THE PREDICTION OF MEMBERSHIP CHURN IN A LARGE GYM CLUB CHAIN

A MACHINE LEARNING APPROACH

MAYRA VAN DER ZANDEN

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

2027859

COMMITTEE

dr. Fred Blain
dr. Dan Stowell

LOCATION

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

DATE

December 4, 2023

WORD COUNT

7358

ACKNOWLEDGMENTS

I want to express gratitude to Dr. Fred Blain for the constructive feedback and support throughout the thesis process. In addition, I want to thank CM.com for allowing me to conduct my research based on their data. In particular, I want to express appreciation to my external supervisor Saskia Dekkers for her guidance and help.

THE PREDICTION OF MEMBERSHIP CHURN IN A LARGE GYM CLUB CHAIN

A MACHINE LEARNING APPROACH

MAYRA VAN DER ZANDEN

Abstract

Engaging in physical activity has numerous physical and psychological benefits. Gym facilities can significantly influence people's activity behavior and promote a healthy lifestyle. However, a great concern of the fitness industry is the lack of long-term commitment. Therefore, it is important to identify which gym members are going to cancel their membership as this creates the possibility to efficiently focus effort on retaining this group. The current study used a Logistic Regression (LR) and a XGBoost algorithm on a dataset from a large gym club chain to predict whether customers cancelled their membership within the timeframe of a year. The results indicated that the LR was able to predict cancellation with a F1-score of 64% and the XGBoost outperformed this by achieving a score of 78%. Additionally, visit frequency was identified as the most important predictor as higher visit frequency decreased the chances of cancellation. This provides practical usefulness for gym facilities, as it aids in establishing effective strategies to increase the satisfaction and experience of gym members at risk of cancelling their membership. This, ultimately, prevents both membership cancellation and the abandonment of physical activity. Furthermore, the limitations of this research and suggestions for future research are discussed.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The data used in this thesis belongs to a gym club chain, which will be refered to as 'GymClubX, for confidentiality reasons. It was provided to me by the company CM.com and I have been given permission to use this data for the purpose of writing my master thesis. The original owner of the data retains ownership of the data during and after the completion of this thesis. I acknowledge that I do not have any legal claim to this data. The obtained data is anonymised. For confidentiality reasons, the data and created code are not publicly accessible or shared. All text, figures, tables

and code in the present research are self-made. For the code, ChatGPT was used as a supervision tool to resolve coding errors (OpenAI, 2023).

2 INTRODUCTION AND SOCIETAL RELEVANCE

Engaging in physical activity has numerous benefits, including the promotion of physical and psychological health and overall well-being (Park et al., 2020). People who are insufficiently engaging in physical activity have a 20 to 30 percent higher increased risk of all-cause mortality compared to people that are active for 30 minutes every day (World Health Organization, 2022). This makes physical inactivity responsible for 6% of global mortality, making it the fourth leading risk factor for mortality rates worldwide (World Health Organization, 2019). Therefore, increasing physical health through motivating physical activity is important to increase overall public health (Park et al., 2020). Gym facilities have the potential to significantly influence people's activity behavior and promote a healthy lifestyle (MacIntosh and Law, 2015). However, globally a growing concern for the fitness industry is the lack of long-term commitment of consumers, as only 50% of gym members continued their membership after one year (MacIntosh and Law, 2015).

Therefore, the cancellation of sport memberships is, besides a business problem, also a problem for physical health (Clavel San Emeterio et al., 2019). To address this issue, ensuring customer satisfaction with the gym experience is essential, as it significantly influences membership retention (Macintosh and Doherty, 2007; MacIntosh and Law, 2015; Yi et al., 2021).

To establish this, firstly it is important to identify and predict which customers are going to churn (Jain et al., 2021). In other words, predict which members are going to cancel their gym membership. With this information it is possible to focus efforts on retaining this group, by improving their satisfaction with the sport facility, in order to maintain their membership, and thus promote their physical activity (Jain et al., 2021).

2.1 *Research Question*

Based on the abovementioned challenge, this thesis aims to answer the following research question:

> *To what extent can sport membership churn be predicted to enhance the experience and satisfaction of members?*

3 RELATED WORK

In academic literature, the prediction of churn with machine learning is a widely investigated topic for different industries (Ahmad et al., 2019; Geiler et al., 2022; Jain et al., 2021; Lalwani et al., 2022; Pamina et al., 2019). Also for the fitness industry, churn has been a topic of interest (Clavel San Emeterio et al., 2019; Sobreiro et al., 2021; Yi et al., 2021).

Study by Clavel San Emeterio et al. (2019) investigated the churn rate of members at three different gyms in Spain, using a Logistic Regression model (LR). Results showed that the model was able to identify the members that would churn within one year of membership with 70% accuracy. Important variables were age and visit frequency, as with both higher age and visit frequency the churn rate decreased significantly.

Another study also found that visit frequency was the most important predictor of churn in gyms, next to the membership duration (Sobreiro et al., 2021). This study focused on one gym, located in Lisbon, for the prediction of churn within a timeframe of approximately one year. Results indicated that the Gradient Boosting Classifier (Friedman, 2001) performed best, with an accuracy of 95%. This study also used the LR algorithm, which obtained an accuracy of 78%.

These studies both identified the importance of visit frequency, and the commonality of these studies is that they investigated a small number of gyms. Study by Yi et al. (2021) addressed the latter by investigating the churn rate for 50 fitness facilities, located in South Korea, with a LR model. Their model obtained an accuracy of 77% in the prediction of dropouts within one year. This study also identified that visit frequency and age had a significant impact on the churn rate.

The abovementioned studies all used a LR model to evaluate the prediction of churn. Geiler et al. (2022) identified that this model performed best at predicting churn, based on the AUC score, compared to other machine learning models such as Random Forest, SVM and k-NN. In this study, the performance of 11 models was evaluated on 16 different churn datasets from various industries. Another algorithm that they identified for its good predictive power is the XGBoost algorithm.

The XGBoost algorithm was identified as best performing in the prediction of churn in the telecom industry (Ahmad et al., 2019; Geiler et al., 2022; Lalwani et al., 2022; Pamina et al., 2019). The previously mentioned study by Geiler et al. (2022) found that, for a telecom dataset, this algorithm obtained an AUC score of 94%. Additionally, Lalwani et al. (2022) obtained an AUC of 84%, hereby outperforming 13 other machine learning methods, including a LR model. However, despite its good performance on churn

prediction in the telecom industry, the XGBoost algorithm has not yet been applied in the fitness industry.

3.1 *Research Strategy and Research Sub-questions*

The current research aims to predict the cancellation of sport membership churn using machine learning methods and identifying important predictors. It will implement LR (Pedregosa et al., 2011) to evaluate its performance in the prediction of churn of sport memberships on a dataset of a large gym club chain. Additionally, the XGBoost algorithm (Chen and Guestrin, 2015) will be implemented to determine whether it outperforms the LR on the dataset of the current research. This research will contribute to the existing literature by evaluating the performance of the XGBoost algorithm and comparing it to LR. Additionally, this research will investigate whether the aspect of a distributed network of gym clubs adds to consumer utility which will be measured by the churn rate. It aims to establish an understanding in the added value of a distributed network of gyms in terms of flexibility and satisfaction, to investigate whether this promotes physical activity. This is formulated in the following sub-questions:

SQ1 *To what extent does having a distributed network of sport facilities increase consumer utility, measured by the churn rate?*

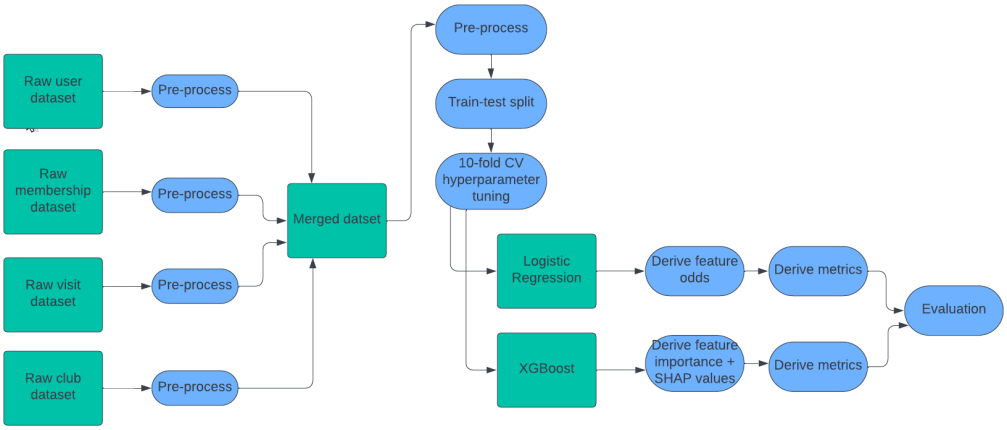
SQ2 *How does XGBoost perform compared to Logistic Regression in predicting sport membership churn?*

4 METHOD

4.1 *Experimental set-up*

The create an overview, Figure 1 illustrates the overall workflow of this thesis. The following Section 4 will elaborate on this.

