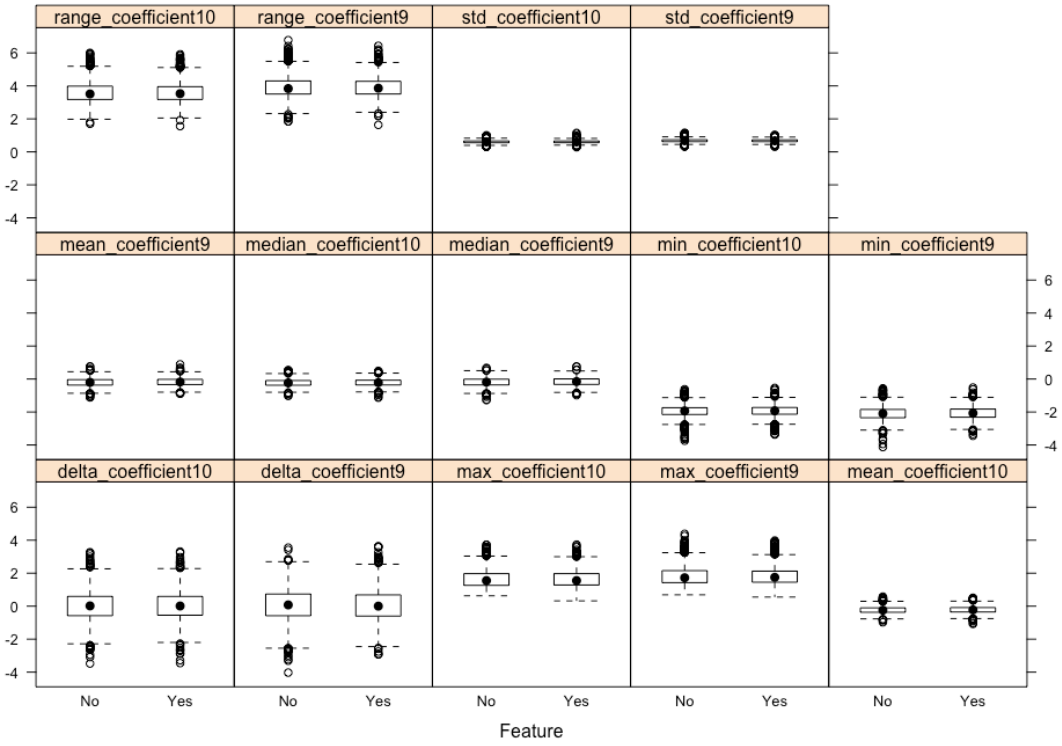

Figure AIV.7 box plots for coefficients 9 and 10 from the MFCC features, for both classes of the target variable.

