PREDICITING CUSTOMER CHURN IN TELECOMUNICATIONS COMPANY

A THESIS SUBMITTED TO

THE FACULTY OF ARCHITECTURE AND ENGINEERING

OF

EPOKA UNIVERSITY

BY

LUTFIE VEISLLARI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR

THE DEGREE OF MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

JUNE, 2024

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled **"Prediciting customer churn in Telecomunications Company"** and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assoc. Prof. Dr. Arban Uka
Head of Department
Date: June, 28, 2024

Examining Committee Members:

Prof. Dr. Gëzim Karapici (Computer Engineering) _____

Dr. Shkëlqim Hajrulla (Computer Engineering) _____

Dr. Florenc Skuka (Computer Engineering) _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Surname: Lutfie Veisllari

Signature: _____

ABSTRACT

PREDICITING CUSTOMER CHURN IN TELECOMUNICATIONS COMPANY

Veisllari, Lutfie

M.Sc., Department of Computer Engineering

Supervisor: Dr. Shkëlqim Hajrulla

Customer churn is one of the most critical issues in telecom companies. As it directly affects the company's revenue, arises the need of finding ways to predict and then prevent this kind of phenomen. Machine learning can highly contribute in developing algorithms that can be used in various companies that can firstly indicate factors affecting and then create patterns. The study aims to emphasize and investigate a conjecturing analysis of most of the predictive algorithms used for customer churn prediction, in telecommunication. Including the key factors affecting this kind of customer behavior, the causes and the consequences for the companies and concluding with how it can be predicted using machine learning algorithms, giving a hand to the companies to take measures before they experience their customer loss. We will use a real dataset obtained by an Albanian telecom company, Vodafone. The algorithm tested will be Logistic Regression and Random Forest. Models will be compared accordind to some evaluation metrics and the outperforming model will our suggestion to the company.The study brings in focus the useful indications for telecommunication companies and suggests some marketing strategies that use algorithmic outcomes to reduce churn rates. Within both of the algorithms Random Forest showed an outstanding performance with an accuracy of 94%, while the Logistic Regression struggled at an accuracy level of 85%. These results indicated that Random Forest is a better practice for this classification. As the dataset obtained consists of mainly categorical data, it is easier for Random Forest to deal with it, while Logistic

Regression struggles when it comes to categorical data. These results will serve as a helpful insight for telecom companies to face the issue of customer churn.

Keywords: *Customer Churn, Telecom Industry, Predictive Modeling, Machine Learning, Logistic Regression, Random Forest, Data Preprocessing Customer Retention*

ABSTRAKT

PARASHIKIMI I LARGIMIT TË KONSUMATORËVE NGA KOMPANITË TELEKOMUNIKATIVE

Veisllari, Lutfie

Master Shkencor, Departamenti i Inxhinierisë Kompjuterike

Udhëheqësi: Dr. Shkëlqim Hajrulla

Shpeshtësia e largimit dhe braktisjes, pas njëfarë kohe e konsumatorëve nga një kompani e caktuar është një nga cështjet me esenciale për jetëgjatësinë e asaj kompanie në treg. Për arsye të efektit të menjëhershë që ky largim ka në të ardhurat e kësaj kompanie, lind nevoja që të bëhen përmirësime në këtë drejtim, në strategjitë e marketingut në menyrë që kjo gjë të përmirësohet. Machine learning mund të kontribuojë në zhvillimin e algoritmave të përshtatshëm për shumë kompani që janë në gjendje të detektojnë arsyet e këtij largimi te konsumatorëve dhe pastaj të krijojnë patterns. Ky studim kërkon të theksojë dhe investigojë një analizë midis disa prej algoritmave më të përdorur ne parashikimin e shpeshtësisë së largimit të konsumatorëve. Duke përfshirë nënvizimin e faktorëve kyc qe ndikojnë mbi këtë dukuri, dhe duke konkluduar ne zgjidhjen e saj nëpërmjet parashikimit të bërë prej algoritmave, duke dhëne kështu një vlerë të shtuar ne cdo kompani për të mbajtur nën kontroll dhe për të rritur vigjilencën kundrejt sjelljes së konsumatorëve.Dataseti që do të përdoret do të jetë një dataset real i marrë prej nga kompanitë më të famshe të telekomunikacionit në Shqipëri, Vodafone. Algoritmat që do të testohen do të jenë Logistic Regression dhe Random Forest. Modelet do të krahasohen bazuar në disa matje të performances, të cilat do të nxjerrin në pah se cili algoritëm funksionon më së miri me këtë dataset. Rezultatet kane treguar se Random Forest ka performuar me mire sesa Logistic Regresion me nje rezultat prej 94% saktesie. Random Forest performon

me mire sesa Logistic Regresion kur behet fjale per te dhena kategorike. Kjo do të shërbejë si një sugjerim i mëtejshem për kompaninë Vodafone, por edhe të gjitha kompanitë e tjera ne vend dhe jashtë.

Fjalët kyçe: *Largimi i klientëve, Industria e telekomunikacionit, Modelimi parashikues, Mësimi i makinerive, Regresioni logjistik, Random Forest, Përpunimi paraprak i të dhënave, Mbajtja e klientëve*

I dedicated this Thesis to my beloved family and friends for their support and encouragement during my studies.

ACKNOWLEDGEMENTS

I would like to express my graditute and special thanks to my advisor Dr. Shkelqim Hajrulla for his constent support, help and continuous concern till my thesis completetion.

I will also like to express my sensierly to all of my academic profesors of Computer Engienering Department that gave me important lessons during my studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction to customer churn

Big businesses always deal with a multitude of issues that are directly related to their level of market success. Customer churn, which is the act of a customer switching from one company to another, is undoubtedly one of them.

The most crucial indicator of a company's success is its income, so creating a system to anticipate customer attrition is a novel task that will directly increase revenue. Particularly in the telecom industry, where daily churn varies dramatically. Additionally, it is well recognized that acquiring new customers is more expensive than keeping existing ones.

Given the significance of controlling customer attrition, the goal of this effort is to assist construct a predictive model that will enable telecom operators to properly identify the customers who are most likely to leave.

1.2 What is customer churn?

Client churn refers to the act of a client deciding to quit or terminate a contract because they are no longer loyal to a certain business. Customer churn is calculated using the customer churn rate. That is the total number of people who stopped being clients within a specified period of time, such as a month, a year, or a fiscal quarter.

$$\frac{customers\ at\ the\ beginning\ of\ the\ TP - customers\ at\ the\ end\ of\ the\ TP}{customers\ at\ the\ beginning\ of\ the\ TP}\ x\ 100 \qquad (1.1)$$

1.3 Why does customer churn matter?

Customer churn is detrimental to the business for a number of reasons, as stated in [1]:

Competitors may overtake you in market share. Your competitors in the market become vital when it comes to retaining customers, especially when they are going through hard circumstances. Given the reduced amount of money available and the increased expectations of consumers for the quality and value of brands, there is more competition between your business and rivals in your industry [2].

Consumer dissatisfaction negatively impacts your brand. Customers who are dissatisfied can easily become churned. You should be worried about more than just your turnover rate; apart from losing their business, you could also experience bad press, low ratings, and a reduction in the overall worth of your brand.

For every lost consumer, you pay more. It's general knowledge that keeping an existing customer is less expensive and more advantageous than acquiring a new one. Based on a range of facts and data, some contend that customer lifetime value is more significant than any one cost dimension, such as acquisition or retention.

Future growth may be impacted by consumer attrition. Your customer churn rate could be a good indicator of your company's potential for future growth—or lack thereof. If you're considering introducing new products and services, your existing customer base, who are already familiar with your brand, is generally the best market to target [3].

1.4 Thesis Objective

The work on this thesis will lead to the development of machine learning algorithms that will improve operator productivity and provide a deeper understanding of customer behavior in the telecommunications sector. A thorough examination of the prior machine learning algorithm used to forecast customer attrition will be carried out, and the most effective approach will be highlighted from the research depending on a few key performance indicators.

