CUSTOMER CHURN WITHOUT PERSONAL INFORMATION IN THE TELECOMMUNICATIONS INDUSTRY

BENTE DE KEIZER

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2023377

COMMITTEE

dr. Cassani
dr. ÖzgödeYigin

LOCATION

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

DATE

December 4th, 2023

WORD COUNT

max 8800

ACKNOWLEDGMENTS

I want to thank Giovanni Cassani for his supervision.

CUSTOMER CHURN WITHOUT PERSONAL INFORMATION IN THE TELECOMMUNICATIONS INDUSTRY

BENTE DE KEIZER

Abstract

Models that predict customer churn in the telecommunications sector use personal information. Due to stricter privacy legislation this is increasingly difficult to maintain. This thesis focuses on researching whether models without personal information will also be sufficient. Furthermore, this thesis will suggest which algorithm companies can use to predict customer churn in the case of minimizing personal information. There are five algorithms that will be compared: AdaBoostClassifier with DecisionTreeClassifier, RandomForestClassifier or ExtraTreeClassifier as the base estimator, CatBoostClassifier and the XGBoostClassifier. The dataset that will be used is the "Sample Telco Customer Churn Dataset" from Kaggle. There are several scenarios that minimize personal information. The preferred scenario is to train the model with the so far collected personal information and then fit the model with uninformative personal information. This is best done with the CatBoostClassifier. If personal information may not be used at all, it is better to make the personal information uninformative and use the AdaBoostClassifier with the ExtraTreeClassifier as the base. This option is slightly better than completely removing personal information from the dataset using CatBoostClassifier. The difference between the models with and without personal information is very small. Thus, the conclusion is that personal information of the customer does not add much value to the model to predict customer churn.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The data has been acquired from Kaggle and can be found here. Work on this thesis did not involve collecting data from human participants or animals. The original owners of the data used in this thesis retains

ownership of the data and during and after the completion of this thesis. The thesis code can be accessed through the GitHub repository following the link. In terms of writing, the author used assistance with the language of the paper.A generative language model was used to improve the author's original content, for paraphrasing, spell checking and grammar. No other typesetting tools or services were used.

2 INTRODUCTION

Due to General Data Protection Regulation (GDPR), companies cannot ask for irrelevant personal information that customers are required to fill in. Variables, such as age and gender, are not necessary to deliver services in the telecommunications industry. Consumers are less likely to agree with disclosing personal information to the company when this information is not required (Chua et al., 2021). Previous literature argues that demographic information of the customer is important in predicting whether a customer will churn (Fujo et al., 2022). As a result, companies may be afraid that customer churn predictions will be less accurate when removing personal information from this prediction and are less likely to deviate from this method of prediction.

Prior research has focused on customer churn prediction in the telecommunications industry using machine learning models when personal information is available (Ahmad et al., 2019, Odusami et al., 2021, Jain et al., 2020, Shumaly et al., 2020, Lalwani et al., 2022). This thesis is about churn prediction without personal information which is increasingly relevant due to introduction of the GDPR.

Churn prediction is relevant to classify customers in possible churners and loyal customers ensuring that companies exert the right marketing activities in each group. For example, companies can convince a possible churner to stay through a discount. How less accurate the churn prediction is due to the lack of personal information, how less effective these marketing activities are. This will result in more churners and thus loss of turnover. Companies lose revenues when customers switch to other companies. It can take up much effort and money to attract new customers so retaining customers is important. Thus, churn prediction without personal information is important for the companies so that they can allocate their marketing strategy in a smart manner.

This thesis investigates churn prediction in the telecommunications industry without personal information using ML. Various ML algorithms were used in the literature to predict customer churn in the telecommunication industry. Therefore, this thesis is limited to five ML models that could best predict customer churn from previous research (Lalwani

et al., 2022). This thesis will focus on the CatBoostClassifier (CatBoost), the AdaBoostClassifier (AdaBoost), the AdaBoostClassifier with RandomForestClassifier as base estimator (AdaBoost RF), the AdaBoostClassifier with the ExtraTreesClassifier as base estimator (AdaBoost Xtree)and the XGBoostClassifier (XGBoost). The following research question follows from the scientific gap:

> *RQ: Out of AdaBoost RF, CatBoost, AdaBoost, AdaBoost Xtree and XGBoost, which machine learning model is the best in predicting customer churn in the telecommunications industry without customer personal information, using the F2-score as the evaluation metric?*

The first sub question focuses on the difference in the F2-score of the machine learning models with personal information and the F2-score of the same machine learning models without personal information.

> *SQ1: Which machine learning model has the least relative deviation between calculated F2-score with and without personal information?*

Fujo et al. (2022) argue that personal information is very important in predicting customer churn. It is given that gender is the most important personal feature when predicting whether a customer will churn. This means that the lack of personal information can impact the customer churn predictions. The second sub question is related to which key features are important to determine whether a customer is likely to churn when personal information is not used in the prediction. This question determines on which features the predictive model should focus on when this kind of information cannot be used. This also raises the question whether ML models make more errors for one particular gender. This will be answered using the third sub-question.

> *SQ2: Which key features are important for the best machine learning model to predict customer churn without personal information?*

> *SQ3: Is there a difference in the error rates of a ML model, without personal information, when applied to female and male customers?*

Shumaly et al. (2020) show that class imbalance impacts the customer churn prediction and therefore it is chosen to take this into account in this thesis. With this subquestion it will be determined whether random undersampling and random oversampling has an impact on the accuracy of customer churn prediction.

> *SQ4: Does class resampling by random undersampling and random oversampling improve the best machine learning model for predicting customer churn without personal information?*

This thesis consists of a review of related literature, succeeded by in-depth descriptions of the research methodologies applied. Following, the obtained results are shown and discussed. Lastly, there is a conclusion that summarizes the key findings and their suggestions.

The main finding is that the differences in training models with and without personal information are minimal. Thus, personal information can be discarded when predicting customer churn. The best strategy is to train CatBoost with personal information but it fit it on data where personal information is masked. To completely discard personal information, it is best to make personal information uninformative and use AdaBoost Xtree as customer churn prediction model.

3 RELATED WORK

3.1 *Machine learning models*

According to Fujo et al. (2022), customer demographics, especially gender, are important in predicting customer churn. The research that already has been done on customer churn prediction in the telecommunication sector has taken into account customer demographics. Thus, papers in literature review used customer demographics for predicting customer churn in the telecommunication sector. There is a noticeable lack of studies that address the specific challenge of customer churn without personal information. This gap becomes particularly interesting in the new reality with strict data privacy legislation. With this thesis personal information is excluded from the data analysis to offer beneficial insights in customer churn prediction while still complying with the GDPR. In Shumaly et al. (2020), XGBoost and Random Forest (RF) were seen as the most accurate models to predict customer churn. Ahmad et al. (2019) found that XGBoost predicts better than RF. However, Odusami et al. (2021) found this conclusion the other way around. Another conclusion was Jain et al. (2020) which shows that the Logistic Regression slightly outperformed these models. Vafeiadis et al. (2015) found that Support Machine Vector with a polynomial kernel with AdaBoost was the most accurate ML model. That AdaBoost improves ML models comes out clearly in the research of Lalwani et al. (2022). They compared thirteen models in which AdaBoost, AdaBoost with extra tree, RF with Adaboost, CatBoost and XGBoost were the most accurately.