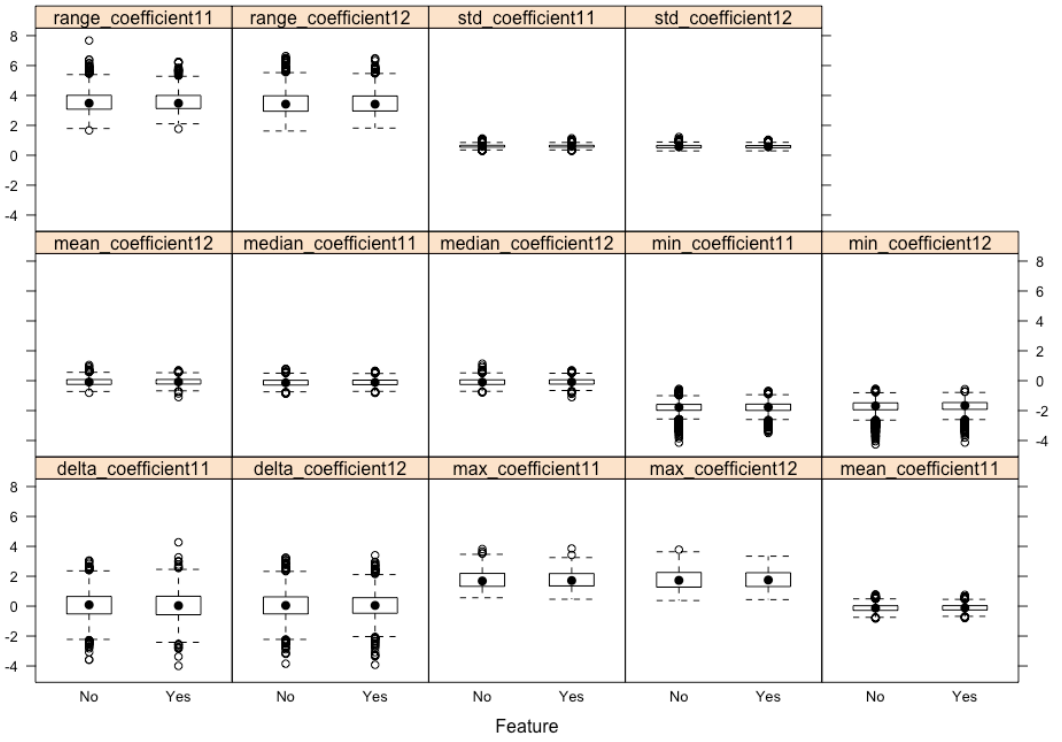

Figure AIV.8 box plots for coefficients 11 and 12 from the MFCC features, for both classes of the target variable.

Appendix V: Switching and non-switching customers, descriptive statistics for every acoustic feature

Table AV.1

Descriptive statistics for acoustic features

	Min.	Median	Mean	Max.
framewise_mean_switched	-6.298	0.262	0.1494	6.312
framewise_mean_not_switched	-6.905	0.2667	0.2112	7.407
framewise_stdev_switched	26.84	72.69	73.71	131.9
framewise_stdev_not_switched	23.24	72.72	74.14	138.6
framewise_min_switched	-359.6	-172.6	-177.9	-44.55
framewise_min_not_switched	-353.2	-175.6	-180.2	-46.85
framewise_max_switched	48.15	177.6	181	339.1
framewise_max_not_switched	53.2	176.7	182.7	365.6
framewise_range_switched	111.7	355	358.9	665.7
framewise_range_not_switched	103.6	355	362.9	699
mean_slopes_switched	-4.745	-0.01506	-0.02945	6.947
mean_slopes_not_switched	-5.741	0.03519	0.00476	5.318
global_IQR_mean_switched	16.65	95.12	98.65	255.8
global_IQR_mean_not_switched	9.5	94.4	98.06	247.5
global_IQR_min_switched	8	68.5	70.51	188.5
global_IQR_min_not_switched	8	68.5	70.22	255.5
global_IQR_max_switched	18.5	170.2	173.9	348.5
global_IQR_max_not_switched	13.5	172	175.8	352.5
delta_F0_switched	-398.5	9.75	9.471	390.5
delta_F0_not_switched	-405.5	7.5	8.044	377
maximum_F0_switched	443	499	496.1	500
maximum_F0_not_switched	422.5	499	496.2	500
minimum_F0_switched	50	84	85.08	176.5
minimum_F0_not_switched	50	82	83.59	167.5
mean_F0_switched	160.5	250.5	251.6	369.1
mean_F0_not_switched	148.2	249.5	250.2	378.3
stddev_F0_switched	39.41	94.92	94.98	148.8
stddev_F0_not_switched	43.6	94.12	94.91	152.2