A final prediction model will be developed when the algorithms are evaluated on a particular and unique dataset that was acquired from an Albanian on a particular telecommunications firm.

1.5 Scope of works

In this project will be tested and demonstrated some of the machine learning algorithms that will emphasize the best working one predicting customer churn using the dataset from Vodafone Company.

The scope is to also create an idea of which are the variables that mostly affects the customer churn in this kind of industry.

1.6 Organization of the thesis

This thesis is divided in 5 chapters. The organization is done as follows:

In Chapter 1, Introduction where it is included the explanation of the topic and is highlighted the importance of it. Chapter 2, presents the literature review Chapter 3, consists of the methodology followed in this study.In Chapter 4, the experimental results. In Chapter 5, conclusions and recommendations for further research are stated.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Through times the concept of customer and the importance of it has changed. It took time and a high evolution in marketing strategy to understand that customer is the most significant element in industry. Earlier, when the production and competition wasn't in this high level as nowdays the customer was obliged to consume whatever was produced because it was a matter of need than a matter of choice for people [4][5].

Later, after 1960 and coming to nowdays, because of market liberalization the competition between different companies began to rise and this emphasized the need to focus on the customer and see him as the prior element in success of the company. According to studies it is a fact that keeping an old customer, or retaining a customer is surely cheaper than trying to acquire new customers.

2.2 Machine learning algorithms

When it comes to predicting a variable, machine learning has proven to be very effective. Since it is fast and accurate in understanding and analyzing large ammounts of data patterns it is often used to built high accuracy predictive models.

In our case, a lot of machine learning algorithms have been tested by previous studies in order to create predictive models for telecom industry. Companies use these kind of models not only to know numbers of people leaving the company but machine learning helps them to create an overview for the variables that affect on this churn.

Consequently, they change their marketing strategy in decrese customer churn [5].

In this literature review it emphasized how far has the studies gone till nowdays based on their prediction accuracy. By writing an overview of these previous

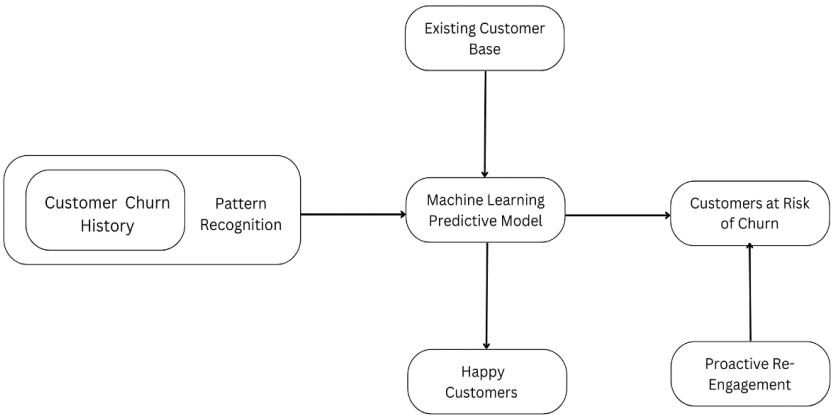

Figure 1. *Churn Rate Prediction Using Machine Learning*

researches we can understand not only how far they have arrived but also the gap for future arrangements and improvements.

2.2.1. Random Forest

This paper [6] deals with the significant issue of customer retention in the telecommunication industry by categorizing customers into two main groups: churn and non-churn. Customer churn refers to the phenomenon where subscribers interrupt their service. Customer churn is a critical issue because it poses considerable financial and operational difficulties for telecom companies. To handle this problem, the study puts forward the implementation of a Random Forest Classifier, which is a highly accurate robust machine learning algorithm that is also famous for its correct interpretability. What is expected by the implementation of this classifier is the

improvement of the prediction performance and obtaining a high efficiency rate, of at least 95 percent.

According to the findings this method can highly contribute in decreasing churn rates, which will derive into a progress in customer satisfaction and will solve the problem of revenue loss.

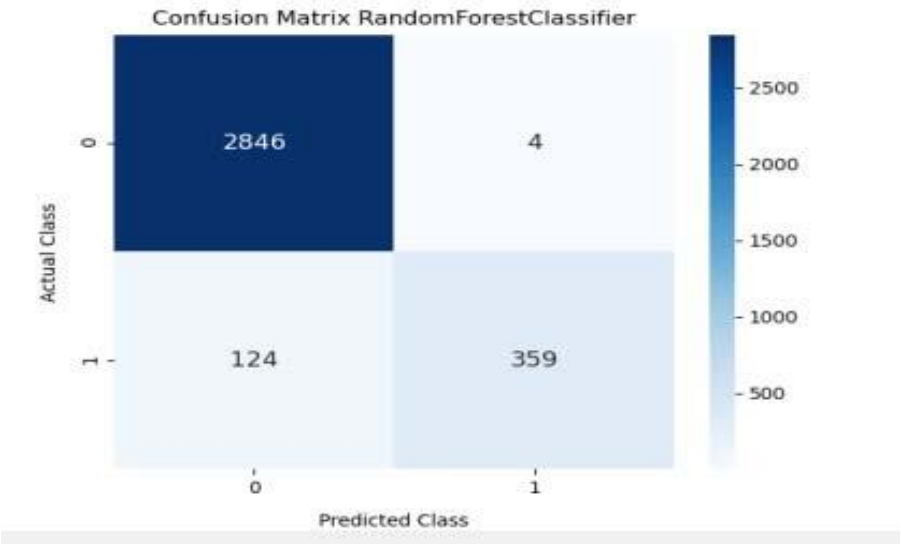

Figure 2. Confusion Matrix Random Forest

- Efficiency:96%

2.2.2 Decision Tree

Tree-shaped structures called Decision Trees (DTs) [7] reflect collections of decisions that can produce categorization rules for a given dataset. The class labels in these tree structures are represented by leaves, and the feature conjunctions that lead to those class labels are represented by branches.

When it comes to capturing intricate and non-linear correlations between the attributes, DT performs poorly. However, depending on the format of the data, a DT may have a high accuracy in the customer churn problem.

Table 1. *Decision Tree Performance*

Precision (%)	Recall (%)	Accuracy (%)	F-measure (%)
DT-C5.0 AdaBoost.M1 77.60	90.07	95.09	83.87

2.2.3 Support Vector Machines (SVM)

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm, as highlighted by Shreyas Rajesh Labhsetwar,[9] that excels in solving both regression and classification problems. The core principle of SVM involves plotting each data point in an n-dimensional space, where 'n' represents the number of features within the dataset. In this space, each data point is positioned such that the value of each feature corresponds directly to the value of each coordinate, creating a comprehensive and multi-dimensional representation of the data.

In the context of data classification, the primary objective of SVM is to identify the most optimal hyperplane that can distinctly differentiate between the various classes present in the dataset. This optimal hyperplane is chosen based on its ability to maximize the margin between the classes, ensuring that the data points belonging to

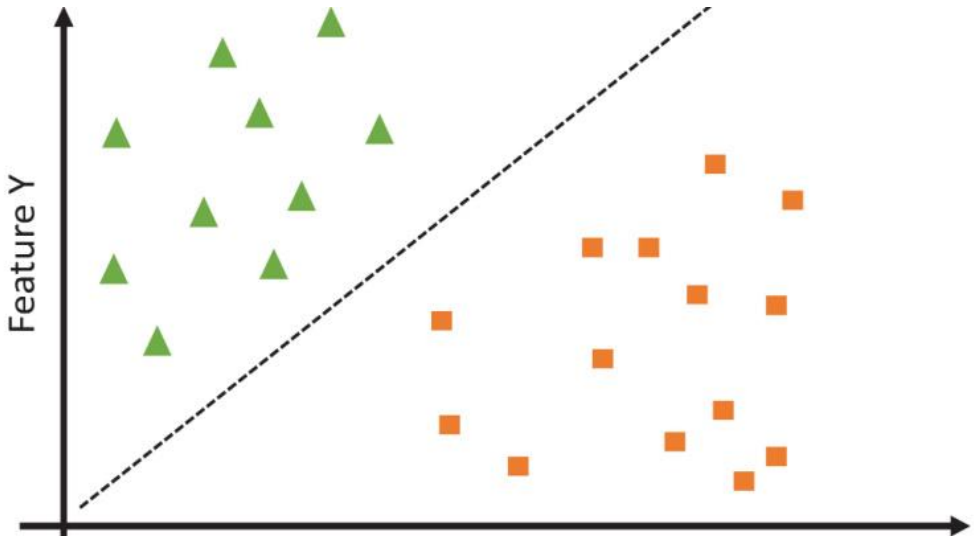

Figure 3. *SVM*

different classes are as far apart as possible. By achieving this, SVM enhances the model's generalizability and accuracy in classifying new, unseen data points [8].