Lalwani et al. (2022) and Odusami et al. (2021) are given more weight in the choice of which ML models we will use in this thesis because these studies used the same dataset. When a standard model, such as Random Forest, is combined with AdaBoost, the model strongly outperforms the standard model. The evaluation metrics of Odusami et al. (2021) are for

these standard models lower than in combination with AdaBoost in the study of Lalwani et al. (2022). CatBoostClassifier has a F2-score of 0.8203, XGBoostClassifier has one of 0.6479. AdaBoost RF has a F2-score of 0.6548. AdaBoost Xtree has a F2-score of 0.6591. AdaBoost has a F2-score of 0.6473.

3.2 *Class imbalance algorithms*

Shumaly et al. (2020) states that machine learning models predict customer churn badly due to class imbalance. To deal with this, some algorithms can be used to balance the training set before data analysis. Ahmad et al. (2019) used random oversampling and undersampling and compared this with the imbalanced dataset. Odusami et al. (2021) used SMOTE for oversampling. Shumaly et al. (2020) compared all these techniques with the imbalanced dataset. They found that randomly resampling performed better than SMOTE and the imbalanced dataset. They also analysed that using Accuracy as an evaluation metric is not optimal when there is class imbalance. According to S. Han et al. (2009) and Weiss (2010), Accuracy is a bad metric for predicting in case of class imbalance because there is a strong bias against the rare class. Recall and Precision are better evaluation metrics. As suggested in Shumaly et al. (2020), this thesis will use random resampling with the F2-score as evaluation metric.

3.3 *Key features*

Fujo et al. (2022) uses deep learning with the dataset that will be used in this research and found that tenure, customer demographics, charges and contract information are important for predicting customer churn in the telecom industry. The importance of tenure is supported by Saba et al. (2017) and Šarić et al. (2018). Contract length appears also an important feature in predicting customer churn (Jain et al., 2020, Christianti et al., 2020. Among customers who have a month-to-month contract, the churn rate is higher than customers who have a yearly contract. Companies need to stimulate customers to conclude a yearly contract to minimize customer churn. Usage patterns and monthly charges are as important to predict customer churn (Saba et al., 2017, Amin et al., 2017)

4 METHOD

Figure 1 shows the flowchart of the data analysis which visualizes the data analysis step-by-step. First the dataset will be imported and preprocessed before the data exploration and feature selection can start. Afterwards the

dataset will be divided into three scenarios. The data analysis is done in three scenarios. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables but fits with and without personal information. The colors in the flowchart refers to the different sample sets. Orange refers to the training set. Yellow is the validation set and green is done with the test set.

4.1 *Dataset and preprocessing*

The dataset is called "Sample Telco Customer Churn Dataset" on Kaggle and is based on the original dataset from IBM Business Analytics community. The dataset consists of information on a fictitious company in the telecommunications industry that offers home phone and internet services to 7043 customers. More information about this dataset can be found on the websites of Kaggle and IBM Business Analytics community.

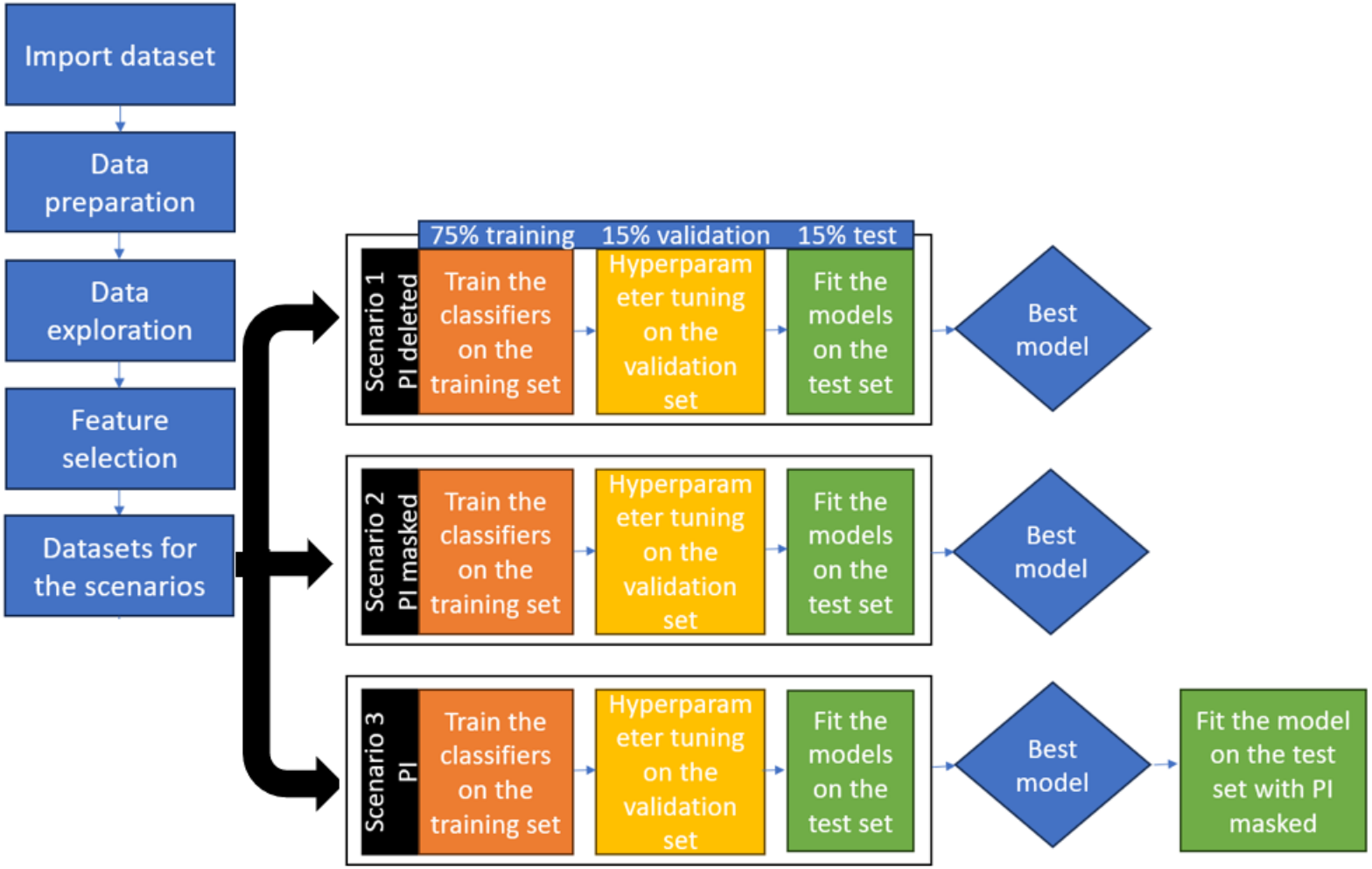

Figure 1: Figure 1 shows the flowchart of the data analysis. First the dataset will be imported and preprocessed before the data exploration and feature selection can start. Afterwards the dataset will be divided into three scenarios. The data analysis is done in three scenarios. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables but fits with and without personal information. The colors in the flowchart refers to the different sample sets. Orange refers to the training set. Yellow is the validation set and green is done with the test set.

When downloading, it is divided into two csv files with 7012 and 21 observations. We import these files and combine them into one dataset. There are 21 variables which relate to personal information, contract, extra services and churn. Personal information that is collected is related to customer identification number, gender, whether the customer is 65+, whether the contract is resold by their partner and if the customer lives with other people. Churn status is provided in which 1 means that the customer churned and moved to another company and 0 relates to customers who are loyal to the company and stayed with them. More detailed descriptions of the variables can be read on Kaggle.

