

**Harbour Space University**

**Capstone Project Report**

**Topic:** “Churn prediction for a fitness coaching ecommerce using machine learning”.

**Student:** Yuri Alcántara Olivero

**Mentor:** MSc. Marius Klages

2022

**Content**

Introduction 3

Dataset 5

Machine learning models and metrics 10

Experimental Results 14

Conclusions 16

References 17

1. **Introduction**

Online personal training is now more relevant than ever. From 2019 to 2021, "home fitness" became a hot-word in startup companies and investments, and online fitness market grows at a high speed, especially after the outbreak of the pandemic.

Customer Relationship Management (CRM) is a comprehensive strategy for building, managing and strengthening loyal and long-lasting customer relationships. It is broadly acknowledged and extensively applied to different fields, e.g., telecommunications, banking and insurance, retail market, etc. One of its main objectives is customer retention. The importance of this objective is obvious, given the fact that the cost for customer acquisition is much greater than the cost of customer retention. Thus, tools to develop and apply customer retention models (churn models) are required and are essential Business Intelligence (BI) applications. Churn models aim to identify early churn signals and recognize customers with an increased likelihood to leave voluntarily. Several, very popular in research community machine learning algorithms have been proposed in order to tackle the churning prediction problem. Such methods include Artificial Neural Networks, Decision Trees learning, Regression Analysis, Logistic Regression, Support Vector Machines and Naive Bayes. We can safely conclude from the existing research in the field of customer churn prediction, that there is not a single model that could give the highest accuracy in all of the cases. Instead, the performance of every algorithm will differ according to the characteristics of the data.

Originality

This study is the first evaluation of different machine learning approaches for churn prediction in a fitness coaching e-commerce. The models studied will be applied for the first time for the kind of data treated, where determination of which customers are likely to leave voluntarily is a problem of special interest.

Economic Justification

In subscription-based businesses, it costs five times more to attract a new customer, than to keep an existing one [1]. Also, a 5% increase in retention can lead to 25% increase in profit [2]. Keeping customers from switching to another service provider becomes especially crucial when the competition in a field is intense. In addition to competition, retention rates in the fitness industry are generally low, sometimes dropping to 60% [3]. This further necessitates the need for online fitness coaching business to apply retention strategies to keep their customers from switching. Customer churn is a critical business metric for online fitness business, and companies often make efforts to minimize churn through a variety of marketing and product development programs. Let be the number of users selected as churners by the machine learning model, the real churn rate within the selected group and the probability of successful retention, the actual number of user retained would be *.* Let *C* be the cost of retention per user and the average revenue per user per month in the company then expected savings due to churn prediction model would be , being *m* the number of months users were retained .

Goal of the project

Evaluate different machine learning approaches  for churn prediction in a fitness coaching ecommerce.

Main hypothesis

Machine learning models are useful to identify customers who are likely to churn in a online fitness coaching ecommerce allowing savings on acquisition cost. Quickly delivering actionable information to operational teams will speed the development of new programs aimed at customer retention.

The remainder of the report organized as follows. In Section 2, we give a brief presentation of the dataset description. Machine learning techniques that were evaluated and evaluation criteria are presented in section 3. The experimental results are given in section 4, and in section 5 we draw our conclusions.

1. **Dataset**

The data used in this study was acquired from an online fitness coaching business source. Data of customers and daily activities were extracted from a PostgreSQL database and the information about contracts from Billwerk, a Subscription Management Software. From the customers we have Customer Id, City and from the completed activities we have Title, Workout Plan and Started Time. From the contracts we have Tax Country, City, Contract Start Date, Contract End Date, Payment Provider, Role and Subscription Plan Variant. All information was cleaned and summarized to extract features using Python. The dataset contains the information of clients whose were either churned or active. Churned customers are those who cancel their contracts. The time horizon of the extracted data is two years, from March 2020 to March 2022.

Feature extraction

Feature selection for predictive modeling has been studied thoroughly in literature [4]. Customer relation management data, CRM, are generally used as features especially when studying abandonment in telecommunication and banking sectors [5][6]. Amount of time spent on a service is a feature that is included in prediction for businesses that rely on customer length of usage, such as mobile games, [7] and phone calls [8]. In this case, fitness coaching retention, we extracted data that reflects the level of commitment and habit formation in the customers. Table 1 illustrates the extracted features, their description and type.