The process begins with SVM attempting to fit multiple hyperplanes and then selecting the one that provides the greatest separation between the classes, which is often referred to as the maximum margin hyperplane. This separation is crucial as it minimizes the risk of misclassification and improves the robustness of the model [9]. Additionally, SVM can employ kernel functions to transform the data into higher dimensions, making it possible to classify data that is not linearly separable in the original feature space.

Table 2. *SVM Results*

Performance Parameters	SVM
Accuracy	0.896513
Sensitivity	0.895615
Specificity or Recall	0.101449
BCR	0.301429
Precision	0.995359
F1-Score	0.176308
MCC	0.465488
AUC	0.73536

2.2.4 Artificial Neural Network

The research aims to create a multilayer perceptron neural network (MLP-NN) model in order to conjecture telecommunication churn by using data taken from the industry. Income, occupation, education call, internet subscribers, voice service ,

monthly bills, unpaid numbers call duration, rates complaints, usage, total calling, and age are some of the variables which the dataset includes.

The way how a neural network works is through moving input data through multiple layers of interconnected neurons, where each neuron operates the data using weighted sums and activation functions in order to apprehend complicated patterns and association. The main reason for the network to adjust these weights during is to be able to minimize error. Other reasons include allowing it to make precise predictions and making new accurate classifications on new data [10][11].

The model has three layers: an input layer, a hidden layer with three nodes, and an output layer with two nodes.While the hidden layer and output layer use tanH activation and softmax, respectively, the error function is cross-entropy. The model structure is delineated in the network diagram, which was taken from the SPSS findings.

It is proven that the model worked very well from the result of a cross-entropy error of 1623.861 and an erroneous prediction rate of 22.7%. 10 subsequent steps with a little error reduction where achieved that made the the model more accurate in the prediction [12][13][14].

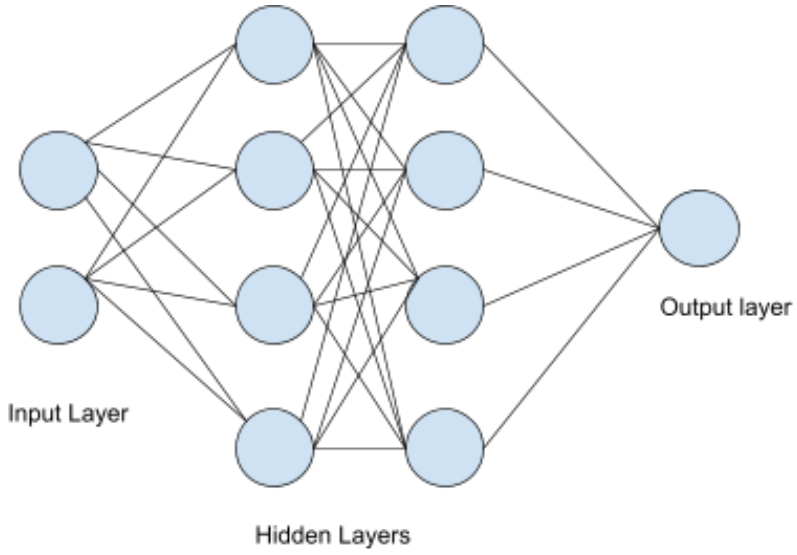

Figure 4. *Artificial Neural Network*

The model has three layers: an input layer, a hidden layer with three nodes, and an output layer with two nodes.While the hidden layer and output layer use tanH activation and softmax, respectively, the error function is cross-entropy. The model structure is delineated in the network diagram, which was taken from the SPSS findings.

It is proven that the model worked very well from the result of a cross-entropy error of 1623.861 and an erroneous prediction rate of 22.7%. 10 subsequent steps with a little error reduction where achieved that made the the model more accurate in the prediction [12][13][14].

CHAPTER 3

METHODOLOGY

3.1 Data Collection and Preprocessing

Vodafone, an outstanding telecommunications company, provided the data set for this research. The data set utilized for this study is comprised of rigorous information on customer engagements and actions, thus providing invaluable insights in order to give an accurate customer churn prediction model. There are 5000 client entries and 21 features that are classified as following:

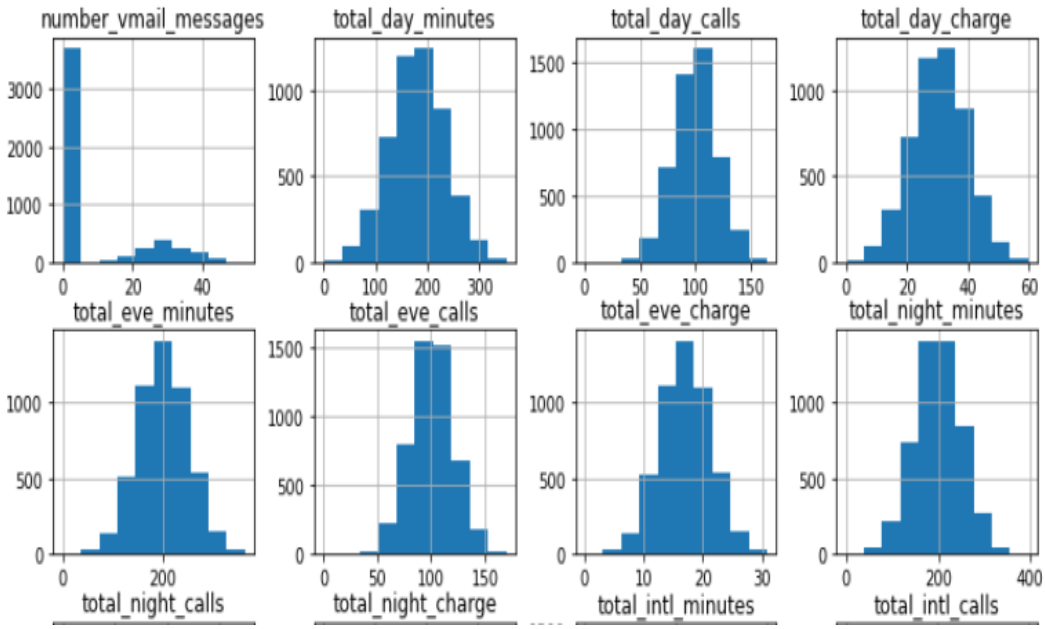

Figure 5. *Churn Prediction Features*

Customer Information: State, phone number, area code

Account Information: Account length, international plan, voice mail plan

Service Usage Metrics: Number of voice mail messages, total day minutes, total day calls, total day charge, total evening minutes, total evening calls, total evening charge,

total night minutes, total night calls, total night charge, total international minutes, total international calls, total international charge, number of customer service calls

Target Variable: Churn (if the customer has left or not)

3.1.1 Description of Variables

There is a variety of characteristics which portray various aspects related to customer behavior and service usage patterns. The variables used in this research are of paramount importance because they help to provide a thorough customer churn comprehension and prediction in the telecom industry.