Figure 1: Architecture of the workflow



4.2 *Data and Study Context*

The data was collected by operation and management systems of GymClubX and obtained from the company CM.com. It contains behavioral and demographic information from a real fitness club chain, that has 1462 locations throughout six countries in the West of Europe: France, The Netherlands, Belgium, Spain, and Luxembourg. All these facilities offer fitness activities, resistance training and cardio activities.

The raw dataset consists of four individual datasets containing anonymous user, visit, club, and membership data (see Appendix A, page 30 for the format). The user data was collected through online administration systems that were filled in by the members at registration and consisted of gender, date of birth, country, and language. The visit data was recorded through turnstiles at the entrance of each gym, and this contained the timestamp at a specified club connected to an anonymized user. The club and membership data were collected through manual administrative systems of GymClubX, and contained the location information of each club, the start and end date and type of the membership. Therefore, all this data is normally recorded in the management of GymClubX and thus were accessed without any interaction with the members.

The membership types consisted of three main categories that are linked to a monthly fee: 'Basic', 'Comfort' or 'Premium'. The 'Basic' subscription offers access to one chosen club for a price of 19,99 euros per month. The 'Comfort' subscription offers access to all the gyms of GymClubX located in a chosen country (€24,99 per month) and the 'Premium', €29,99 per month, offers access to all facilities of GymClubX throughout Europe. The

dataset consisted of all the members that had a membership at some point between August 2022 and ending with July 2023, for a minimum of 30 days. This timeframe was chosen with consideration of the COVID-19 virus which initiated government restrictions regarding gym visitations and this data would thus not be representative in terms of underlying and, within the scope of the current dataset, unmeasurable reasons to cancel a gym membership. The last restrictions, in the West- Europe, concerning public gym club usage were abolished in January of 2022 (Ministerie van Algemene Zaken, 2023). To minimize the impact of COVID-19 on the data and include a transitioning period, the timeframe starting from August 2022 is chosen.

The churn prediction analysis focused whether a member retained membership within the 12 months' timeframe. Churn was defined based on the contractual conditions of membership, meaning that it occurred when the predetermined length of a contract expired or when a member provided notice of their intention to cancel their subscription. In other words, when an end date was registered in the system occurring within the timeframe. The design of this prediction analysis was inspired by earlier studies conducted on the prediction of gym membership churn (Clavel San Emeterio et al., 2019; Sobreiro et al., 2021, see Section 3).

4.3 *Data Processing*

Data processing was done using Python (version 3.11.5, van Rossum and Drake, 2009) and libraries, such as such as Sklearn, Pandas and Numpy (McKinney et al., 2010; Pedregosa et al., 2011; Van Der Walt et al., 2011). The raw datasets (see Subsection 4.2) were retrieved from Microsoft Azure Cloud in .parquet files which were transformed into datasets usable for analysis using the Python (van Rossum and Drake, 2009) library Pandas (McKinney et al., 2010). From these datasets, important information is selected, engineered, and merged into one dataset that acts as input for the machine learning models: Logistic Regression and XGBoost. Considering the size of the sample size, the data was analysed on an online accessible computer (i.e. virtual machine), provided by the company 'CM.com', operating on a Windows system (version 11 Pro), with 4 vCPUs and 128 GiB memory capacity.

4.3.1 *Pre-processing of the User and Membership Dataset*

The first part of the data analysis consisted of the pre-processing of the user dataset merged with the membership dataset. This merge was conducted with an inner join, based on 'userId', to ensure that for each member there

is complete information about their demographics and membership. For this merged dataset, the data processing involved the following steps: (1) initial user dataset ($n = 14661631$); (2) specifying timeframe ($n = 4617895$); (3) removing missing values ($n = 4343357$); (4) removing incorrect values ($n = 3786776$); (5) final dataset ($n = 3786776$). The removing of missing values mostly consisted of deleting the users that had no membership, which indicates that these members were in the database because of a trail membership containing a certain number of visits. The removing of incorrect values consisted of deleting the instances where people took a consecutive membership, as this indicates that people only changed their type of membership and thus this cannot be seen as churn. Additionally, for the age variable the instances outside of the range 10 to 90 were deleted. Also, the users that contained incorrect administration errors were removed by, for instance, deleting the rows where the end date was before the start date.

This cleaned dataset was used for feature engineering for the variables: registration month and membership duration. The registration month was determined by analysing the original 'start date' variable and identifying the month of this date. The duration was calculated, according to existing literature, from start date till dropout date or, in absence of an end date, till the end of the recorded timeframe (i.e. August 1st 2023) (Clavel San Emeterio et al., 2019; Sobreiro et al., 2021). Churn was coded based on whether the member terminated their contract between August 2022 and August 2023.

4.3.2 *Pre-processing of the Visit and Club Datasets*

Further pre-processing involved the merging of the visit dataset with the club dataset. This merge was also conducted with an inner join, based on 'clubId', to ensure that the data for each visit consists of complete information about the location of the visit. This merged dataset was used for feature engineering for the variables: visit frequency, the number of distinct gyms (club count) and the use of international gyms (country count). The visit frequency variable was engineered by counting the number of unique visitId's per unique userId within the predefined timeframe. The total sum of the visits was stored in a new variable per member. By counting the number of unique clubId's within the visits per unique userId, the club count variable was engineered. Lastly, the country count was determined by counting the number of unique country codes within the visits per unique userId. Thus, visit frequency, club count and country count were all determined in terms of total values per user during the whole length timespan of interest.

For the club count and country count variables, further pre-processing consisted of binary transforming these variables, by determining the instances were the number of distinct gyms visited was more than one (a '1' reflecting more than one unique gyms, $n = 1156089$) and the number of distinct countries visited was more than one (a '1' reflecting more than one country, $n = 46335$). Consequently, the resulting variables reflect the use of multiple unique gym clubs and the use of international gyms, respectively. In the continuation of the analysis, these variables are referred to as: 'club > 1' and 'international'.

4.3.3 *Pre-processing of the Merged Dataset*

After the pre-processing of the user and membership dataset, this dataset was merged with the new features engineered from the visit and club datasets. This was done using a left join, based on the user and membership date, and adding the new variables to these members. This was done so that all the users that were a member within the specified timeframe were included, regardless of whether they went to the gym or not. Consequently, for the members without gym visits ($n = 702227$), the missing values from the new engineered features were transformed into the value '0'.

The last step of the pre-processing consisted of transforming all the variables into binary, also referred to as 'dummy', variables (resulting in k-1 columns per variable). The categorical variables were divided into exclusive categories based on their coded values. The scale variables were divided into exclusive categories based on the values their quartiles. This variable transformation was done with consideration of the large sample size, thus providing better interpretation and understanding of the results (Clavel San Emeterio et al., 2019). With the finalization of this step, the dataset was suitable to be fitted to the LR and XGBoost models.

4.4 *Dataset*

After pre-processing, the dataset consisted of 3786776 users (mean age = 30.21, std = 11.42) from France ($n = 2042372$), The Netherlands ($n = 806240$), Belgium ($n = 638893$), Spain ($n = 247964$), Luxembourg ($n = 48146$), and Germany ($n = 3161$). It comprised 1508847 (40%) females and 2277929 (60%) males that were members for at least one month between August 2022 and ending with July 2023. Throughout the year, 1375358 people cancelled their membership and 2411418 kept their membership throughout this period. In other words, 36,3% cancelled their membership during the year of analysis. The dataset consisted of 9 variables, stored as a Pandas dataframe (Pedregosa et al., 2011). This selection of variables is based on

data availability, meaning that it was limited to the collected and stored data from CM.com, and literature about informative predictors for churn (Clavel San Emeterio et al., 2019; Sobreiro et al., 2021; Yi et al., 2021). See Table 1 for a definition of the variables and Table 4 for the descriptive statistics of the variables.

Table 1: Variable definitions

Variable	**Definition**
Churn	Whether the member cancelled membership or not (0 = active, 1 = dropout)
Gender	Gender (0 = female, 1 = male)
Age	Age, in years (divided into quartiles, categorized by dummy)
Membership type	The type of membership per member (values = Basic (0), Comfort (1), Premium (2), categorized by dummy)
Country	The country of residence per member (values = Netherlands (0), Belgium (1), Luxembourg (2), France (3), Spain (4), Germany (5), categorized by dummy)
Duration	The duration of membership per member, from registration date till end date (in case of dropout) or till August 1st, 2023 (divided into quartiles, categorized by dummy)
Start month	The membership registration month per member (values = 1 to 12, categorized by dummy)
Visit frequency	The number of times the member visited a gym during, cumulative for the one year timeframe (divided into quartiles, categorized by dummy)
Club > 1	Whether the member visited more than one unique gym club (0 = no, 1 = yes)
International	Whether the member visited more than one country (0 = no, 1 = yes)

Note: Table layout was inspired by Clavel San Emeterio et al., 2019.

