ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF MANAGEMENT

CHURN PREDICTION USING MACHINE LEARNING ALGORITHMS

B.Sc. THESIS

Ali Özcan KÜREŞ
070180022
Furkan Kaymaz
070180144

Department of Industrial Engineering

JANUARY 2023

ISTANBUL TECHNICAL UNIVERSITY ★ FACULTY OF MANAGEMENT

CHURN PREDICTION USING MACHINE LEARNING ALGORITHMS

B.Sc. THESIS

Ali Özcan KÜREŞ
070180022
Furkan Kaymaz
070180144

Department of Industrial Engineering

Thesis Advisor: Assist. Prof. Dr. Mehmet Ali ERGÜN

JUNE 2023

ISTANBUL TEKNİK ÜNİVERSİTESİ ★ İŞLETME FAKÜLTESİ

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE MÜŞTERİ KAYIP TAHMİNİ

LİSANS TEZİ

Ali Özcan KÜREŞ
070180022
Furkan KAYMAZ
070180144

Endüstri Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr.Öğr.Üyesi Mehmet Ali ERGÜN

HAZİRAN 2023

To our families,

FOREWORD

We would like to express my gratitude to Mehmet Ali ERGÜN for his guidance and support throughout this process. We would also like to thank our families and friends for their invaluable input and assistance. We hope that this thesis will serve as a useful resource for researchers in the field, and it will inspire further study and exploration.

June 2023

Ali Özcan KÜREŞ
Furkan KAYMAZ

TABLE OF CONTENTS

Page

ABBREVIATIONS

ANN **:** Artificial Neural Network
CRM **:** Customer Relationship Management
SVM : Support Vector Machines
AUC : Area Under the Curve
SMOTE : Synthetic Minority Oversampling Technique
SVM : Support Vector Machine
PRC : Precision Recall Curve
ADASYN : Adaptive Synthetic
ROS : Random Over Sampling
RUS : Random Under Sampling
PCA : Principal Component Analysis
SVD : Singular Value Decomposition
TFS : Tree-based Feature Selection
CFS : Correlation-based Feature Selection
PR : Precision-Recall

LIST OF TABLES

Page

LIST OF FIGURES

Page

MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE MÜŞTERİ KAYIP TAHMİNİ

ÖZET

Kayıp müşteri tahmini, işletmelerin müşterilerinin gelecekteki davranışlarını önceden tahmin edebilmeyi, olası müşteri kayıplarının sebeplerini gösterebilmesi ve bu bilgilere dayanarak şirketlerin önlem alabilmesini sağladığı için şirketler için önemli bir araçtır. Günümüzün rekabetçi piyasa koşullarında sadık müşterilere sahip olmak şirketlerin rekabet gücünü arttırırken aynı zamanda yeni müşteriler edinebilmelerini ve dolayısıyla market paylarını genişletmelerini sağladığı için kayıp müşteri analizinin öneminin gittikçe arttığı görülmektedir. Fakat, rekabetçi şirketlerde genel olarak müşteri kaybetme durumu nadir olarak gerçekleştiği için kayıp müşteri analizi yapılacak veri setleri genelde dengesiz olmaktadır. Bu dengesizliğin ortaya çıkarabileceği düşük performans ve yanlılık sorununu ortadan kaldırmak için literatürdeki çalışmalarda olduğu gibi bu çalışmada da oversampling, undersampling ve resampling gibi teknikler kullanılmıştır. Öncelikle veriseti farklı özniteliklerin dağılım grafikleri, korelasyon matrisleri oluşturulması ile görselleştirilmiştir. Eksik ve aykırı gözlemler giderilmiş ve veriseti standardize edilmiştir. Standardize edilen veriset üzerinde AdaBoost, Logistik Regresyon, Random Forest, XGBoosting, GradientBoosting, Decision Trees, CatBoost, LightGBM, Support Vector Machine gibi makine öğrenmesi algoritmaları kullanılmıştır ve çeşitli metriklere (Fbeta skoru, recall) göre en başarılı temel modeller seçilmiştir. Metrik seçiminde azınlık sınıfındaki recall'u arttırmak için F Beta skoru kullanılmıştır. Daha sonra seçilen temel modeller üzerinde çeşitli öznitelik seçimi ve örnekleme algoritmaları kullanılarak farklı senaryolar için en başarılı algoritmalar ve bu algoritmaların hiperparametreleri bulunmuştur. Sonuçlar, örnekleme ve öznitelik seçimi yapılmış verisetinde performansın önemli ölçüde arttığını ve CatBoost'un diğer makine öğrenmesi algoritmalarına göre daha başarılı olduğunu göstermiştir.

CHURN PREDICTION
USING MACHINE LEARNING ALGORITHMS

SUMMARY

Customer churn prediction is an important tool for businesses as it allows them to forecast their customers' future behaviors, identify possible reasons for customer attrition, and take preventive measures based on this information. In today's competitive market conditions, having loyal customers enhances companies' competitiveness, enabling them to acquire new customers and expand their market share. Therefore, the importance of customer churn analysis is increasing. However, in competitive companies, the occurrence of customer attrition is generally rare, resulting in imbalanced datasets for customer churn analysis. To address the performance and bias issues that may arise from this imbalance, techniques such as oversampling, undersampling, and resampling have been used in previous studies, as well as in this study. Initially, the dataset was visualized by creating distribution graphs and correlation matrices for different features. Missing and outlier observations were addressed, and the dataset was standardized. Machine learning algorithms such as AdaBoost, Logistic Regression, Random Forest, XGBoosting, GradientBoosting, Decision Trees, CatBoost, LightGBM, and Support Vector Machine were applied on the standardized dataset. The best-performing base models were selected based on various metrics (F-beta score, recall). F-beta score was used for metric selection to increase recall in the minority class. Subsequently, various feature selection and sampling algorithms were applied on the selected base models to determine the most successful algorithms and their hyperparameters for different scenarios. The results showed that performance significantly improved with sampling and feature selection on the modified dataset, and CatBoost outperformed other machine learning algorithms.

1. INTRODUCTION

In industries with high competition such as telecommunication, digital marketing and banking, the importance of loyal customers is increasing. Many companies have begun to develop CRM strategies to maintain or improve relationships with loyal and long-term customers. Customers who have been with a company for a long time are less likely to be influenced by the marketing efforts of competitors and they can also be effective in gaining new customers, as long-term customers will make positive comments to potential new customers. In addition, churned customers will have a negative effect on changing the preferences of potential customers or existing customers, and the credibility of the company will decrease. Also, instead of finding new customers, customer retention is more profitable and cheaper for many companies (Sharma and Panigrahi, 2011). Because losing customers has such significant negative effects, companies must constantly monitor customer satisfaction, create predictive models that can predict potential customers which can stop their transactions.

Churn management procedures fall under the customer relationship management (CRM) framework. Two main analytical modeling tasks can be considered. The first task is identifying customers who are about to leave, and the second is managing risky customers in two ways as proactive and reactive. In reactive management, the customer waits until the cancellation request comes, and then tempting offers are made to win the customer back. In proactive management, from the moment a risky customer is identified, necessary offers and transactions are made to retain this customer. (Lalwani, Mishra and Chadha, 2021).

According to Mitkees et al., churn prediction models allow us to understand the customer's behavior and anticipate potential customer churns. By recognizing this potential lost earlier, companies may satisfy the needs of customer and can prevent potential losses (2017). Iranmanesh et al. states that losing an existing customer not only leads to lower revenue but also the cost of attracting a new customer (2019). This cost can be 5 to 6 times more expensive than customer retention (Kaur et al., 2013). It is a well-known problem for many industries but it is more important in markets where

competition is higher and finding new customers are harder, like telecommunication sector. Thus, we would like to focus on churn rates of customers in telecommunication industry.

2. LITERATURE REVIEW

Since acquiring new customers is difficult and costly nowadays, many companies attach importance to not losing their existing customers. Long-term customers are generally not affected by the campaigns of rival companies, and since they will express their satisfaction with the company they are in to other potential customers, it can greatly increase the probability of acquiring potential customers. In sectors such as banking, telecommunications, and insurance, where competition is high, companies strive to establish long relationships with their customers. For this reason, making predictions about possible customer losses and when these losses will occur can help businesses take precautionary measures to protect their customers, prevent financial losses and regain lost customers. However, due to the complexity of customer behavior and the large number of variables that need to be examined, it is difficult to accurately predict customer churn. Classification algorithms such as K-nearest neighbor, SVM, naive bayes classifier, decision tree, random forests are common machine learning algorithms that is used to make these predictions (Kaur and Kaur, 2020; Hemalatha & Amalanathan, 2019; Olaniyi et al., 2020; Coussement & Van den Poel**,** 2008; Kumar & Ravi, 2008).