3.1.2 Data cleaning and Preparation

The two most demanding steps in the preprocessing stage are data cleaning and preparation which guarantee the data set is accurate and in fit condition for immediate analysis and modeling. Numerous techniques are used in this process in order to deal with missing values, improve inconsistencies, and make data alteration, thus ensuring more appropriate formats for machine learning algorithms. The following will provide a comprehensive overview of the steps included in both data cleaning and preparation to make customer churn forecasts for the telecom industry.

1. Handling Missing Values

Finding Missing Values: The na_cols=churn_database.isna().any() print(na_cols) function was used to search the dataset for any missing values. Luckily, there were no missing values in the dataset, therefore imputation was not required.

Table 3. N*umber of missing values*

Variable	Missing Values
state	FALSE
account_length	FALSE
area_code	FALSE
phone_number	FALSE
international_plan	FALSE
voice_mail_plan	FALSE
number_vmail_messages	FALSE
total_day_minutes	FALSE
total_day_calls	FALSE
total_day_charge	FALSE
total_eve_minutes	FALSE
total_eve_calls	FALSE
total_eve_charge	FALSE
total_night_minutes	FALSE
total_night_calls	FALSE
total_night_charge	FALSE
total_intl_minutes	FALSE
total_intl_calls	FALSE
total_intl_charge	FALSE
number_customer_service_calls	FALSE
churn	FALSE

2. Balancing the Dataset

Handling Class Imbalance: It was found that just 14% of the customers in the dataset had churned. Techniques like undersampling the majority class or oversampling the minority class (using the SMOTE approach) were taken into consideration to correct this imbalance and make sure the model does not become biased in favor of the majority class.

from imblearn.over_sampling import SMOTE

smote = SMOTE()

X_resampled, y_resampled = smote.fit_resample(X, y)

6. Exploratory Data Analysis (EDA)

Understanding the distribution of features and their link to the target variable was accomplished through the use of descriptive statistics and visualizations. To find patterns and correlations between variables, methods like correlation matrices, box plots, and histograms were used.

7. Normalization and Scaling

Numerical feature scaling: To guarantee that each feature contributes equally to the model, features with varying scales were normalized. Using the StandardScaler from Scikit-learn, standard scaling (divided by the standard deviation and subtracting the mean) was applied to continuous data.

11. Dimensionality Reduction

The dataset's dimensionality was decreased by using Principal Component Analysis (PCA) or feature selection methods like SelectKBest. As a result, noise is decreased, only the most informative characteristics are kept, and model performance is enhanced.

3.2 Data Visualization

To fully comprehend the distribution and connections between distinct characteristics in the data set, data visualization is an essential part of the procedure. Data visualization is important when recognizing patterns, trends and possible outliers that might affect the model’s performance. The following description provides a detailed explanation on the usage of data visualization for the telecom customer churn data set.

Key numerical data including total call minutes, total charges, and number of customer support calls were plotted as histograms.To see the quartiles and outliers of numerical features, box plots were made.

Density plots were employed to decipher the distribution and pinpoint the feature's mode or modes.

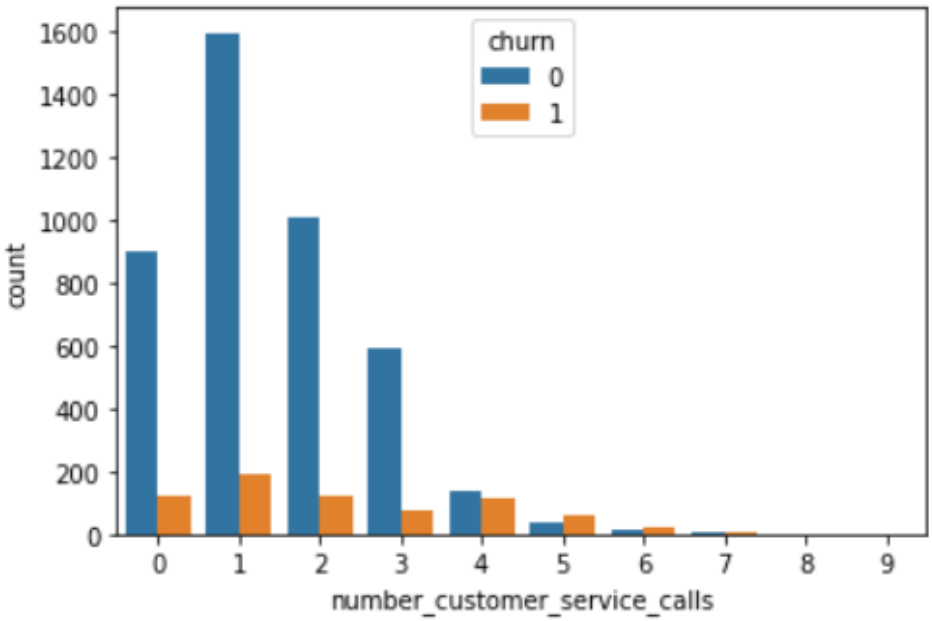

Figure 6. *Histogram showing number of service calls*

Through the utilization of diverse visualization methodologies, we can acquire a thorough comprehension of the dataset. Simple descriptive statistics might not be able to reveal patterns, trends, or anomalies that these visualizations might aid with. Among the visuals' most important lessons are:

Distribution of Usage Metrics: A small percentage of consumers have extremely high usage, whereas the majority of usage metrics have a skewed distribution.

Churn Patterns: greater customer service engagement levels are associated with greater churn rates, which may be a sign of discontent.

Strong correlations between some features (total minutes and total charges, for example) imply that some features can be deduced from others, which would simplify the model. In order to ensure that we have a thorough knowledge of the data before

developing predictive models, these visualizations offer a strong platform for additional research and modeling.

3.3 Training the Model

Once the data is completely preprocessed and adapted, training the machine learning models is the other essential step. The aim is to create a churn prediction model which provides precise customer churn details. The following sections provide information.

3.3.1 Splitting the Data

The dataset is split into training and testing sets to evaluate the model's performance on unseen data. This is a crucial step to ensure that the model generalizes well and does not overfit to the training data.

```
from sklearn.model_selection import train_test_split

X = churn_database.drop('churn', axis=1)

y = churn_database['churn']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

3.3.2 Training the Logistic Regression Model

Logistic Regression is a commonly used algorithm for binary classification problems such as churn prediction. It models the probability of a binary outcome based on one or more predictor variables.

```
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

```
# Train the Logistic Regression model

logmodel = LogisticRegression()

logmodel.fit(X_train, y_train)

# Predict the test set results

y_pred = logmodel.predict(X_test)

# Evaluate the model

print(f'Accuracy: {accuracy_score(y_test, y_pred):.2f}')

print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))
```

3.3.3 Training the Random Forest Model

Part of ensemble leanring methods is Random Forest that is built as a joining of numerous decision trees when the training occurs and as an output brings out the sum of the features.

```
from sklearn.ensemble import RandomForestClassifier

# Train the model - RF

clf_rf = RFClassifier(n_estimators=100, random_state=42)

clf_rf.fit(X_train, y_train)

# Predict the test set results

y_pred_rf = clf_rf.predict(X_test)

# Evaluate the model
```

```
print(f'Accuracy: {accuracy_score(y_test, y_pred_rf):.2f}')

print(confusion_matrix(y_test, y_pred_rf))

print(classification_report(y_test, y_pred_rf))
```

3.4 Model Architecture

In this section, we will briefly explain how our selected models work, how their structure looks and the importance of some key concepts does. This analysis will help us better understand and explain why we choose these two models.

3.4.1 Logistic Regression architecture

When it comes to binary classification Logistic Regresssion is one of the most famous algorythms used. It will use one or more variables that are called predictors in order to give a specific output which it is a binary one, meaning (in our case) churn or non churn. Logistic Regression architecture it is explained in detailed in the below section [15][16].