range_F0_switched	289.5	413.5	411.1	450
range_F0_not_switched	298.5	414.5	412.6	450
median_F0_switched	105.8	234	231.9	406.5
median_F0_not_switched	107.5	231	230.1	440
skewness_switched	-1.125	0.6694	0.6883	2.621
skewness_not_switched	-0.8411	0.6796	0.7068	4.439
kurtosis_switched	-1.56	-0.2292	0.04829	12.7
kurtosis_not_switched	-1.533	-0.2006	0.1064	22.14
SNR_switched	0	0.9996	0.9665	1
SNR_not_switched	0	0.9997	0.9656	1
delta_db_switched	-4.431	49.37	49.92	97.13
delta_db_not_switched	-18.15	48.52	49.21	98.27
maximum_db_switched	28.72	78.32	77.4	91.45
maximum_db_not_switched	30.94	78.2	77.18	90.08
minimum_db_switched	-12.33	16.92	16.25	56.86
minimum_db_not_switched	-12.33	17.5	16.63	59.18
stddev_db_switched	1.097	7.502	8.131	25.41
stddev_db_not_switched	1	7.46	7.953	21.6
range_db_switched	23.53	59.67	61.15	102.4
range_db_not_switched	22.35	59.6	60.56	101.4
mean_coefficient1_switched	-3.537	-0.5421	-0.5024	3.983
mean_coefficient1_not_switched	-7.404	-0.5422	-0.5118	4.075
min_coefficient1_switched	-11.87	-7.136	-7.141	-1.677
min_coefficient1_not_switched	-11.37	-7.202	-7.202	-1.195
max_coefficient1_switched	-0.7259	5.392	5.236	10.45
max_coefficient1_not_switched	-3.561	5.448	5.288	10.45
median_coefficient1_switched	-3.825	-0.7844	-0.5831	4.714
median_coefficient1_not_switched	-7.497	-0.7842	-0.578	4.574
delta_coefficient1_switched	-12.78	-0.1597	-0.213	11.59
delta_coefficient1_not_switched	-11.37	-0.1159	-0.1689	11.56
std_coefficient1_switched	0.3956	2.405	2.413	6.17
std_coefficient1_not_switched	0.461	2.41	2.42	4.018
range_coefficient1_switched	2.41	12.38	12.38	17.82

range_coefficient1_not_switched	3.35	12.52	12.49	19.49
mean_coefficient2_switched	-3.81	-1.338	-1.321	2.018
mean_coefficient2_not_switched	-3.581	-1.272	-1.271	2.053
min_coefficient2_switched	-9.535	-6.227	-6.175	-0.6931
min_coefficient2_not_switched	-8.902	-6.207	-6.154	-0.9537
max_coefficient2_switched	-0.3533	3.651	3.657	8.202
max_coefficient2_not_switched	0.2359	3.814	3.819	7.357
median_coefficient2_switched	-3.908	-1.17	-1.183	2.125
median_coefficient2_not_switched	-3.778	-1.128	-1.147	2.335
delta_coefficient2_switched	-8.658	-0.07761	-0.1401	9.476
delta_coefficient2_not_switched	-9.999	-0.01691	-0.06677	8.748
std_coefficient2_switched	0.3604	1.934	1.919	2.96
std_coefficient2_not_switched	0.4013	1.955	1.939	2.975
range_coefficient2_switched	2.598	9.828	9.833	14.8
range_coefficient2_not_switched	2.405	9.997	9.973	15.28
mean_coefficient3_switched	-3.166	-1.181	-1.18	1.19
mean_coefficient3_not_switched	-3.423	-1.225	-1.194	0.6742
min_coefficient3_switched	-8.875	-5.129	-5.138	-1.182
min_coefficient3_not_switched	-8.378	-5.184	-5.183	-1.062
max_coefficient3_switched	-0.1717	1.946	2.063	5.223
max_coefficient3_not_switched	-1.14	2.013	2.105	5.335
median_coefficient3_switched	-3.313	-1.055	-1.064	0.8117
median_coefficient3_not_switched	-3.433	-1.096	-1.077	0.9415
delta_coefficient3_switched	-5.578	0.03675	0.1114	6.137
delta_coefficient3_not_switched	-7.458	0.04061	0.06739	6.654
std_coefficient3_switched	0.3507	1.265	1.267	2.208
std_coefficient3_not_switched	0.4057	1.286	1.286	2.218
range_coefficient3_switched	2.162	7.148	7.2	11.31
range_coefficient3_not_switched	2.527	7.263	7.288	11.88
mean_coefficient4_switched	-3.286	-1.458	-1.444	0.4995
mean_coefficient4_not_switched	-2.998	-1.43	-1.423	0.2593
min_coefficient4_switched	-7.22	-5.149	-5.135	-0.9357
min_coefficient4_not_switched	-7.7	-5.108	-5.102	-1.163