There are 1520 missing values in the binary variables related to offering an extra service which are OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingMovies and StreamingTV. These data points are related to each other which means that if a customer misses one of the variables there are also missing the five other variables. Thus, missing values are not missing at random and will therefore not be imputed. It is not sure why these variables are missing because there are no relations found with other variables such as contract type. An explanation is that it is not applicable for these customers due to the type of device. For example, the television that the customer has does not support the streaming service or online security service from the provider. Another explanation is that the company knows this but the customer does not agree with storing this information.

There is class imbalance with 5163 customers who made the choice to stay with the company and 1869 customers who decided to churn. This makes the churn rate 26.58% and is comparable to churn rates of Vodafone in the European markets which are 15.5 to 31.5%.

All variables, except MonthlyCharges and Tenure, are falsely classified as integers but there are no relations between these different values. Therefore, the data types of these variables are changed to categories.

4.2 *Exploratory data analysis (EDA) and feature selection*

Appendix A shows the tables and figures belonging to the EDA. EDA is used to indicate to what extent personal information is important in customer churn prediction. First, the difference in churn percentage in the variables with personal information needs to be calculated. If a customer resides with any dependents, the customer is less likely to leave the company with a churn percentage of 15.5312%. Customers who live alone have a churn percentage of 31.2791%. Looking at gender, there is not a large difference between females and males. Females just have a slightly higher churn rate than males which is so small that it can be discarded. If their

contract is resold by their partner (now or earlier), the churn percentage is 32.9761% which is much higher than customers who have a contract that is never resold by their partner who have a churn percentage of 19.7171%. Seniors are more likely to churn with a churn percentage of 23.6503% than customers who are younger than 65 years who have a churn percentage of 41.6813%. These differences are also seen within the categories of the variables. The most noticeable difference in the churn percentage of seniors and non-seniors is the category mailed cheque within the variable PaymentMethod. The churn percentage of the customers is 19.02020% in this category. But when it becomes known that the customer is a senior, the churn percentage doubles to 46.8085%. It is also important to look at the difference between the longevity of the customer's journey with the company. On average a churned customer stays with the company for 17.9791 months. Customers who have a partner resold their contract are more likely to stay longer with the company. On average a churned customer pays monthly 74.4413 to the company. On average seniors pay 8 euros more than non senior customers. Customers who have a partner resold their contract are more likely to pay more than customers who did not have this. The difference for Dependents and gender is much smaller in terms of monthly payment.

In the EDA, some variables were not useful such as TotalCharges and customerID. Therefore, these variables are deleted. TotalCharges is the total amount of money paid by the customer and thus depends on Tenure (in-contract term) and MonthlyCharges.

4.3 *Experimental procedure*

4.3.1 *Scenarios*

The goal of this thesis is to find which machine learning model is the best for predicting customer churn when personal information is not available. The first scenario is to optimize without personal information to indicate which model would be the best if personal information is deleted from the beginning of the process. The best model in this scenario would be the model with the highest F2-score on the test set.

The second scenario is the scenario in which the models will be trained upon a dataset in which the variables with personal information are uninformative. This means there is a second scenario in which we impute a general value replacing the values of the personal information variables. In this case, "-99" is used because this number is outside of the distribution of the dataset. Real personal information will be masked from the model.

The third scenario is to find which model is the best when it is trained on personal information but tested when personal information is masked. In this step, the best model is indicated as the model with the least deviation in F2-score from going from a test set with personal information to a test set with masked personal information. First, the models would be optimized on the complete dataset. Then these models will be tested on the test set with personal information and the test set with masked personal information to calculate the deviation in the F2-score.

Appendix B shows a complete oversight of every variable in each scenario. Here it is seen which variable is in each training, validation and test set. Idem means that it is the same list as above.

4.3.2 *Model comparison*

We chose the models from Lalwani et al. (2022). They used the same dataset and came to the conclusion that these models are the best in predicting customer churn with personal information. Therefore, it seems likely that these models would predict customer churn the most accurately without personal information. The following explanations are based on the study of Lalwani et al. (2022) and J. Han et al. (2022).

In the AdaBoost classifier, training observations are initially given equal weights. For every variable a stump (node + 2 leaves) is created. With each stump it is determined how likely it is that the customer will churn. Stumps predict customer churn, and the one with the lowest error becomes the first stump. Misclassified observations are given more weight than others. After this stump, sampling with replacement of size n weighted by observation weights is done. This means that the second stump focuses more on misclassified observations. Stumps are also given weights in terms of prediction quality. Then, whether a customer is predicted to churn depends on majority voting according to those weights. Now it uses the DecisionTreeClassifier as a base estimator but it can also use the RandomForestClassifier and the ExtraTreesClassifier as base estimators. Just like AdaBoost, XGBoost is an ensemble boosting technique in which multiple decision trees are trained sequentially. XGBoost uses more difficult weak learners and does not give weights to training data but optimizes model parameters with a loss function. CatBoost is the same as XGBoost but builds symmetric trees which are more balanced.

4.3.3 *Evaluation metric*

The goal is to find a machine learning model in which the total number of churners is predicted correctly. Customers who are predicted to stay but churned cost the companies the most money. But it is also costly to

spend money on trying to retain customers who never planned to leave. In short, minimizing false negatives is more important than false positives. Recall focuses on minimizing the customers that are predicted to stay loyal to the company but eventually churned. Precision focuses on minimizing the amount of customers that are predicted to churn but were actually loyal customers. The main evaluation metric is the F2-score because it favors Recall more than Precision. Furthermore, Recall and Precision are calculated because the F2-score depends on these metrics.

4.3.4 *Out-of-sample generalization and resampling*

The dataset is split into 70-15-15 training-validation-test. This means that the largest part of the dataset will be used for training, namely 4922 observations. The validation and test split have each 1055 observations. The untuned models fit on the training set. Then, the hyperparameters are tuned on the validation set and evaluated. For each ML model, the hyperparameter selection with the highest F2-score on the validation set is chosen as the best selection for that model. Last step is running these tuned models on the test set and choosing which ML model is the best in terms of F2-score.

The training set is resampled using RandomOversampler and RandomUndersampler. Random sampling resamples the training set by making both classes, churned and loyal customers, equal. Random oversampling focuses on randomly duplicating instances of the churned class so that the amount of churned observations are the same as the loyal customers. With random undersampling, the focus lies on randomly removing loyal customers until the quantity of both classes are equal.

4.3.5 *Hyperparameter tuning*

The five classifiers are tuned on different hyperparameters. The hyperparameters that are tuned are the number of estimators, maximal depth of the tree and the learning rate of the classifier. There is chosen to do a grid search in which all combinations of hyperparameters are tested. n_estimators refers to the maximal number of estimators at which boosting is terminated in the AdaBoostClassifier, CatBoostClassifier and the XGBoostClassifier while in terms of the RandomForestClassifier and the ExtraTreesClassifier, this refers to the number of trees. It is chosen to only tune the n_estimators in the AdaBoostClassifier because otherwise it took too much time. The default option of the number of estimators is 50 for the AdaBoostClassifier. In this thesis, the options that are compared are 25, 50 and 75. To avoid overfitting the maximal depth of the tree is tuned on 5, 10 and 20 but there is also an option to indicate that the maximal depth of the

tree should not be tuned which is the default option. The different learning rates that are compared are 0.5, 1 and 1.5 with 1 indicating as the default option. The learning rate refers to the weight applied to each classifier at each boosting iteration. When the learning rate is increased, this increases the contribution of each classifier. These hyperparameters are chosen for different reasons. The first one being that these hyperparameters have the same meaning in all of the five classifiers which makes the data analysis easier and equal. The second is that the maximal depth of the tree will make sure that each decision tree is not pruned to overfitting. Learning rate and the maximal number of boosting options focus on minimizing overfitting of the entire classifier. Another reason is that these hyperparameters are tuned in the literature so are proven to be the most important to tune. The hyperparameter that is mentioned in the literature is the minimal samples in an internal node to split the decision tree. This hyperparameter is not available in the XGBoostClassifier and the CatBoostClassifier, at least not in the same form as in the other classifiers. Therefore, it is chosen to not tune these hyperparameters. The tuned hyperparameters for each classifier in each scenario can be found in table 7.