Although these classification algorithms generally give high accuracy rates, different algorithms such as ANN's are used on different data sets. Sharma and Panigrahi used a neural network to perform a predictive model. Even though they experimented with multiple hidden layers containing three to seven hidden layers, best result was obtained having one hidden layer with three neurons. The output layer has two neurons that correspond to the output field's two values. They also created a table which shows the relative importance of input fields in descending order and show some patterns about their data (2011). Lemmens and Croux used both bagging and stochastic gradient boosting algorithms in order to predict churns (2006). Guliyev and Tatoğlu used the XgBoost algorithm in addition to other algorithms. They used cross validation (5-fold validation) and grid search methods to calculate hyperparameters. They found that the XGboost algorithm worked better than other algorithms with an AUC of %96.97

(2021). Mishra & Rani established machine learning algorithms such as Adaboost, decision tree-based classification, partial tree-based classification, bagged tree-based classification and boosted classification trees. They evaluated the models using performance metrics such as accuracy, sensitivity, specificity, and AUC (area under the curve). As a result of all these performance evaluations, they saw that the best performance was in the Adaboost algorithm (2017). Dias, Godinho & Torres implemented a data collection process that involved creating rolling time window datasets for predicting customer churn at different time horizons to predicting when will customer churn (2020). Some applications used ensemble method which is the integration of multiple classifiers to achieve better accuracy performance than can be achieved by using a single classifier (Buckinx & Van den Poel, 2005).

Recently, researchers have been studying how various aspects of the data and preprocessing can affect the accuracy of a machine learning model, regardless of the specific algorithm being used (Gür Ali & Arıtürk,2014). Factors such as the quality and quantity of data, as well as any imbalances or biases present in the dataset, can all affect the model's ability to learn and make accurate predictions. For example, it has been seen that most of the data used for churn prediction are imbalanced data. To improve the performance of machine learning algorithms on these data, it is necessary to reduce this imbalance and there are two common approaches in handling imbalanced data; first is sampling approach and second is cost-sensitive approach (Chen, Liaw & Breimenn, 2004). The basic sampling methods for balancing class imbalance in a dataset are under-sampling and over-sampling. Under-sampling involves removing some observations from the majority class, while over-sampling involves adding more observations from the minority class. These techniques can help to balance the class distribution, making the minority class less rare (Burez & Van den Poel, 2009). Both approaches can be effective in reducing class imbalance in a dataset. Mishra & Rani performed operations such as sampling with the SMOTE (synthetic minority oversampling) technique, selecting the most suitable features with the gain ratio method, and replacing the missing data so that the built models could work efficiently and without errors (2017). In addition to these, some advanced methods are also used. As an example, we can give the CUBE method that allows the selection of approximately balanced samples (Deville & Tillé, 2004).

Different methods are used for feature selection, which is another preprocessing process. Gunay & Ensari used the data of a telecommunication company, which includes 16 numerical and 4 categorical features. In order to find the most suitable ones for their data among these 20 features, they used the highly preferred correlation feature calculator by using WEKA software, and they reduced the number of features from 20 to 10 by looking at these correlation values (2018). Rahman & Kumar made their feature selections using the Relief and mRMR (Minimum Redundancy Maximum Relevance) methods (2020). While making predictions, we compare the algorithms used according to performance metrics when choosing the right model for our data set. Performance metrics such as accuracy, AUC score, precision, recall, true positive rate, true negative rate and F-measure are generally used in the literature. Models created on the same dataset may yield different results in different performance metrics. For example, Sisodia et al. used metrics such as accuracy, precision, recall, true positive rate, true negative rate, and F-measure to evaluate the performance of the results of these models. Looking at these metrics, they saw that the random forest classifier had the highest accuracy, with an accuracy as high as 98.97%. They also observed that the random forest classifier method gave good results in metrics such as sensitivity, recall, F-Measure and specificity (2017).

3. METHODOLOGY

3.1 Problem Definition

Telecommunication companies typically have a system composed of components and sub-systems such as physical infrastructure like fiber optic cables, satellites, and wireless transmission towers, customer premises equipment like phones, modems, and routers, data centers and servers for storing and processing data, network security systems, and customer management systems. The purpose of telecommunication companies is to provide their customers with reliable, high-quality, price-planned service for every need and preference. They have to invest in the infrastructure and technology necessary to provide this quality service, to ensure customer satisfaction and prevent churn of customers.

Churning of customers is a major problem for telecommunications companies. Losing customers can reduce the company's revenues and deteriorate its financial performance. In addition, losing customers can reduce the company's power and popularity, and other customers may lose confidence in the company or even leave the company. Since telecommunications companies also operate in a competitive environment, losing customers can increase the superiority of other companies and reduce the company's market share. In addition, the loss of customers can also reduce the motivation of the company's employees. In order to prevent these losses, the company needs to create a machine learning model that can predict the potential customers to leave the company using the personal data of the customers and the transaction data between the company and the customer.

Churn can be caused by a variety of factors, such as dissatisfaction with the product or service, a change in personal circumstances or the influence of competitors. It's important to identify the key variables that are contributing to churn before build an effective predictive model. This can be done through data analysis, customer surveys, and other methods of gathering information. Once you have a good understanding of the factors that causing customers to churn, forecasting models should be established

to identify patterns and trends that may indicate the likelihood of churn using data of past churning customers. Once you have a churn prediction model in place, you can use it to identify at-risk customers and identify factors that best describes whether the customer is churn or not. After building the prediction model, you can focus on addressing those factors in order to reduce the churn rate and can make attractive offers such as promotions or discounts for long-term customers, improving the product or service to retain those customers. By taking proactive actions to reduce churn, businesses can improve customer retention and ultimately drive long-term growth.

In most of the customer churn projects, it is known that predicting a churned customer as not churned can lead to a loss of revenue and potentially harm the overall business. By placing more emphasis on recall, we can better capture the number of true positives (i.e., correctly identifying a churned customer) and minimize the number of false negatives (i.e., incorrectly identifying a non-churned customer as churned). That is why, although accuracy is used in the most of the classification problems, this metric is not appropriate for churn prediction. As mentioned in the literature review part: precision, recall, F1Score, AUC and PRC are more advised metrics for churn prediction. F Beta Score with beta=2 is also a suitable metric for evaluating the performance of the model in this context. This metric balances the trade-off between precision and recall, with the beta parameter controlling the degree of emphasis placed on recall. A higher beta value, such as beta=2, places more weight on recall than precision. This is appropriate in cases where recall is more important than precision, such as in predicting customer churn. Therefore, the F Beta Score with beta=2 is a good choice for evaluating the performance of the model in customer churn project.

3.2 Data and Variables:

Data has been taken from IBM Datasets. There are 7032 observations and 21 features in the dataset. 17 features are numerical and 4 features are categorical. 5163 observations in the dataset belong to the majority class, whereas 1869 observations belong to the minority class, making up %27 of the dataset's contributions. The definition of the attributes for predicting the customer’s churn status are selected and summarized in the Table 3.1.

Table 3.1: Definition of Attributes

Variable	Explanation	Data Type
CustomerID	Unique ID for every customer	Categorical
Gender	%50 male - %50 female	Categorical
SeniorCitizen	Senior citizen or not (1,0)	Boolean
Partner	Customer has a partner or not (1,0)	Boolean
Dependents	Customer has dependents or not (1,0)	Boolean
Tenure	The length of months the consumer has been a customer	Numerical
PhoneService	Customer has a phone service or not (1,0)	Boolean
MultipleLines	Customer has Multiple Lines or not (1,0)	Boolean
InternetService	Customer's internet service provider (DSL, Fiber optic, No)	Categorical
OnlineSecurity	Customer's has online security or not (Yes, No, No internet service)	Categorical
OnlineBackup	Customer has online backup or not. (Yes, No, No internet service)	Categorical
DeviceProtection	Customer has device protection or not (Yes, No, No internet service)	Categorical
TechSupport	Customer has tech support or not(Yes ,No ,No internet service)	Categorical
StreamingTV	Customer has Streaming TV or not(Yes ,No ,No internet service)	Categorical
StreamingMovies	Customer has Streaming Movies or not(Yes ,No ,No internet service)	Categorical
Contract	The type of contract customer has (Month-to-month , One year, Two year)	Categorical
PaperlessBilling	Customer has paperless billing or not (Yes, No)	Categorical
PaymentMethod	The type of customer's payment method (Electronic check, Mailed check, Bank transfer- automatic, Credit card - automatic)	Categorical
MonthlyCharges	The monthly fee assessed to the customer	Numerical
TotalCharges	The total fee charged to the customer	Numerical
Churn	Whether the customer churn or not	Categorical

3.3 Descriptive Statistics, Correlation and Exploration

In order to have a first insight about numerical features in our dataset, descriptive statistics of data is shown in Table 3.2. pd.describe function of Python's Pandas library was used to get this table. According to the table, it is seen that the customers stay with the company for an average of 32 months and the total payment they pay is around $ 2283.