Table 1. Extracted features

|  |  |  |
| --- | --- | --- |
| Feature | Description | Type |
| NumberOfDaysActive | Total number of days a user has been active | Quantitative, discrete (Integer) |
| AmountTasksPerMonth | Average number of completed tasks per month | Quantitative, continuous (Float) |
| AmountTasksLastMonth | Total number of completed tasks last month | Quantitative, discrete (Integer) |
| FirstPlanVariant | Subscription plan | Qualitative, nominal (Text) |
| DaysFromLastCompletedTask | Total number of days since the last completed task | Quantitative, discrete (Integer) |
| Churned | Either 1 for customers who already canceled their contract (churn) or 0 for active ones (non-churn) | Qualitative, ordinal, (Binary) |

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to describe key characteristics of each attribute including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

After preprocessing the data the obtained dataset contains 1432 rows in the following format:

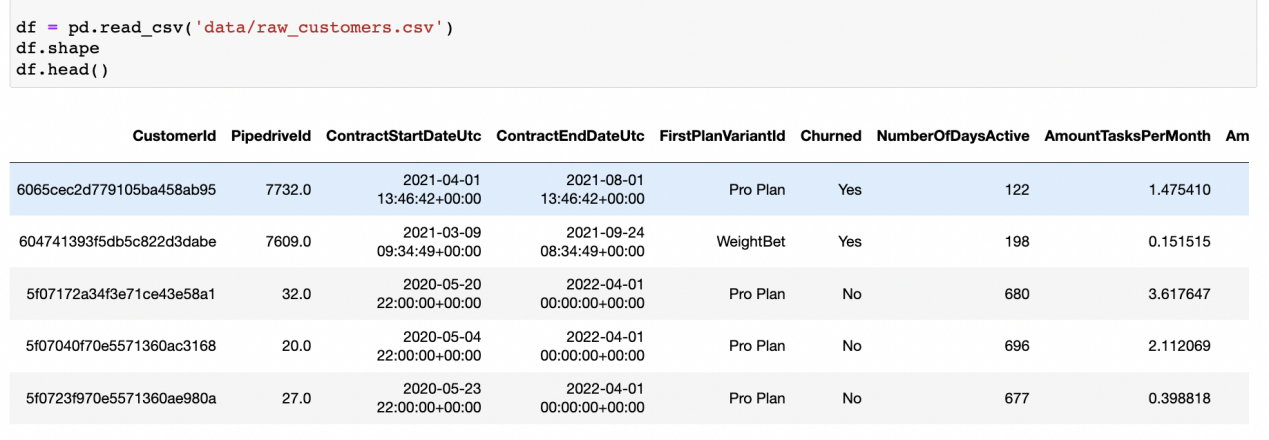


Fig. 1. Dataset format

There are no faulty records, nor null values or duplicated records. 6 features are numeric and 3 are categorical. As shown in Fig. 2, 32.6% of customers (in red color) did not churn and 67.4% of them (in blue color) churned. The dataset is highly biased towards the churned class. An oversampling method was used to attempt to resolve this issue.

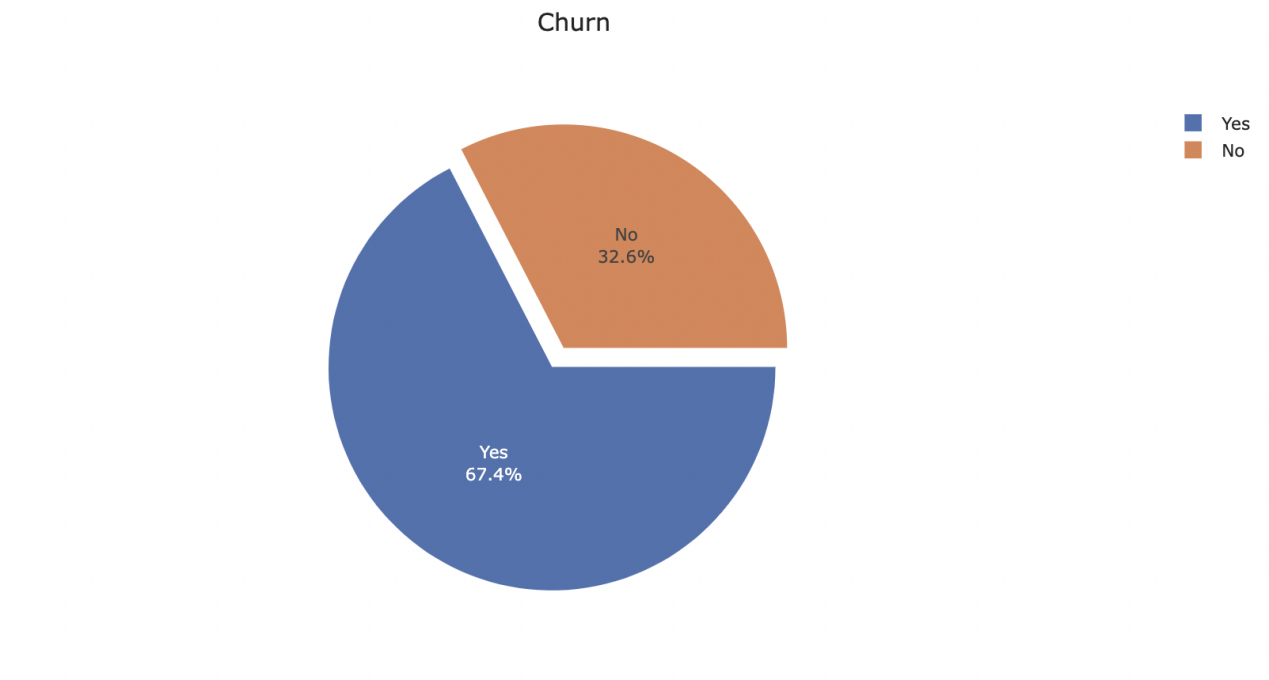


Fig. 2. Churners vs non-churners

Table 2 and Fig. 3 show how churn interact with the amount of days active . The average amount of days active of churned customers is lower. It's possible to conclude that 75% of clients who discontinue their subscriptions do so within the first 7 months of service, half of them leave until the 4th month.

