ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF MANAGEMENT

CHURN PREDICTION WITH MACHINE LEARNING

B.Sc. THESIS

Bahri Selçuk EŞKİL

Department of Industrial Engineering

Thesis Advisor: Dr. Mehmet Yasin Ulukuş

JUNE 2023

ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF MANAGEMENT

CHURN PREDICTION WITH MACHINE LEARNING

B.Sc. THESIS

Bahri Selçuk EŞKİL
(070190032)

Department of Industrial Engineering

Thesis Advisor: Dr. Mehmet Yasin ULUKUŞ

JUNE 2023

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ İŞLETME FAKÜLTESİ

MAKİNE ÖĞRENMESİ İLE MÜŞTERİ KAYBI TAHMİNLEME

LİSANS TEZİ

Bahri Selçuk EŞKİL
(070190032)

Endüstri Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Mehmet Yasin ULUKUŞ

HAZİRAN 2023

To my family,

FOREWORD

In this study, the importance of predicting customer churn using machine learning algorithms and taking actions to reduce churn rate is emphasized.

I would like to thank my esteemed teacher, Dr. Mehmet Yasin ULUKUŞ, who helped me at every opportunity by not sparing his knowledge, experience and valuable time in determining the subject of the study and at every other stage of the preparation process of the study.

June 2023 Bahri Selçuk EŞKİL

TABLE OF CONTENTS

Page

FOREWORD ..... **vii**
TABLE OF CONTENTS ..... **ix**
ABBREVIATIONS ..... **xi**
LIST OF FIGURES ..... **xiii**
SUMMARY ..... **xiii**
ÖZET ..... **xiii**
1. INTRODUCTION **1**
2. CONCEPTUAL STUDY **3**
3. SYSTEM ANALYSIS **5**
4. LITERATURE REVIEW **9**
4.1 General Review 9
4.2 Reasons of Churn and Ways to Prevent It 10
4.3 Data Set Selection 11
4.4 Explanation of Other Churn Prediction Focused Studies 12
5. METHODOLOGY **17**
5.1 Data Preperation 17
5.2 Modelling 18
5.2.1 Extreme gradient boosing (XGBoost) 18
5.2.2 Logistic regression 18
5.2.3 Other machine learning algorithms 19
5.3 Performance Evaluation of Models 19
5.3.1 Confusion matrix 19
5.3.2 Accuracy 19
5.3.3 Recall 20
5.3.4 Precision 20
5.3.5 F-Score 20
5.4 Designing and Monitoring the Process to Prevent Customer Churn 20
5.4.1 Designing the process 20
5.4.2 Monitoring the process 21
6. IMPLEMENTATION **23**
6.1 About the Company and It's Departments Related to the Implementation 23
6.2 Data Set to Be Used in This Study and the Approach to the Problem 24
6.3 Data Preperation 25
6.4 Modelling 31
6.5 Overall Analysis and Interpretation of Findings 32
6.6 Evaluation of Alternative Solutions 33
6.7 Design of Process and Monitoring Method 34
7. CONCLUSION **37**
8. REFERENCES **39**

ABBREVIATIONS

AUC **:** Area Under Curve
CRM **:** Customer Relationship Management
mRMR **:** Minimum Redundancy Maximum Relevance
SMOTE **:** Synthetic Minority Over-sampling Technique
XGBoost **:** Extreme Gradient Boosting

LIST OF FIGURES

Page

Figure 5.1 : Churn prediction modeling. **17**
Figure 5.2 : Structure of confusion matrix **19**
Figure 6.1 : Validation results of different models **31**
Figure 6.2 : Hold-out test results **32**
Figure 6.3 : Feature importances for XGBoost. **32**
Figure 6.4 : Designed process flow. **34**

CHURN PREDICTION WITH MACHINE LEARNING

SUMMARY

Companies that want to be competitive and ensure customer loyalty have to be aware of churn prediction studies. In this paper, the main goal is to reduce the churn rate of KKBox, one of the biggest music streaming companies in Asia. Prior to the development of the machine learning model, which is aimed to be the main element of the process that will enable this goal to be achieved; conceptual study, system analysis, literature review and methodology research were emphasized. Afterwards, datasets to be used in the development of the machine learning model were prepared and 156 new features were created with a feature generation study that could set an example for studies in similar sectors. After increasing the number of minority class samples in the dataset with SMOTE and solving the data imbalance problem, the modeling phase was started. In the modeling phase, Extreme gradient boosing (XGBoost), Logistic Regression and Random Forest algorithms were used. In order to determine the optimum hyperparameter values of these algorithms, grid search was used in the validation part. As a result, the optimum algorithm and hyperparameter values found were XGBoost with tree number 80, max depth 9, learning rate 0.05 and subsample 0.8. This algorithm achieved approximately 71% recall, 42% precision and 94% accuracy score. Then, the process in which this developed machine learning model will be used is designed and how it will be monitored is explained. As a result of this designed process being carried out efficiently by KKBox's data scientists and CRM experts, it is aimed to increase the satisfaction of the company's customers, decrease the churn rate and increase the competitive power of the company.

MAKİNE ÖĞRENMESİ İLE MÜŞTERİ KAYBI TAHMİNLEME

ÖZET

Rekabetçi olmak ve müşteri sadakatini sağlamak isteyen şirketler, müşteri kayıplarını tahmin etme çalışmalarından haberdar olmak zorundadır. Bu yazıda asıl amaç, Asya'nın en büyük müzik akışı şirketlerinden biri olan KKBox'ın kayıp oranını azaltmaktır. Bu amaca ulaşılmasını sağlayacak sürecin ana unsuru olması hedeflenen makine öğrenmesi modelinin geliştirilmesi öncesinde; kavramsal çalışma, sistem analizi, literatür taraması ve metodoloji araştırmasına ağırlık verilmiştir. Daha sonra makine öğrenmesi modelinin geliştirilmesinde kullanılacak veri setleri hazırlanmış ve benzer sektörlerdeki çalışmalara örnek olabilecek bir değişken yaratma çalışması ile 156 yeni değişken oluşturulmuştur. SMOTE ile veri setindeki azınlık sınıfın örneklem sayısı artırılıp veri dengesizliği sorunu çözüldükten sonra modelleme aşamasına geçilmiştir. Modelleme aşamasında XGBoost, Logistic Regresyon ve Random Forest algoritmaları kullanılmıştır. Bu algoritmaların optimum hiperparametre değerlerini belirlemek için grid search tekniği kullanılmıştır. Sonuç olarak, bulunan optimum algoritma ve hiperparametre değerleri; ağaç sayısı 80, maksimum derinlik 9, öğrenme oranı 0,05 ve alt örneklem 0,8 olan XGBoost'tur. Bu algoritma yaklaşık %71 recall, %42 precision ve %94 accuracy oranlarını elde etmiştir. Daha sonra geliştirilen bu makine öğrenmesi modelinin hangi süreçte kullanılacağı tasarlanmış ve sürecin nasıl sürekli olarak monitörize edileceği anlatılmıştır. Tasarlanan bu sürecin KKBox'ın veri bilimcileri ve CRM uzmanları tarafından verimli bir şekilde yürütülmesi sonucunda, şirket müşterilerinin memnuniyetinin artırılması, müşteri kaybı oranının düşürülmesi ve şirketin rekabet gücünün artırılması amaçlanmaktadır.