Table 3.2: Descriptive Statistics of Numerical Data

Variable	Mean	Std	%25	%50	%75	Min	Max
Tenure	32.421	24.545	9	29	55	1	72
MonthlyCharges	64.798	30.085	35.587	70.350	89.8625	18.25	118.75
TotalCharges	2283.300	2266.771	401.450	1397.475	3794.7375	18.80	8684.80

The correlation was calculated to demonstrate the magnitude of the statistical relationship between the target column and other columns. Since our data has many categorical columns, dummy variables were used for these columns before calculating the correlation. We used pd.get_dummies function of Pandas library for dummy variables and dropped the CustomerID column because unique ID of customers does not affect whether the customers are going to churn or not. As mentioned in the problem definition part, this study focuses on finding customers who might churn so feature correlation for "Yes Churn" column is used for correlation graph in the Figure 3.1.

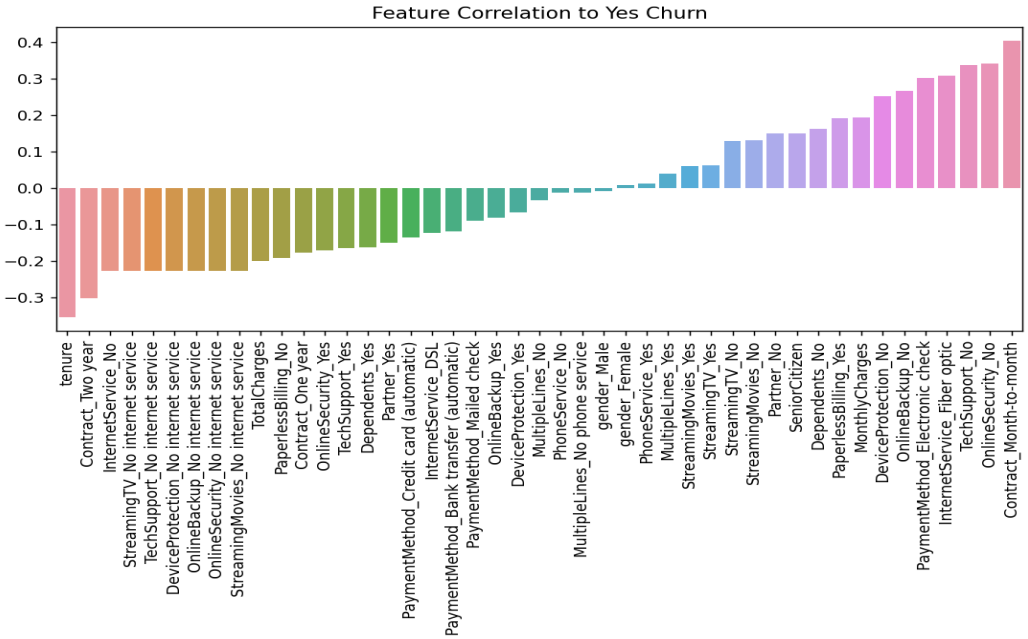

Figure 3.1: Feature Correlation Graph for Yes Churn.

Based on this graph, it can be interpreted that customers with month-to-month contract are more likely to churn. On the other hand, customers who have higher tenure, or the number of months they stayed with the company, and longer contract seem to be "loyal" customers for the company. The abovementioned features (tenure and contract type) should be more thoroughly investigated because they seem to have a significant impact on the target column.

To demonstrate tenure's impact on the customers, Figure 3.2 which shows Churn Ratios by tenure graph was created.

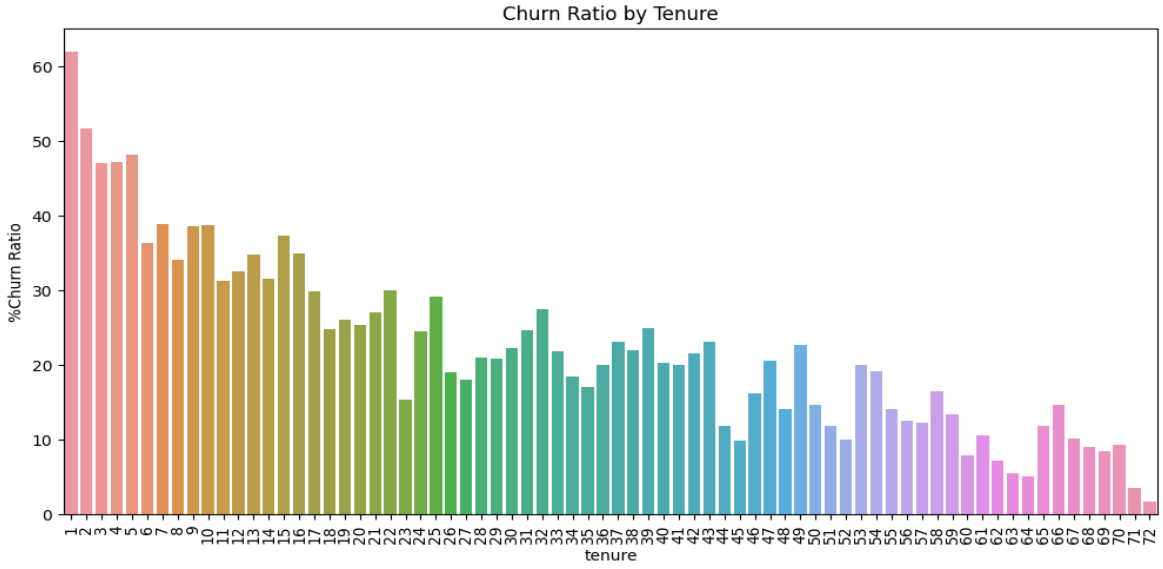

Figure 3.2: Churn Ratio by Tenure.

This graph shows that, as expected, the churn rate of customers who have not been customers of the company for a long time is higher. This may mean that short-term fees are found expensive by customers for the company. Also, the sudden increase in the 13th, 25th and 49th months may be dangerous for the company. Since the company offers one year and two years contracts, increases in these months may indicate that there are problems other than the fee.

Lastly the relation between total charges, contract type and churn are showed in Figure 3.3. There is a common pattern in One-Year-Contract and Two-Year-Contract. According to this pattern, customers with high TotalCharges are more likely to churn.

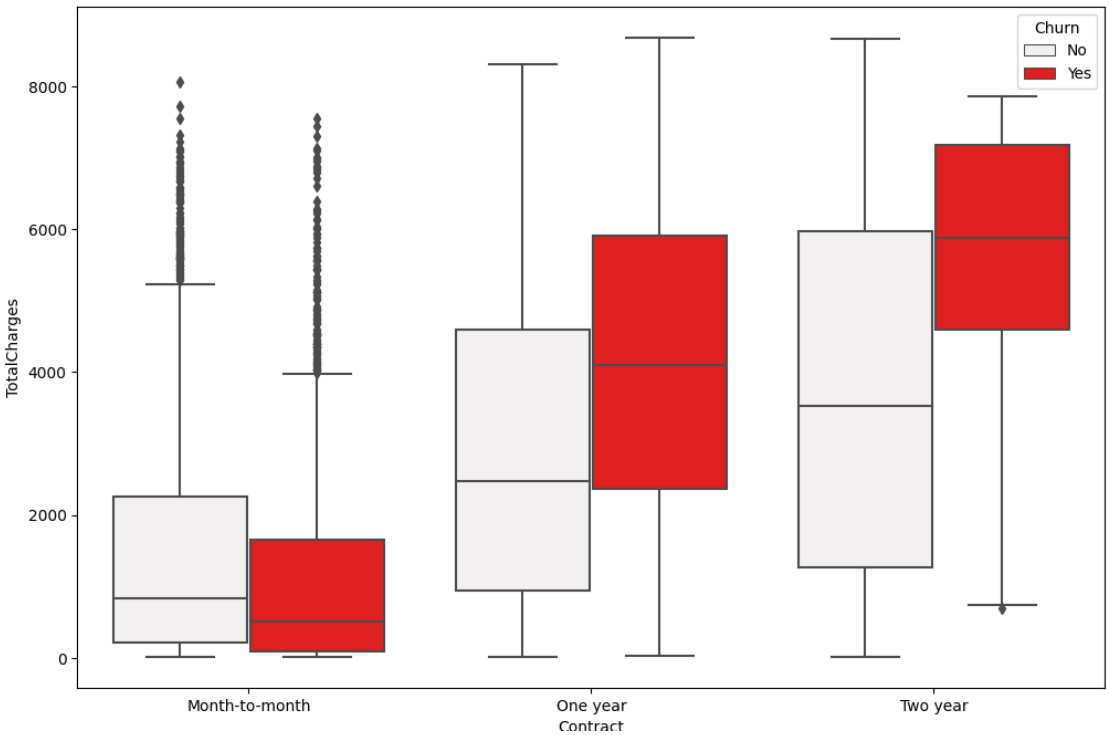

Figure 3.3: TotalCharges vs Contract Type Boxplot

Also, in order to detect outliers and visualize the distribution, we have prepared boxplots and kdeplots of the numerical features in our dataset. There don't seem to be any outliers in our data when we look at the boxplot. These graphs can be seen in the Figure 3.3.1 and Figure 3.3.2 below.