4.5 *Data Analysis*

The goal of the analysis is to determine to what extent sport membership churn can be predicted and what variables have a significant effect on this. All data analysis was also conducted using Python software and Sklearn, Matplotlib and XGBoost packages (Chen and Guestrin, 2016; Hunter, 2007; Pedregosa et al., 2011). Firstly, exploratory and statistical analysis was conducted to inspect the descriptive values of the variables.

After this, the dependent (i.e. target) variable was separated from the independent variables (i.e. predictor variables). After this, the dataset

was split into a training (80%) and test set (20%) following a stratified procedure. This stratified technique was chosen to ensure that the class imbalanced dependent variable, churn, was equally presented in both the training and test set (Prusty et al., 2022).

For both of the models implemented, LR and XGBoost, hyperparameter tuning was done using a stratified 10-fold cross-validation technique. The 10-fold was chosen as this is a very common use in machine learning literature, and as it often is the optimal value for k in a k-fold cross-validation (Nti et al., 2021). To find the best hyperparameters, a gridseach was conducted investigating various options for the maximization of the F1-score. Both of the gridsearches tuned three different hyperparameters and the values within these gridsearches were chosen with consideration of the literature, see Table 2 and 3 for an overview of the hyperparameter space. These hyperparameters determined by the gridsearch define the optimal performance of the model, in terms of F1-score, while creating a balance of capturing complex patterns in the data and countering the risk of overfitting (Hastie et al., 2009).

4.5.1 *Logistic Regression Model*

Logistic Regression (LR) is a statistical technique that is used to estimate the association between a dependent variable and one or more predictor variables (Stoltzfus, 2011). It is useful for modeling the probability of a binary event, such as dropout or non-dropout. The model estimates the probability of the event occurring using a logistic function, that transforms linear combinations of input features into probabilities. These predictions fall within the range of 0 to 1 and therefore the model offers great interpretability (Stoltzfus, 2011).

For the Logistic Regression model, the binary variables that were divided into the ranges of their quartiles require a reference group for interpretation. To create this, for all variables the lowest quartile (<25%) was taken as a reference group.

The Correlation Matrix was generated and Variance Inflated Factors (VIF) were calculated to inspect the assumption of no multicollinearity. This means that, for the input of the LR, the independent variables have to be independent of each other to be able to see the individual effects of the variables on the dependent variable (Daoud, 2017). The VIF score indicates how much of the variance in the dependent variable is explained by the independent variables, and scores higher than 10 are an indication of the presence of problematic multicollinearity (Vittinghoff et al., 2006).

After this, the Logistic Regression model was fitted to the data. For the implementation of this model, hyperparameter tuning was done using a gridsearch for several hyperparameters. These parameters were chosen

based on compatibility with each other, and for their good fit to large dataset (Pedregosa et al., 2011). See Table 2 for the hyperparameter space that was used for the gridsearch.

Table 2: Hyperparameter Space for Logistic Regression

Hyperparameter	Values
classifier_c	0.001, 0.01, 0.1, 1, 10, 100
classifier_penalty	l2, none
classifier_solver	lbfgs, sag

The analysis of the feature importance was conducted by investigating the Regression Coefficient (B), its p-value, the Standard Error, and the Wald Statistic. The Regression Coefficient represents the expected change in the log odds of the outcome for a one-unit increase in the predictor variable (Kleinbaum and Klein, 2010). Positive values indicate an increase in the log odds of the dependent variable, the cancellation of gym membership, negative values indicate the opposite. This coefficient is accompanied by an indicator of significance, the p-value, which indicates that whether the effect of the variable is statistically significant (Kleinbaum and Klein, 2010). The Standard Error determines the precision of the estimated Regression Coefficient and the Wald Statistic illustrates the role of the variables in the prediction of churn in the LR, as higher values indicate a more significant role (Agresti, 2007; Snedecor and Cochran, 1989).

The use of this analysis is chosen as it creates a direct measure of how each individual feature alters the odds of churn, which offers the most direct, usable and easy interpretation, that is only possible with a LR.

4.5.2 *XGBoost Model*

XGBoost, which refers to Extreme Gradient Boosting, is a powerful and scalable machine learning algorithm that belongs to the ensemble learning category (Chen and Guestrin, 2016). It builds a series of decision trees sequentially, each correcting errors of the previous ones, and hereby creating a robust predictive model. It incorporates regularization techniques to prevent overfitting and offers flexibility in handling different types of data (Chen and Guestrin, 2016).

The XGBoost model was fitted to the data with an implemented hyperparameter gridsearch. These values were set within a range that is not too high, thus prevent overfitting, and not too low, thus prevent underfitting (Shibao et al., 2021). See Table 3 for the hyperparameter space that was used for the gridsearch.

Table 3: Hyperparameter Space for XGBoost

Hyperparameter	Values
classifier_learning_rate	0.01, 0.1, 0.2, 0.3
classifier_max_depth	3, 4, 5, 6, 7, 8
classifier_n_estimators	50, 100, 150, 200, 250

To analyze the contributions of the features that lead to the predictions of the XGBoost model, the feature importance plot and the SHAP-value plot are examined (Chen and Guestrin, 2016; Lundberg and Lee, 2017). The former illustrates the frequency and usefulness of the feature in the XGBoost model, meaning that variables with a high feature importance score are frequently used to make key splits in the XGBoost trees (Chen and Guestrin, 2016). The SHAP (SHapley Additive exPlanations) values explain the impact of having a certain value for a given feature, in comparison to the prediction of a baseline value of that feature (Lundberg and Lee, 2017). These values measure the impact of each feature on each individual prediction of the model. It indicates the direction of the prediction, as it contains blue (representing the binary value 0) and red (representing the binary value 1) dots that are divided by a zero line in the middle. The position of these dots relative to this middle line indicates the impact of the model's prediction for that particular instance. Features with a majority of positive SHAP values, which are dots to the right of the zero line, are associated with an increased likelihood of churn. Contrary, dots on the left of the middle line indicate a negative SHAP value which indicate an increase in the probability of the prediction of the model to no cancellation of membership (Lundberg and Lee, 2017).

The use of this analysis is chosen as it provides insights into how different variables affect the prediction, including a direction of the influence.

4.5.3 *Evaluation metrics*

The performance of the LR, as a baseline, was compared to the performance of the XGBoost. For both these models, the primary metric of evaluation is the F1-score. This metric is chosen with consideration of the class imbalance that is present in the current dataset. The F1-score is a suitable metric for such datasets, as it defines the harmonic mean between precision and recall (Davis and Goadrich, 2006). The score is calculated by the following formula:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

Precision is the ratio of true positive observations to the total predicted positives, and recall is the ratio of true positive observations to the actual

positives (Davis and Goadrich, 2006). Additionally, precision, recall, accuracy metrics and AUC-scores (based on the ROC-curve) were analysed and corresponding confusion matrices were printed (Davis and Goadrich, 2006; Fawcett, 2006; Japkowicz and Shah, 2011).

5 RESULTS

5.1 *Descriptive statistics*

Table 4, shows the descriptive statistics of the variables.

Table 4: Descriptive Statistics

Variables	**Value**	**%**
Churn	Yes	36
	No	44
Gender	Female	40
	Male	60
Country	France	54
	Netherlands	21
	Belgium	17
	Spain	6,6
	Luxembourg	1,3
	Germany	0,1
Membership type	Basic	10,7
	Comfort	48
	Premium	41,3
>1 club	Yes	30,5
	No	69,5
International	Yes	1,2
	No	98,8

					Percentile	
	Min	**Max**	**Mean**	**Std**	**25**	**50**	**75**
Age (years)	13	90	31.21	11.42	22	27	36
Start month	1	12	5.9	3.4	3	6	9
Duration (days)	35	11535	436.3	333.3	203	392	519
Visit frequency	0	487	16.8	26.7	2	9	31
Club count	0	73	1.4	1.3	1	1	2
Country Count	0	5	0.8	0.4	1	1	1

Note: Table layout was inspired by Clavel et al., 2016.