1. INTRODUCTION

Churn is a concept that is generally used for companies that sell goods/services periodically. It means that customers stop buying products/services due to competitor companies or dissatisfaction. In today's world, where competition has increased significantly and the concept of customer relationship management has gained importance, it has become very important for companies to be able to predict customer churn. The most common way to create processes to solve this problem is machine learning. A machine learning algorithm is a computational process that uses input data to perform a desired task. These tasks are usually related to the prediction of a dependent variable according to other variables or the classification of various instances. These algorithms change their architecture using various mathematical methods through iteration (i.e. experience) to become progressively better at accomplishing the task (Naqa and Murphy, 2015). So, various machine learning algorithms are also used for churn prediction. In this thesis, it is aimed to predict churn using customer data of KKBox, a major music streaming company, and to make this model the main tool for a process that will reduce the churn rate. In this paper; a conceptual study, literature review and methodology study were carried out for the mentioned subject.

The aim of the thesis is to develop a process for predicting the churn of customers and taking actions to prevent it, which is a more effective and less costly approach, instead of an aggressive strategy of attracting potential customers and increasing the number of customers. As the main element of this process; one of the machine learning models created with a data set containing customers' transactions, user logs and demographic data will be selected using the necessary performance metrics.

2. CONCEPTUAL STUDY

Most simply; customer churn is defined as; a customer's decision not to continue receiving services from a company. It can also be said that customer churn is the phenomenon of a customer stopping to buy goods or services from the company within a time period that varies according to the sector or company. Depending on these definitions; analytical examination of the probability of a customer leaving a service or product can be defined as customer churn prediction. The main purpose is to perform some preventive actions after predicting that the customer will leave the product or service (Çelik and Osmanoğlu, 2019). Customer churn is a common problem in many industries, especially businesses that periodically sell a product or provide a service. Acquiring new customers requires a lot of resources and effort, and every business has invested in acquiring existing customers. Based on this, it can be said that there is a loss of investment every time a customer leaves. In addition, the churn of a customer creates an opportunity cost due to the money that may be earned from that customer in the future. Predicting which customers will churn and trying to give them reasons to stay can be very rewarding for any business.

The level reached in technology, the emergence of new companies, the regular exposure of customers to advertisements and the increase in customers' market awareness and expectations lead to increased competition. This leads to an increase in the concern of creating stronger customer ties, especially among companies in sectors where the customer can easily change the company from which they buy the goods or services. In addition, it has become much more important to ensure customer loyalty in sectors such as telecommunications, where it is very difficult to increase the total number of customers. Customer loyalty is the desire of a customer to continue working with the company he is currently working with as long as he can reach the same quality product/service, and the loyal customer is generally not affected by the strategies and advertisements of other companies. Building loyalty bonds requires a well-established process and a systematic approach to its management. At this point, the concept of the customer lifecycle should be mentioned. The customer lifecycle refers to the process of customers becoming aware of a product/service, making a purchase from a firm,

and ideally becoming a company's longtime customer. Companies should define and implement policies that will enable their core assets, their customers, to reach and prolong their maximum commercial potential. In other words, they must implement appropriate business actions for each stage of the lifecycle and extend the life of their client portfolio as much as possible. In this way, it is possible to provide high efficiency in terms of value (García et al, 2015). After all, it is about churn rate concept to extend the life of the customer portfolio as much as possible. The churn rate means the percentage of customers who stopped using your company's product or service in a given time period. It is estimated that with just a 5% increase in customer retention rates, the average net present value of a customer could be increased by 35% for software companies and 95% for advertising agencies. Therefore, a company needs to shift its strategic focus from acquiring customers to retaining customers by reducing the churn rate (Ahn et al, 2006).

In this project, a model will be built to predict whether a subscription user will churn or not, using a data set from KKBox. KKBox is Asia's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 90 million tracks. After the necessary data preprocessing steps to make this raw data ready for modelling, various machine learning models will be developed and the most efficient model will be selected. With this model, it will be possible for KKBox to take special actions for customers that are predicted to be churn.

3. SYSTEM ANALYSIS

In this study, customer data of a company named KKBox, which competes in the music streaming industry, was used. KKBOX is a large company that offers subscription-based music streaming service and has a music library of around 90 million songs. The main goal of the system is to provide a music listening experience that meets the expectations of customers. It can be said that the components used by the system to achieve this goal are music files, application, servers, user accounts, payment systems etc. In today's world, where competition has increased significantly and the concept of customer relationship management has gained importance, building the loyalty of customers has become very valuable for companies. Especially in music streaming industry, since it is very easy for customers to change the company from which they buy services, if the customer churn rate of the firm is reduced, the profit and thus the competitiveness of the company can be increased. At this point, the system environment can be summarized as; customers who have not entered the system, competitors, legal regulators, and technological developments. So, the boundary of this system can be determined as creating an account for KKBox's application and purchasing a subscription. This means that the customers who have chosen KKBox should be counted in the system and reducing the churn rate of the customers in the system is the focus of this study. The opportunities for this system are as follows:

- The increase in demand in the global music market
- The possibility of increasing customer satisfaction and thus profits by retaining customers already in the system
- Technological innovations that improve the music listening experience
- The system already has a large data set containing customers' transactions, user logs and demographic data

In addition, threats for this system are as follows:

- Increasing competition in the industry and the strengthening of existing competitors
- Due to licensing costs, conflicts with the companies holding the music broadcasting rights

- Possibility of changing music listening habits
- Potential to incur large costs in terms of time and money due to the constant storage and use of large customer data

The area for improvement in this system is that the company does not currently have a process focused on reducing the churn rate. As explained later in this study, such systems are essential for volatile industries (where customers can easily switch providers) such as the music streaming industry. The aim of the thesis is to design a process for predicting the churn of customers and taking actions to prevent it, which is a more effective and less costly approach, instead of an aggressive strategy of attracting potential customers and increasing the number of customers. It is necessary to be aware of system constraints that may affect the project. This project requires the firm to have customer data. So, while creating and using such a system, there will be legal constraints that must be followed continuously due to data privacy. Additionally, the size of customer data for such a large firm would be very large, and storing and using this data for a model can be costly in terms of time and money. Secondly, maintaining the use of the machine learning model in the process requires a team of data scientists with high knowledge of statistics, learning algorithms and Python programming language. In addition, for the process designed to be efficient, CRM managers and data scientists will need to work together in coordination. Lack of cooperation between departments can greatly reduce efficiency. Lastly, it can be said as a constraint that the predictive performance of the learning model used in the process must be above a certain level for the designed process to be successful.

The system requirements related to the design made in this study are generally due to the steps in the flow chart of the designed process. According to the process designed in the project; the company should have the data of its current customers, this data should be able to be pulled from the CRM system or other databases, this dataset should be used in the trained learning model and the customer churn should be predicted, the actions to be taken for the customers predicted to churn should be listed by the CRM managers, and finally, the selected preventive strategies should be implemented. In addition, the process designed in this project can be monitored using various performance metrics, and if the performance of the process falls below a threshold, the model will be retrained. So, in summary, it is necessary to constantly have and update the customer data in the system, to have CRM managers who can

decide and implement the actions to be taken for the customers who are predicted to churn, and to have data scientists who can monitor and update the learning model used. As mentioned before, a process focused on reducing the churn rate was designed and a machine learning model was used as the main tool of this process. It is necessary to mention the impact of this design on the stakeholders of the system. As the expected result of the project is to design a new process to increase profit and customer satisfaction, it will be easier for the top management to ensure that the company targets are achieved. The company's data scientists and CRM (Customer Relationship Management) department will also be affected as they will be responsible for the implementation and continuity of this project. It will be easier for the company's CRM (Customer Relationship Management) department to identify customers who need more support, service or information. The company's competitors will be negatively affected by struggling to attract new customers. Lastly, the quality of service received by the firm's current customers will increase. In addition, it can be said that the company's investors, business partners and artists with songs in the application will be indirectly affected by the benefits of such a project.