Table 2. Statistics of amount of days active

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Churned | mean | std | min | 25% | 50% | 75% | max |
| No | 281.0 | 156.0 | 31.0 | 170.0 | 269.0 | 394.0 | 696.0 |
| Yes | 148.0 | 105.0 | 31.0 | 56.0 | 122.0 | 209.0 | 518.0 |

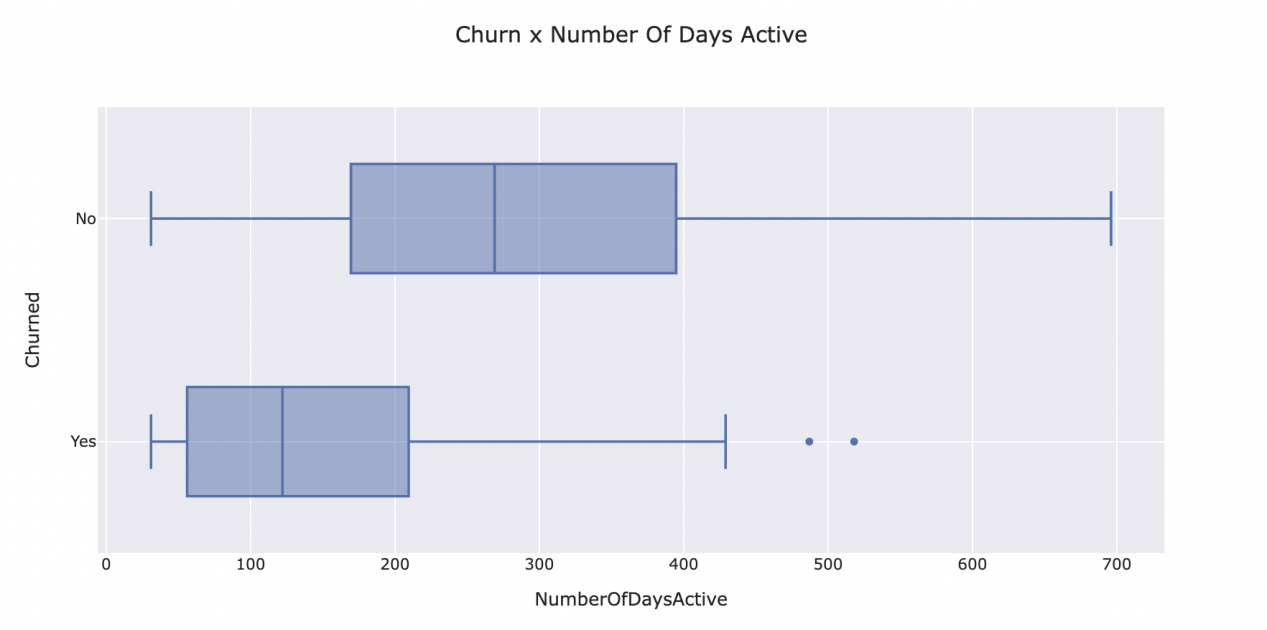


Fig. 3 Boxplot of amount of days active separated by churn status

Table 3 and Fig. 4 show how churn interact with the amount of tasks per month. Through the boxplot, it's clear to see that customers who leave are those who complete less tasks per month. Averagely, the amount of tasks completed per month of these clients was lower than the average number of tasks per month for clients who stayed in the company.

Table 3. Statistics of amount of completed tasks per month

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Churned | mean | std | min | 25% | 50% | 75% | max |
| No | 9.9 | 17.8 | 0.1 | 0.2 | 1.4 | 7.9 | 73.8 |
| Yes | 5.9 | 12.2 | 0.1 | 0.6 | 1.7 | 5.8 | 131.5 |