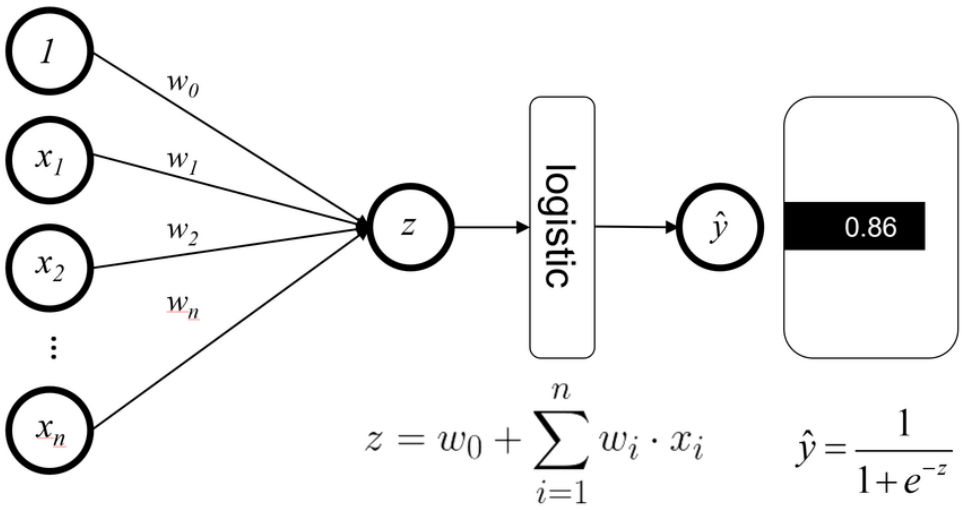

Figure 7. *Logistic Regression Architecture.*

1. Input Layer: So, in the input layer we will see the the features that are used in order to make this prediction, the predicted variables. In our case these features contain of service plans, account information etc.
2. Linear Combination of Inputs:

   The logic behind this, it's easy, the weighted sum of the features is computed.

$$Z = w_1\, x_1 + w_2\, x_2 + \cdots + w_n x_n + b \tag{3.1}$$

3. Sigmoid Activation Function:

   The weighted sum z is passed through the sigmoid function to convert it into a probability value between 0 and 1. In order to have an output that only appears as 0 or 1, the weighted sum z it is passed through the sigmoid function.

$$\sigma(z) = 1/1 + e - 2 \tag{3.2}$$

4. Output Layer:

   After generating a single probability value from the output layer, which is typically thresholded at 0.5, a binary classification (such as churn or no churn) can be made.

5. Optimization and Loss Function:

   The binary cross-entropy loss function is used in logistic regression to calculate the discrepancy between the actual class labels and the projected probability.

$$L(y, \hat{y}) = -[ylog(\hat{y} + (1 - y)\log(1 - \hat{y}) \tag{3.3}$$

6. Interpretability:

   To understand the influence of each feature on the probability of our called target outcome we will have to chech the coefficients of our model.

3.4.2 Random Forest Architecture

Classified as an essemble learning method, Random Forest makes the usage of multiple decision trees and take in consideration the output of each of them. It is commonly used for these kind of classsication issues because of some specific benefits. Handling large datasets and dealing with overfitting are two of the steps that this algorithm does successfully [17][18].

Input Layer: As mentioned in Logistic Regression, this layer consists off the predictor variable.

Bootstrap Sampling: To create many subsets of the training data, Random Forest employs bootstrap sampling, which combines random sampling and replacement. Each subset is used to train a different decision tree..

Decision Trees: The forest's decision trees are all trained separately. The data at each node of a decision tree is divided depending on the feature that, when measured against a selected criterion (such as entropy or Gini impurity), yields the best separation.

Random Feature Selection: A random collection of features is taken into consideration for splitting at each decision tree split. The overall performance of the ensemble is enhanced by this randomness since it increases variety among the trees and decreases correlation between them [19][20].

Output Layer: The output layer uses the majority vote from each individual tree to determine the final class label (such as churn or no churn).

Feature importance:

A measure of feature relevance is provided by Random Forest, which shows how much each characteristic influences decision-making across all trees. This aids in determining which characteristics have the greatest bearing on churn prediction.

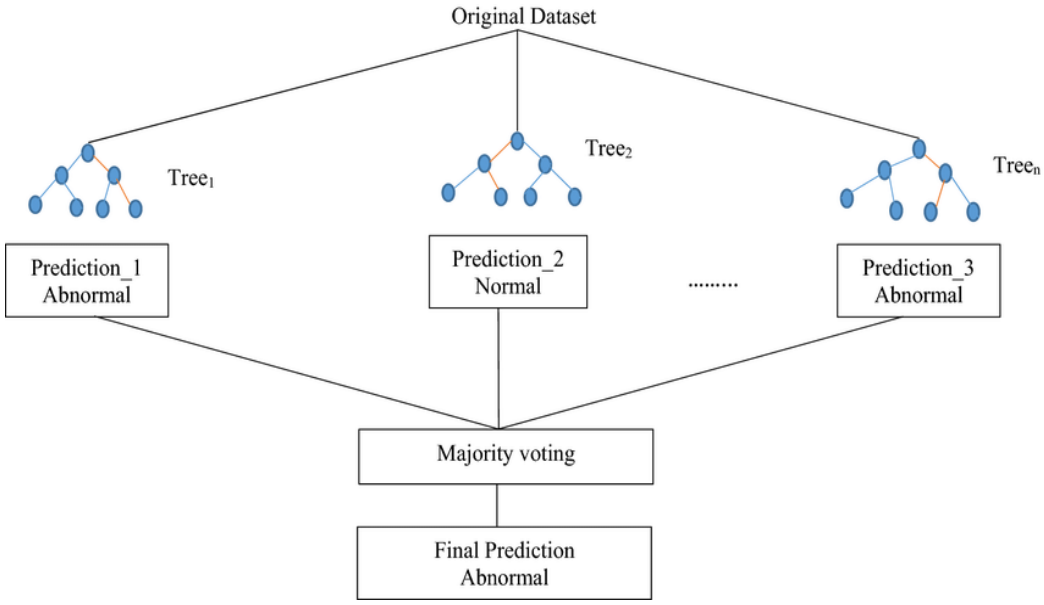

Figure 8. *Random Forest Architecture*

3.4.3 Neural Network Architecture

Neural networks are computational models inspired by the human brain, designed to recognize patterns and solve complex problems through interconnected nodes called neurons. A typical neural network has an input layer, one or more hidden layers, and an output layer, with each layer consisting of numerous neurons that process and transmit information. Activation functions like ReLU, Sigmoid, and Tanh introduce non-linearity into the model, allowing it to learn and represent complex patterns.

During training, neural networks use forward propagation to calculate outputs and backpropagation to adjust weights and biases, minimizing the error between predicted and actual outcomes. Convolutional Neural Networks (CNNs) are specialized for image and video processing, capturing spatial hierarchies in data through convolutional layers. Recurrent Neural Networks (RNNs) are designed for sequence prediction tasks, maintaining information across time steps, making them suitable for language modeling and time series forecasting.

Regularization techniques, such as dropout and L2 regularization, help prevent overfitting, ensuring the model generalizes well to unseen data. Hyperparameter

tuning involves optimizing parameters like learning rate, batch size, and the number of neurons in each layer to enhance model performance. Transfer learning leverages pre-trained neural networks on large datasets, allowing models to be fine-tuned for specific tasks with limited data, significantly reducing training time and improving accuracy.

Neural networks have diverse applications, including image and speech recognition, natural language processing, autonomous driving, and financial forecasting, demonstrating their versatility and efficacy in solving various problems.