After the implementation of this project in the system, the main criterion that can be used to monitor the success is the churn rate. Moreover, customer satisfaction rate and average customer lifetime metrics should also be used. In addition, metrics such as accuracy and precision will be used to monitor the machine learning model used. This part is explained in more detail in the following sections.

4. LITERATURE REVIEW

4.1 General Review

The popularity of customer churn analysis studies using machine learning has increased considerably in recent years; especially due to the increased competition, the understanding of the importance of CRM studies for companies, and the advancement of techniques. Churn studies have extended from the telecommunications industry, where it is already widespread, even to the gaming industry (Kawale et al, 2009). This paper will study the data set provided by a company from music streaming industry, and will set an example as churn studies are not very common in this industry.

In the literature, the telecommunications sector is more active in terms of customer churn predictive modeling research (46% of the publications are related to this sector), followed by the banking sector with 23%. The high rate of churn studies in the telecommunication companies is understandable since it is a very volatile sector (García et al, 2015). It should also be noted that the annual churn rate is between 20-40% for the telecommunications industry (Xu et al, 2021). Acquiring new customers is not easy in a competitive banking market and banks are generally customer oriented, so their primary goal is to retain existing customers (Guliyev and Tatoğlu, 2021). Additionally, churn prediction is an important issue for many sectors such as streaming, e-commerce, insurance, medical, retailing, where periodic sales or subscription sales are targeted. The definition of churn differs slightly in data sets depending on the firm (and thus the industry). However, most of the data sets are collected over a period of three to six months and is checked to see if customers are subscribed in the month following the data collection period; in general, a customer who has unsubscribed is considered to be churn. But, a gaming company may consider it churn if a player does not play the game for 2 months, or for an e-commerce company, a customer not shopping for 3 months can be considered churn.

According to a research; the cost of churn prediction is 16 times less than acquiring new customers, and reducing the churn rate by 5% increases profits by 25-85% (Xu et al, 2021). In another source, it is said that acquiring new customers in today's

competitive conditions is 10 times more costly than retaining existing customers, according to researchers (Çelik and Osmanoğlu, 2019). Although the rates given in the literature are not exactly the same, it is clear that it is true that acquiring new customers is more expensive.

4.2 Reasons of Churn and Ways to Prevent It

It has already been mentioned that churn is a major problem for many industries. But a proper understanding of the causes of churn will be useful for feature selection in the later stages of the study. Possible causes of churn are listed as follows (Jain et al, 2020):

- Customers may purchase services/products and do not know all of their content/benefits, are not interested in learning, or do not want to use them. It can be summarized as a lack of participation.
- The customer may lose interest in the product/service because the company does not offer customer-specific benefit packages or new offers. In this case, the customer usually starts working with other companies that better promote their products/services.
- Consumers may not be getting satisfactory answers from the customer service department.
- The customer may be changing his location, so he may be changing the service provider.

Two more possible reasons can be added as follows (Johny and Mathai, 2017):

- Customers may want newer technology/service that cannot be offered by their current service/product providers.
- The customer may want to stop buying the product/service due to unexplained social or psychological factors.

After predicting the customers who will churn, a process should be followed, which includes selecting and implementing a few of the following strategies:

- Information about the contents and offers of products/services can be provided.

- Offer incentives such as special offers and discounts, after making sure that the costs of the retention program do not outweigh the profits from the customers you intend to save.
- By establishing customer-specific communication networks; make the customer feel special and ensure that the customer receives fast and quality service from the customer service and support department.
- Free trainings, seminars or product demos can be offered on current products and services.
- Try to understand the customer's complaints and make him feel that the service/product provider is trying to solve the problem.

4.3 Data Set Selection

The first step in the process of developing a churn prediction model consists of collecting relevant data and selecting candidate explanatory variables. Raw data is the initial form of data collected from various sources. That is, it has not yet been processed, cleaned and organized. On its own, raw data doesn't make much sense but has the potential to be processed for analysis. If the data is obtained from different sources, it is necessary to collect the data in tabular format in one place and decide which part of the data to keep in which data file. Parts containing errors and uncertainties should be removed from the raw data. Additionally, the columns within the data files must be properly and precisely labeled (Jain et al, 2020). It is a known fact that the quality of the data set to be used in model development and its suitability for the problem determine the accuracy and predictive power of the resulting model. So, one of the important issues of churn prediction studies has been to identify and obtain the best data set. It should also be noted that in the search for the best data set, the cost of acquiring data must also be considered.

Data sets that contain columns such as socio-demographic data (for example; gender, age, or zip code), customer behavior data, and number of calls to the customer help desk are suitable for churn prediction (Verbeke et al, 2012). In addition, features such as customer account information (for example; first date of using the services, types of service packages), complaint information, and historical information of bills and

payments are also suggested for customer churn prediction in the resources (Huang et al, 2012).

In the literature, there is a disagreement about the required data set size for a successful churn prediction model, due to the performance-accuracy trade-off. There is a source that shows using 4 times larger training data increases the success of the model by 15%, and in the same source, it is argued that every feature that provides even a little information about the target value should be added to the model (Huang et al, 2015). Contrary to this idea; there is also a paper that argue that the model speed may decrease if there are many features in the data set, and therefore, the number of features should be reduced in each churn prediction model during the data preparation phase. However, it is noted in this paper that converting the data to binary form may increase the model speed (Jain et al, 2020). In one source; from an economic point of view, it has been argued that investing in data quality is more efficient than collecting a comprehensive range of features that gathers all available information about customers. In the same source it is stated that usually six to eight variables are sufficient to predict churn with high accuracy (Verbeke et al, 2012). In addition, it has been observed that the data set sizes used in the literature vary between approximately 2000 and 1 million (Jain et al, 2020).

4.4 Explanation of Other Churn Prediction Focused Studies

Vafeiadis et al. (2015) have published a study with the goal of comparing the most popular machine learning methods applied to the demanding customer churn prediction problem in the telecommunications industry. In this study, artificial neural network, support vector machines, decision trees learning, Naïve Bayes, regression analysisis and logistic regression analysis methods, which can be used for classification in machine learning, are examined in terms of reliability, efficiency and popularity. It has been deduced that logistic regression gives good results if a successful data preparation is made, that Naïve Bayes can give successful results for the wireless telecommunications industry, and that the artificial neural network approach can outperform decision trees or logistic regression. In the study, precision, recall, accuracy and F-measure were defined and used to compare the performance of classifiers.

Qureshi et al. (2013), used data mining techniques to predict customers who will churn. The dataset provided from the Customer DNA website used in this study includes data of 106,000 customers and their usage behavior for 3 months. In addition, in the data set used in this study, there were only 5.9% churn customers versus 94.1% active users. So, the use of resampling methods to solve the problem of data imbalance is also discussed in the study. It was decided to use the oversampling method to solve the imbalance problem, and thus the churners to active ratio was brought closer to 40:60. Logistic regression, artificial neural networks and decision trees methods were used in the study and as a result, it has shown that decision trees are the most accurate algorithm for churn prediction. In this study, precision, recall and F-measure were used to compare the performance of different prediction models.

Ahmad et al. (2019), have published a study that aims to develop a churn prediction model that helps SyriaTel telecom company to predict customers who will churn. The dataset in the study included the information of all customers for 9 months. In this study, the standard measure of Area Under Curve (AUC) was used to measure the performance of the models, and the resulting AUC value was 93.3%. The data set was split by 70% for training and 30% for testing. For validation and hyperparameter optimization, 10-fold cross-validation was performed. During the data preparation phase, feature engineering was done and the data imbalance problem was solved by undersampling. Decision tree, random forest, gradient boosting machines and extreme gradient boosting (XGBoost) techniques were used in the study, and the best results were obtained by XGBoost algorithm. The gradient boosting machine algorithm came in the second place and the random forest algorithm came in the third place.