5 RESULTS

5.1 *Data analysis*

Appendix B shows a large table of the data analysis. Table 1 is a summary of the three scenarios with the five classifiers whereas the best classifiers are shown in bold. Figure 1 is a visualization of this table.

Table 1: F2-scores of the five classifiers (after hyperparameter optimization) on the test set for the three models. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables, but personal data variables are masked in scenario 3, while scenario 3 (PI) is tested on the entirety of variables.

	Scenario 1	Scenario 2	Scenario 3	Scenario 3 (PI)
Adaboost	0.5456	0.5152	0.5432	0.5629
Adaboost RF	0.5372	0.5453	0.5597	0.5510
Adaboost Xtree	0.5485	**0.5527**	0.5572	0.5584
CatBoost	**0.5515**	0.5515	**0.5722**	**0.5808**
XGBoost	0.5487	0.5487	0.5361	0.5124

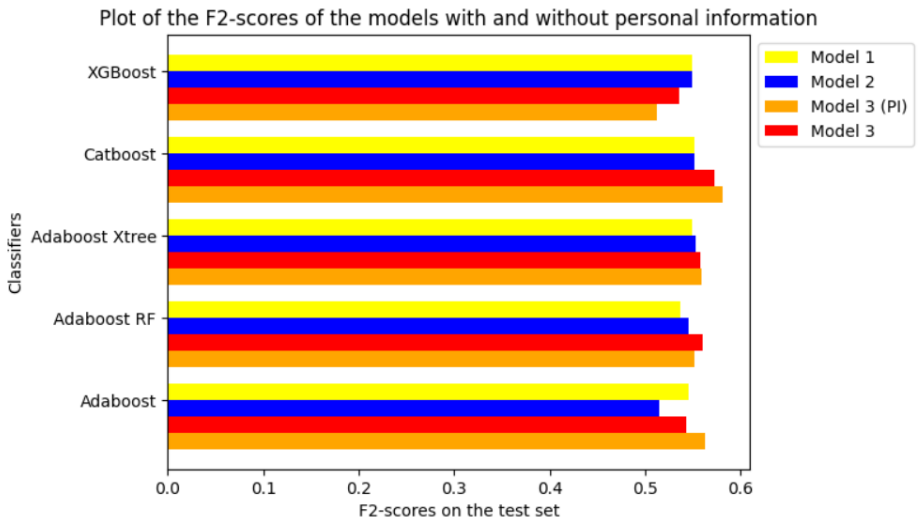

Figure 2: Plot of the F2-scores of the five classifiers (after hyperparameter optimization) on the test set for the three models. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables, but personal data variables are masked in scenario 3, while scenario 3 (PI) is tested on the entirety of variables..

5.1.1 *Scenario 1: Personal information is deleted*

For AdaBoost, the F2-score on the test set is 0.5456. AdaBoost RF reaches a F2-score on the test set of 0.5372. When AdaBoost switches to a base estimator of the ExtraTreeClassifier, the F2-score on the test set increases to 0.5485. CatBoost has a F2-score of 0.5515. The last classifier is the XGBoost classifier and ensures a F2-score of 0.5487.

The best classifier in this scenario is CatBoost because it has the highest F2-score out of the five classifiers.

5.1.2 *Scenario 2: Personal information is imputed by -99*

When adding uninformative personal variables, AdaBoost ensures a F2-score of 0.5152 on the test set. But this score is seen as a decrease when it is compared with the first model. Changing the base estimator to the RandomForestClassifier, the F2-score increases to 0.5453. AdaBoost Xtree reaches a F2-score of 0.5527 on the test set, which is a decrease in comparison with the first model. The F2-scores of the CatBoostClassifier and XGBoostClassifier did not change when personal information was imputed with the number "-99". This means that these classifiers are less sensitive to adding uninformative information than the AdaBoostClassifier.

In this scenario, it would be best to use AdaBoost Xtree. Thus, adding uninformative information changes which ML model is the best.

5.1.3 *Scenario 3: Model is optimized for personal information*

With personal information, AdaBoost responds differently to this dataset. The classifier is optimized with 0.5629 as a F2-score on the test set. This decreases with a percentage of 3.4955 to a F2-score of 0.5432 when the classifier is tested on the test set with uninformative personal information. AdaBoost- RF has a F2-score on the test set with personal information is 0.5510. This increases when personal information is imputed in the test set to 0.5597. This means that there is a 1.5833% increase when personal information is masked. When optimizing the model for AdaBoost Xtree,the F2-score on the test set with personal information is 0.5584 while this decreases to a score of 0.5572 when personal information is concealed. This results in a decrease of 0.2174%. On the test set with personal information, the F2-score of CatBoost is 0.5808. When personal information is masked, this decreases to 0.5722 which ensures a percentage decrease of 1.4829. If XGBoost is optimized, the F2-score on the test set with personal information is 0.5124. When personal information is unusable, this increases with 4.6171% to a F2-score of 0.5361.

When optimizing the model with personal information, the CatBoost-Classifier has the highest F2-score for scenario 3 and scenario 3 (PI). This indicated that this classifier is the best when fitting the model with and without personal information. Looking at the relative differences, the best classifier is XGBoost. This classifier has a positive percentage when going from a test set with personal information to a test set where personal information is imputed. This positive difference is not expected because the model is optimized with personal information.

5.2 *Feature importance*

The most important characteristics are tenure, monthly payment, whether the customer has technical support and which contract the customer has. Tenure was far out the most important of them with a feature importance. This means that how long the customer is with the company matters in the question whether the customer stays or leaves. Whether the customer also has a subscription on internet service matters the least in predicting customer churn. Furthermore, a subscription to streaming tv and movies are not that important in evaluating customer churn.

For the second scenario, tenure and monthly charges are again the most important features in predicting customer churn. But the payment method

Table 2: Feature importance per model in percentages. This is calculated on the basis of Gini importance. More information about the calculations can be found on scikit-learn contributors, 2022. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables, but personal data variables are masked in scenario 3, while scenario 3 (PI) is tested on the entirety of variables..

Variable	Scenario 1	Scenario 2	Scenario 3
Personal information			
Dependents		0.0	0.0023
gender		0.0	0.0
Partner		0.0	0.2745
SeniorCitizen		0.0	1.0917
Contract information			
Contract	10.2635	8.7816	16.5402
DeviceProtection	5.3794	5.2528	1.1352
InternetService	0.8002	3.9565	0.7675
MonthlyCharges	13.1122	13.4761	28.9468
MultipleLines	4.7880	3.9128	2.3877
OnlineBackup	4.6835	5.7756	3.7898
OnlineSecurity	4.4236	5.5314	8.5519
PaperlessBilling	6.7382	7.0444	2.8164
PaymentMethod	7.4698	10.1605	3.3607
PhoneService	3.5111	1.7772	2.8993
StreamingMovies	2.6521	5.5569	1.1497
StreamingTV	2.2485	4.6624	0.0
Techsupport	12.4030	5.8137	2.2977
Tenure	21.5270	18.2981	23.9886

is also relevant when predicting if a customer will leave the company. Subscription to the home phone service is not that relevant in predicting customer churn.