max_coefficient4_switched	0.244	1.799	1.862	5.629
max_coefficient4_not_switched	-0.2229	1.837	1.913	4.959
median_coefficient4_switched	-3.405	-1.328	-1.312	0.3258
median_coefficient4_not_switched	-3.057	-1.312	-1.293	0.3209
delta_coefficient4_switched	-6.657	0.0127	0.001924	6.555
delta_coefficient4_not_switched	-6.138	0.05091	0.031	8.766
std_coefficient4_switched	0.357	1.359	1.352	2.202
std_coefficient4_not_switched	0.369	1.358	1.351	2.166
range_coefficient4_switched	2.101	6.961	6.997	10.85
range_coefficient4_not_switched	2.232	6.963	7.015	10.69
mean_coefficient5_switched	-2.129	-0.7062	-0.7131	0.8166
mean_coefficient5_not_switched	-1.992	-0.7048	-0.7135	0.9971
min_coefficient5_switched	-5.243	-3.327	-3.367	-0.697
min_coefficient5_not_switched	-5.892	-3.323	-3.356	-0.5765
max_coefficient5_switched	0.1855	1.659	1.729	4.049
max_coefficient5_not_switched	0.4217	1.669	1.736	4.278
median_coefficient5_switched	-2.167	-0.6569	-0.6649	0.5423
median_coefficient5_not_switched	-2.049	-0.6577	-0.6678	0.9285
delta_coefficient5_switched	-4.602	0.01117	0.01072	5.213
delta_coefficient5_not_switched	-4.356	0.03329	0.04396	4.231
std_coefficient5_switched	0.3416	0.9179	0.9234	1.576
std_coefficient5_not_switched	0.3674	0.916	0.9199	1.457
range_coefficient5_switched	2.003	5.045	5.096	8.049
range_coefficient5_not_switched	2.336	5.039	5.092	7.933
mean_coefficient6_switched	-1.398	-0.4628	-0.4506	0.7545
mean_coefficient6_not_switched	-1.38	-0.4962	-0.4768	0.6434
min_coefficient6_switched	-5.506	-3.075	-3.087	-0.549
min_coefficient6_not_switched	-5.43	-3.088	-3.103	-0.7594
max_coefficient6_switched	0.641	1.898	1.934	4.354
max_coefficient6_not_switched	0.5546	1.886	1.915	4.18
median_coefficient6_switched	-1.424	-0.4016	-0.3962	0.9416
median_coefficient6_not_switched	-1.452	-0.4483	-0.426	0.6221
delta_coefficient6_switched	-5.418	0.01315	0.0439	4.822

delta_coefficient6_not_switched	-4.728	0.02652	0.01265	4.441
std_coefficient6_switched	0.3393	0.9015	0.8999	1.533
std_coefficient6_not_switched	0.3803	0.9021	0.9004	1.452
range_coefficient6_switched	1.854	4.977	5.021	8.191
range_coefficient6_not_switched	2.167	4.992	5.018	8.257
mean_coefficient7_switched	-1.149	-0.1824	-0.1761	0.8774
mean_coefficient7_not_switched	-1.108	-0.2057	-0.2057	0.9147
min_coefficient7_switched	-4.838	-2.35	-2.364	-0.5845
min_coefficient7_not_switched	-4.138	-2.344	-2.366	-0.5901
max_coefficient7_switched	0.6798	1.863	1.916	3.86
max_coefficient7_not_switched	0.6638	1.862	1.909	4.196
median_coefficient7_switched	-1.202	-0.1485	-0.1488	0.8101
median_coefficient7_not_switched	-1.113	-0.1828	-0.1852	0.8244
delta_coefficient7_switched	-3.993	-0.05745	-0.03853	4.603
delta_coefficient7_not_switched	-4.502	-0.05144	-0.04401	3.511
std_coefficient7_switched	0.3384	0.7406	0.7445	1.221
std_coefficient7_not_switched	0.3427	0.744	0.7484	1.211
range_coefficient7_switched	2.027	4.246	4.28	7.48
range_coefficient7_not_switched	2.215	4.226	4.275	7.074
mean_coefficient8_switched	-1.343	-0.5055	-0.5007	1.137
mean_coefficient8_not_switched	-2.079	-0.5081	-0.5047	0.9745
min_coefficient8_switched	-4.68	-2.549	-2.548	-0.6715
min_coefficient8_not_switched	-4.233	-2.546	-2.549	-0.6388
max_coefficient8_switched	0.4622	1.458	1.518	2.992
max_coefficient8_not_switched	0.4577	1.473	1.536	3.374
median_coefficient8_switched	-1.45	-0.4749	-0.4766	0.8446
median_coefficient8_not_switched	-2.401	-0.4825	-0.4827	0.6623
delta_coefficient8_switched	-3.513	0.03243	0.0006819	3.973
delta_coefficient8_not_switched	-4.287	0.05219	0.0575	3.728
std_coefficient8_switched	0.3009	0.7413	0.745	1.18
std_coefficient8_not_switched	0.3436	0.7414	0.7466	1.318
range_coefficient8_switched	1.896	4.025	4.066	6.369
range_coefficient8_not_switched	2.101	4.032	4.086	6.489