Lalwani et al. (2022), published a study consisting of 6 phases. In the first two phases, data preprocessing and feature analysis were performed. The third phase was about feature selection using the gravitational search algorithm. In the next phase, they split 80% of the dataset for train and 20% for testing. Then, logistic regression, naive bayes, support vector machine, random forest and decision trees algorithms were used to create models, and in addition, boosting and ensemble techniques were used to see the effect on the accuracy of the models. It should also be noted that K-fold cross validation is used for hyperparameter tuning and preventing overfitting. As a result, confusion matrix and AUC curve were used to compare the performances of the

models, and it was observed that the XGBoost algorithm gave the best result with the AUC score of 84%.

Rahman and Kumar (2020), published a churn prediction study focused on the banking sector. They stated that, as the level of competition within the banking industry has grown, the importance of predicting and preventing customer churn has become a key focus for organizations in this sector. In the dataset used in this study, there were features such as customer ID, surname, credit score, location, gender, age, tenure, bank account balance, number of bank products used by the customer, whether a credit card is used or not, and salary. The dataset was taken from Kaggle and in this study, K-nearest neighbor, support vector machine, decision tree and random forest algorithms were used to create machine learning models. The dataset they used had information about 10000 customers and of this, 7963 were positive class (maintained) samples and 2037 were negative class (exited) samples. After the data preprocessing step, they used the Minimum Redundancy Maximum Relevance (mRMR) method and relief method for feature selection. Additionally, they solved the problem of data imbalance by oversampling, since the size of the dataset would be greatly reduced if undersampling was preferred. As a result of the performance comparison, it was seen that the algorithm with the highest predictive performance was random forest.

Tékouabou et al. (2022), published a study on the importance of solving the data imbalance problem in improving the performance of churn prediction models. In this study, it was stated that the diversity of data collected today increases the problem of heterogeneous variables, which are generally not suitable for classification algorithms. Additionally, it has been said that despite significant improvements in the efficiency of machine learning-based analytics for classification in customer relationship management systems (CRM), their performance remains limited due to heterogeneous data processing and data imbalance. Additionally, this situation is more effective for simple machine learning methods that often suffer from overfitting. So, this study proposes to create a machine learning model using SMOTE (synthetic minority oversampling technique) to solve the data imbalance problem. Undersampling method is not recommended in this study either, as it can lead to data information loss. As a result of the study, it has been shown that the predictive performance can be increased considerably by using SMOTE. For example, the F1-Score of the model created with the K-nearest neighbor algorithm was 0.12 without SMOTE, while it was 0.70 with

SMOTE. In another example, the F1-Score of the model created with the random forest algorithm was 0.58 without SMOTE and 0.86 with SMOTE.

Malyar et al. (2020), published a study in which ensemble tree methods (random forest, XGBoost, LightGBM) were chosen as learning algorithms. They begin by determining how many days it would be considered a churn for a customer not to purchase a product/service based on the dataset they are working on. Then, they labeled the dataset based on this information. After the modeling phase; the best result was shown by the XGBoost algorithm with an AUC score of 0.8322. The worst result was shown by the usual decision tree with an AUC score of 0.65 and the LightGBM algorithm also performed very closely on XGBoost, with an AUC score of 0.818. As a result, it was stated that it is more effective to use a collection of trees rather than a single tree.

5. METHODOLOGY

The churn prediction modeling process is visualized as in figure 5.1 (Eria and Marikannan, 2018).

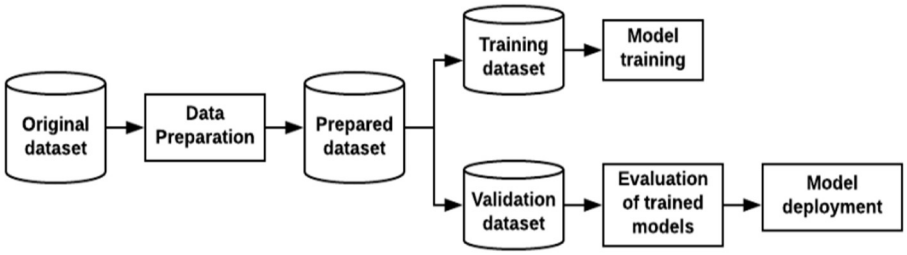

Figure 5.1 : Churn prediction modeling.

5.1 Data Preperation

Data preparation is a process aimed at converting categorical and continuous independent variables into a suitable form for further analysis. A properly done data preparation is very important for overall churn prediction performance. Because it can lead to improvements of up to 14.5% measured by AUC (Coussement et al, 2017). Data preprocessing is an important step in the data mining and data analysis process, generally related to the conversion of raw data into a format suitable for machine learning models. Machine learning models need clean and uniform data. However, regardless of the context, real-world data contain errors and inconsistencies, and are often incomplete. At this stage, data quality assessment will be made first. In other words; whether there are different identifiers for the features or the existence of outliers or missing values etc. will be checked. Then, filling in the missing data and correcting, repairing or removing the incorrect data in the data set, that is, data cleaning will be performed. Next, data transformation methods such as normalization, feature selection and aggregation will be applied if deemed necessary. Finally, data reduction methods will be applied depending on the performance targets. In addition, churn datasets often have data imbalances. This imbalance is due to the relatively low rate of churning customers in the data. In a different study that used an imbalanced dataset, it was observed that the use of SMOTE (Synthetic Minority Over-sampling Technique)

could increase the predictive performance of the machine learning model (Safitri and Muslim, 2020). SMOTE is a method of oversampling that involves generating new minority class samples by combining multiple minority class cases that are similar to each other (Joshi, 2019). Therefore, in this study, implementing SMOTE during data preprocessing can increase performance.

5.2 Modelling

The following machine learning algorithms are planned to be used in the modeling phase.

5.2.1 Extreme gradient boosing (XGBoost)

The extreme gradient boosting algorithm has started to become very popular in recent years as it pushes the limits of computational success (Çelik and Osmanoğlu, 2019). Although its interpretability is expected to be lower depending on the conditions, it was preferred for this study because it is a new algorithm that stands out with its exceptional predictive capacity (Pesantez-Narvaez et al, 2019). In addition, advantages such as being faster than other boosting algorithms as it can use the power of parallel processing and being able to process missing values itself, were also factors in the preference of this algorithm.

5.2.2 Logistic regression

Although its accuracy largely depends on variable selection, logistic regression which is a linear model, is a strong and effective way of interpreting the effects/importance of independent variables based on a binary target, as in churn estimation (Stoltzfus, 2011). Interpretability, is a very important aspect of a churn prediction model as it enables a firm's marketing department to extract valuable insights from the model so that they can design effective strategies and retention actions (Verbeke et al, 2012). In addition, it has been observed that the accuracy of logistic regression can be 85.2385 % in churn prediction studies (Jain et al, 2020). To summarize, this algorithm was chosen because of its high interpretability and possibility of successful prediction.

5.2.3 Other machine learning algorithms

It should also be noted that one of the random forest, decision tree or support vector machine algorithms is planned to be used in model building for comparison and selection of the best model. Algorithm selection will be made depending on future research and the needs that will arise during the study.

5.3 Performance Evaluation of Models

5.3.1 Confusion matrix

The confusion matrix is a matrix output that shows the overall performance of the model. Its structure can be seen in figure 5.2.

	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Figure 5.2 : Structure of confusion matrix.