When optimizing the model with personal information, how much the customer pays monthly is very important in predicting whether this customer will switch to another company. Tenure is also an important feature in determining customer churn. Which contract a customer chooses is important too. Less relevant for predicting customer churn are the variables with personal information. Whether the customer is subscribed to streaming services, home phone service, internet service, multiple phone lines, device protection and technical support does not matter as much

in predicting churn. Furthermore, choosing paperless billing is not that relevant in predicting whether a customer will leave the company.

5.3 *Class balancing*

Table 3: Resampling the best algorithms to determine whether class balancing the training set will improve the F2-scores. This table shows the F2-scores of in the normal situation, oversampling the training set and undersampling the training set. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables, but personal data variables are masked in Scenario 3, while scenario 3 (PI) is tested on the entirety of variables.

Model	Normal	Oversampling	Undersampling
Scenario 1: CatBoost	0.5515	0.6036	0.6751
Scenario 2: AdaBoost Xtree	0.5527	0.6658	0.6973
Scenario 3: CatBoost	0.5722	0.7190	0.7313
Scenario 3 (PI): CatBoost	0.5808	0.7157	0.7336

For all scenarios, random sampling improves the F2 scores of the models. In the first scenario, the CatBoostClassifierr has the highest F2-score on the test set. Therefore, this model is resampled. When random oversampling the training set, the F2-score improved to 0.6036. Also random undersampling boosted the score to 0.6751. The model with the highest F2-score on the test set is the AdaBoost Xtree in the second scenario thus this model is resampled. When random oversampling the training set, the F2-score improved to 0.6036. But also random undersampling increased the score to 0.6751. The CatBoostClassifier also excels in the last scenario, securing the highest F2-score on the test set. When random oversampling the training set, the F2-score on the test set with personal information improved to 0.7157 and on the test set without personal information this score increased to 0.7190. Random undersampling boosts up the F2-scores on the test sets to 0.7336 and 0.7313.

5.4 *Disparate impact*

Disparate impact is measured by confusion matrices. True negatives are customers who are predicted to be loyal and stay loyal to the company. False negatives are those who are predicted to be loyal but eventually churned. False positives refer to customers who are predicted to churn but stay loyal to the company. True positives are customers who are predicted

to churn and switch to another company. First it is important to know how high the overall error rate is. This can be seen in figure 2. This error rate is calculated by summing the misclassified labels by the total number of observations in the test set which is 1055. In other words, false positives and false negatives are summed together and divided by 1055.

Figure 3 shows the confusion matrices with the error rates. When personal information is deleted there is an error rate of 22.7488%. If personal information is imputed, this decreases to 21.1374%. When the model is trained with personal information, the error rate decreases to 18.6730%. This means that with uninformative personal information the model makes less mistakes in comparison to when personal information is deleted. The best option is to train the models with personal information but test them without personal information.

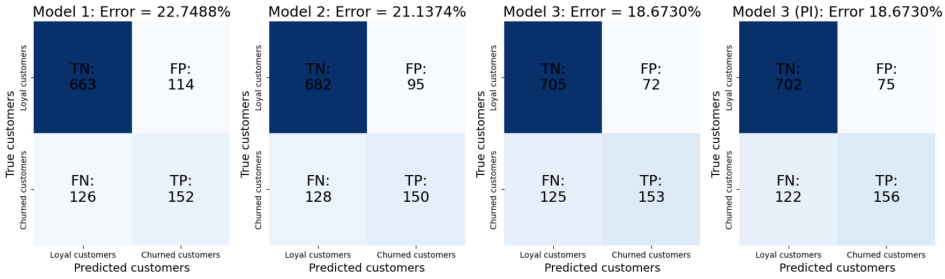

Figure 3: Confusion matrices: Error rates for each model in the test set

The confusion matrices for females and males can be found in figures 3 and 4. There are 532 females and 523 males in the test set. For the first model, this error rate is 25.3759% for females and 20.0765% for males. When personal information is imputed, the female error rate changes to 23.1203% and the male error rate decreases to 19.1205%. If the models are trained on personal data but tested without, the error rates decrease to 18.6090% and 18.7380%, respectively. If this is tested with personal information, these rates change to 18.9850% and 18.3556%. For scenario 1, 2 and 3 (PI), the model makes more mistakes for females than for males. For scenario 3, the model slightly makes less mistakes for females. Appendix 3 shows the gender differences in variables. There are no clear gender differences between variables. Thus we are currently unable to identify why the model performs better for males in scenario 1, 2 and 3 (PI) and performs worse in scenario 3. .