mean_coefficient9_switched	-0.889	-0.18	-0.177	0.8895
mean_coefficient9_not_switched	-1.111	-0.2036	-0.1993	0.7648
min_coefficient9_switched	-3.443	-2.07	-2.073	-0.5161
min_coefficient9_not_switched	-4.125	-2.094	-2.092	-0.5674
max_coefficient9_switched	0.5525	1.748	1.842	3.981
max_coefficient9_not_switched	0.6884	1.724	1.845	4.386
median_coefficient9_switched	-0.9802	-0.1543	-0.163	0.7629
median_coefficient9_not_switched	-1.277	-0.187	-0.1894	0.6755
delta_coefficient9_switched	-2.923	0.001306	0.03542	3.629
delta_coefficient9_not_switched	-4.029	0.07564	0.05258	3.55
std_coefficient9_switched	0.3157	0.6712	0.6766	1.025
std_coefficient9_not_switched	0.3333	0.6753	0.6796	1.16
range_coefficient9_switched	1.635	3.86	3.915	6.43
range_coefficient9_not_switched	1.847	3.838	3.937	6.774
mean_coefficient10_switched	-1.077	-0.224	-0.2228	0.5085
mean_coefficient10_not_switched	-0.9767	-0.2421	-0.2338	0.5774
min_coefficient10_switched	-3.364	-1.929	-1.939	-0.5389
min_coefficient10_not_switched	-3.741	-1.935	-1.955	-0.621
max_coefficient10_switched	0.3214	1.55	1.668	3.729
max_coefficient10_not_switched	0.6288	1.55	1.657	3.724
median_coefficient10_switched	-1.131	-0.2188	-0.2221	0.4914
median_coefficient10_not_switched	-1.021	-0.2328	-0.2337	0.5315
delta_coefficient10_switched	-3.451	0.01035	0.004228	3.312
delta_coefficient10_not_switched	-3.479	0.01692	0.01748	3.27
std_coefficient10_switched	0.2838	0.6034	0.6131	1.15
std_coefficient10_not_switched	0.3052	0.6033	0.614	1.005
range_coefficient10_switched	1.561	3.53	3.607	5.914
range_coefficient10_not_switched	1.702	3.51	3.613	6.002
mean_coefficient11_switched	-0.7985	-0.1154	-0.1059	0.7553
mean_coefficient11_not_switched	-0.8261	-0.1328	-0.1206	0.7846
min_coefficient11_switched	-3.497	-1.767	-1.796	-0.6851
min_coefficient11_not_switched	-4.143	-1.776	-1.808	-0.5593
max_coefficient11_switched	0.4626	1.71	1.797	3.854