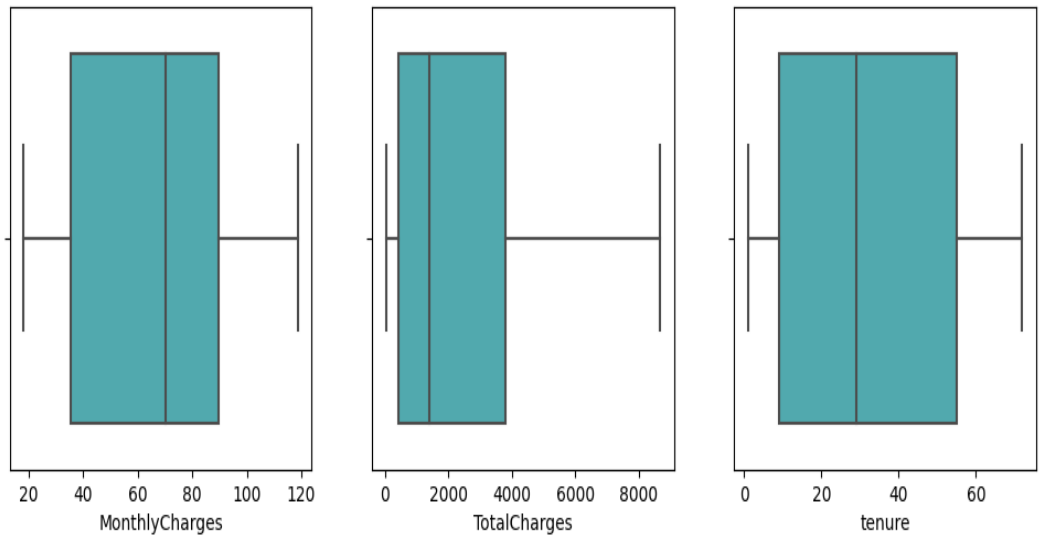

Figure 3.3.1: Boxplots of Numerical Features

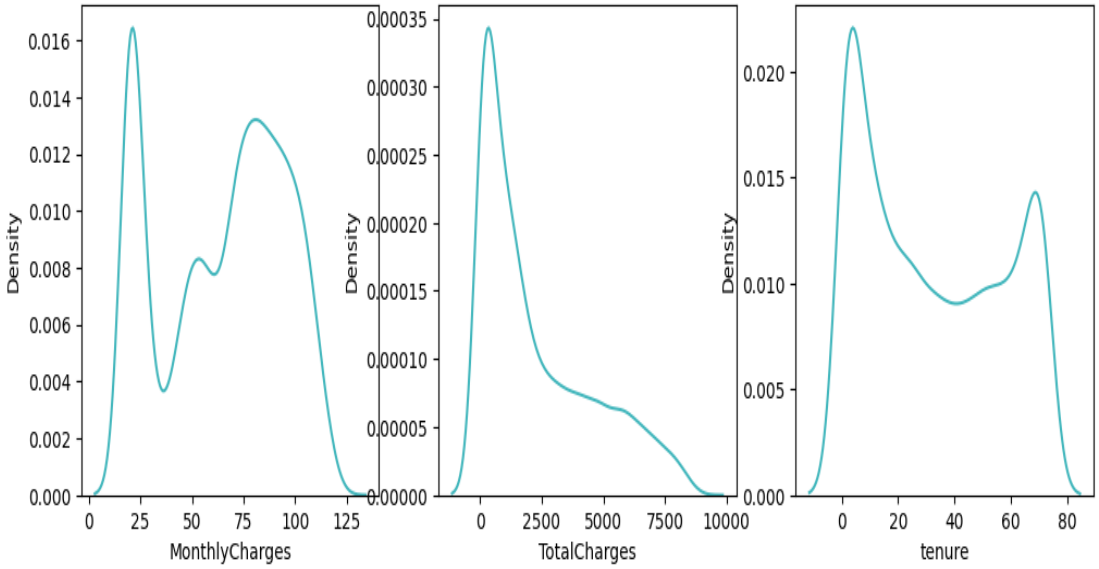

Figure 3.3.2: Kdeplots of Numerical Features

3.4 Data Preprocessing

Data preprocessing is a data mining technique used to turn the raw data into a format that is both practical and effective (Ahmad and Aziz,2018). Generally, a data preprocessing process includes data cleaning, data transformation and data reduction. In Telco dataset, there are not any missing values but as discussed in the Exploration Section there is a noisy data problem in the dataset. Data points that are duplicated or semi-duplicated, data segments that are useless for a certain research method, outliers or undesirable information fields are all examples of noise. In our case outliers can be identified by using a boxplot (Figure 3).

After cleaning the data, data was scaled it in order to compere the attributes easily, reduce the computational expenses and improve the performance of the model. Although there are several scaling methods such as centering, standardization and normalization we used Z-score standardization in preprocessing. Z-score standardization formula is defined in the Figure 3.4.

$$X_{std} = \frac{X - \overline{X}}{s_X}$$

Figure 3.4: Z-Score Standardization Formula.

Where $\overline{X}$ and s_X are sample mean and standard deviation of the any given X variable.

3.5 Imbalanced Learning

According to Li et al., an imbalanced dataset is one in which one or more classes contain significantly more examples than the other classes. The majority class is the most common, while the minority class represents the most infrequent (2016).

Because of their rarity and casualness, unusual events are challenging to detect, but misclassifying rare occurrences can have expensive consequences (Haixiang et al.,2017). In churn prediction case, a misclassification of a customer who is going to churn might cost thousands of dollars for Telco. Since the minority class is approximately %27 of the dataset, it will be more difficult to find patterns in it unless some strategies are implemented. The visualization of class imbalance problem can be seen in the Figure 3.5.

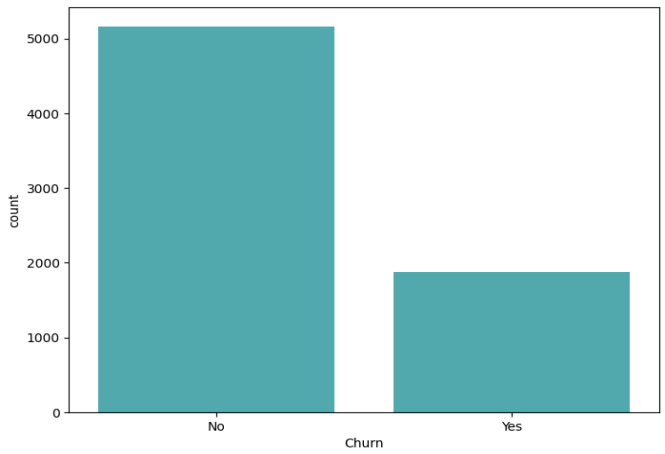

Figure 3.5: Class Frequency for Churn Column.

In the next section, several strategies for dealing with imbalanced data will be discussed.

3.5.1 Resampling

To counteract the negative effects of a skewed class distribution on learning, resampling techniques are performed to rebalance the sample space for an unbalanced dataset**.** According to the technique used to balance the class distribution, resampling techniques can be divided into three groups:

3.5.1.1 Oversampling

Random over-sampling, a nonheuristic technique that balances the class distribution by the random replication of positive samples, is the easiest way to increase the size of the minority class. Yet, overfitting is more likely to happen since this strategy copies existing instances from the minority class (Ganganwar,2012). Random selection or picking samples from the margins of the majority and minority classes are used for repetition. This causes the classifier to give the minority sample class such borderline spaces. Oversampling methods are often criticized because these algorithms tend to automatically rebalances the minority and majority samples without adding any additional data (Fotouhi et al, 2018).