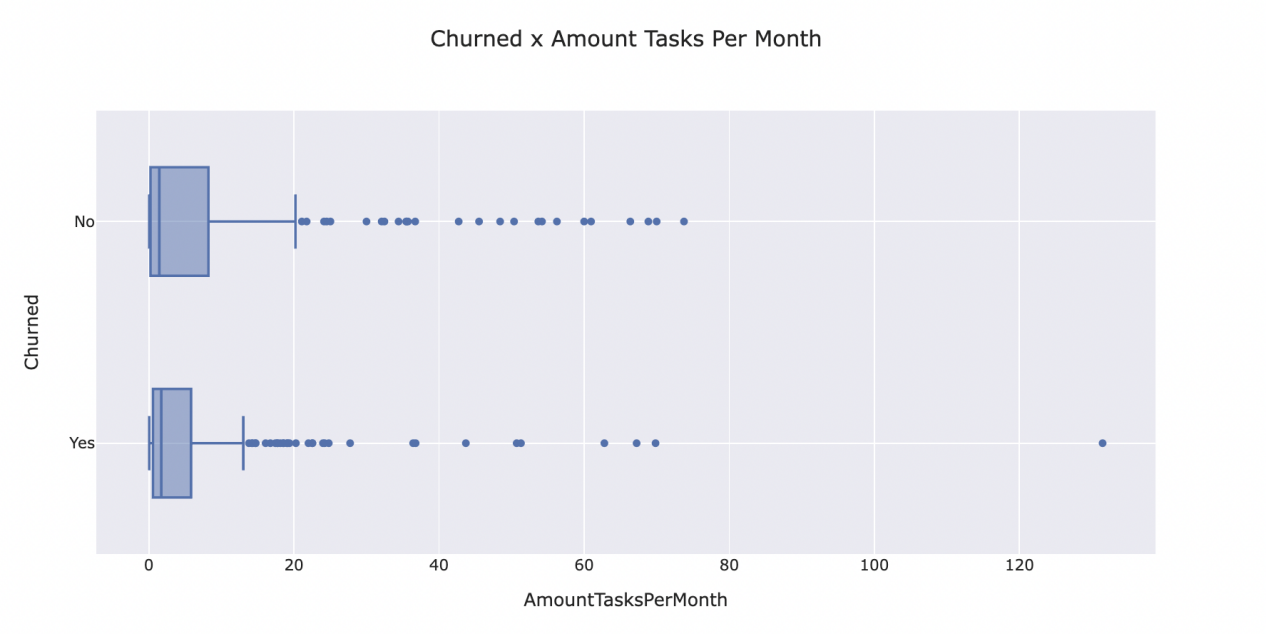


Fig. 4 Boxplot of amount of completed tasks per month separated by churn status

Fig. 5 shows a correlation matrix indicating the correlations between numerical variables. The “NumberOfDaysActive”, “AmountTasksPerMonth”, and “AmountTasksLastMonth” columns are negatively correlated with customer churn (“Churned”). The feature that shows the largest absolute value of correlation with the target is “NumberOfDaysActive”.

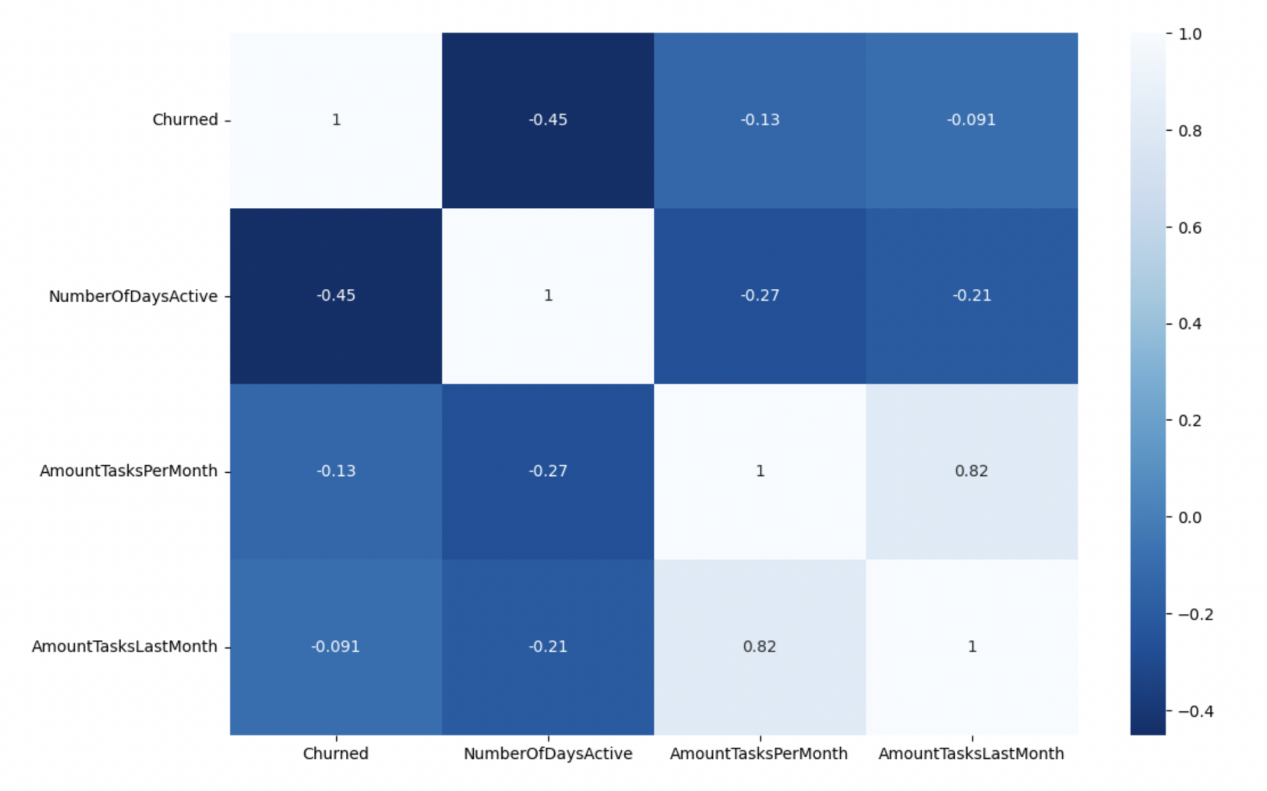


Fig. 5 Correlation matrix between numerical variables

Fig. 6 and Fig. 7 show the distribution of customers among the subscription plan variants separated by churn status.