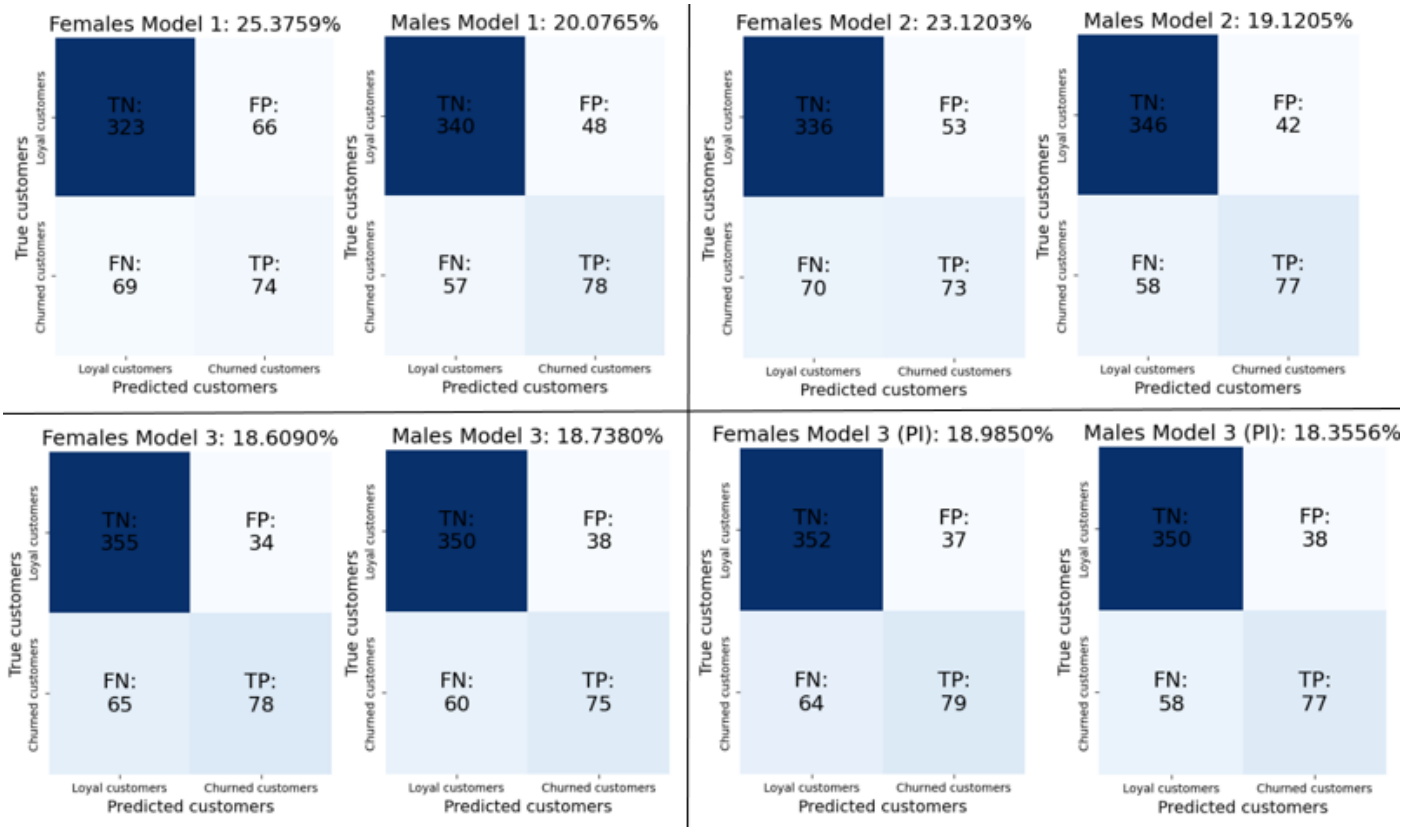

Figure 4: Confusion matrices: Error rates for each model in terms of 532 female customers in the test set and 523 male customers in the test set. True positives are customers who are predicted by the model to leave the company and also churned. False positives are customers who are predicted by the model to leave the company but stay loyal to the company. True negatives are customers who are predicted by the model to stay with the company and also stay loyal to the company. False negatives are customers who are predicted to stay loyal to the company but churned anyway. Scenario 1 refers to the situation in which the variables with personal information are removed. The second scenario deals with the scenario where the variables with personal information are masked. The final scenario is trained on all variables, but personal data variables are masked in scenario 3, while scenario 3 (PI) is tested on the entirety of variables. Error rate is calculated by the amount of errors that the model makes divided by the number of females or males in the test set.

6 DISCUSSION

There is a lot of discussion on whether a company can ask for personal data. Companies can better predict customer churn and can allocate their marketing resources better. But, collecting more personal data can have a negative impact on the customers due to their lack of privacy. The goal of this thesis is exploring whether personal information can be deleted from models that predict customer churn in the telecommunications industry.

In this thesis, there are three scenarios. The first one is the scenario in which personal information is deleted. In the second scenario, the personal information is imputed with "-99" which makes this information useless. The last scenario is to optimize the model for personal information but fit the model with personal information masked.

There are five classifiers that are compared: AdaBoostClassifier, AdaBoostClassifier with RandomForestClassifier, AdaBoost with ExtraTreesClassifier, CatBoostClassifier and the XGBoostClassifier. From the results, it is clear that deleting personal information from the model has just a minor negative effect on the predictive model. The best option would be to first train the CatBoostClassifier with personal information and then test the model with a dataset with uninformative personal information. If a company does not want to use personal information to train the model, it would be best to make the personal information uninformative and use the AdaBoostClassifier with ExtraTreesClassifier as base. Deleting personal information in the first scenario is the worst option. This is also seen in the study of Lalwani et al. (2022). In this study, the CatBoostClassifier is the best classifier when models are optimized with personal information.

Resampling the training set by random undersampling and random oversampling improves the best models. The best model for each scenario achieves a higher F2-score when it is compared to the normal situation. This indicated that class balancing resulted in better F2-scores than in the normal scenarios. This is also seen in the research of Shumaly et al. (2020).

When the models are optimized for missing personal information, the error rates are higher. This means that the first and second model makes more mistakes than when the model is trained upon personal information. When looking at the difference in mistakes for females and males, the first and second scenario make more mistakes for females than males, For the third model this is only the case when the model is tested with personal information. If the model is tested without personal information, the model makes more mistakes for males than females. It is not clear why there are gender differences in error rates.

The biggest limitation is that directly comparing the test set with personal information and the test set with deleted personal information

is not possible. Therefore, the thesis needed to add a second situation in which the personal information becomes uninformative. Besides this, the models and train-validation-test split are set by a random state. The models can behave differently when another random state is used. Furthermore, there are only 27 combinations of hyperparameters trained due to lack of time. The correct pair of hyperparameters does not have to be in one of the 27 combinations. Another limitation is that this thesis only focuses on five different classifiers thus ignorant to other classifiers. This means that not all possible classifiers are not trained and thus there is a chance that another classifier performs even better. The last limitation is that the outcome depends on the dataset that is used in this thesis. Other datasets can have other variables or have customers that are very different from the customers in this dataset.

For further research, it is better to look at the difference in performance in other personal information.

7 CONCLUSION

There is a lot of discussion on whether a company can ask for personal data. Companies can better predict customer churn and can allocate their marketing resources better. But, collecting more personal data can have a negative impact on the customers due to their lack of privacy. The goal of this thesis is exploring whether personal information can be deleted from models that predict customer churn in the telecommunications industry. Models that predict customer churn in the telecommunications sector use personal information. Due to stricter privacy legislation this is increasingly difficult to maintain. This thesis focuses on researching whether models without personal information will also be sufficient. Furthermore, this thesis will suggest which algorithm companies can use to predict customer churn in the case of minimizing personal information. There are five algorithms that will be compared: AdaBoostClassifier with DecisionTreeClassifier, RandomForestClassifier or ExtraTreeClassifier as the base estimator, CatBoostClassifier and the XGBoostClassifier. The dataset that is used is the "Sample Telco Customer Churn Dataset" from Kaggle. There are several scenarios that minimize personal information. The preferred scenario is to train the model with the so far collected personal information and then fit the model with uninformative personal information. This is best done with the CatBoostClassifier. If personal information may not be used at all, it is better to make the personal information uninformative and use the AdaBoostClassifier with the ExtraTreeClassifier as the base. This option is slightly better than completely removing personal information from the dataset using CatBoostClassifier. The difference between the

models with and without personal information is very small. Thus, the conclusion is that personal information of the customer does not add much value to the model to predict customer churn. In this case, companies can make better use of models without personal information.

REFERENCES

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning and social network analysis in big data platform. *arXiv preprint arXiv:1904.00690*.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. *Neurocomputing*, *237*, 242–254.

Christianti, D., Abdullah, S., & Nurrohmah, S. (2020). Bayes risk post-pruning in decision tree to overcome overfitting problem on customer churn classification.

Chua, H. N., Ooi, J. S., & Herbland, A. (2021). The effects of different personal data categories on information privacy concern and disclosure. *Computers & Security*, *110*, 102453.

Fujo, S. W., Subramanian, S., Khder, M. A., et al. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 24.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan kaufmann.

Han, S., Yuan, B., & Liu, W. (2009). Rare class mining: Progress and prospect, 1–5.

Jain, H., Yadav, G., & Manoov, R. (2020). Churn prediction and retention in banking, telecom and it sectors using machine learning techniques, 137–156.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: A machine learning approach. *Computing*, 1–24.

Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Sharma, M. M. (2021). A hybrid machine learning model for predicting customer churn in the telecommunication industry, 458–468.

Saba, M., Shaikh, Z., & Javed, S. (2017). Autonomous toolkit to forecast customer churn. *International Journal of Current Research*, *9*, 62999–63006.

Šarić, M., Hubana, T., & Begić, E. (2018). Fuzzy logic based approach for faults identification and classification in medium voltage isolated distribution network, 44–54.

scikit-learn contributors. (2022). *Scikit-learn documentation* [Accessed: 2024-01-09]. https://scikit-learn.org/stable/

Shumaly, S., Neysaryan, P., & Guo, Y. (2020). Handling class imbalance in customer churn prediction in telecom sector using sampling techniques, bagging and boosting trees, 082–087.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9.

Weiss, G. M. (2010). Mining with rare cases. *Data mining and knowledge discovery handbook*, 747–757.