Several methods for creating fresh synthetic samples are developed in order to overcome this problem such as Random Over Sampling (ROS), Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Tecnique (SMOTE). Random over-sampling, a nonheuristic technique that balances the class distribution by the random replication of positive samples, is the easiest way to increase the size of the minority class (Ganganwar,2012)**.** The ADASYN method is based on adaptively creating minority data samples in accordance with its distributions. The algorithm's main process is to apply weights to various minority class samples in order to generate various quantities of synthetic data for each sample. SMOTE algorithm is designed to solve the overfitting problem in oversampling methods and improve the accuracy of models. With this method, synthetic minority instances are created along the line segments connecting the minority samples and their "k" nearest neighbors. The neighbors from the "k" nearest neighbors are selected at random according to the preferred rate of oversampling (Gosain and Sardana,2017). Distribution of classes after oversampling using ADASYN can be seen in the Figure 3.6:

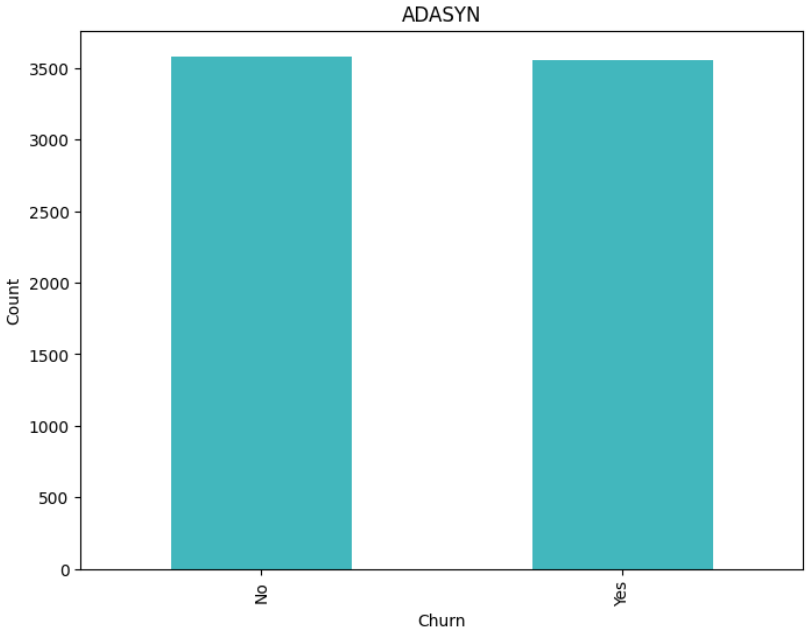

Figure 3.6: Class Frequency for Churn Column after ADASYN.

3.5.1.2 Undersampling

Ganganwar (2012) states that undersampling is a useful tecnhnique for classifying imbalanced datasets. In this approach, the classifier is trained using only a subset of the majority class, thus the training set becomes more balanced and the training process is accelerated. Random majority under-sampling (RUS), the most popular preprocessing method, randomly removes instances of the majority class from the dataset. The distribution of classes after RUS can be seen in the Figure 3.7.

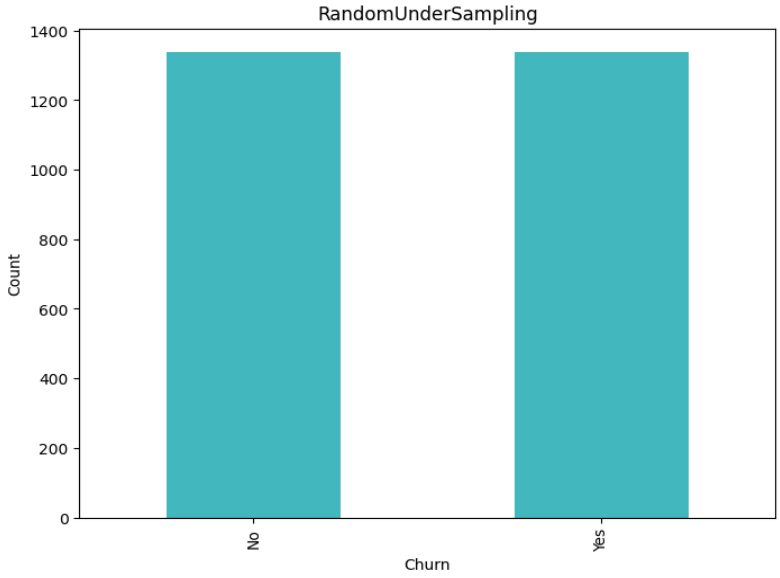

Figure 3.7: Class Frequency for Churn Column after UnderSampling.

Yet, the primary disadvantage of under-sampling is the ignoring of potentially relevant information included in these disregarded cases. Lin et al. (2017) suggests using clustering instead of random undersampling to get over undersampling's drawbacks. The goal of clustering analysis is to place comparable objects (i.e., data samples) into the same clusters, where the items in each cluster differ in terms of their feature representation.

3.5.2 Feature Selection and Extraction

Haixiang et al., defines the objective of feature selection as, choosing a subset of k features from the complete feature space that enables a classifier to perform at its best, where k is a user-specified or adaptively selected parameter (2017).

Feature extraction is another approach to dealing with dimensionality. Dimensionality reduction, which converts data into a low-dimensional space, is also related to feature extraction. Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and f Non-negative Matrix Factorization (NMF) are a few of the approaches available for feature extraction (Haixiang et al., 2017).

PCA is a linear dimensionality reduction technique that transforms the data into a lower-dimensional space while retaining most of the variance in the original data. PCA has been used in facial recognition (Turk and Pentland, 1991) and bioinformatics (Wold et al., 1993). The Chi-Squared Test is a statistical test used to determine whether two categorical variables are independent of each other. The Chi-Squared Test has been used in image retrieval (Swain and Ballard, 1991) and text classification (Yang and Pedersen, 1997). RFE is an iterative method that recursively eliminates features based on their importance. RFE has been used in cancer classification (Golub et al., 1999) and gene expression analysis (Alvarez et al., 2007). CFS is a filter method that selects features based on their correlation with the target variable. CFS has been used in heart disease diagnosis (Balakrishnan and Raja, 2011) and stock prediction (Chen et al., 2016). TFS is another filter method that selects features based on their importance in a decision tree. TFS has been used in credit scoring (Kohavi and John, 1997) and fraud detection (Kotsiantis et al., 2006). SVD is a matrix factorization technique that decomposes a matrix into its constituent parts. SVD has been used in image compression (Szeliski, 1999) and collaborative filtering (Koren et al., 2009).

3.5.2.1 Ensemble Methods

By integrating many base classifiers that perform better together than individually, ensemble-based classifiers are known to increase the performance of a single classifier (Lopez et al.,2013). According to their training processes, ensemble methods are divided into 2 groups (Haixiang et al., 2017).

3.5.2.2 Iterative Based Ensembles

Most of the iterative based ensembles are based on Adaboost, the first practical boosting algorithm. The advantage of Adaboost is that data that are incorrectly classified are given heavier weights, pushing a future classifier to concentrate more on learning these incorrectly classified samples. These boosting algorithms are frequently used with resampling and cost-sensitive learning methods**.**

3.5.2.3 Parallel Based Ensembles

Ensemble models known as "parallel based ensembles" allow each base classifier to be trained simultaneously. Feature selection-based ensembles, resampling based ensembles, and bagging are examples of parallel based ensemble technique

3.5.3 Performance Metrics for Imbalanced Datasets

It is known that there are several evaluating metrics such as ROC (Receiver Operating Characteristics) , PRC(Precision-Recall Curve) , F1-Score, F Beta-Score etc. are commonly used in classifying models. The most common evaluation method for binary classifiers is ROC; however according to Berrar and Peter(2010) , when applied with unbalanced datasets, interpreting ROC curves calls for extra caution. While ROC curves have traditionally been the preferred method, recent studies have shown that PR curves offer several advantages over ROC curves, particularly when dealing with imbalanced datasets (Japkowicz & Stephen, 2002; Davis & Goadrich, 2006).

PR curves are particularly useful when the positive class is rare or when the cost of misclassification is asymmetric between positive and negative instances. In these cases, ROC curves can be misleading, as they are not sensitive to the distribution of positive and negative instances (Saito & Rehmsmeier, 2015). On the other hand, PR curves provide a more intuitive representation of the trade-off between precision and

recall, which are arguably more relevant in many real-world scenarios (Flach & Kull, 2015). The baseline in the context of ROC and PRC curves refer to the diagonal line that depicts the effectiveness of a random classifier. The baseline for PRC is defined by the ratio of positives (P) and negatives (N) as y = P / (P+N), whereas the baseline for ROC is set in (y = x) (Saito & Rehmsmeier, 2015). In our case , the baseline of the PRC is approximately 0.27 (1869 / (5163+1869)).

PR curves allow for a more direct comparison between different classifiers, as they are not affected by the overall class distribution. This is particularly important in applications such as information retrieval, where the goal is to retrieve as many relevant documents as possible while minimizing the number of irrelevant ones (Liu et al., 2018). In some cases , adding many “negative class” (customers who are not going to churn in our case) may significantly improve the ROC curve without increasing the sensitivity or the positive predictive value of the parameter under consideration. Nonetheless, the inclusion of the negative class, or a imbalanced distribution of data in binary classification, had no effect on the precision-recall curves (Ekelund,2017). All in all , it can be said that PR curves offer several advantages over ROC curves, particularly in scenarios where the class distribution is imbalanced or where precision and recall are more relevant metrics. As such, PR curves should be considered as a valuable tool for evaluating imbalanced classification models (Saito & Rehmsmeier, 2015).