5.2 *Logistic Regression Model*

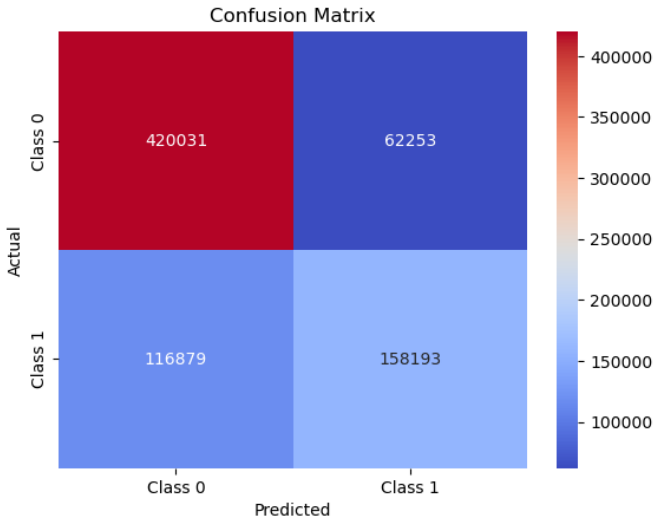
The inspection for multicollinearity involved the examination of the Correlation Matrix (Appendix B, page 30, for the individual variables and Appendix F, page 7, for the quartile variables). These matrices indicate that there is high correlation between the different quartiles of the variables, especially membership type. However, the VIF scores were also examined and did not indicate a score higher than 10 for all the variables, which means that the level of multicollinearity is tolerated for the LR model (Vittinghoff et al., 2006) (see Appendix C, page 31 and Appendix G, page 7 for VIF scores). Additionally, it was chosen to keep all the variables and quartiles in order to preserve information and interpretability.

The hyperparameter gridseach concluded that the optimal hyperparameter settion was:

Optimal hyperparameters: {'classifier_C': 1, 'classifier_penalty': l2, 'classifier_solver': 'lbfgs'}

These settings for the hyperparameters were applied for fitting the LR to the data and predictions were made on the test set of the data. The performance of the model on the test set showed that the F1-score of the LR was 64%, the precision 72%, recall 58%, and accuracy 76%. Figure 2 shows the distribution of predicted and actual values, represented in a confusion matrix. It shows that the model was sensitive to the majority class, members without churn, as there is a high number for the false negatives. This indicates that the LR predicted that members would continue their membership, while these members actually cancelled their membership. The sensitivity to the majority class is also visible in the high accuracy, which suggests that the model mostly predicts the majority class.

Figure 2: Confusion Matrix Logistic Regression



The analysis of the LR feature importance is presented in Table 5. The Regression coefficient (B) indicated that the all variables had a significant effect ($p < 0.05$) on the prediction of churn.

Concluding from these coefficients, the most influential variable is visit frequency. The greatest reduction in the probability of the cancellation of membership is for a visit frequency higher than 31, as this decreases the chances of churn by 89% compared to members that have a visit frequency of less than 2. Additionally, going to the gym between 2 to 31 times a year reduces the chances of cancellation by approximately 79% compared to going less than 2 times a year. Thus, as visit frequency increases the likelihood of the membership cancellation decreases.

This is also the case for age, as all the coefficients for the age variables are negative. This indicates that as age increases, the chances of cancellation decrease. There is a small difference between the reference group of members till 22 years old and the members with as age till 27 years, as the odds of belonging to the latter and churning are 0.9663 times the odds of the reference group (a decrease of 3.8%). However, between the reference group and the upper quartile (>75%) of the ages in the dataset, there is great difference of 36.4% decreases likelihood of membership cancellation.

The results for the variables of duration are more complex, as having a membership between 203 and 519 days increases the odds of cancellation compared to having a membership for less than 7 months (OR 2.334, p < .001 for the 25%-50% quartile, and OR 4.903, p < .001 for the 50%-75% quartile), while having a membership for over 519 days decreases the probability of churn (OR 0.764, p < .001)

The coefficients of the different countries indicate that a member in Spain has the highest probability of churning, compared to a member in The Netherlands, with an OR of 2.3249 ($p < .001$), followed by members from France (OR 1.0436, $p < .001$). Members from Belgium, Luxembourg and Germany have a lower probability of membership cancellation, with an OR of 0.9417 ($p < .001$), 0.8295 ($p < .001$), and 0.6835 ($p < .001$) respectively.

For the registration months, starting a sport membership between March and June decreases the chances of membership cancellation compared to starting in January. It reduces the chances around 40% for March, April, and May and 17% for June. On the other hand, all other months as start month for a sport membership increase the odds of churn compared to starting in January. Starting a sport membership in September increases the odds of cancellation the most, as the odds are 1.438 times the odds of cancellation in January, which is an increase of around 44%.

Overall, female members from Germany of 36 years or older, that have a visit frequency over 31 times during a year and a membership duration of 519 days or more, with a registration date in the month March, are least likely to cancel their membership. On the other hand, being a male from Spain younger than 22 years that has a 'Comfort' membership type that started in September have the highest probability of churning.

Then for the variables that measure the impact of the distributed network of GymClubX, the use of more than one gym ('>1club') and the use of international gyms ('International'), the results are that using more than one gym decreases the probability of churning with an OR of 0.9596 ($p < .001$) which reflects a decrease of around 4% in the chances of cancelling, compared to only visiting one gym. However, the use of international gyms illustrates an increases probability of churn, with an OR of 0.0288 ($p = .035$) which is an increase of around 3% compared to using gyms within one country.

Table 5: Logistic Regression Variables Results

Variable	B	S.E.	Wald	Sig.
Constant	-0.5857	0.008	4950.699	0.000
Gender	-0.1022	0.003	1249.605	0.000
>1 club	-0.0410	0.004	121.496	0.000
International	0.0288	0.014	4.460	0.035
Visit frequency (base) (<2)				
Visit frequency (2.01-9)	-1.5846	0.004	165319.021	0.000
Visit frequency (9.01-31)	-1.5584	0.004	137014.209	0.000
Visit frequency (>31)	-2.1906	0.005	215714.421	0.000
Age (base) (<22 years)				
Age (22.01-27 years)	-0.0343	0.004	80.220	0.000
Age (27.01-36 years)	-0.2310	0.004	3443.495	0.000
Age (>36 years)	-0.4528	0.004	12922.725	0.000
Netherlands				
Belgium	-0.0596	0.005	164.331	0.000
Luxembourg	-0.1874	0.013	200.723	0.000
France	0.0423	0.004	133.184	0.000
Spain	0.8432	0.007	15911.823	0.000
Germany	-0.3819	0.067	32.936	0.000
Basic				
Comfort	1.3817	0.006	53640.877	0.000
Premium	0.5213	0.006	8436.239	0.000
Duration (base) (<203 days)				
Duration (203.01-392 days)	0.8457	0.005	3433.598	0.000
Duration (392.01-519 days)	1.5913	0.004	128110.935	0.000
Duration (>519 days)	-0.2681	0.005	2932.771	0.000
Start month 1				
Start month 2	0.0241	0.006	15.101	0.000
Start month 3	-0.5545	0.006	7863.121	0.000
Start month 4	-0.4885	0.007	5074.727	0.000
Start month 5	-0.5190	0.006	6441.670	0.000
Start month 6	-0.1915	0.006	987.660	0.000
Start month 7	0.3234	0.007	2088.135	0.000
Start month 8	0.2959	0.007	1763.300	0.000
Start month 9	0.3636	0.006	3613.775	0.000
Start month 10	0.2409	0.007	1321.692	0.000
Start month 11	0.1629	0.007	544.596	0.000
Start month 12	0.0867	0.008	127.654	0.000

Note: This table presents the LR variables results. B represents the Regression Coefficient, S.E. stands for standard error, Wald is the Wald-statistic, and Sig. denotes the p-value. Table layout was inspired by Clavel San Emeterio et al., 2019.

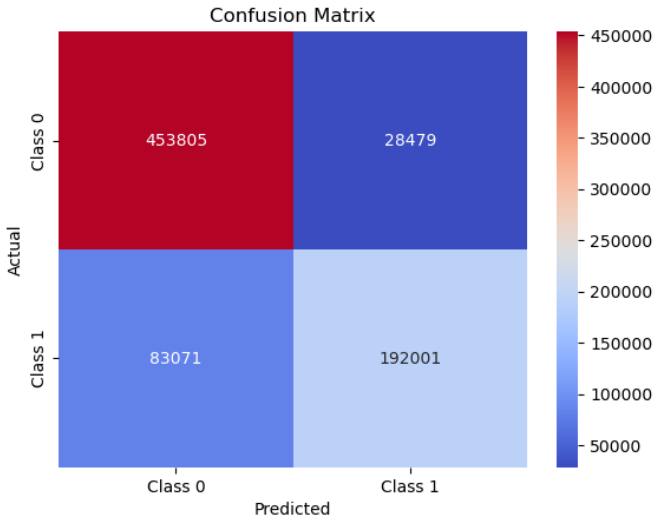
5.3 *XGBoost Model*

The hyperparameter gridsearch concluded that the optimal hyperparameter setting was:

Optimal hyperparameters: {'classifier_learning_rate': 0.3, 'classifier_max_depth': 6, 'classifier_n_estimators': '100'}