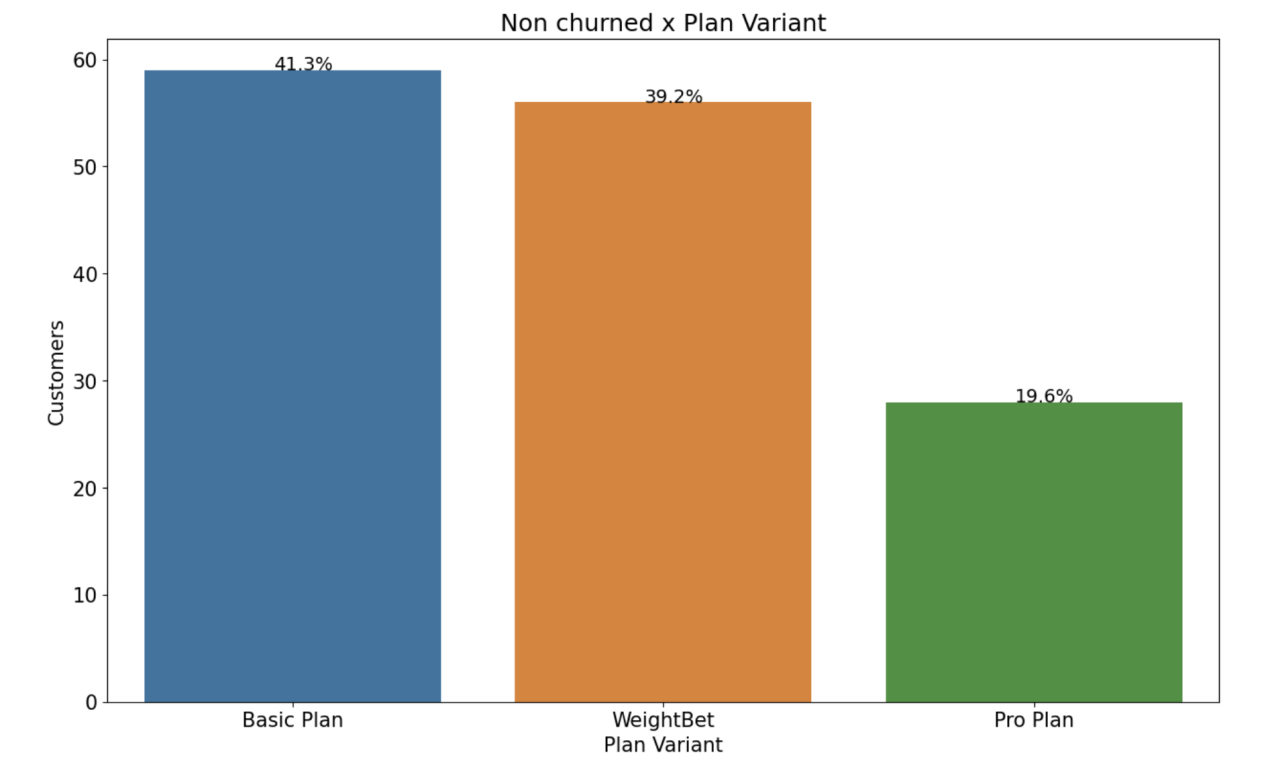


Fig. 6 Active customers x subscription plan variant

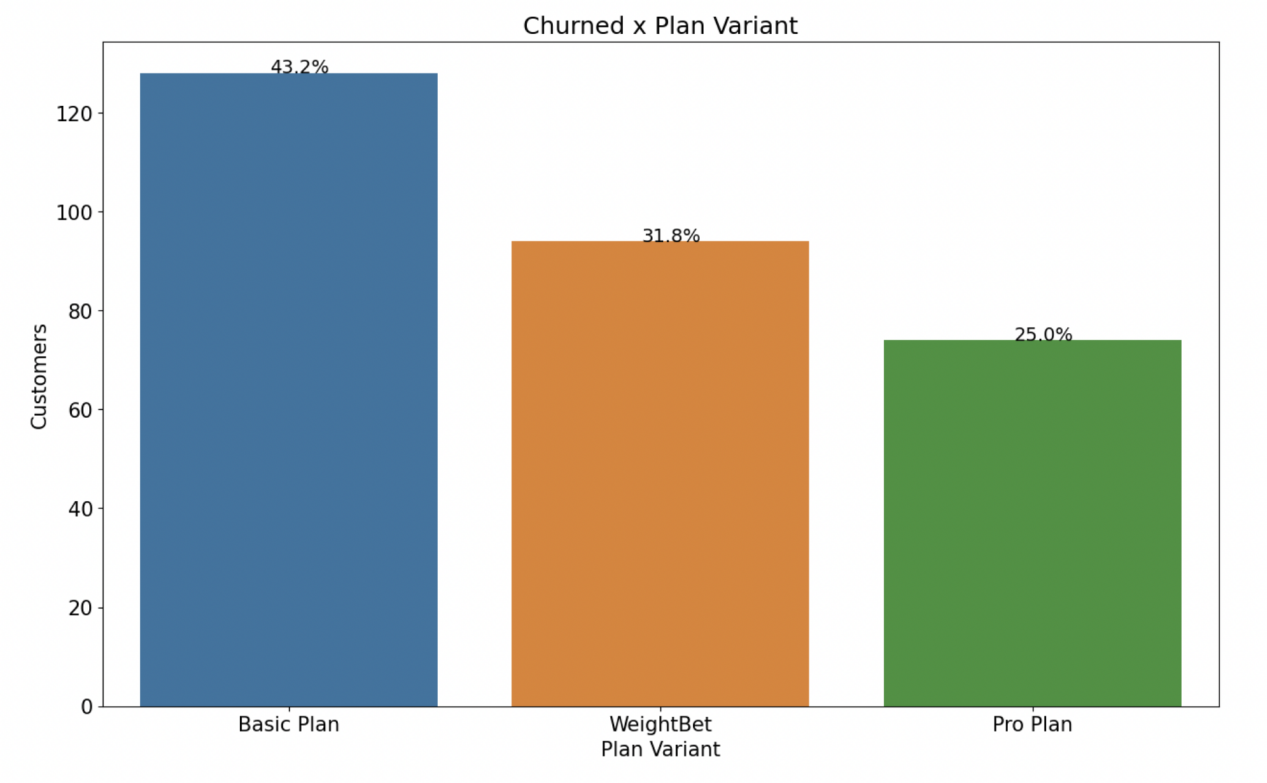


Fig. 7 Churned customers x subscription plan variant

1. **Machine learning techniques-classification methods**

In the following, we briefly present 3 well established and popular techniques used for churn prediction, taking into consideration reliability, efficiency and popularity in the research community.

Support Vector Machines

Support Vector Machines (SVMs), also known as Support Vector Networks, introduced by Boser, Guyon, and Vapnik [9], are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. SVM is a machine learning technique based on structural risk minimization. Kernel functions have been employed for improving performance. Research on selecting the best kernels or combinations of kernels is still under way. For instance, an SVM can learn to recognize fraudulent credit card activity by examining hundreds or thousands of fraudulent and non fraudulent credit card activity reports. Alternatively, an SVM can learn to recognize handwritten digits by examining a large collection of scanned images of handwritten zeroes, ones and so forth. SVMs have also been successfully applied to an increasingly wide variety of biological applications. In essence, an SVM is a mathematical entity, an algorithm (or recipe) for maximizing a particular mathematical function with respect to a given collection of data. The basic ideas behind the SVM algorithm, however, can be explained without ever reading an equation. Indeed, It’s claimed that, to understand the essence of SVM classification, one needs only to grasp four basic concepts: (i) the separating hyperplane, (ii) the maximum-margin hyperplane, (iii) the soft margin and (iv) the kernel function.

In the churn prediction problem, SVM outperform DT and sometimes ANN, depending mainly on the type of data and data transformation that takes place among them [10].

Decision trees learning