max_coefficient11_not_switched	0.561	1.683	1.79	3.82
median_coefficient11_switched	-0.8172	-0.112	-0.1122	0.6544
median_coefficient11_not_switched	-0.8696	-0.1359	-0.1302	0.7964
delta_coefficient11_switched	-3.997	0.03382	0.04445	4.272
delta_coefficient11_not_switched	-3.588	0.0857	0.06315	3.051
std_coefficient11_switched	0.3036	0.5987	0.611	1.136
std_coefficient11_not_switched	0.3147	0.5975	0.6126	1.121
range_coefficient11_switched	1.764	3.478	3.593	6.24
range_coefficient11_not_switched	1.665	3.481	3.597	7.673
mean_coefficient12_switched	-1.094	-0.07772	-0.0657	0.6927
mean_coefficient12_not_switched	-0.8076	-0.08748	-0.08107	1.051
min_coefficient12_switched	-4.138	-1.663	-1.727	-0.5659
min_coefficient12_not_switched	-4.259	-1.687	-1.762	-0.5433
max_coefficient12_switched	0.4353	1.75	1.785	3.343
max_coefficient12_not_switched	0.376	1.726	1.777	3.779
median_coefficient12_switched	-1.099	-0.08121	-0.08105	0.6901
median_coefficient12_not_switched	-0.7901	-0.0978	-0.09861	1.139
delta_coefficient12_switched	-3.923	0.04893	0.05121	3.404
delta_coefficient12_not_switched	-3.85	0.04724	0.04893	3.23
std_coefficient12_switched	0.296	0.5623	0.5835	1.029
std_coefficient12_not_switched	0.2909	0.5644	0.5888	1.221
range_coefficient12_switched	1.817	3.417	3.511	6.485
range_coefficient12_not_switched	1.622	3.42	3.54	6.635

Appendix VI: Cross validation results.

Table AVI.1, Cross validation results expressed in accuracy and ROC.

Feature set	Pre-processing	Naïve Bayes		k-NN		LDA		SVM	
		accuracy	ROC	accuracy	ROC	accuracy	ROC	accuracy	ROC
meta	none	0.621	0.683	0.645	0.651	0.649	0.673	0.651	0.667
meta	z-score	0.624	0.683	0.647	0.648	0.657	0.672	0.651	0.672
meta	min-max	0.610	0.685	0.627	0.625	0.651	0.675	0.653	0.671
meta	z-scores + PCA	0.588	0.681	0.649	0.646	0.655	0.674	0.651	0.668
meta	min-max + PCA	0.599	0.659	0.626	0.623	0.657	0.673	0.651	0.671
meta + pitch + intensity	none	0.614	0.632	0.632	0.617	0.658	0.671	0.650	0.662
meta + pitch + intensity	z-score	0.608	0.647	0.625	0.602	0.652	0.667	0.650	0.669
meta + pitch + intensity	min-max	0.610	0.640	0.615	0.602	0.649	0.671	0.648	0.665
meta + pitch + intensity	z-scores + PCA	0.606	0.639	0.627	0.600	0.654	0.671	0.652	0.668
meta + pitch + intensity	min-max + PCA	0.609	0.636	0.610	0.603	0.653	0.671	0.649	0.664
all features	none	0.602	0.632	0.629	0.608	0.642	0.648	0.641	0.655
all features	z-score	0.616	0.631	0.597	0.571	0.634	0.656	0.647	0.650
all features	min-max	0.611	0.640	0.608	0.607	0.634	0.655	0.641	0.658
all features	z-scores + PCA	0.612	0.635	0.601	0.568	0.650	0.655	0.655	0.665
all features	min-max + PCA	0.615	0.631	0.619	0.619	0.654	0.658	0.649	0.660
pitch+ intensity	none	0.569	0.512	0.559	0.499	0.595	0.479	0.595	0.515
pitch+ intensity	z-score	0.559	0.512	0.555	0.510	0.594	0.485	0.595	0.523
pitch+ intensity	min-max	0.557	0.510	0.557	0.514	0.595	0.502	0.595	0.514
pitch+ intensity	z-scores + PCA	0.588	0.498	0.563	0.509	0.595	0.493	0.595	0.514
pitch+ intensity	min-max + PCA	0.583	0.497	0.550	0.501	0.596	0.489	0.595	0.516
full acoustic	none	0.547	0.552	0.558	0.491	0.568	0.518	0.595	0.526
full acoustic	z-score	0.547	0.544	0.575	0.516	0.570	0.535	0.595	0.519
full acoustic	min-max	0.543	0.547	0.564	0.531	0.564	0.524	0.595	0.532
full acoustic	z-scores + PCA	0.565	0.540	0.562	0.533	0.581	0.532	0.595	0.540
full acoustic	min-max + PCA	0.562	0.518	0.571	0.536	0.570	0.534	0.595	0.535