5.3.2 Accuracy

Accuracy is the ratio of total right prediction to the total number of samples. Accuracy shows the correct prediction rate, but sometimes it gives a false sense of high achievement in data sets with an unbalanced target variable ratio. For example, in churn-related data sets, the rate of customers who churn usually does not exceed 10%. And if a data set consists of 90% non-churn customers and 10% churn customers, then the model can easily achieve 90% accuracy. However, 90% accuracy may not be an indication of real success, as this accuracy may also have occurred because the model predicted that all customers will not churn. For this reason, accuracy is less important than other metrics in this study.

5.3.3 Recall

Recall is the ratio of positive samples correctly classified to the all positive samples. This metric is important for churn prediction. Because, the more customers we predict that will actually churn, the more the company's earnings will increase.

5.3.4 Precision

In brief, it is what percentage of customers that are predicted to leave actually do. If most of the customers that we predict will churn, do not actually churn, this will create an extra expense. However, the fact that this ratio is a little low is not an overly important problem, it may even be beneficial, as long as the cost does not increase excessively, as it will create the perception that our customer communication is successful.

5.3.5 F-Score

It is defined as the harmonic mean of the model’s precision and recall. F-Score value comes near the lowest values of precision and recall.

5.4 Designing and Monitoring the Process to Prevent Customer Churn

5.4.1 Designing the process

In previous sections, based on the available literature it was stated that, it is less costly for companies to retain existing customers than to attract new ones. Therefore, in recent years modern companies are increasingly emphasizing a customer-oriented approach, rather than primarily product-oriented. This means that companies must understand their customers' needs, preferences and purchasing habits in order to effectively manage their relationships with them. And processes related to customer management are typically the focus of the Customer Relationship Management (CRM) department within a company (Emtiyaz and Keyvanpour, 2012). So, as it will involve implementing strategies for retaining customers who have been identified by a machine learning model as likely to leave, the process to be designed in this study is likely to fall within the responsibility of a team consisting of CRM managers and data scientists working at KKBox. The machine learning model that is chosen after

evaluating the performance of trained models will be used to make predictions of customer churn in the process that is being designed.

5.4.2 Monitoring the process

Machine learning models that are being actively used actually make some statistical assumptions. As the model remains in use, it is common for the statistical properties of the data (e.g. mean, standard deviation or distribution) to change from the values that were present in the training data, potentially affecting the model's ability to accurately make predictions. These changes may cause the parameters for the evaluation of the model to fall below the required threshold, that is, negatively affect the predictive performance of the model and make it unusable (Ackerman et al, 2021). So, it is a fact that it is important to detect such a decrease in the quality of the machine learning model used. This can be detected by constantly monitoring the performance of the machine learning model in use. It is possible to use a combination of the metrics such as accuracy, recall, precision etc. to monitor model performance. When these metrics fall below a predetermined threshold, the model needs to be retrained with up-to-date data.

6. IMPLEMENTATION

6.1 About the Company and It's Departments Related to the Implementation

This study will use the data set provided by KKBox, Asia's leading music streaming company, which holds the world's most comprehensive Asian-Pop music library with over 90 million tracks. It is a Taiwan-based company targeting East and Southeast Asia in general. It received investment from Japanese firm KDDI in 2011, which now owns 76 percent of KKBox. KKBox operates on Windows, Windows Media Center, Mac OS X, iOS, Android, Symbian, Bada, partly Java and since 2015 Apple Watch. KKBox offers service supported by ads and paid subscriptions to millions of people and the firm emphasizes that this delicate business model depends on accurately predicting the churn of its paid users. Currently, the company uses survival analysis techniques to determine the residual membership life time for each subscriber. However, instead of this method, it may be much more effective to acquire a process that includes up-to-date machine learning techniques with a very high potential for success.

As previously stated, the process to be designed in this study involves predicting churn customers and selecting and implementing necessary actions for these customers to stay. Such a process will require the collaboration of CRM managers and data scientists in the company. The reasons for this are as follows:

- CRM managers are qualified to manage and analyze customer data. This data has an important role in the development and continued use of the churn prediction model.
- CRM managers are knowledgeable about customer behaviors and preferences. Therefore, they will be effective in deciding and implementing the actions to be taken for the customers who are expected to churn.
- Data scientists are qualified in statistics, Python, and machine learning techniques. These techniques are necessary to correctly implement and monitor the developed prediction model and update it when necessary.

- Collaboration of CRM managers and data scientists will help to understand the data properly, use the learning model correctly, and achieve the efficiency of the designed process.

6.2 Data Set to Be Used in This Study and the Approach to the Problem

When users sign up for KKBox's service, they can choose to renew the service manually or automatically. Users can actively cancel their membership at any time. For this business model, the prediction of the paying user churn is very important. Even small changes in customer churn can have a huge impact on profits.

The data set in question consists of millions of rows. This data is kept in 9 files named train.csv, train_v2.csv, sample_submission_zero.csv, sample_submission_v2.csv, transactions.csv, transactions_v2.csv, user_logs.csv, user_logs_v2.csv and members_v3.csv. Train and sample_submission files contain the user ids (column named "msno") and whether they have churned (column named "is_churn"). The sample_submission files contain the test set. "is_churn" column is the target variable. is_churn = 1 means churn, is_churn = 0 means renewal. Transactions files contain information such as user id, payment method, length of membership plan in days, actual paid amount, transaction date, membership expiration date, whether auto-renewal is made, whether or not the user canceled the membership in that transaction (column named "is_cancel"). user_logs files contain log data describing a user's listening behaviors. Members file contains users' demographics, service sign-up methods, and location.

To understand the data set and the problem to be solved in this project, it is necessary to explain how "churn" should be defined according to KKBox's business model. Since most of KKBox's subscription period is 30 days, many users resubscribe every month. Key features used for churn labeling are; transaction date, membership expiration date, and is_cancel. The is_cancel feature indicates whether a user has actively unsubscribed, and unsubscribing does not mean that a user is churn. Because, a user may unsubscribe from a service due to changes in service plans or other reasons. As a result, for this study "churn definition" is that there is no new and valid subscription to the service within 30 days after the expiration of the current subscription.

The train_v2 dataset includes customers whose subscription expires in February and who are labeled according to whether there is a churn or not in March. The sample_submission_v2 file includes customers whose subscription expires in March and who are labeled according to whether there is a churn or not in April. It should be noted that splitting operations by date in the datasets are done using the transaction date column. The features in the members_v3 file can be merged directly to the train_v2 and sample_submission_v2 data sets, but user log files and transaction files will be used to generate new features. Then, the generated features will be used by merging them to the train and test data. 2 dataframes will be created with the data of user logs and transactions before March (including March), and other 2 dataframes will be created with data before February (including February). Therefore, 4 dataframes will be created and these 4 dataframes of user logs and transactions will be used in the feature generation part. To illustrate, for example, when KKBox wants to predict customers who will churn in December at the end of November, they will have November data and previous months's data of their customers. For this reason, when generating new features, the activities of the customers in the month they were labeled and in the previous months will be used. In this way, it will be ensured that the churn prediction model reflects the reality and can be used in the designed process.

6.3 Data Preperation

At first, the train_v2 dataframe and the members_v3 dataframe were merged with left join using the "msno" column (containing the user ids) as the key. Left join was preferred as it returns all records from the left table (train_v2) and matching records from the right table (members_v3). As a result, a dataframe was created with 907960 rows and 7 columns. The sample_submission_v2 dataframe and the members_v3 dataframe were merged with left join using the "msno" column as the key. So, a dataframe was created with 907941 rows and 7 columns. These 2 dataframes (named train_v2_with_members_v3 and sample_submission_v2_with_members_v3) can be considered as base dataframes. Because the features that will be generated using user logs and transactions will be merged with these dataframes.