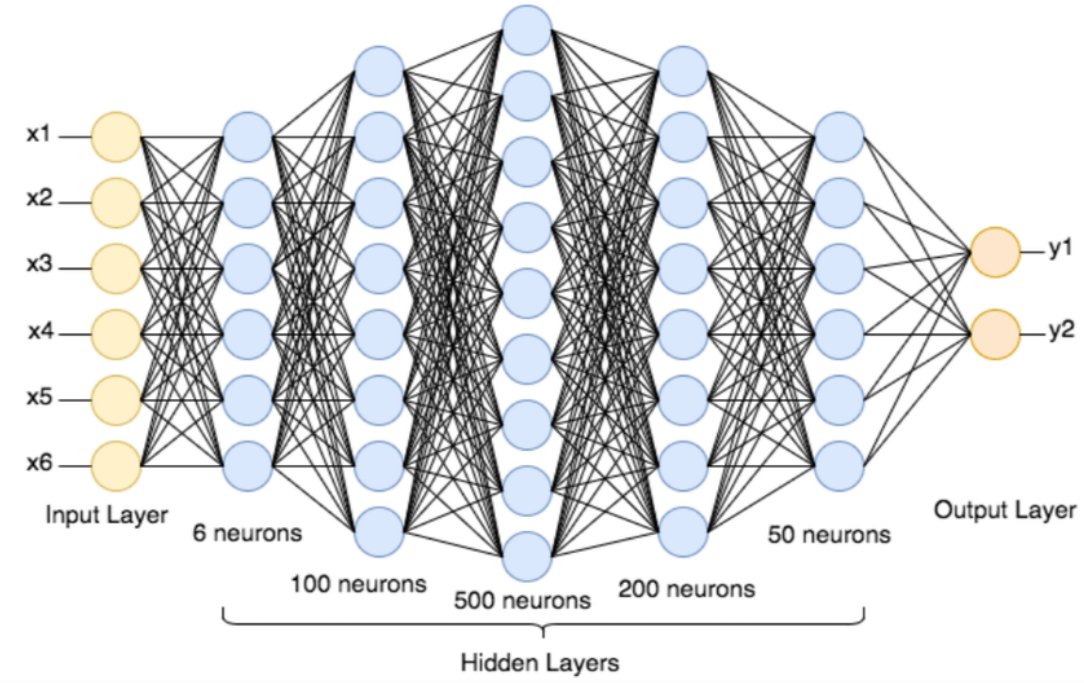

Figure 9. *Neural Network Architecture*

3.4.4 Support Vector Machine Architecture

Support Vector Machine (SVM):

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. It works by finding the hyperplane that best divides a dataset into classes. In a high-dimensional space, SVM constructs a hyperplane or set of hyperplanes that can be used for classification,

regression, or other tasks. The goal is to find the maximum-margin hyperplane that divides the classes from each other.

In this project, we used a linear kernel SVM to classify customer churn. We chose the linear kernel for its simplicity and effectiveness with the high-dimensional dataset. The dataset was split into training and testing sets, and then standardized before training the model. We also applied upsampling to address class imbalance.

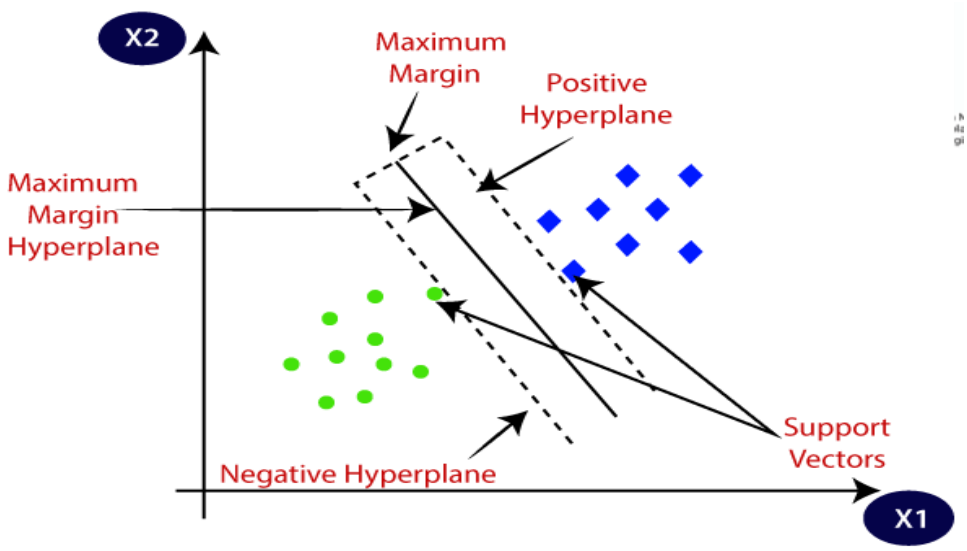

Figure 10. *SVM architecture*

3.5 Evaluation Metrics

To find out how well a machine learning model predicts the target variable, it is essential to evaluate its performance. Several criteria are used to evaluate the efficacy and accuracy of the customer churn prediction algorithms. These metrics—accuracy, precision, recall, F1-score, and ROC-AUC—offer insights into many facets of model performance. This is a thorough breakdown of every metric and its calculation method:

Accuracy is the ratio of correctly predicted instances to the total instances. It is a simple metric that provides an overall measure of model performance.

The Confusion Matrix is a table that summarizes the performance of a classification model by showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions.

```
conf_matrix = confusion_matrix(y_test, y_pred)

print('Confusion Matrix:')

print(conf_matrix)
```

True Positives (TP): Correctly predicted positive cases.

True Negatives (TN): Correctly predicted negative cases.

False Positives (FP): Incorrectly predicted positive cases.

False Negatives (FN): Incorrectly predicted negative cases.

The ratio of actual positive predictions to all expected positives is known as precision. It gauges how accurate the optimistic forecasts were.

```
precision = precision_score(y_test, y_pred)

print(f'Precision: {precision:.2f}')
```

The ratio of true positive forecasts to all actual positives is called recall, sometimes referred to as sensitivity or true positive rate. It assesses how well the model can recognize good examples.recall = recall_score(y_test, y_pred)

```
print(f'Recall: {recall:.2f}')
```

The harmonic mean of recall and precision is known as the F1-Score. It offers a solitary metric that harmonizes recall and precision, which is particularly helpful when there are unequal classes.

```
f1 = f1_score(y_test, y_pred)

print(f'F1-Score: {f1:.2f}')
```

The receiver operating characteristic - area under the curve, or ROC-AUC, gauges how well the model can discriminate between positive and negative scenarios. The true positive rate is plotted against the false positive rate in the ROC curve, and an overall performance metric across all classification thresholds is provided by the AUC.

CHAPTER 4

RESULTS AND DISCUSSIONS

In this section we present the findings and the results from the predictive models that were tested and compared for their precision achieved. The aim of this study was to build models that will help companies to find and predict in real time the customers that were likely to churn and to adopt their marketing strategies according to that.

The results evaluation it is done based on some specific metrics. Also it is discussed in detail the difficulties that each of the algorithm faced on the specific evaluation metric. The study brings out the strengths and weaknesses of each model and proposes improvements for each implication. In the end, we discuss the limitation of this study rising the issue of doing further research in this topic as it will improve a lot the marketing strategy of Vodafone Company, and general in any telecom company.

The Logistic Regression model, achieved a very good accuracy and high precision, but struggled with recall. The high false negatives number makes this model unable to identify a high number of customers that where churned and consequently makes it unable to complete the task.