Decision Trees (DTs) are tree-shaped structures representing sets of decisions capable to generate classification rules for a specific dataset [11], or as Berry and Linoff noted ‘‘a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules’’ [12]. More descriptive names for such tree models are classification Trees or Regression Trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. DT have no great performance on capturing complex and non linear relationships between the attributes.

Decision trees trace their origins to the era of the early development of written records. This history illustrates a major strength of trees: exceptionally interpretable results which have an intuitive tree-like display which, in turn, enhances understanding and the dissemination of results. The computational origins of decision trees—sometimes called classification trees or regression trees—are models of biological and cognitive processes. This common heritage drives complementary developments of both statistical decision trees and trees designed for machine learning. The unfolding and progressive elucidation of the various features of trees throughout their early history in the late 20th century is discussed along with the important associated reference points and responsible authors. Statistical approaches, such as a hypothesis testing and various resampling approaches, have co-evolved along with machine learning implementations. This had resulted in exceptionally adaptable decision tree tools, appropriate for various statistical and machine learning tasks, across various levels of measurement, with varying levels of data quality. Trees are robust in the presence of missing data and offer multiple ways of incorporating missing data in the resulting models. Although trees are powerful, they are also flexible and easy to use methods. This assures the production of high quality results that require few assumptions to deploy. The treatment ends with a discussion of the most current developments which continue to rely on the synergies and cross-fertilization between statistical and machine learning communities.

Yet, in the customers churn problem, the accuracy of a DT can be high, depending on the form of the data [13].