In order to summarize this section and to assist the modelling, the workflow chart of our methodology can be seen in Figure 3.8

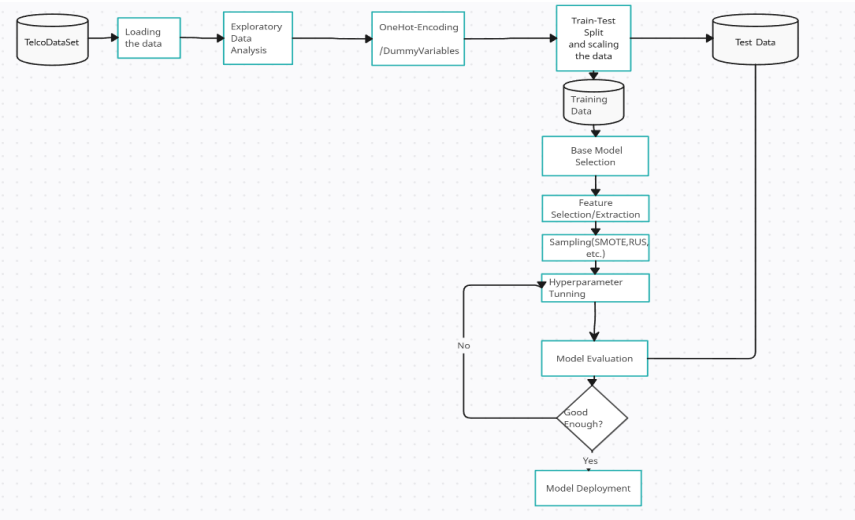

Figure 3.8: Workflow Chart of Modelling

4. MODELLING

4.1 Default Model Selection

After exploring, cleaning, scaling and manipulating the data the last step of a typical machine learning process is training the data and testing it with relevant metrics. The default model forms the core of the machine learning pipeline, and choosing the wrong default model might result in subpar performance, resource waste, and, ultimately, a defective model (Raschka & Mirjalili, 2021). Firstly, we started to evaluate performance of models without any hyperparameter tunnig to see performances of default models. The performances of default models are evaluated standart train and test process without any sampling, feature selection or hyperparameter tunning and compared by F Beta Score. F Beta scores of several machine learning algorithms can be seen in Table 4.1. The best score for every metric is highlighted in each table.

Table 4.1: F Beta-score (beta=2) of Default Models

Algorithm	F Beta-score
LogisticRegression	0.5803
AdaBoostClassifier	**0.5853**
RandomForestClassifier	0.5442
DecisionTreeClassifier	0.5117
SupportVectorClassifier	0.5510
LightGMB	0.5714
CatBoost	0.5717
GradientBoosting	0.5830
XGBClassifier	0.5332

4.2 Feature Selection

In this study, we used six different feature selection techniques to reduce the number of features and improve the performance of the default logistic regression algorithm. The techniques we used are Principal Component Analysis (PCA), Chi-Squared Test, Recursive Feature Elimination (RFE), Correlation-based Feature Selection (CFS), Tree-based Feature Selection (TFS), and Singular Value Decomposition (SVD).

In our study, we applied each feature selection technique separately to the dataset and evaluated the performance of the default AdaBoost algorithm using F Beta-score with beta=2 as metric. As can be seen in the Table 4.2 RFE technique gave us the best F Beta-score, indicating that it was the most effective technique for reducing the number of features and improving the performance of the AdaBoost algorithm on our dataset. In the continuation of the process, we reduced the number of variables to 10 by using the RFE technique to increase the accuracy and efficiency of the machine learning models we will build, reduce the complexity and dimensionality of the dataset, and reduce the processing power required for model training and prediction.

Table 4.2: F Beta-score of Each Feature Selection Technique on AdaBoost

Feature Selection Technique	F Beta-score
Principal Component Analysis	0.5052
Chi-Squared Test	0.4756
Recursive Feature Elimination	**0.5410**
Correlation-based Feature Selection	0.4883
Tree-based Feature Selection	0.5306
Singular Value Decomposition (SVD)	0.5385

4.3 Sampling

In the process of developing a predictive model, it is essential to have a balanced dataset. However, in some cases, the dataset may be imbalanced, i.e., one class may be underrepresented compared to the other. In such cases, the performance of the model may be biased towards the majority class, resulting in poor performance for the minority class. Some sampling techniques such as SMOTE, ADASYN, oversampling and undersampling can be implement to address this problem to prevent data leakage and resample data to handle the imbalanced class problem.

SMOTE technique creates synthetic samples for the minority class by using k-nearest neighbors. SMOTE helps to increase the number of minority class samples in the dataset, thus balancing the dataset. Undersampling technique reduces the number of samples in the majority class to balance the dataset. Oversampling technique increases the number of samples in the minority class to balance the dataset. ADASYN

(Adaptive Synthetic Sampling) technique is similar to SMOTE, but it focuses on regions that are difficult to learn by the model.

Previous studies have shown that these sampling techniques can improve the performance of the model for imbalanced datasets. For example, Balakrishnan and Raja (2011) used SMOTE and undersampling techniques to balance a heart disease dataset and improve the performance of a Naive Bayes classifier. Alvarez et al. (2007) used undersampling and oversampling techniques to balance a microarray dataset and improve the performance of a decision tree classifier.

In this study, we used these sampling techniques with default AdaBoost algorithm to evaluate their performance on the top 10 features obtained from the RFE technique. We found that sampling techniques had a huge impact on the performance of the model which can be seen in the Table 4.3. We used GridSearch with F Beta score (beta=2) scorer on training data and then evaluate the performance on the test data. F Beta scores of sampling methods can be seen in Table 4.3 below.

Table 4.3: F Beta-score of Each Sampling Technique

Sampling Technique + RFE	F Beta-score
ADASYN	0.7113
SMOTE	**0.7199**
Random Oversampling	0.7127
Random Undersampling	0.7111

4.4 Hyperparameter Tuning

After performed feature selection and sampling techniques. We corrected our dataset with relevant features and balanced samples, we then proceeded with the crucial step of hyperparameter tuning. This optimization process was undertaken to fine-tune the model's hyperparameters, such as learning rate, regularization strength, or maximum depth, with the objective of achieving the best possible performance. By performing hyperparameter tuning, we aimed to strike the optimal balance between model complexity and generalization ability, ultimately enhancing the model's predictive power. The iterative nature of hyperparameter tuning allowed us to systematically explore different parameter combinations and select the configuration that maximizes the performance of our customer churn prediction models. Several performance

metrics, hyperparameter set and the best parameters for every algorithm can be seen in the Table 4.4, Table 4.5 and Table 4.6 below.

Table 4.4: Performance of ML Algorithms for Different Metrics

ML Algorithms	F Beta-score	AP	ROC AUC	Recall
LightGBM Classifier	0.7312	0.6400	0.8307	0.9358
CatBoost Classifier	**0.7332**	0.6437	0.8353	0.9394
Logistic Regression	0.7289	0.6369	0.8288	0.8253
Random Forest Classifier	0.6840	0.6305	0.8307	0.7415
Decision Tree Classifier	0.6840	0.6035	0.8099	0.7594
AdaBoost Classifier	0.7214	0.6404	0.8372	0.8057
GradientBoosting	0.7219	0.6398	0.7513	0.8253
XGBoost Classifier	0.7161	0.5964	0.8173	0.8164
SupportVectorClassifier	0.7128	0.6381	0.8241	0.8717

Table 4.5: Hyperparameter Sets of Algorithms

ML Algorithms	Hyperparemeters	Number of Parameters
AdaBoost	MaxDepth, Criterion, MaxFeatures, NumberofEstimators, MaxLeafNodes	288
Random Forest	NumberofEstimators, MaxDepth, MinSamplesSplit, MinSamplesLeaf	81
Decision Tree	Criterion, MaxDepth, MinSamplesSplit, MinSamplesLeaf, MaxFeatures	216
SVC	C, Gamma, Kernel	18
Logistic Regression	C, Penalty	12
LightGBM	LearningRate, NumberofEstimators, MaxDepth, Subsample, Colsample	243
CatBoost	LearningRate, Iterations, Depth, Subsample, Colsample	243
Gradient Boosting	NumberofEstimators, MaxDepth, MinSamplesSplit, MinSamplesLeaf, MaxLeaf	243
XGBClassifier	MaxDepth,NumberofEstimators,MinChildWeight,Gamma,Subsample,Colsample	2187