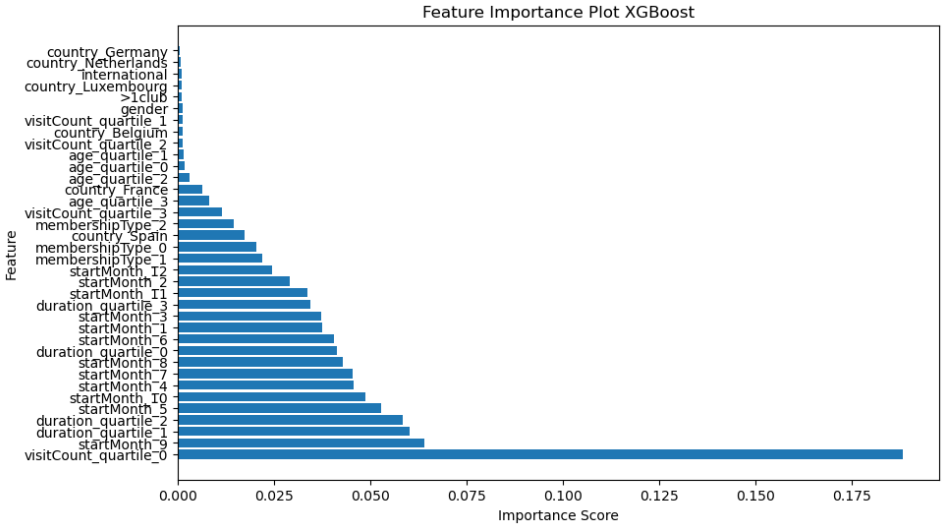
These settings for the hyperparameters were applied for fitting the XGBoost to the data and predictions were made on the test set of the data. The performance of the model on the test set showed that the F1-score of the XGBoost was 78%, the precision 87%, recall 70%, and accuracy 85%. Figure 3 shows the distribution of predicted and actual values, represented in a confusion matrix. It shows that the model was more sensitive to the majority class, members without churn, than to the minority class, members that churned. This is suggested by the higher rate of false negatives than false positives. The sensitivity to the majority class is also visible in the high accuracy, which suggests that the model mostly predicts the majority class: that members will continue their membership.

Figure 3: Confusion Matrix XGBoost



The feature importance plot, see Table 4, shows that having a visit frequency of less than two times is the most used feature in making splits in the XGBoost trees in the prediction of gym membership churn, followed by starting a membership in September and having a membership duration between 203 and 392 days. The least important variables are being a member in Germany or The Netherlands and the usage of international gyms.

Figure 4: Feature Importance Plot XGBoost



The SHAP plot indicates that members that have a visit frequency of less than two times are more likely to churn, as this variable has a greater stretch of red dots to the right of the zero line on the SHAP graph's horizontal axis. Thus according to the models's learned patterns, lower gym visitation is positively correlated with the likelihood of membership churn. The opposite effect is seen for a having a visit frequency of more than 31 times, as this decreases the likelihood of membership cancellation.

For the registration months, there are varied effects on the prediction of churn. However, a positive effect on the probability of churning can be seen in the months: September, October, November and December.

For the duration quartiles, the SHAP graph indicates that having a membership duration between 203 and 519 days strongly increases the likelihood of churn. This is illustrated by the great stretch of red dots on the right side of the zero line, indicating that this feature is a significant predictor of churn in the XGBoost model. However, having a membership of less than 203 days has a varied influence of the prediction of churn, but leaning more to a positive impact on the prediction of churn. In other words, belonging in this duration quartile increase the chances of membership cancellation. The opposite effect can be viewed for having a membership duration for over 519 days, as this decreases the likelihood of membership cancellation.

For the variables that display the effect of a distributed network of gyms, '>1club' and 'international', both of these variables have a very low feature importance score. These features are both in the top five least important features in the usefulness of the XGBoost in the splitting of the

trees. The SHAP values for these variables, see Figure 6, indicate that the effect of these variables increases the probability of churning. However, the effect of both of these variables, especially the use of international gyms, is very low.

Figure 5: SHAP values XGBoost

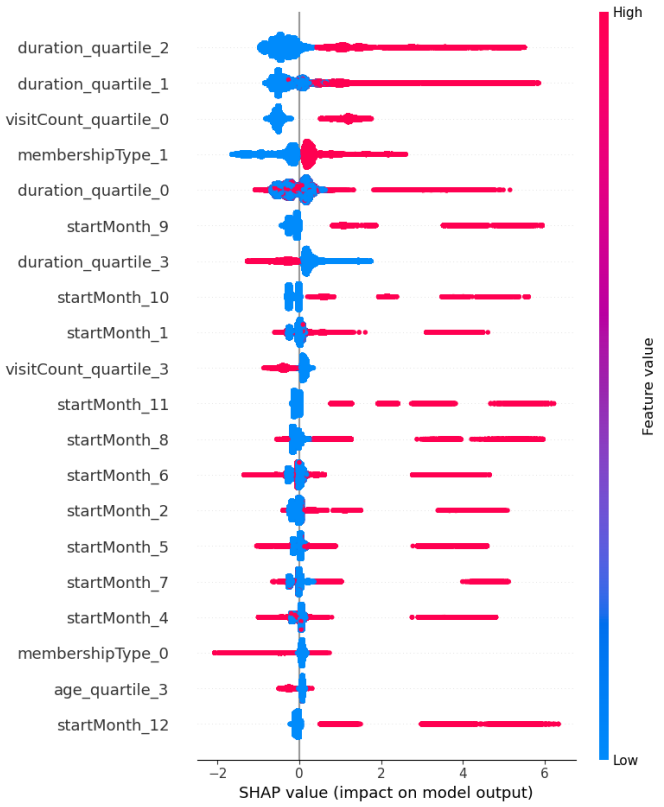
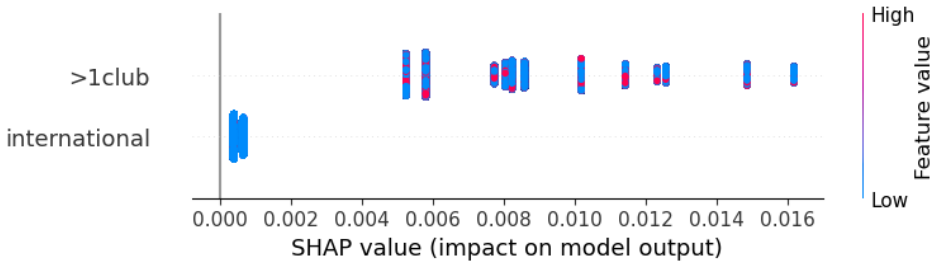


Figure 6: SHAP values for the '>1club' and 'international' variables for the XGBoost



5.4 *Comparison of Logistic Regression and XGBoost*

Based on all the evaluation metrics, the XGBoost outperformed the baseline model, Logistic Regression, in the prediction of churn. In Table 6 are performances of both the LR and XGBoost models of the different evaluation metrics evaluated on the test set of the data (see Appendix D, page 32, for ROC curves, and Appendix E for precision-recall curves, page 32). A comparison of the Confusion Matrices, Table 2 and 3, shows that especially the false negatives rate decreased for the XGBoost model. This indicates that the XGBoost performed better at distinguishing the minority class of members that would cancel their membership.

Table 6: Performance of LR and XGBoost

Models	**F1-score**	**Precision**	**Recall**	**Accuracy**	**AUC**
LR	0.638	0.718	0.575	0.763	0.798
XGBoost	**0.775**	0.871	0.698	0.853	0.913

In terms of the important features for both models, there are some similarities and some discrepancies.

The most important similarity is that both models indicate visit frequency as an important predictor, specifically having a visit frequency of less than 2 times or more than 31 times in a year. As having a visit frequency of less than 2 times increases the probability of membership cancellation, while having a visit frequency of more than 31 times decreases it. Additionally, both models have similar feature results for the membership duration variables. As both models indicate that having a duration between 203 and 519 days increases the likelihood of churning, while having a duration of more than 519 days decreases it.

The discrepancies mostly consist of the fact the LR identified having an age of over 36 years old as an important variable, while the XGBoost attributed a low feature importance it. Controversy, the XGBoost assigned great feature importance to the different registration months of membership while for the LR these had less impact on the prediction of churn. Especially for the months November and December, the XGBoost indicated an increase in the probability of membership cancellation and the LR indicated an increase but not as impactful as the XGBoost.

6 DISCUSSION

The aim of this thesis is to determine to what extent sport membership churn can be predicted in order to enhance the experience and satisfaction

of members. To answer this question, a machine learning approach was taken by implementing Logistic Regression and XGBoost models. The results indicate that the cancellation of membership churn can be predicted reasonably good, with a F1-score of 78% for the XGBoost algorithm. This finding can be used to predict gym members with a high probability of churning in order to focus efforts on them to increase their experience and satisfaction.