Transactions and transactions_v2 datasets are imported and appended. By using that dataframe containing all transactions of customers; transaction data from February and before February are transferred to a new dataframe named

transaction_before_february, and transaction data from March and before March into another dataframe named transaction_before_march. The user_logs dataset and the user_logs_v2 dataset together are about 400 million rows. A random sample of 19 million lines was taken from all user logs, as using all user log data available would not provide any information worth the computational cost it caused to the learning model. By using that sample dataframe containing user logs of customers; user log data from February are transferred to a new dataframe named user_logs_february_random_sampled, and transaction data from March into another dataframe named user_logs_march. Efforts were made to reduce the memory consumption of imported dataframes in order to increase the speed of the system where the project will start to be used. A function is defined for this purpose. In this function, numeric columns are converted to data types that consume as little memory as possible. Thanks to this function, the memory consumption of all dataframes was reduced by approximately 23.8 percent on average. Then, it was checked whether the transaction and user log dataframes were ready for the feature engineering part. In other words, for this purpose, it was checked for null, infinite or illogical values (anomalies) in all 4 dataframes, but no problem was found. Next, train_v2_with_members_v3 and sample_submission_v2_with_members_v3 dataframes (those previously described as base dataframes) were also checked for null, infinite, and illogical values. Approximately 100000 null values were found in the city, bd (age), gender, registered_via, registration_init_time columns in both dataframes. The null values in the city column and the registered_via column were replaced with 0 to mean "unknown", the null values in the bd (age) column were replaced with 28, the null values in the gender column were replaced with "unknown" (string data type), and the null values in the registration_init_time column were replaced with the median. In addition, the bd (age) column was found to have illogical values; greater than 75 or less than 0. These values are also equaled to 28.

Then, the feature generation part was started, using 2 transaction dataframes. The features created using transaction dataframes and their explanations can be seen below:

- discount_amount: This column is equal to the difference between the plan_list_price column and the actual_amount_paid column. (Negative values are set to 0.)

- is_discount: It is equal to 1 where the plan_list_price column value is greater than the actual_amount_paid column value, 0 otherwise.
- list_price_per_payment_plan_day: It is equal to the value of the plan_list_price column divided by the value of the payment_plan_days column.
- amount_paid_per_payment_plan_day: It is equal to the value of the actual_amount_paid column divided by the value of the payment_plan_days column.
- membership_duration_in_days: It is equal to the difference (in days) of the membership_expire_date column and the transaction_date column.
- auto_renew_not_cancel: In the case where the is_auto_renew column has a value of 1 and the is_cancel column has a value of 0, it equals 1, otherwise 0.
- not_auto_renew_is_cancel: In the case where the is_auto_renew column has a value of 0 and the is_cancel column has a value of 1, it equals 1, otherwise 0.
- transaction_count: It contains how many transactions customers have made so far.

Then, all these features created on the transaction dataframes were aggregated using max, min, sum, std, mean, mode and nunique. The dataframe formed as a result of the aggregation was merged with the base dataframes and the process of creating features from the transactions was completed. As a result, 38 features were created using transaction datasets.

Next, features were generated using 2 user log dataframes. The features created using these dataframes and their explanations can be seen below:

- total_login_number: It contains how many times a user has logged into the application so far.
- num_25_per_num_unq: It is equal to the num_25 column value (number of songs played less than 25% of the song length) divided by the num_unq (number of unique songs played) column value.
- num_50_per_num_unq: It is equal to the num_50 column value (number of songs played less than 50% of the song length) divided by the num_unq (number of unique songs played) column value.

- num_75_per_num_unq: It is equal to the num_75 column value (number of songs played less than 75% of the song length) divided by the num_unq (number of unique songs played) column value.
- num_985_per_num_unq: It is equal to the num_985 column value (number of songs played less than 98,5% of the song length) divided by the num_unq (number of unique songs played) column value.
- num_100_per_num_unq: It is equal to the num_100 column value (number of songs played less than 100% of the song length) divided by the num_unq (number of unique songs played) column value.
- sec_per_num_25: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_25 (number of songs played less than 25% of the song length) column value.
- sec_per_num_50: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_50 (number of songs played less than 50% of the song length) column value.
- sec_per_num_75: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_75 (number of songs played less than 75% of the song length) column value.
- sec_per_num_985: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_985 (number of songs played less than 98,5% of the song length) column value.
- sec_per_num_100: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_100 (number of songs played less than 100% of the song length) column value.
- secs_per_num_unq: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the num_unq (number of unique songs played) column value.
- total_not_unq_songs: It is equal to the sum of the columns num_25, num_50, num_75, num_985, num_100.

- repeated_song_number: It is equal to the difference between the total_not_unq_songs column and the num_unq column.
- avg_time_on_a_song: It is equal to the total_secs column value (how many seconds the customer listened at that login) divided by the total_not_unq_songs column value.

Then, all these features created on the user logs dataframes were aggregated using max, min, sum, std, mean, mode and nunique. The dataframe formed as a result of the aggregation was merged with the base dataframes and the process of creating features from the user logs was completed. As a result, 65 features were created using user logs datasets.

After these steps, the features we generated were merged into 2 dataframes; containing customers of February and customers of March. Various features were generated on these 2 resulting dataframes as well. The features created directly on these dataframes and their explanations can be seen below:

- registiration_duration_in_days: It is the difference (in days) between the membership_expire_date_max column and the registration_init_time column.
- long_time_user: It is equal to 1 if the registiration_duration_in_days column value is greater than 300, and 0 otherwise.
- membership_duration_in_days_registiration_duration_in_days_difference: It is equal to the difference between the membership_duration_in_days_mean column and the registration_duration_in_days column.
- activity_period_in_days: It is the difference (in days) between the date_max (the date of the last day logged in in that month) column value and the date_min (the date of the first day logged in in that month) column value.
- inactive_days: It is equal to the difference between the activity_period_in_days column and the active_day_number (the number of days logged in that month) column.
- avg_listening_time_per_day: It is equal to the total_secs_sum column value divided by the active_day_number (the number of days logged in that month) column value.

- unq_song_per_day: It is equal to the num_unq_sum (number of unique songs played in that month) column value divided by the active_day_number column value.
- days_left_for_membership_expiration: It is the difference (in days) between the membership_expire_date_max column and today's date. (Negative values are set to 0.)
- loyalty_range_in_days: It is equal to the difference (in days) between the transaction_date_max column and the registration_init_time column. (Negative values are set to 0.)
- days_since_last_transaction: It is equal to the difference (in days) between today's date and the transaction_date_max column. (Negative values are set to 0.)

After all, the feature generation part was finished, and as a result; a dataframe was obtained for customers with or without churn in March and their feature values, and another dataframe for customers with or without churn in April and their feature values. These dataframes are appended and again randomly split into two dataframes named train and test (with a test size ratio of 0.3).

As mentioned earlier, machine learning algorithms need clean and uniform data. So, the process of preparing the train and test dataframes for modeling began. First of all; msno, registration_init_time, transaction_date_max, membership_expire_date_max, date_max,membership_expire_date_max_formatted,registration_init_time_formatted , date_min, date_max_formatted, transaction_date_max_formatted columns are dropped because they cannot be used in modelling. One hot encoding was used for the categorical features: city, gender, registered_via, payment_method_id_mode. Rows with null target variable column value are dropped. Columns with more than 28 percent null were dropped, and the remaining null values were filled with the median value in their columns. Train and test data were normalized with Min-Max Scaling and in this process, attention was paid to normalize the test data with the train data parameters and to avoid data leakage.