Regression analysis-logistic regression analysis

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. Logistic regression is another powerful supervised ML algorithm used for binary classification problems (when target is categorical). The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function defined below to model a binary output variable. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

As opposed to linear regression where MSE or RMSE is used as the loss function, logistic regression uses a loss function referred to as “maximum likelihood estimation (MLE)” which is a conditional probability. If the probability is greater than 0.5, the predictions will be classified as class 0. Otherwise, class 1 will be assigned. Before going through logistic regression derivation, let's first define the logit function. Logit function is defined as the natural log of the odds. A probability of 0.5 corresponds to a logit of 0, probabilities smaller than 0.5 correspond to negative logit values, and probabilities greater than 0.5 correspond to positive logit values.

In terms of customer churning, regression analysis is not widely used, and that is because linear regression models are useful for the prediction of continuous values. On the other hand, Logistic Regression or Logit Regression analysis (LR) is a type of probabilistic statistical classification model. It is also used to produce a binary prediction of a categorical variable (e.g., customer churn) which depends on one or more predictor variables (e.g., customers’ features). In the churn prediction problem, LR is usually used after proper data transformation is applied on initial data, with quite good performance [14] and sometimes performs as well as DT [15]

* 1. **Model metrics**

In order to evaluate classifiers performance in churn prediction for different schemes with their appropriate parameters, we use the measures of precision, recall, accuracy and F-measure, calculated from the contents of the confusion matrix, shown in Table 4. True positive and false positive cases are denoted as TP and FP, respectively, while true negative and false negative cases are denoted as TN and FN.

Precision is the proportion of the predicted positive cases that were correct and is calculated from the equation:

Recall is the proportion of positive cases that were correctly identified and is calculated from the equation:

Accuracy is the proportion of the total number of predictions that were correct and is calculated from the equation:

Precision or recall alone cannot describe the efficiency of a classifier since good performance in one of those indices does not necessarily imply good performance on the other. For this reason, F-measure, a popular combination is commonly used as a single metric for evaluating classifier performance.

F-measure is defined as the harmonic mean of precision and recall

A value closer to one implies that a better combined precision and recall is achieved by the classifier [16].

Table 4. Confusion matrix for classifier evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted Class | |
| Churner | Non-churner |
| Actual Class | Churner | TP | FN |
| Non-churner | FP | TN |

1. **Experimental Results**

We conducted comparative experiments on the Support Vector Machine, Decision Tree and Logistic Regression models. Accuracy, recall, precision and f-measure values for each model were calculated according to the confusion matrix to evaluate the performance of the three models.

The best hyper-parameters for each model were selected using the python library Hyperopt, cross validation score was used to compare the performance.

Table 5 show the experimental results on the validation tests of the three models with the best hyper-parameters selected.

Table 5. Experimental results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
| SVM | 71.6 | 51.2 | 71.6 | 59.7 |
| Decision Tree | 85.2 | 84.9 | 85.2 | 85.0 |
| Logistic Regression | 79.5 | 78.6 | 79.5 | 78.6 |

To confirm the prediction quality, accuracy alone is sometimes misleading [17]. Therefore, when evaluating the prediction performance of the model, not only accuracy but also recall, precision and f-measure should be observed, and the performance of the prediction model should be comprehensively determined according to the four performance indicators.

The three models evaluated in this study showed accuracy and recall scores over 70%, not being so in the case of precision and f-measure where SVM showed values below 60%. In general the Decision Tree Classifier got better results than SVM and Logistic Regression, it reached the highest score in all the metrics evaluated.

Risks of the project

Poor Data

As we know, a machine learning model only works on the data that we provide to it, or we can say it completely depends on human-given training data to work. What we will be input that we will get as an output, so if we will enter the poor data, the ML model will generate abrupt output. Poor data or dirty data includes errors in training data, outliers, and unstructured data, which cannot be adequately interpreted by the model.

Class imbalance

In the data we used, one of the major problems we faced was class imbalance. In case of class imbalance, the ratio of the output categories is one-sided to the extent that the learning algorithm only predicts the majority class. For example in case of our data, there were 961 churned users, whereas there were only 471 active, presenting a typical case of class imbalance. As a result, whenever we ran algorithms such as logistic regression, decision tree or support vector machine; all of the predictions made were in favor of the majority class (churned class in this case).

Project Draft

The data engineering solutions and the exploratory data analysis were implemented using Python 3. The experiment tracking was conducted using Mlflow, an open source platform for the machine learning lifecycle, see Fig. 8. Also was implemented a simple web page application as a tool to obtain the predictions. It was created using Flask, a web application framework written in Python.