APPENDIX A: EXPLORATORY DATA ANALYSIS

Table 4: Churn Percentage Based on Various Personal Variables

				Churn % for personal information							
Variables	Categories	Count	Churn %	Female	Male	No senior	Senior	No partner	Partner	No dependents	Dependents
Churn	No / loyal customer	5163									
	Yes / churned customer	1869	26.5785								
Personal information											
Dependents	No dependents	4933	31.2791								
	Dependents	2099	15.5312								
gender	Female	3483	26.9595								
	Male	3549	26.2046								
Partner	No partner	3639	32.9761								
	Partner	3393	19.7171								
SeniorCitizen	No senior	5890	23.6503								
	Senior	1142	41.6813								
Contract information											
Contract	Month	3875	42.7097	43.7403	41.6923	39.5698	54.6468	44.6894	39.1304	45.2366	32.8264
	Year	1472	11.2772	10.4457	12.0690	10.6864	15.2632	10.5754	11.8215	12.4204	9.2453
	Two Years	1685	2.8487	2.6190	3.0769	2.7273	4.1379	3.3932	2.6182	3.3149	2.3077
DeviceProtection	No	3094	39.1403	40.0520	38.2391	36.0729	51.2821	44.9111	30.4770	43.0316	26.2865
	Yes	2418	22.5393	22.2686	22.8056	20.1844	32.4034	28.3417	18.4821	26.0299	14.6703
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
InternetService	No	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
	DSL	2416	18.9983	18.4966	19.4805	17.6634	30.1158	25.6390	11.8557	22.4629	12.0000
	Fiber	3096	41.8928	42.7560	41.0240	39.9117	47.2924	49.6875	33.5562	44.9877	30.5136
MultipleLines	No	4065	25.0677	26.1762	23.9961	22.9376	41.0901	30.5402	17.5613	29.3992	15.3355
	Yes	2967	28.6485	28.0135	29.2848	24.7611	42.1053	37.4224	21.9178	33.7736	15.8205
OnlineBackup	No	3087	39.9417	40.8190	39.0973	36.7570	52.7687	45.2563	32.1628	43.4382	28.0627
	Yes	2425	21.5670	21.6680	21.4642	19.2919	30.8824	28.0943	16.8444	25.3453	13.2895
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
OnlineSecurity	No	3497	41.7787	42.0899	41.4798	39.1967	50.3713	47.1281	34.3151	44.7349	30.8210

Table 4 – Continued from previous page

				Churn % for personal information							
Variables	Categories	Count	Churn %	Female	Male	No senior	Senior	No partner	Partner	No dependents	Dependents
	Yes	2015	14.6402	15.8203	13.4208	13.3295	22.6950	19.1411	11.5833	17.4383	9.5967
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
PaperlessBilling	No	2864	16.3757	17.1674	15.6207	15.0500	29.3233	21.0562	11.5220	19.6730	10.4956
	Yes	4168	33.5893	33.5252	33.6534	30.4374	45.4338	40.9445	25.5159	38.1536	20.3738
PaymentMethod	Bank	1542	16.7315	17.2808	16.1589	15.6608	22.7468	21.0692	13.6865	19.4664	11.5094
	Credit	1521	15.2531	17.4434	13.1169	13.0769	28.0543	21.1078	10.6682	18.9655	8.4112
	Electronic cheque	2365	45.2854	44.6154	45.9414	42.5748	53.3670	50.8076	37.7866	48.6214	32.1503
	Mailed cheque	1604	19.2020	19.3548	19.0591	17.4834	46.8085	23.9466	11.8859	23.0696	11.8919
PhoneService	No	680	25.0000	24.3161	25.6410	21.8750	42.3077	29.3801	19.7411	30.8824	11.2745
	Yes	6352	26.7475	27.2353	26.2664	23.8427	41.6185	33.3843	19.7147	31.3215	15.9894
StreamingMovies	No	2781	33.7289	34.8208	32.6733	31.1024	45.8586	39.7319	25.0877	37.5603	22.4894
	Yes	2731	29.9524	29.7080	30.1984	26.8727	41.0084	38.3154	23.2895	34.3623	18.4106
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
StreamingTV	No	2809	33.5351	34.7042	32.3963	30.5543	46.7181	39.1957	25.2843	37.0308	22.6138
	Yes	2703	30.1147	29.7557	30.4734	27.4519	40.0350	39.0388	23.1378	34.8595	18.4379
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956
TechSupport	No	3472	41.6475	41.9148	41.3872	38.8342	50.6024	47.4346	33.8949	45.0018	29.6978
	Yes	2040	15.1961	16.1133	14.2717	14.5506	19.6154	20.0231	11.6497	17.7745	10.2710
	NA	1520	7.4342	7.5067	7.3643	7.3569	9.6154	10.6734	3.9563	9.6263	4.3956

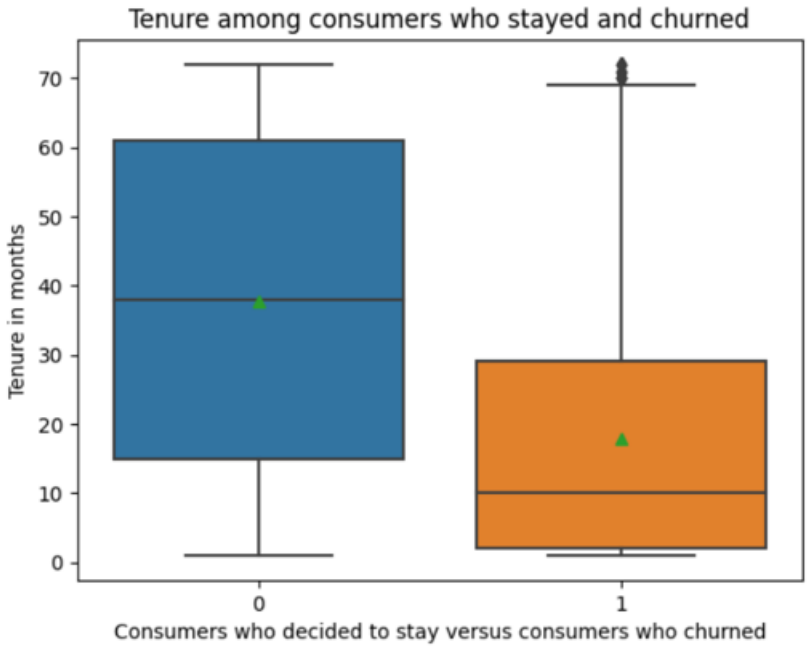

Figure 5: Exploratory data analysis of the variable Tenure. Box plots with the difference between loyal and churned customers in terms of Tenure. Tenure is how long the customer has been a customer at the company.

	Minimum	1st quartile	Mean	Median	3rd quartile	Maximum
Female	1	2	17.0043	9	27.5	72
Male	1	2	18.9634	10	31	72
No senior	1	2	16.9354	9	25	72
Senior	1	3	21.0336	15	35	72
No partner	1	1	13.1767	6	18	71
Partner	1	7	26.5934	21	43	72
No dependents	1	2	17.1231	9	27	72
Dependents	1	4	22.0307	16	35	72

Table 5: Exploratory data analysis of the variable Tenure. This table shows the summary statistics of Tenure which is how long the customer has been a customer with the company. This table is only focused on the churned customers. The loyal customers have been disregarded, as it is important to clarify the differences due to personal information.