Table 4. *Logistic Regresion*

	precision	recall	f1-score	support
0	0.87	0.98	0.92	1298
1	0.32	0.07	0.11	202
accuracy			0.85	1500
macro avg	0.59	0.52	0.52	1500
weighted avg	0.80	0.85	0.81	1500

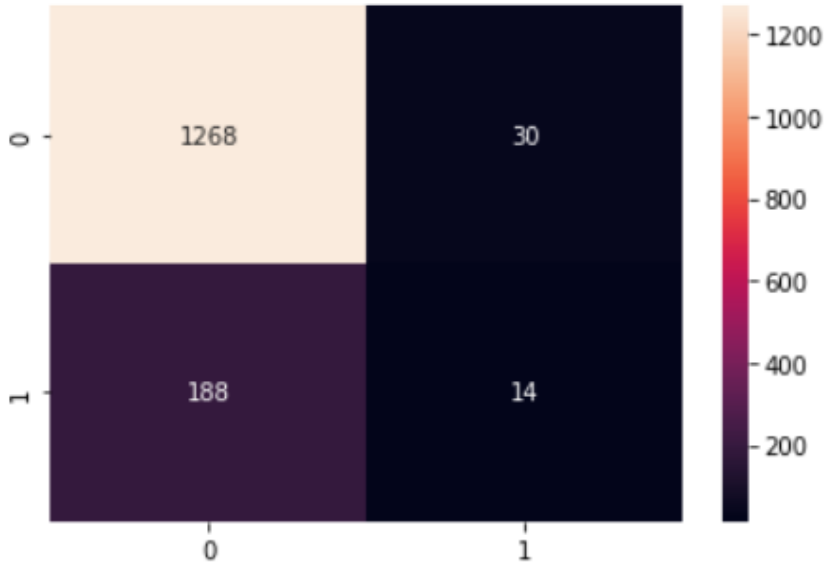

Figure 11. *Confussion Matrix Logistic Regression*

The model was very good and providing interpretability but this was at the cost of reducing overall performance of the algorithm.

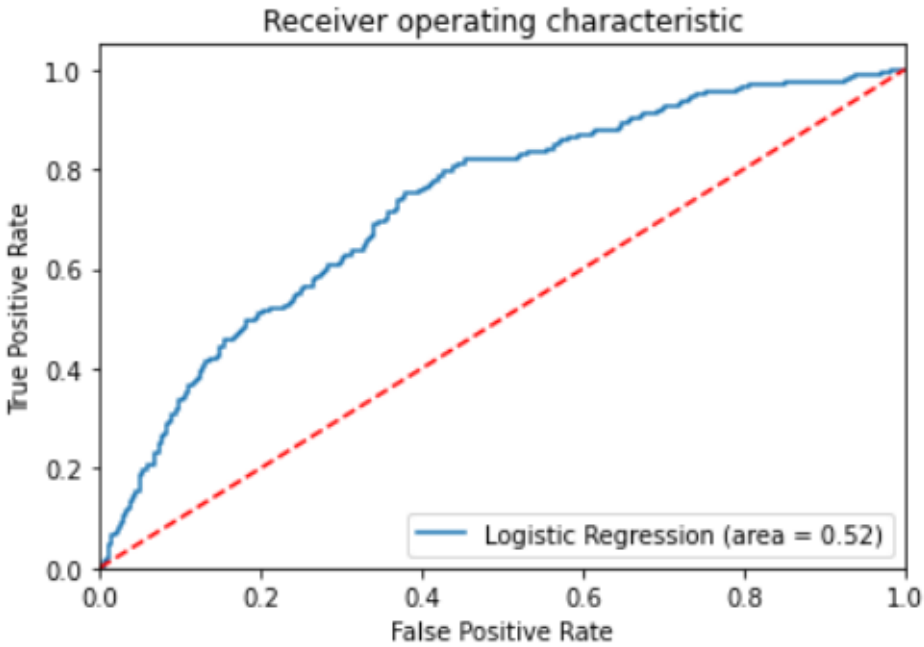

Figure 12. *Prediction Accuracy*

All metrics showed that the Random Forest model performed better than the Logistic Regression model, including accuracy, precision, recall, F1-Score, and ROC-AUC.

Table 5. *Random Forest Results*

	precision	recall	f1-score	support
0	0.94	0.99	0.97	862
1	0.95	0.64	0.76	138
accuracy			0.94	1000
macro avg	0.95	0.82	0.87	1000
weighted avg	0.95	0.94	0.94	1000

The balance that the algorithm made between recall and precision, was a good point to classify it as a good algorithm in recognizing customer churn.

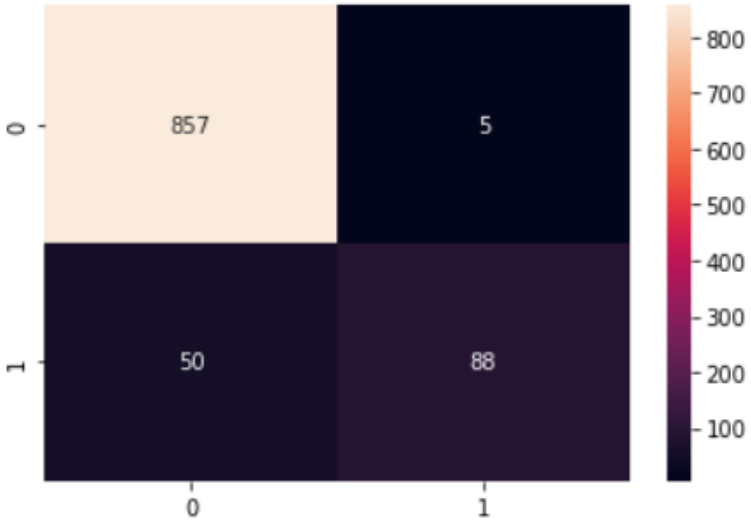

Figure 13. *Confussion Matrix Random Forest*

Vodafone Company became able to focus on the most important elements by using feature importance analysis from the Random Forest model, which gave useful insights into the main causes of churn.

Relevance in Practice

In the telecom sector, the comparison between the algorithms determines that the Random Forest model is the preferable option for prediciting customer churn. It is a powerful tool for identifying clients who are at danger because of its improved accuracy and well-balanced performance measures. The telecom company can improve client retention tactics, prevent revenue loss, and proactively handle churn by putting the Random Forest model into practice.

The neural network model for predicting customer churn showed strong performance during the evaluation phase. Trained over ten epochs, the model achieved a training accuracy of 94.34% and a validation accuracy of 90.50%, indicating robust learning and generalization capabilities.

On the test set, the neural network achieved an accuracy of 92.2%, demonstrating its effectiveness in predicting unseen data. The classification report revealed a precision of 0.94, recall of 0.97, and an F1-score of 0.96 for the non-churn class. For the churn class, the precision was 0.77, recall was 0.63, and the F1-score was 0.69, highlighting the model's reasonable balance between precision and recall.

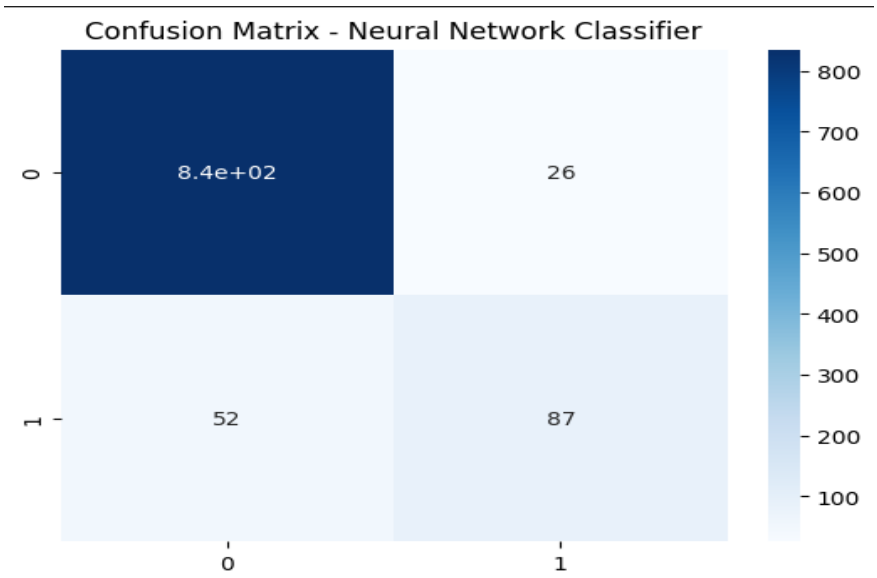

Figure 14. *Neural Network Confusion Matrix*

Table 6. *Neural Network Results*

Class	Precision	Recall	F1-Score	Support
0	0.94	0.97	0.96	861
1	0.80	0.75	0.77	139
Accuracy			0.92	1000
Macro Avg	0.86	0.8	0.82	1000
Weighted Avg	0.92	0.92	0.92	1000

Support Vector Machine (SVM) Results:

After training the SVM model on the upsampled dataset, we evaluated its performance using a confusion matrix and classification report.