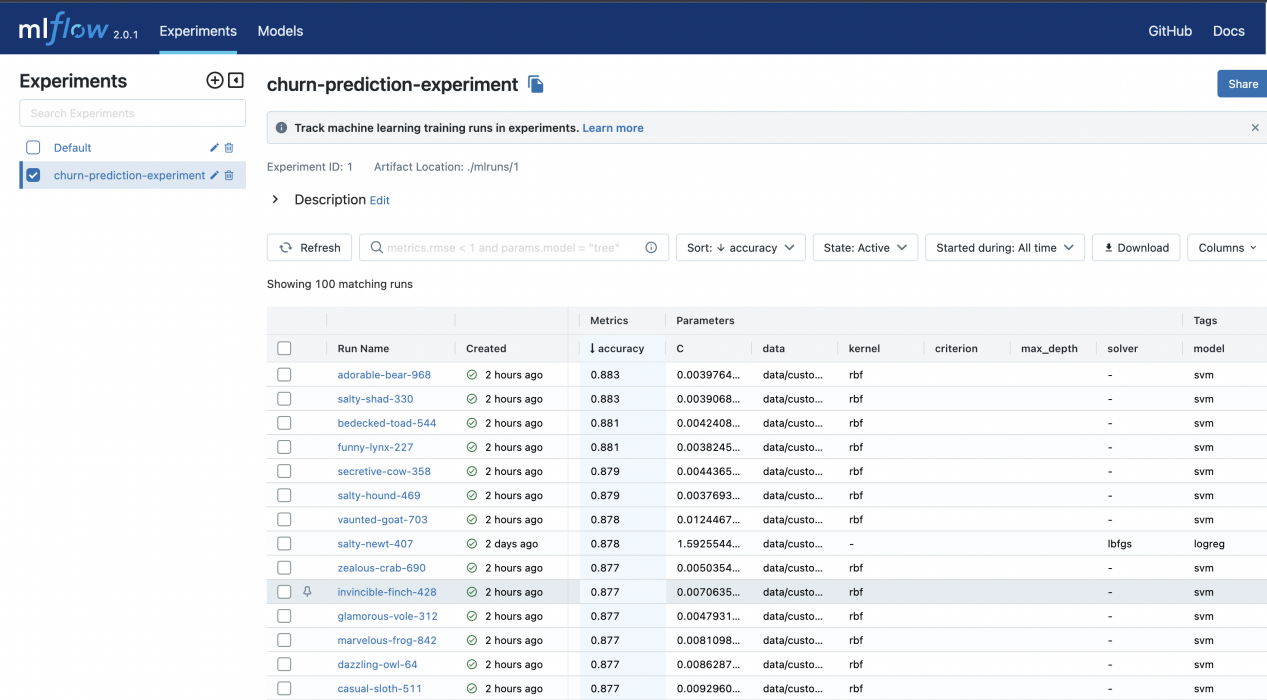


Fig. 8 Mlflow experiment tracking

Here the link to the repository containing all the code implementation: https://github.com/YuriAlcantara1995/fitness-ecommerce-churn-prediction.git

1. **Conclusions**

Customer churn predictions are very important in e-commerce. To maintain market

competitiveness, fitness enterprises should make full use of machine learning in customer relationship management to predict the potential loss of customers and devise new marketing strategies and customer retention measures according to the prediction results. This will help establish efficient and accurate loss prediction for fitness ecommerce enterprises.

In the experimental results of this study can be seen that the prediction performance of the Decision Tree model was better than that of the Logistic Regression and Support Vector Machine models. The results of this study also have some limitations. We used a real data set containing 1432 customers under the online fitness coaching environment, ideally, the research results should be verified by a bigger amount of data.

For future work should be considered conduct the experiments on top of a bigger amount of data and also test the effect of the best churn prediction model in a production environment.

**References**

1. Jackson, B. B., & Bund, B. (1985). Winning and keeping industrial customers: The dynamics of customer relationships. Jossey-Bass.
2. Reichheld, F. (2001). Prescription for cutting costs. Harvard Business School Publishing.
3. Watts, H. (2012). A psychological approach to predicting membership retention in the fitness industry (Doctoral dissertation, University of Worcester).
4. García, D. L., Nebot, À., & Vellido, A. (2017). Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, 51(3), 719-774.
5. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.
6. Hassan, M. M., & Mirza, T. (2018). Churn Prediction in Banking Sector using Bayesian Neural Networks. International Journal of Research in Advent Technology, 6(12), 3343-3346.
7. Viljanen, M., Airola, A., Heikkonen, J., & Pahikkala, T. (2017). Playtime measurement with survival analysis. IEEE Transactions on Games, 10(2), 128-138.
8. Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. International Journal of Engineering Research and Applications, 4(5), 165-171.
9. Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory (pp. 144-152).
10. Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. International Journal of Engineering Research and Applications, 2(4), 693-697.
11. Lee, S. J., & Siau, K. (2001). A review of data mining techniques. Industrial Management & Data Systems.
12. Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.
13. Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2008). Churn prediction: Does technology matter?. International Journal of Industrial and Manufacturing Engineering, 2(4), 524-536.
14. Gürsoy, U. Ş. (2010). Customer churn analysis in telecommunication sector. İstanbul Üniversitesi İşletme Fakültesi Dergisi, 39(1), 35-49.
15. Radosavljevik, D., van der Putten, P., & Larsen, K. K. (2010). The impact of experimental setup in prepaid churn prediction for mobile telecommunications: What to predict, for whom and does the customer experience matter?. Trans. Mach. Learn. Data Min., 3(2), 80-99.
16. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
17. Sturm, B. L. (2013). Classification accuracy is not enough. Journal of Intelligent Information Systems, 41(3), 371-406.