6.1 *Discussion of the Overall Results*

The overall results in terms of the model's performances are similar to the findings of previously conducted studies on the prediction of gym membership churn. As Clavel San Emeterio et al. (2019), Sobreiro et al. (2021), and Yi et al. (2021) all found accuracy scores ranging from 70% to 78% for their LR models, and this study found an accuracy of 76% for the LR. Moreover, the XGBoost model used in the current study, achieving an accuracy of 85%, outperformed not only the LR baseline of this study but also previous research using LR.

Additionally, the current study reinforces the importance of member engagement in enhancing adherence to sport practices (Clavel San Emeterio et al., 2019). This is pointed out by the importance of the visit frequency variable in the prediction of membership cancellation. As with a higher visit frequency the probability of churning decreases, which is in line with the findings of Clavel San Emeterio et al. (2019), Sobreiro et al. (2021), and Yi et al. (2021) (see Section 3). This emphasizes the consistent role of engagement that proves pivotal in the fitness industry and is related to the satisfaction of members with the sport facility (Ferrand et al., 2010).

This coincidences with an interesting effect, that was pointed out by Ferrand et al. (2010), as higher visit frequency increases satisfaction of members and satisfaction increases visit frequency. Therefore, an effective strategy that increases engagement should be established and efforts should be focused on engagement of members by providing increasing their adherence. This can be created by offering personal training sessions (Yi et al., 2021, or with professionally semi-supervised exercise programmes (Bottorff et al., 2014), or by offering group training sessions (Kováčová et al., 2011). This will increase adherence as members are going to the gym and this will increase their satisfaction with the gym, and consequently it will increase overall engagement (Ferrand et al., 2010). Therefore, as visit frequency is a pivotal aspect to decrease churn, efforts should be focused on increasing adherence and engagement to the gym facility.

Furthermore, increasing engagement proves to be effective because it increases employee engagement with the members (MacIntosh and

Law, 2015; Yi et al., 2021). Employee engagement can be defined as the ability to help the company's success in terms of efforts, having a positive attitude towards the company and working with involvement and enthusiasm (Yi et al., 2021). It has a great influence on the satisfaction and experience of members, and this has a positive effect on membership retention (Macintosh and Doherty, 2007; MacIntosh and Law, 2015; Yi et al., 2021).

Additionally, the current study confirmed the finding of Sobreiro et al. (2021) and Clavel San Emeterio et al. (2019) by identifying membership duration as influential predictors. The results indicate that the membership duration displays a complex effect, as shorter duration increases the probability of churning and being in the upper quartile of membership duration decreases the chances of churn. This pattern was also found by Clavel San Emeterio et al. (2019) and a proposed reason for this is that satisfaction with the facility is significantly more important for members with a long membership time, than for new customers (Avourdiadou and Theodorakis, 2014). This suggests that it is very important to keep non novice members, especially members with a membership between 203 days and 519 days, satisfied with the experience of the gym facility.

Furthermore, the practical strength of this research is that the variables used for this research do not require interaction with customers. As all the variables used in this study were already being recorded for operating and management uses by gym facilities. Additionally, the number of variables in this study is relatively low (9) and divided into different interpretable categories based on ranges. This allows for an easy and useful application in new settings and gives gym facilities the opportunity to conduct a churn analysis on their member database (Hosmer et al., 2013). Therefore, this study has extensive practical implications.

Therefore, this thesis demonstrates that it is possible to effectively predict gym membership churn. These findings have practical implications for gym facilities, as it creates a possibility to focus efforts on members with a high risk of churning. Practical strategies should be investigated and established to effectively increase the satisfaction and experience of these members and thus decrease their chances of churning.

6.2 *Discussion of the Sub-questions*

6.2.1 *Sub-question 1*

This thesis investigated to what extent having a distributed network of sport facilities increases consumer utility, which was measured by the churn rate. To investigate this, two variables of interest were engineered:

utilizing more than one distinct gym and utilizing gyms internationally. It was examined what the effect, magnitude and direction, of these variables was on the prediction of churn.

The LR indicated that the variables of interest were significant for the prediction of churn, but the magnitudes of these probabilities were very low. For the direction of influence, visiting more than one gym slightly decreases the probability of churning. This suggests that having a network contributes positively to the decision of maintaining a gym membership. However, contrary, using gyms internationally slightly increased the probability of churning. This demonstrates the opposite effect, as it suggests that having an internationally distributed network of gym facilities negatively affects members in their decision to maintain or cancel their membership.

The XGBoost model also assigned a very low feature importance to these variables. This is also indicated by the SHAP values for these variables, see Figure 6, as the results indicate that using more than one distinct gym and using gyms internationally increases the chances of churning. However, the effect of both of these variables, especially the use of international gyms, is very low.

Thus, both the LR and the XGBoost indicated that the effect of having a distributed network of gyms is very minimal in magnitude. The results suggest that using more than one gym nationally decreases the chances of churning. Contrary, having an internationally distributed network of gyms increases the probability of churn.

6.2.2 *Sub-question 2*

This thesis investigated the performance of XGBoost compared to Logistic Regression in the prediction of sport membership churn. The results indicate that the XGBoost outperforms the LR baseline model on the metric of interest, namely the F1-score. Moreover, it outperforms the LR on all other examined metrics in this study.

This finding supports the findings of previous literature, that stated that the XGBoost model outperformed the LR on the prediction of churn (Ahmad et al., 2019; Geiler et al., 2022; Lalwani et al., 2022; Pamina et al., 2019). Thus, the current study extends these findings by proving that this is also the case for a churn analysis in the fitness industry. This finding suggests that while traditional models like LR provide valuable insights, especially in terms of interpretability, models like XGBoost offer a more nuanced understanding of complex consumer behaviors.

The superiority of XGBoost over LR suggest that this model is better at distinguishing complex patterns for the prediction of churn, which suggests a potential shift in the model preference for future studies. However,

interpretabilty is also essential to establish valuable practical implications which is where the LR proves to have its strengths.

6.3 *Limitations*

This study has several limitations that should be acknowledged. First, while it does supply practical implications, using only superficial behavioral and demographic variables is not sufficient in the prediction of churn. It is important to note that gym membership churn is an interplay of numerous factors, including physical, psychological and external factors (Ferrand et al., 2010; Gonçalves and Diniz, 2015). As pointed out in Section 6.1, employee engagement and satisfaction also play important roles (Macintosh and Doherty, 2007; Yi et al., 2021). Therefore, it would be valuable to collect the reasons of abandonment in order to create a greater understanding of the decision to maintain or cancel gym membership.

Furthermore, there is a limitation in the comparison of the features of the LR and XGBoost. This is because the LR allows for a direct interpretation of its features and their influence on churn in terms of the odds ratios, while the XGBoost model requires a more complex interpretation. It cannot directly derive odd ratios, but instead the feature importance's and SHAP values give an indication of the effect on the direction and magnitude of the variable. Therefore, the format of comparison is different and this should be noted as a limitation.

6.4 *Future research*

Following from Sections 6.1 and 6.3, several suggestions for future research can be formulated.

Firstly, additional research is needed on the effect of a distributed network of gym facilities. The results from this study indicate inconclusive results, as the LR indicated a significant effect but with a low magnitude and XGBoost indicated that these variables have a very low feature importance. Therefore, further research should establish a better understanding into the role of a distributed network of gyms in the decision to maintain or cancel a gym membership.

Additionally, this study did not include the actual reason of membership cancellation. However, to fully understand the decision to churn, qualitative variables are needed in conjunction with quantitative variables. This would give insight into the direct causes of churn and create an opportunity to more effectively adapt churn prevention strategies. Additionally, further research is needed in the establishment of an effective strategy to increase the satisfaction and experience of members with a high probability

of churning. As this study does not provide information on the direct influences of the satisfaction of members.

7 CONCLUSION

The aim of this thesis is to determine to what extent sport membership churn can be predicted in order to enhance the experience and satisfaction of members. The results indicate effective prediction of membership churn, as the XGBoost model obtained a F1-score of 78%. The LR model, which achieved a F1-score of 64%, was outperformed by the XGBoost model. Consequently, the current research contributes to the existing literature by determining that the XGBoost model also has a better performance in the prediction of churn in the fitness industry. These results have practical implications for gym facilities, as they aid in establishing effective strategies to increase the satisfaction and experience of gym members at risk of cancelling their membership. A proposed approach for this strategy is to focus on increasing adherence with the gym facility and increase employee engagement. This approach aims at increasing satisfaction and visit frequency, and as discussed in Section 6.1, visit frequency in turn increases satisfaction. Therefore, this study has practical usefulness for gym facilities as it creates the possibility to effectively identify members at risk of churning. Consequently, this study has high societal relevance as the retention of gym members promotes physical activity and overall public health.

REFERENCES

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley-Interscience.