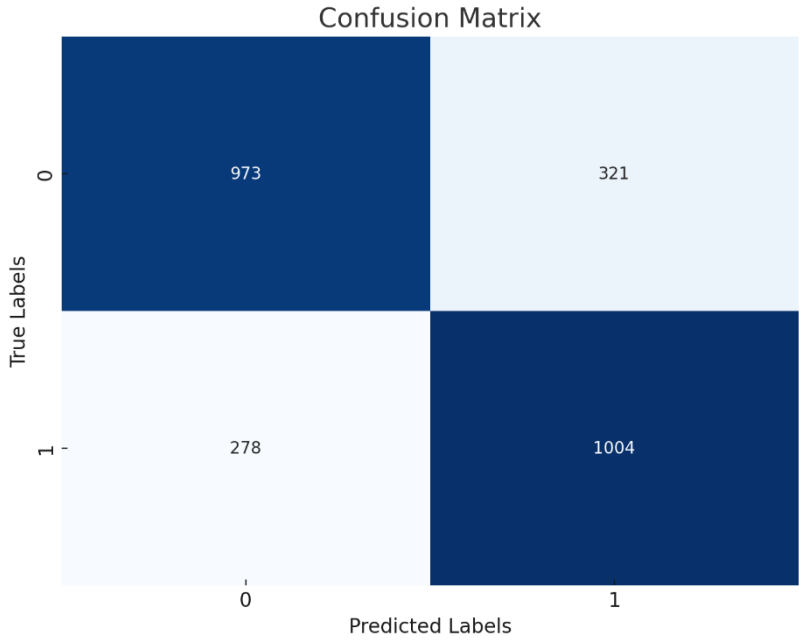

Figure 15. *SVM confussion matrix*

Table 7. SVM results

	Precision	Recall	F1-score	Support
Class 0	0.778	0.752	0.7646	1294
Class 1	0.758	0.783	0.7702	1282
Accuracy			0.7675	2576
Macro Avg	0.768	0.768	0.7674	2576
Weighted Avg	0.768	0.767	0.7674	2576

CHAPTER 5

CONCLUSIONS

5.1 Conclusions

The study's main objective was to anticipate customer turnover by comparing and implementing two of the most well-known predictive algorithms, Random Forest and Logistic Regression. Based on the aforementioned results, the Random Forest algorithm outperformed the Logistic Regression algorithm when all evaluation criteria were taken into account, including accuracy, recall, precision, F1-score, and ROC-AUC.

Logistic Regression had a very good performance in the interpretability but it showed a very high number of false negatives when it came to the recall metric, that means that the algorithm cannot be considered as applicable for this case.

Random Forest was able to emphasize the most important factors that were affecting customer churn in this company. According to the results the high expensive charges and the bad customer service was a key factor that drive the customer to churn. By being conscious of these factors the company can now focus its marketing strategy in creating some personalize offers and increasing the customer satisfaction.

The Random Forest showed how effective can predictive modeling be in providing proactive customer relationship management. Helping to identify customers that are likely to churn before they take the decision helps the company to make immediate changes and not only prevent but increseing the customer loyalty.

5.2 Recommendations for future research

The topic has space for more research and further improvement in order to achieve better results for prediciting customer churn, specifically in Vodafone Company. Testing the dataset in other predictive algorithms especially in deep learning

and also adding other features that can be taken in consideration such as the interaction in Vodafone social media.

More work needs to be done also with the issue of class imbalance. Systems that will be helpful in the immediate churn prediction needs to be implemented and also a daily upadate of the data will help in achieving real results in no time. The faster the company gets conscious about the customer churn they will be able to respond with a faster strategy, driven by the features predicted by the algorithm that were the main reason for this churn.

References

[1] A. Caldwell, "Losing customers? Calculate why.," Oracle NetSuite., 27 01 2021. [Online]. Available: https://www.netsuite.com/portal/resource/articles/human-resources/customer-churn-analysis.shtml. [Accessed 02 05 2024].

[2] "Customer churn analysis: Why analyzing churn is so important.," 17 01 2022. [Online]. Available: https://www.paddle.com/resources/customer-churn-analysis. [Accessed 25 04 2024].

[3] K. Skurikhin, "8 reasons why customers churn. NetHunt Blog | Sales, Marketing, and CRM.," 26 09 2023. [Online]. Available: https://nethunt.com/blog/why-customers-churn/.

[4] "What is customer churn? learn to measure & prevent it. Qualtrics.," 15 05 2024. [Online]. Available: https://www.qualtrics.com/experience-management/customer/customer-churn/.

[5] A. G. a. R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," in *International Conference on Advanced Computation and Telecommunication (ICACAT)*, India, 2018.

[6] Y. X. a. Y. T. Y. He, "Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company," in *Proceedings of the 2020 Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, 2020.

[7] L. Feng, "Research on Customer Churn Intelligent Prediction Model based on Borderline-SMOTE and Random Forest," in *4th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Shenyang, China, 2022.

[8] Y. H. a. P. T. S. Arya, "Telecom Sector Churn Prediction Using Decision Tree and Random Forest Model," in *3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, 2023.

[9] Y. Y. W. a. C. G. Vung, "Churn Prediction Models Using Gradient Boosted Tree and Random Forest Classifiers," in *Conference on Computer Applications (ICCA)*, Yangon, Myanmar, 2023.

[10] X. Z. a. Y. X. H. Zhao, "Customer Churn Prediction by Classification Models in Machine Learning," in *9th International Conference on Electrical and Electronics Engineering (ICEEE)*, Alanya, Turkey, 2022.

[11] J. H. e. al., "A Recurrent Neural Network based Approach for Customer Churn Prediction in Telecommunication Sector," in *International Conference on Big Data (Big Data)*, 2018.

[12] P. Pulkundwar, "A Comparison of Machine Learning Algorithms for Customer Churn Prediction," in *6th International Conference on Advances in Science and Technology (ICAST)*, Mumbai, India, 2023.

[13] C. Zhang, "Customer churn model based on complementarity measure and random forest," in *International Conference on Computer, Blockchain and Financial Development (CBFD)*, Nanjing, China, 2021.

[14] L. Chen, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," in *5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, Chengdu, China, 2020.

[15] R. Yahaya, "An Enhanced Bank Customers Churn Prediction Model Using A Hybrid Genetic Algorithm And K-Means Filter And Artificial Neural Network," in *2nd International Conference on Cyberspac (CYBER NIGERIA)*, Abuja, Nigeria, 2021.

[16] R. K. Peddarapu, "Customer Churn Prediction using Machine Learning," in *6th International Conference on Electronics, Communication and Aerospace Technology*, Coimbatore, India, 2022.

[17] A. F. Ramadhan, "The Comparison of Random Forest and Artificial Neural Network for Customer Churn Prediction in Telecommunication," in *3rd International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS)*, Bali, Indonesia, 2023.

[18] A. K. Ahmad, "Customer churn prediction in telecom using machine learning in Big Data Platform - Journal of BigData," vol. II, pp. 6-20, 20 03 2019.

[19] K. S. Wagh, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization,* vol. 14, 2024.

[20] K. Soundarapandiyan, "Comparative Study of Customer Churn Prediction Based on Data Ensemble Approach," in *Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, India, 2023.