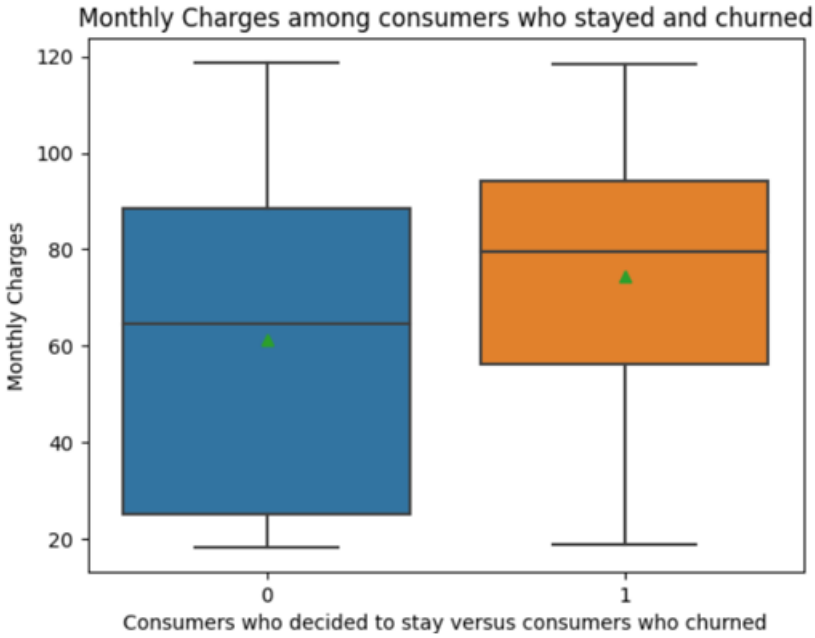

Figure 6: Exploratory data analysis of MonthlyCharges. Box plots with the difference between loyal and churned customers in terms of MonthlyCharges. MonthlyCharges is how much the customer pays monthly to the company.

	Minimum	1st quartile	Mean	Median	3rd quartile	Maximum
Female	19	63.1250	74.8121	79.65	93.75	117.45
Male	18.85	54.45	74.0670	79.625	94.6375	118.35
No senior	18.85	53.85	72.2981	78.1	91.85	118.35
Senior	19.45	73.625	80.7134	84.8250	95.7125	117.45
No partner	18.85	53.475	71.4510	75.85	90	118.35
Partner	19	70.1	79.8052	84.95	98.75	117.80
No dependents	18.85	59.425	74.7741	79.65	94.275	118.35
Dependents	19	55.2625	72.8661	79.5	93.375	114.2

Table 6: Exploratory data analysis of MonthlyCharges. This table shows the summary statistics of MonthlyCharges which is how much a customer needs to pay each month to the company. This table is only focused on the churned customers. The loyal customers have been disregarded, as it is important to clarify the differences due to personal information.

APPENDIX B: SUMMARY DATA ANALYSIS

	Scenario 1	Scenario 2	Scenario 3	Scenario 3 (PI)
Training set	Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges	gender (-99), SeniorCitizen (-99), Partner (-99), Dependents (-99), Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges	gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges	gender, SeniorCitizen, Partner, Dependents, Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges
Validation set	idem	idem	idem	idem
Test set	idem	idem	gender (-99), SeniorCitizen (-99), Partner (-99), Dependents (-99), Tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges	idem

Figure 7: Variables in each training, validation and test set

Table 7: Summary of the classifiers for each model

		Evaluation metrics and Hyperparameter selection						
Models	Classifiers	F2-score validation set	F2-score test set	Precision	Recall	n estimators	learning rate	max depth
Model 1	AdaBoost	0.4937	0.5456	0.5409	0.5468	75	0.5	5
	AdaBoost RF	0.4914	0.5372	0.5911	0.5252	50	1.5	5
	AdaBoost Xtree	0.4864	0.5485	0.6447	0.5288	50	0.5	5
	CatBoost	**0.5355**	**0.5515**	**0.5714**	**0.5468**	**50**	**1.5**	**10**
	- Oversampling		**0.6036**	**0.5389**	**0.6223**			
	- Undersampling		**0.6751**	**0.4794**	**0.7518**			
	XGBoost	0.5185	0.5487	0.5720	0.5432	25	1	5
Model 2	AdaBoost	0.4934	0.5152	0.5338	0.5108	25	1	5
	AdaBoost RF	0.4989	0.5453	0.6041	0.5324	50	1.5	5
	AdaBoost Xtree	**0.4868**	**0.5527**	**0.6122**	**0.5396**	**75**	**1**	**5**
	- Oversampling		**0.6658**	**0.5280**	**0.7122**			
	- Undersampling		**0.6973**	**0.4887**	**0.7806**			
	CatBoost	0.5355	0.5515	0.5714	0.5468	50	1.5	10
	XGBoost	0.5185	0.5487	0.5720	0.5432	25	1	5
Model 3	AdaBoost	0.5357	0.5432	0.6100	0.5288	25	1	10
	AdaBoost RF	0.4756	0.5597	0.6804	0.5360	25	0.5	5
	AdaBoost Xtree	0.4951	0.5572	0.6622	0.5360	25	0.5	5
	CatBoost	**0.5182**	**0.5722**	**0.6800**	**0.5504**	**25**	**0.5**	**5**
	- Oversampling		**0.7190**	**0.5396**	**0.7842**			
	- Undersampling		**0.7313**	**0.5159**	**0.8165**			
	XGBoost	0.5250	0.5361	0.5676	0.5288	25	1.5	5
Model 3 (PI)	AdaBoost	0.5357	0.5629	0.6194	0.5504	25	1	10
	AdaBoost RF	0.4756	0.5510	0.6407	0.5324	25	0.5	5
	AdaBoost Xtree	0.4951	0.5584	0.6292	0.5432	25	0.5	5
	CatBoost	**0.5182**	**0.5808**	**0.6753**	**0.5612**	**25**	**0.5**	**5**
	- Oversampling		**0.7157**	**0.5371**	**0.7806**			
	- Undersampling		**0.7336**	**0.5158**	**0.8201**			
	XGBoost	0.5250	0.5124	0.5512	0.5036	25	1.5	5

APPENDIX C: DIFFERENCE IN GENDER IN TERMS OF CATEGORIES

Table 8: Differences in gender

Variables	Categories	Count	Female	Male
Contract	Month	3875	49.6774	50.3226
	Year	1472	48.7772	51.2228
	Two Years	1685	49.8516	50.1484
DeviceProtection	No	3094	49.7091	50.2909
	Yes	2418	49.5864	50.4136
	NA	1520	49.0789	50.9211
InternetService	No	1520	49.0789	50.9211
	DSL	2416	49.0066	50.9934
	Fiber	3096	50.1615	49.8385
MultipleLines	No	4065	49.1513	50.8487
	Yes	2967	50.0506	49.9494
OnlineBackup	No	3087	49.0444	50.9556
	Yes	2425	50.4330	49.5670
	NA	1520	49.0789	50.9211
OnlineSecurity	No	3497	48.9848	51.0152
	Yes	2015	50.8189	49.1811
	NA	1520	49.0789	50.9211
PaperlessBilling	No	2864	48.8128	51.1872
	Yes	4168	50.0240	49.9760
PaymentMethod	Bank	1542	51.0376	48.9624
	Credit	1521	49.3754	50.6246
	Electronic cheque	2365	49.4715	50.5285
	Mailed cheque	1604	48.3167	51.6833
PhoneService	No	680	48.3824	51.6176
	Yes	6352	49.6537	50.3463
StreamingMovies	No	2781	49.1549	50.8450
	Yes	2731	50.1648	49.8352
	NA	1520	49.0789	50.9211
StreamingTV	No	2809	49.3414	50.6586
	Yes	2703	49.9815	50.0185
	NA	1520	49.0789	50.9211

Table 8 – Continued from previous page

Variables	Categories	Count	Female	Male
	No	3472	49.3376	50.6624
TechSupport	Yes	2040	50.1961	49.8039
	NA	1520	49.0789	50.9211