Table 4.6: Best Hyperparameter Combinations for Every Algorithm

ML Algorithms	Hyperparemeters for Best Model
AdaBoost	MaxDepth:1, Criterion: entropy, MaxFeatures: auto, NumberofEstimators:50, MaxLeafNodes:6
Random Forest	NumberofEstimators:200, MaxDepth:10, MinSamplesSplit:2, MinSamplesLeaf:1
Decision Tree	Criterion: entropy, MaxDepth: 5, MinSamplesSplit:2, MinSamplesLeaf:1, MaxFeatures: auto
SVC	C:0.1, Gamma:0.01, Kernel: linear
Logistic Regression	C:100, Penalty: l2
LightGBM	LearningRate:0.05, NumberofEstimators:100, MaxDepth:7, Subsample:0.8, Colsample:1
CatBoost	LearningRate:0.05, Iterations:200, Depth:7, Subsample:1 Colsample by level:0.8
Gradient Boosting	NumberofEstimators:350, MaxDepth:3, MinSamplesSplit:2, MinSamplesLeaf:3, MaxLeafNodes:4
XGBClassifier	MaxDepth:3, NumberofEstimators:50, MinChildWeight:1, Gamma:0, LearningRate: 0.01, ColsamplebyTree:1

After an extensive hyperparameter tuning process and evaluating multiple models, we assessed their performance using various metrics including the F Beta score (with beta=2), Average Precision (AP) score, ROC AUC score, and Recall. Upon reviewing the results, it became evident that the CatBoost algorithm outperformed the other models, consistently achieving the highest F Beta score. The F Beta score, with its emphasis on recall, is particularly relevant in the context of customer churn prediction, as it prioritizes the identification of customers at risk of churn. Additionally, the CatBoost model demonstrated strong performance across other evaluation metrics, further affirming its suitability for our dataset. Based on these findings, we can confidently conclude that CatBoost is the most appropriate algorithm for accurately predicting customer churn in our telecommunication dataset. Its ability to effectively leverage the selected features and optimize model parameters makes it a powerful tool for informing targeted retention strategies and minimizing customer churn.

4.5 Feature Importance

In our customer churn prediction project, we employed a feature selection technique called Recursive Feature Elimination (RFE) to identify the most important predictors from our dataset. By iteratively eliminating less relevant features, we aimed to enhance the performance and interpretability of our models. After applying RFE, we arrived at a refined set of 10 key features that exhibited strong associations with customer churn. These features encompassed various aspects, including customer tenure, financial interactions, demographic attributes, service subscriptions, contract durations, and payment preferences. By leveraging RFE, we ensured that our models were focused on the most influential features, enabling us to gain deeper insights into the factors that drive customer churn in the telecommunication domain.

In our telecommunication dataset, several features play a crucial role in predicting customer churn. Firstly, 'tenure' represents the number of months a client has stayed with the company. It serves as a measure of customer loyalty and engagement, with longer tenures potentially indicating a lower likelihood of churn. Secondly, 'MonthlyCharges' refers to the monthly amount received from the client, reflecting their spending behavior. Higher monthly charges might imply a higher level of service utilization, potentially impacting churn probabilities. Additionally, 'TotalCharges'

represents the yearly amount received from the client, providing a broader perspective on the customer's financial commitment to the company.

Moving on to categorical features, 'gender_Male' indicates the client's gender, while 'Partner_Yes' denotes whether the client has a partner. These attributes help capture variations in churn propensity based on demographics and social factors. 'InternetService_Fiber optic' signifies whether the client has internet service, offering insights into the impact of this particular service on churn rates. The duration of the client's contract is also a significant feature. 'Contract_One year' indicates a one-year contract, while 'Contract_Two year' represents a two-year contract. These features shed light on how contract length influences churn, as customers with longer-term commitments may be less likely to churn compared to those with month-to-month contracts.'PaperlessBilling_Yes' indicates whether the client receives paperless invoices, reflecting their preference for digital communication. Lastly, 'PaymentMethod_Electronic check' describes the client's payment method, specifically highlighting the use of electronic checks. This feature captures the impact of different payment methods on churn probabilities, as customer satisfaction with the payment process can influence their decision to churn. By considering these important features, we can gain valuable insights into the various factors that contribute to customer churn in the telecommunication industry. Analyzing and understanding these attributes can help in developing effective churn prediction models and formulating targeted strategies to mitigate churn and enhance customer retention.

The feature importance table provides insights into the relative importance of each feature in predicting customer churn. The importance values are based on the CatBoost algorithm, which was identified as the best-performing model. According to the feature importance scores, "tenure" emerges as the most influential predictor, with an importance score of 19.4868. This suggests that the number of months a customer has stayed with the company is a crucial factor in determining churn. Following closely is "MonthlyCharges" with an importance score of 18.8535, indicating that the monthly amount received from customers also plays a significant role in churn prediction. Among the categorical variables, "Contract_Two year" holds considerable importance with a score of 13.4977. This suggests that customers with longer contract durations are more likely to churn compared to those with shorter-term contracts.

"TotalCharges" also carries substantial importance (11.2433), implying that the yearly amount received from customers has a significant impact on churn prediction. Other features that contribute to churn prediction, albeit to a lesser extent, include "InternetService_Fiber optic" (8.349586), indicating that customers with fiber optic internet service may be more prone to churn. "Contract_One year" (8.2232) implies that customers with one-year contract durations also exhibit a higher likelihood of churn. Moreover, "PaymentMethod_Electronic check" (6.1785) suggests that customers who opt for electronic check as their payment method may have a higher propensity to churn. The remaining features, "gender_Male" (5.1629), "PaperlessBilling_Yes" (4.8276), and "Partner_Yes" (4.177), also contribute to churn prediction, but to a lesser degree compared to the top-ranking features.

While it is expected that the increase in total/monthly charges strongly affect the churn, the fact that the fiber optic internet service, electronic payment method and long-term contracts have a negative effect on customer satisfaction may give us strong insights about customer behavior. Also, mentioned in Data Exploration part and Figure 3.3, in one- and two-year contracts, customer with higher tenure are more likely to churn. Regarding this information, company might start customer-loyalty program for its customers who use these payment methods. By understanding the relative importance of these features, businesses can prioritize their retention strategies accordingly. For instance, they can focus on providing incentives to long-tenured customers, addressing issues related to monthly charges, and tailoring offerings to customers with specific contract types or payment methods. Feature importance graph for CatBoost algorithm can be seen below in Figure 4.1.

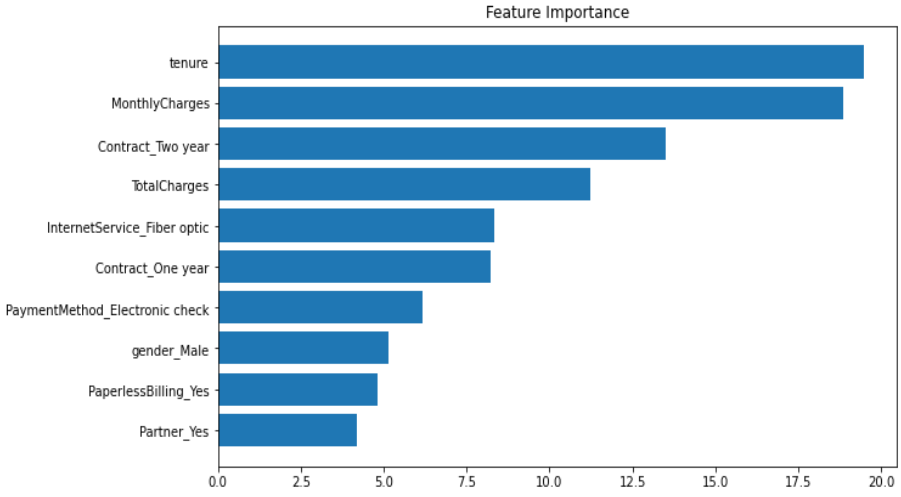

Figure 4.1: Feature Importance Graph of Best Model

4.6 Performance Metrics for Best Model

The analysis of the confusion matrix in Figure 4.2, which captures the performance of our churn prediction model, provides valuable insights when viewed in the context of customer churn. 1 represented churned customers and 0 represents non-churned customers. Among the results, we observed 726 true negatives and 527 true positives, indicating the accurate identification of non-churned and churned customers, respectively. However, it is important to recognize that misclassifications occurred, with 823 false positives and 34 false negatives. While both types of misclassifications have implications, as we mentioned in introduction part, it is crucial to note that misclassifying churned customers as non-churned (false negatives) can be more costly and detrimental for businesses. By misidentifying churned customers, we risk losing revenue and valuable opportunities for retention strategies, personalized engagement, and addressing their concerns. Furthermore, misclassified churned customers who are treated as non-churned may include high-value customers, exacerbating the impact on revenue and customer satisfaction. To mitigate these consequences, it is imperative to refine our churn prediction model, placing a greater emphasis on minimizing false negatives and accurately identifying customers at risk of churn. As mentioned in the first part, the algorithm was optimized in this way because the first focus was on the correct prediction of the customers which will churn. While evaluating the algorithms, FBeta score (beta=2), a metric that gives importance to recall and false negatives, was preferred. By prioritizing recall and reducing false negatives, we can enhance our ability to identify and retain churned customers effectively, leading to improved customer retention rates, revenue preservation, and enhanced customer experiences.