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*.

Avourdiadou, S., & Theodorakis, N. D. (2014). The development of loyalty among novice and experienced customers of sport and fitness centres. *Sport Management Review*, *17*(4), 419–431. https://doi.org/10.1016/j.smr.2014.02.001

Bottorff, J. L., Seaton, C. L., Johnson, S. R., Caperchione, C. M., Oliffe, J. L., More, K. R., Jaffer-Hirji, H., & Tillotson, S. M. (2014). An Updated Review of Interventions that Include Promotion of Physical Activity for Adult Men. *Sports Medicine*, *45*(6), 775–800. https://doi.org/10.1007/s40279-014-0286-3

Chen, T., & Guestrin, C. (2015). Xgboost: Reliable large-scale tree boosting system. *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA*, 13–17.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Clavel, I., Iglesias-Soler, E., Gallardo, L., Rodríguez-Cañamero, S., & García-Unanue, J. (2016). A prediction model of retention in a Spanish fitness centre. *Managing Sport and Leisure*, *21*(5), 300–318. https://doi.org/10.1080/23750472.2016.1274675

Clavel San Emeterio, I., García-Unanue, J., Iglesias-Soler, E., Luis Felipe, J., & Gallardo, L. (2019). Prediction of abandonment in spanish fitness centres. *European journal of sport science*, *19*(2), 217–224.

Daoud, J. I. (2017). Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, *949*(1), 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning*, 233–240.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, *27*(8), 861–874.

Ferrand, A., Robinson, L., & Valette-Florence, P. (2010). The Intention-to-Repurchase paradox: a case of the health and fitness industry. *Journal of Sport Management*, *24*(1), 83–105. https://doi.org/10.1123/jsm.24.1.83

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. *International Journal of Data Science and Analytics*, *14*(3), 217–242.

Gonçalves, C., & Diniz, A. (2015). Analysis of member retention in fitness through satisfaction, attributes perception, expectations and well-being. *Revista Portuguesa de Marketing*, *38*, 65–76.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013, March). *Applied Logistic Regression*. https://doi.org/10.1002/9781118548387

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: A review. *Telecommunication Systems*, *76*, 613–630.

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Kleinbaum, D. G., & Klein, M. (2010). *Logistic regression*. Springer.

Kováčová, L., Stejskal, P., Neuls, F., & Elfmark, M. (2011). Adherence to the aerobics exercise program in women aged 40 to 65. *Acta Gymnica*, *41*(2), 55–63. https://doi.org/10.5507/ag.2011.013

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://arxiv.org/abs/1705.07874

Macintosh, E., & Doherty, A. (2007). Reframing the service environment in the fitness industry. *Managing Leisure*, *12*(4), 273–289.

MacIntosh, E., & Law, B. (2015). Should i stay or should i go? exploring the decision to join, maintain, or cancel a fitness membership. *Managing Sport and Leisure*, *20*(3), 191–210.

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*(1), 51–56.

Nti, I., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different k values in k-fold cross-validation. *International Journal of Information Technology and Computer Science*, *6*, 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

OpenAI. (2023). *Chatgpt* [Large language model (version 3.5)]. https://chat.openai.com/chat

Pamina, J., Raja, B., SathyaBama, S., Sruthi, M., VJ, A., et al. (2019). An effective classifier for predicting churn in telecommunication. *Jour of Adv Research in Dynamical & Control Systems*, *11*.

Park, J. H., Moon, J. H., Kim, H. J., Kong, M. H., & Oh, Y. H. (2020). Sedentary lifestyle: Overview of updated evidence of potential health risks. *Korean journal of family medicine*, *41*(6), 365.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.5555/1953048.2078195

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in nanotechnology*, *4*. https://doi.org/10.3389/fnano.2022.972421

Shibao, W., Zhuang, J., Zheng, J., Fan, H.-Y., Kong, J., & Zhan, J. (2021). Application of Bayesian Hyperparameter optimized Random Forest and XGBOOST model for landslide susceptibility mapping. *Frontiers in Earth Science*, *9*. https://doi.org/10.3389/feart.2021.712240

Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods*. Iowa State University Press.

Sobreiro, P., Guedes-Carvalho, P., Santos, A., Pinheiro, P., & Gonçalves, C. (2021). Predicting fitness centre dropout. *International journal of environmental research and public health*, *18*(19), 10465.

Stoltzfus, J. (2011). Logistic Regression: A brief primer. *Academic Emergency Medicine*, *18*(10), 1099–1104. https://doi.org/10.1111/j.1553-2712.2011.01185.x

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in science & engineering*, *13*(2), 22–30.

van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2006). Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models.

World Health Organization. (2019). *Physical activity: Fact sheet on sustainable development goals (sdgs): Health targets* [Copenhagen: WHO Regional Office for Europe]. https://iris.who.int/handle/10665/340892

World Health Organization. (2022). *Physical activity*. https://www.who.int/news-room/fact-sheets/detail/physical-activity

Yi, S., Lee, Y. W., Connerton, T., & Park, C.-Y. (2021). Should i stay or should i go? visit frequency as fitness centre retention strategy. *Managing Sport and Leisure*, *26*(4), 268–286.

APPENDIX A

Figure 7: Raw user data

gender	age	isEmployee	language	country
Male	22	0	Vlaams	BE

Figure 8: Raw membership data

	userId	membershipId	endDate	startDate	membershipCancel	membershipType	consecutiveMembership	clubId
1	0000723B-5EFE-42B5-80DE-E2A445DC93FF	1CE9773F-CD4E-4BE8-A26C-98790F0507A0	2022-11-15	2022-02-22	2022-09-21	premium	0	092B7B65-579D-4E90-BD4E-6CA67F961772

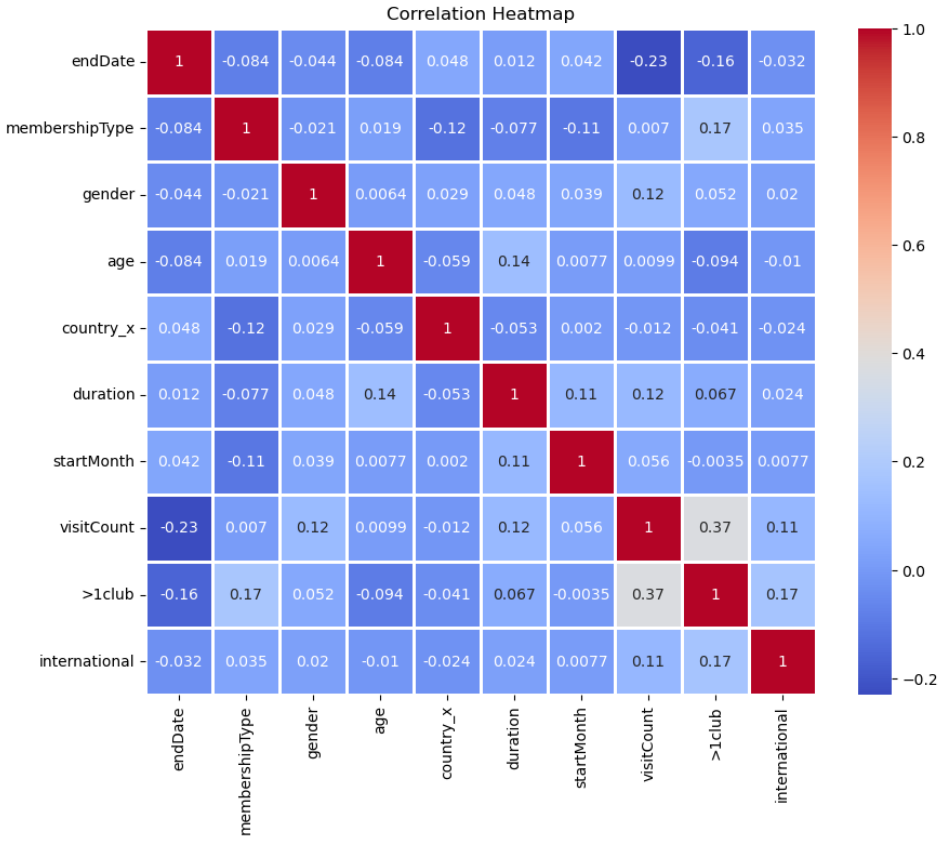
Figure 9: Raw visit data

	visitId	dateTime	userId	clubId	membershipId
0	FB493898-63C0-4A93-85FF-4CC9C95376D7	2022-06-07 19:41:00	0000972C-8A28-4D93-986F-3B4EA7FF41F4	402FD1A7-C6FE-42DB-9967-B4CC65C5103C	9B6FAD71-10EC-455A-8294-E66AD6DB3A77

APPENDIX B

The following table illustrates the correlations of the individual variables. Note:'country_x' refers to the collection of countries.