It was observed that the total churn rate in the train and test data was 4.88%. To solve this data imbalance problem, SMOTE (Synthetic Minority Over-sampling Technique) was used as stated in the methodology section. In this way, 1192851 new minority

class samples were generated and the churn rate was artificially equalized to 50% in the train data.

6.4 Modelling

In the modeling phase, as stated in the methodology, Extreme gradient boosing (XGBoost), Logistic Regression and Random Forest algorithms were used. In order to determine the optimum hyperparameter values of these algorithms, grid search was used in the validation part. The hyperparameter values used in grid searches for XGBoost and Random Forest algorithms are given below:

- For XGBoost: The values of 64, 80, and 96 were tried for the number of trees parameter. For the max depth parameter, values of 3, 6, 9, and 12 were experimented with. The learning rate parameter was tested with values of 0.05, 0.1, and 0.2. And, the subsample parameter was varied with values of 0.8 and 1.
- For Random Forest: The number of trees parameter was experimented with values of 64, 80, and 96. The max depth parameter was tested with values of 4, 8, and 12.

Notable validation results and hyperparameter values can be seen in the table below:

Algorithm	Parameter Values	Recall	Precision	F-Score	Accuracy
XGBoost	Tree number=80, Max depth=9, Learning Rate=0.05, Subsample=0.8	0,7004	0,43	0,53	0,942
XGBoost	Tree number=96, Max depth=12, Learning Rate=0.2, Subsample=1	0,522	0,641	0,576	0,963
Random Forest	Tree number=80, Max depth=12	0,774	0,329	0,461	0,915
Logistic Regression	-	0,73	0,18	0,3	0,83

Figure 6.1 : Validation results of different models.

The hold-out test method was used to ensure that the models that were successful in validation could also be successful in unseen data. The results of this test can be seen in figure 6.2.

Algorithm	Parameter Values	Recall	Precision	F-Score	Accuracy
XGBoost	Tree number=80, Max depth=9, Learning Rate=0.05, Subsample=0.8	0,709	0,42	0,53	0,942
XGBoost	Tree number=96, Max depth=12, Learning Rate=0.2, Subsample=1	0,53	0,636	0,578	0,964
Random Forest	Tree number=80, Max depth=12	0,78	0,325	0,459	0,915
Logistic Regression	-	0,74	0,19	0,3	0,84

Figure 6.2 : Hold-out test results.

6.5 Overall Analysis and Interpretation of Findings

First of all, at first glance, it is striking that the accuracy values are high in both validation and hold-out stages. As mentioned earlier, accuracy sometimes gives a false sense of high achievement on datasets with an imbalanced target variable ratio. For this reason, accuracy is not one of the primary metrics that we will use in choosing a model. In the project, the most important metric should be recall. Because in this study, recall shows what percentage of customers who will be churn are predicted. And that is the primary purpose of this study. The second most important metric is precision. Because in this study, it refers to how many of the customers predicted to be churn by the model actually churn. During the modeling process, it was observed that there is a trade-off between recall and precision.

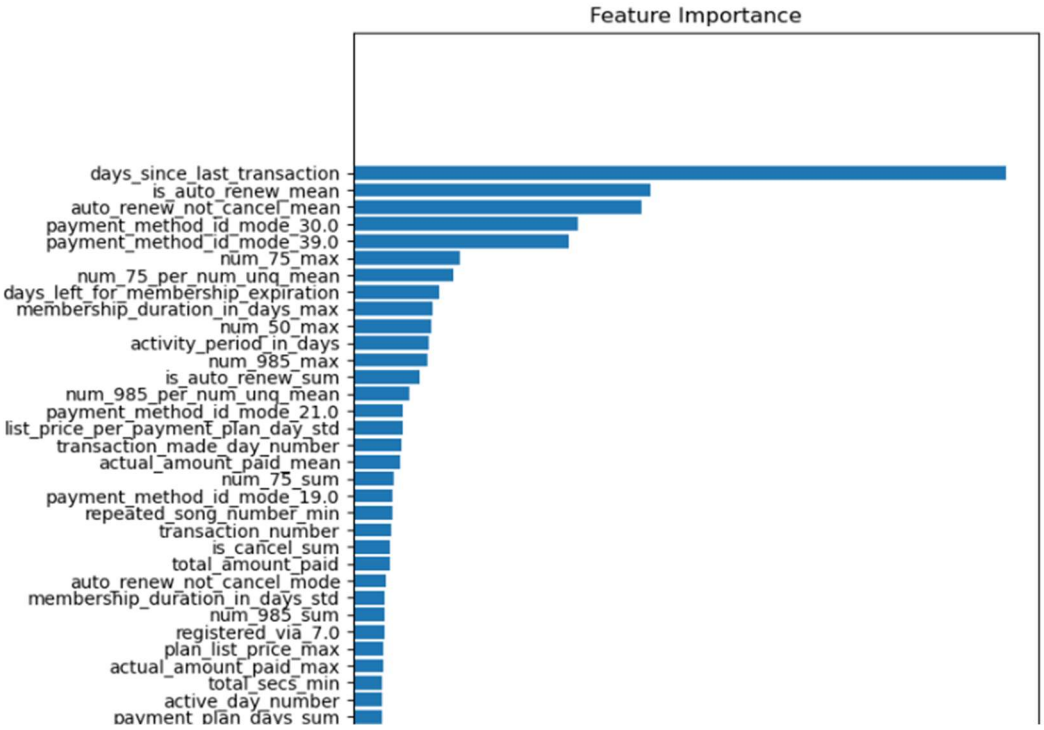

Figure 6.3 : Feature importances for XGBoost.

The goal in this case is to opt for a slightly better model in terms of the recall metric than precision. The reason for this will be explained later. Additionally, the top of the feature importance graph of one of the models developed with XGBoost can be seen in figure 6.3.

It has been observed that many designed features can have a great impact on the model. Additionally; it makes sense that features such as days_since_last_transaction, actual_amout_paid_mean, active_day_number etc. are most relevant to the churn situation. As a result of the findings, it has been seen that it is possible to make churn predictions and benefit from the data that the company can currently access.

6.6 Evaluation of Alternative Solutions

First, although the logistic regression has a high recall value in both the validation and hold-out test part, the precion value is too low. Precision is too low when recall is high, indicating that the model is overestimating churn to be able to predict most of the customers that will churn. For example; assuming 10000 customers will churn, to predict 7400 of them, the model claims 39000 people will churn. As we have stated in the study before; the fact that precision is a little low is not an overly important problem, it may even be beneficial as it will create the perception that firm's customer communication is successful. However, this does not mean that an extremely low precision value is also acceptable. Extremely low precision will increase costs by increasing unnecessary churn predictions. Therefore, only a slightly lower precision is acceptable. According to the results, logistic regression was able to give a very low precision value and therefore it can be said that it will increase costs. For this reason, it should be eliminated.

F-score is defined as the harmonic mean of the model's precision and recall. The XGBoost model, which was created with 96 trees, obtained the highest f-score value. It may be thought that this model should be preferred, but it should be noted that the f-score value in question comes from 0.63 precision and 0.53 recall values. For the process designed in this study, it is preferred that the recall be higher than precision. Because the priority is to predict customers who will churn with a high success rate; as explained earlier, it is preferable to have a slightly lower precision. Therefore, this model should also be eliminated.