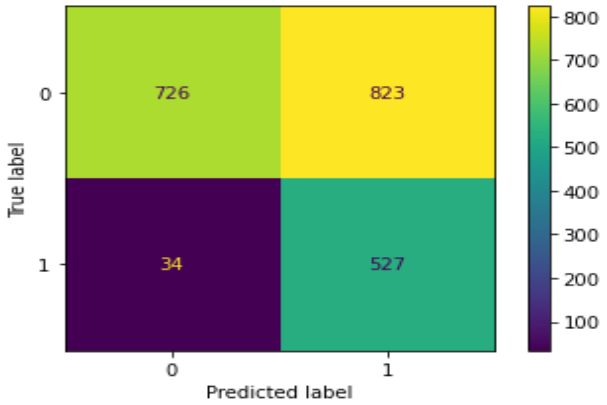

Figure 4.2: Confusion Matrix of Best Model

5. CONCLUSION

In this article, we explored the crucial task of churn prediction and investigated various techniques to address the challenges posed by imbalanced data. Churn prediction plays a pivotal role in understanding customer behavior and enables businesses to proactively retain their valuable customers.

One of the significant challenges encountered in churn prediction is imbalanced data, where the number of churned instances is significantly lower than the non-churned instances. To tackle this issue, we used sampling methods, which allowed us to balance the dataset and mitigate the bias towards the majority class. By oversampling the minority class or undersampling the majority class, we were able to create a more representative dataset that improved the performance of our churn prediction models. Furthermore, we also used various feature selection techniques to identify the most informative and relevant features for churn prediction. Feature selection helped us to reduce dimensionality and remove redundant or irrelevant features, which not only improved model performance but also enhanced interpretability. By focusing on the most influential factors, businesses can gain valuable insights into customer behavior and identify key drivers that contribute to churn. Throughout our analysis, we employed a number of machine learning algorithms, such as boosting methods, logistic regression, ensemble methods etc. and utilized grid search to optimize their hyperparameters. This allowed us to find the best configuration for our models and improve their predictive accuracy. By comparing different criteria, maximum depths, minimum samples splits, minimum samples leafs, and maximum features, we were able to identify the optimal combination of parameters that yielded the highest performance. All in all, Catboost with combination of SMOTE and RFE outperformed the other algorithms while scoring 0.7332 F Beta score and 0.9394 recall on minority class. The purpose of using the F Beta score is to increase the power of the model to identify churned customers (True Positives on confusion matrix) by giving more importance to recall. As mentioned in the literature section, importance was attached to increasing the number of True Positives by risking a decrease in precision, where it

is known that a possible loss of customer is much more costly for the company in terms of money and reliability.

In conclusion, this article demonstrated the importance of addressing imbalanced data in churn prediction and highlighted the effectiveness of sampling methods in achieving balanced datasets. Additionally, we showcased the value of feature selection in identifying the most influential factors contributing to churn. By leveraging these techniques and optimizing our models, businesses can make more accurate predictions, gain valuable insights into customer behavior, and take proactive measures to retain their customers, ultimately leading to improved customer satisfaction and business success.

REFERENCES

Alvarez, M. J., Hernandez-Perez, R., & de la Fuente, C. (2007). Feature selection and classification of microarray data using goal programming. Computers in Biology and Medicine, 37(4), 529-538.

Balakrishnan, N., & Raja, R. (2011). Heart disease diagnosis using feature selection and naive bayes classifier. International Journal of Computer Applications, 20(1), 37-42.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research, 164(1), 252–268.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626–4636.

C. Chen, A. Liaw and L. Breimenn. (2004). "Using Random Forest to Learn Imbalanced Data," Statistics Department of University of California, Berkeley.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, Y., Xie, Z., Liu, X., & Yin, L. (2016). Stock price prediction based on feature selection using random forests. Expert Systems with Applications, 46, 199-206.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313–327.

Deville, J.-C., Tillé, Y. (2004). Efficient balanced sampling: The cube method. Biometrika, 91(4), 893–912.

Dias, J., Godinho, P., & Torres, P. (2020). Machine learning for customer churn prediction in retail banking. In International Conference on Computational Science and Its Applications (pp. 576-589). Springer, Cham.

Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of biomedical informatics, 90, 103089.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42-47.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286(5439), 531-537.

Gosain, A., & Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: A review. In 2017 international conference on advances in computing, communications and informatics (ICACCI) (pp. 79-85). IEEE

Guliyev, H., Tatoglu, ˘ F.Y. (2021) Customer churn analysis in banking sector: evidence from explainable machine learning models. J. Appl. Microecon. 1 (2), 85–99.

Gunay, M., & Ensari, T. (2018). Predictive churn analysis with machine learning methods. 2018 26th Signal Processing and Communications Applications Conference (SIU).

Gür Ali, Ö., Arıtürk, U. (2014). Dynamic churn prediction framework with more effective use of rare event data: The case of private banking. Expert Syst. Appl. 41, 7889–7903

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications, 73, 220-239.

Hemalatha, P., & Amalanathan, G. M. (2019). A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector. 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN).

Iranmanesh, S.H.; Hamid, M.; Bastan, M.; Hamed Shakouri, G.; Nasiri, M.M. (2019). Customer churn prediction using artificial neural network: An analytical CRM application. In Proceedings of the International Conference on Industrial Engineering and Operations Management, Pilsen, Czech Republic, 23–26 July 2019.

Kaur, I., & Kaur, J. (2020). Customer Churn Analysis and Prediction in Banking Industry Using Machine Learning (PDGC).

Kaur, M., Singh K., & Sharma, N. (2013) Data Mining as a Tool to Predict the Churn Behavior among Indian Bank Customers. International Journal on Recent and Innovation Trends in Computing and Communication, 720-725.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30-37.

Kotsiantis, S. B., Kanellopoulos, D. N., & Pintelas, P. E. (2006). Feature selection using genetic algorithms and decision tree for large datasets. Expert Systems with Applications, 31(2), 197-206.

Kumar, D. A., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies, 1(1), 4. doi:10.1504/ijdats.2008.020020

Lalwani, P., Mishra, M. K., Chadha, J. S., and Sethi, P. (2021). Customer churn prediction system: a machine learning approach, Computing.

Lemmens, A., & Croux, C. (2006). Bagging and Boosting Classification Trees to Predict Churn. Journal of Marketing Research, 43(2), 276–286.

Lima, R. F., & Pereira, A. C. M. (2015). A fraud detection model based on feature selection and undersampling applied to web payment systems. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 3, pp. 219-222). IEEE.

Mishra, K., & Rani, R. (2017). Churn prediction in telecommunication using machine learning. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).

Mitkees, I. M. M., Badr, S. M., & ElSeddawy, A. I. B. (2017). Customer churn prediction model using data mining techniques. 2017 13th International Computer Engineering Conference (ICENCO).

Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303

Olaniyi, A.S., Olaolu, A.M., Jimada-Ojuolape, B. and Kayode, S.Y. (2020). Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms. International Journal of Multidisciplinary Sciences and Advanced Technology, 1(1), pp.48-54.

Rahman, M., & Kumar, V. (2020). Machine Learning Based Customer Churn Prediction in Banking. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA).

Raschka, S., & Mirjalili, V. (2021). Python machine learning (3rd ed.). Packt Publishing.

Sharma, A., & Kumar Panigrahi, P. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. International Jou rnal of Computer Applications, 27(11), 26–31.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. 2017 International Conference on Inventive Computing and Informatics (ICICI).

Swain, M. J., & Ballard, D. H. (1991). Color indexing. International Journal of Computer Vision, 7(1), 11-32.

Szeliski, R. (1999). Image compression using the singular value decomposition. IEEE Transactions on Image Processing, 8(9), 1217-1221.

Tahir, M. A., Kittler, J., Mikolajczyk, K., & Yan, F. (2009). A multiple expert approach to the class imbalance problem using inverse random under sampling. In International workshop on multiple classifier systems (pp. 82-91). Springer, Berlin, Heidelberg.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1), 71-86.

Wei-Chao, L., Chih-Fong, T., Ya-Han, H., & Jing-Shang, J. (2021). Clustering-based undersampling in class-imbalanced data. Expert Systems with Applications, 184, 115556.

Wold, S., Esbensen, K., & Geladi, P. (1993). Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1-3), 37-52.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, 412-420.

CURRICULUM VITAE

Name Surname **:** Furkan Kaymaz

Place and Date of Birth **:** 25.03.2000

E-Mail **:** kaymazf18@itu.edu.tr

EDUCATION **:**

- **B.Sc.** **:** 2023, Istanbul Technical University, Faculty of Management, Industrial Engineering

CURRICULUM VITAE

Name Surname **:** Ali Özcan Küreş

Place and Date of Birth **:** 04.10.2000

E-Mail **:** kures18@itu.edu.tr

EDUCATION **:**

- **B.Sc.** **:** 2023, Istanbul Technical University, Faculty of Management, Industrial Engineering