Figure 10: Correlation Heatmap of the Variables



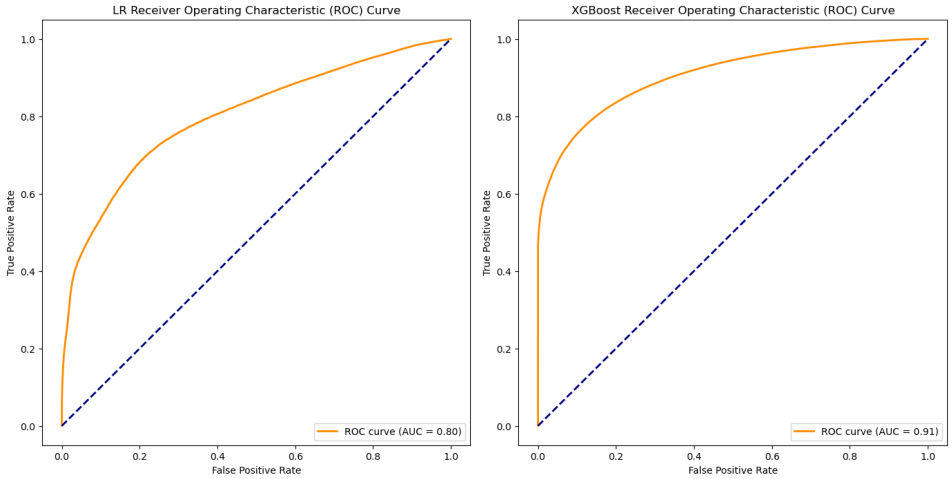
APPENDIX C

The following table shows the VIF scores of the individual variables.

Feature	**VIF score**
Membership Type	5.301048
Gender	2.556541
Age	8.265722
Country	1.000049
Duration	3.452524
Start month	2.930002
Visit count	1.863274
>1 club	1.778928
International	1.044861

APPENDIX D

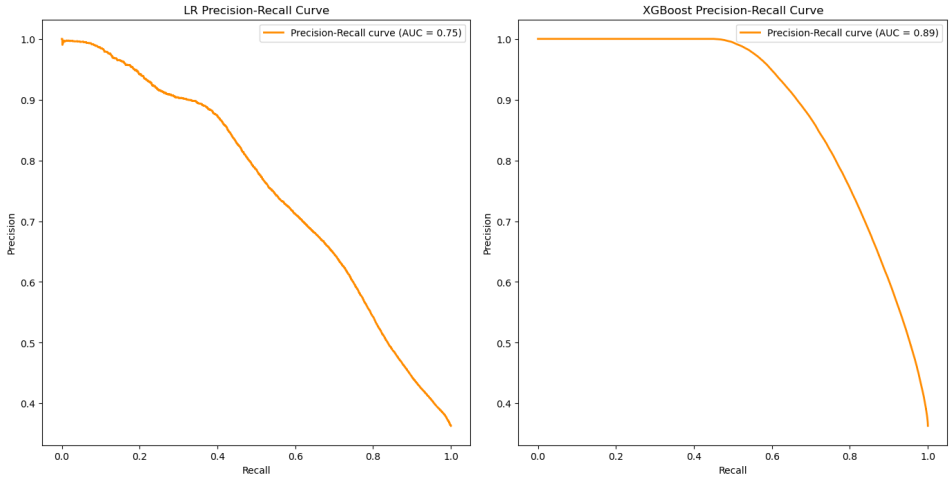
Figure 11: Comparison of ROC Curves for LR and XGBoost



(a) LR ROC Curve (b) XGBoost ROC Curve

APPENDIX E

Figure 12: Comparison of Precision-Recall Curves for LR and XGBoost



(a) LR Precision-Recall Curve (b) XGBoost Precision-Recall Curve

APPENDIX F

The following figure illustrates the correlations of the variables divided into quartiles.

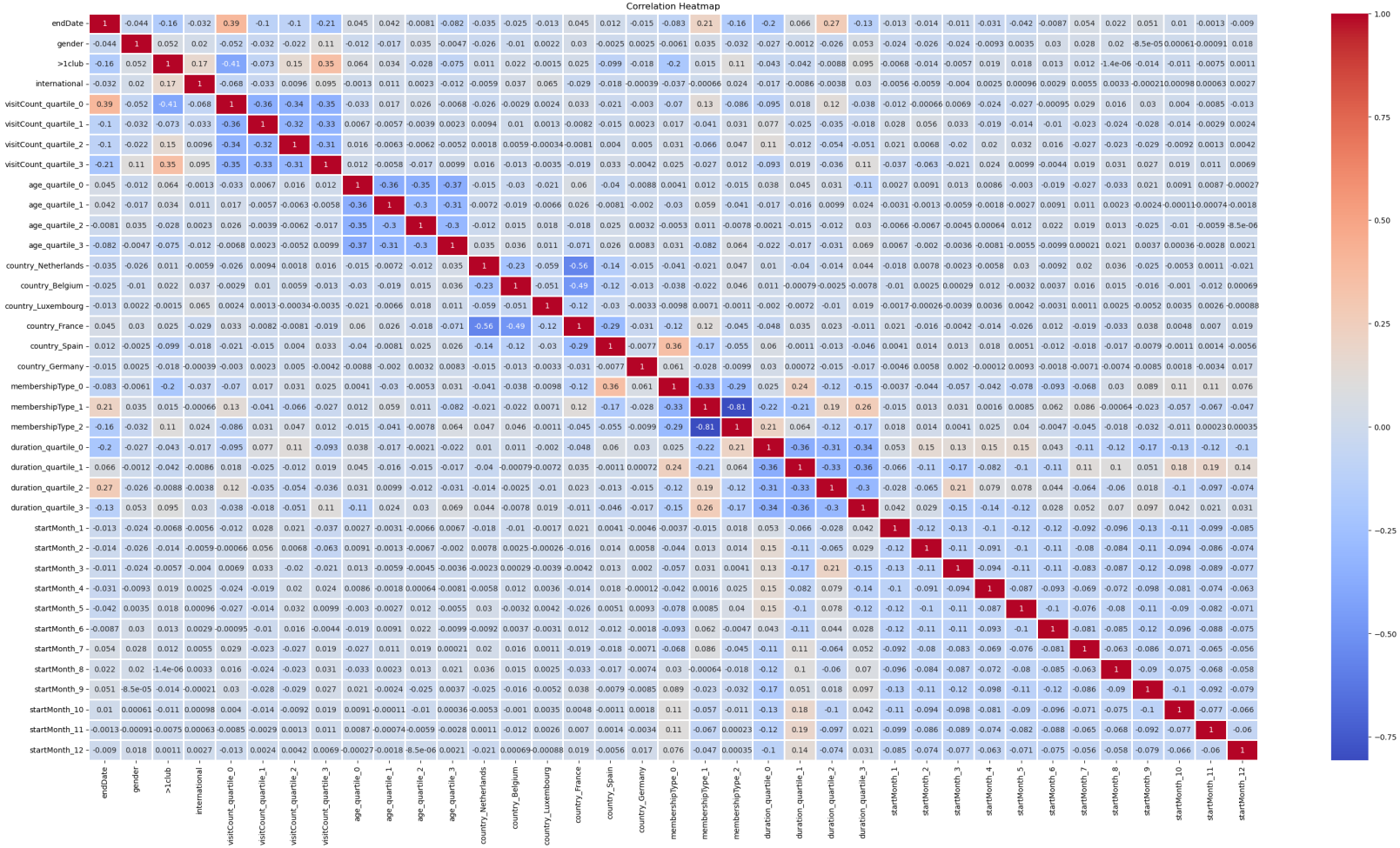


Figure 13: Correlation Matrix of the Quartile Variables

APPENDIX G

The following table shows the VIF scores for the variables divided into quartiles.

Feature	VIF
gender	1.0198
clubCount	2.6249
>1club	2.3581
international	1.0585
age_quartile	1.0428
visitCount_quartile_0	1.5609
visitCount_quartile_1	**8.1781**
visitCount_quartile_2	2.0608
visitCount_quartile_3	2.4542
country_x_0	1.4831
country_x_1	1.0457
country_x_2	4.5036
country_x_3	3.9845
country_x_4	2.5926
country_x_5	1.8534
membershipType_0	7.4670
membershipType_1	3.5102
membershipType_2	**9.9819**
duration_quartile_0	3.8842
duration_quartile_1	5.1575
duration_quartile_2	4.9103
duration_quartile_3	**8.0613**
startMonth_1	6.1527
startMonth_2	**9.6810**
startMonth_3	5.0590
startMonth_4	4.0658
startMonth_5	1.2171
startMonth_6	2.3214
startMonth_7	5.3052
startMonth_8	**8.6804**
startMonth_9	1.7851
startMonth_10	2.1385
startMonth_11	7.4794
startMonth_12	2.0746