The XGBoost model created with 80 trees, 9 depth and the Random Forest model created with 80 trees, 12 depth remained as options. Scenarios can be compared to choose one of these 2 models:

- Choosing the XGBoost model: The XGBoost model achieved approximately 0.71 recall and 0.43 precision. In other words, when this model is used, it will be possible to predict customer churns with high success without making too many wrong churn predictions. In addition, choosing this model will significantly reduce the time cost of the designed process compared to other algorithms, as the XGBoost algorithm can be easily run on the GPU and can process the missing values itself.
- Choosing the Random Forest model: The Random Forest model achieved approximately 0.78 recall and 0. 32 precision. With this model, customers who will churn can be predicted with very high success (78%), but slightly more churn will be predicted. This situation will cause unnecessary churn preventive actions to be taken and create additional costs.

As a result, the optimum algorithm and hyperparameter values found were XGBoost with tree number=80, max depth=9, learning rate=0.05, subsample=0.8.

6.7 Design of Process and Monitoring Method

The designed process flow can be seen in the figure 6.4.

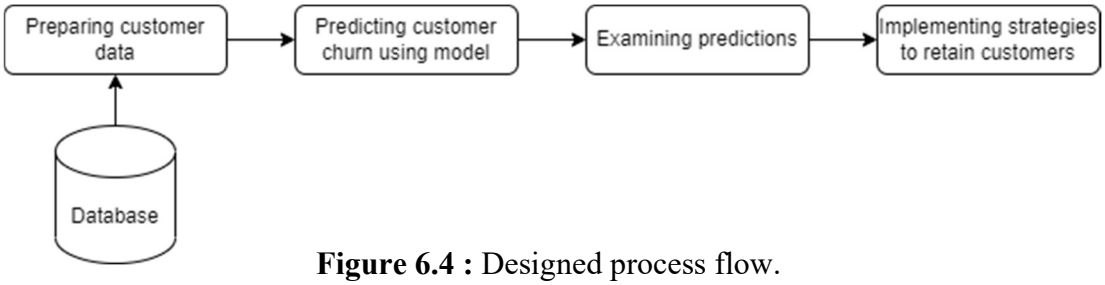

Figure 6.4 : Designed process flow.

Detailed explanation of the steps of the process can be seen below:

- Preparing customer data: To be used in the model, customers' data should be pulled from the CRM system or other database systems of KKBox. Necessary

filters should be applied and the dataset should be ready to be used in the model.

- Predicting customer churn using model: The machine learning model that has been trained will be used to predict which customers will churn in the upcoming period. This step will be the responsibility of the data scientists.

- Examining predictions: The output of the model should be examined and the customers expected to churn should be identified. A report/list containing the preventive strategies selected for these customers should be created by the CRM managers. Preventive strategies, as previously discussed in the study, can involve actions to keep customers engaged with the company. These may include offering special discounts, promotions, or additional services at no cost.

- Implementing strategies to retain customers: Strategies determined in the previous step will be applied for the customers that are predicted to churn. This step will mostly be the responsibility of the CRM department.

In addition, as mentioned earlier, it is important to detect a decrease in the quality of the machine learning model used. This can be detected by constantly monitoring the performance of the machine learning model in use. Threshold values for the selected machine learning model can be determined as 0.65 for recall, 0.35 for precision and 0.45 for F-Score. When one these metrics fall below its threshold, the model needs to be retrained with up-to-date data.

7. CONCLUSION

In this paper, the main goal is to reduce the churn rate of KKBox, one of the biggest music streaming companies in Asia. A machine learning model was developed as the main element of the process designed for this purpose. The first step of the development process was to perform various manipulations to prepare the data presented by KKBox for use in modelling. The feature generation involved in this step set an example for other companies in the music streaming industry that will plan a churn prediction project (156 features were generated). Then, modeling was started after using SMOTE to solve the data imbalance problem, which is quite common in churn prediction projects. In the modeling phase; XGBoost, Logistic Regression and Random Forest algorithms were used. In order to determine the optimum hyperparameter values of these algorithms, grid search was used in the validation part. After the hold-out test was performed, the best model was selected by focusing on the recall and precision metrics. In the last step of the development of the project, the process in which the developed machine learning model will be used was designed and the method of monitoring was explained. The applicability of the project is quite high, as the KKBox is already storing the necessary data. After this study, customer satisfaction will increase, the churn rate will decrease and therefore the competitiveness of the company will increase. As seen in the literature, such a project is very valuable for the management to achieve its short and long term goals. Predicting customer churn improves businesses' chances of communicating with customers, offering special offers and increasing their satisfaction. Such projects contribute to increasing the quality of service available to the society. And of course, the most important effects of this project are economic. The main focus of the work is on reducing lost revenues and retaining customers. In this way, it is aimed for KKBox to gain a competitive advantage. Such projects contribute to the growth of companies and thus the chance to create employment. The project is suitable for the duty of increasing the quality of a company's product and service. In the project, special emphasis was placed on protecting the rights of customers such as data privacy. This

study has the potential to benefit KKBox and similar companies and offers them the potential to strengthen customer relationships and increase customer satisfaction.

8. REFERENCES

Ackerman, S., Raz, O., Zalmanovici, M., & Zlotnick, A. (2021). Automatically detecting data drift in machine learning classifiers. *arXiv preprint arXiv:2111.05672*.

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.

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10-11), 552-568.

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27-36.

Çelik, O., & Osmanoğlu, U. O. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, *4*(1), 30-38.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In *machine learning in radiation oncology* (pp. 3-11). Springer, Cham.

Emtiyaz, S., & Keyvanpour, M. (2012). Customers behavior modeling by semi-supervised learning in customer relationship management. *arXiv preprint arXiv:1201.1670*.

Eria, K., & Marikannan, B. P. (2018). Systematic review of customer churn prediction in the telecom sector. *Journal of Applied Technology and Innovation*, *2*(1).

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.

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *Journal Of Applied Microeconometrics*, *1*(2), 85-99.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425.

Huang, Y., Zhu, F., Yuan, M., Deng, K., Li, Y., Ni, B., ... & Zeng, J. (2015, May). Telco churn prediction with big data. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 607-618).

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, 101-112.

Jain, H., Khunteta, A., & Srivastava, S. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. *Telecommunication Systems*, *76*(4), 613-630.

Johny, C. P., & Mathai, P. P. (2017). Customer churn prediction: A survey. *International Journal of Advanced Research in Computer Science*, *8*(5).

Joshi, P. (2019). Predicting customers churn in telecom industry using centroid oversampling method and KNN classifier. *Int. Res. J. Eng. Technol.*

Kawale, J., Pal, A., & Srivastava, J. (2009, August). Churn prediction in MMORPGs: A social influence based approach. In *2009 international conference on computational science and engineering* (Vol. 4, pp. 423-428). IEEE.

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294.

Malyar, M., Robotyshyn, M. M., & Sharkadi, M. (2020, October). Churn Prediction Estimation Based on Machine Learning Methods. In *2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC)* (pp. 1-4). IEEE.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks*, *7*(2), 70.

Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., & Rehman, A. (2013, September). Telecommunication subscribers' churn prediction model using machine learning. In *Eighth international conference on digital information management (ICDIM 2013)* (pp. 131-136). IEEE.

Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (pp. 1196-1201). IEEE.

Safitri, A. R., & Muslim, M. A. (2020). Improved accuracy of naive bayes classifier for determination of customer churn uses smote and genetic algorithms. *Journal of Soft Computing Exploration*, *1*(1), 70-75.

Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.

Tékouabou, S. C., Gherghina, Ș. C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. *Mathematics*, *10*(14), 2379.

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

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European journal of operational research*, *218*(1), 211-229.

Xu, T., Ma, Y., & Kim, K. (2021). Telecom churn prediction system based on ensemble learning using feature grouping. *Applied Sciences*, *11*(11), 4742.