University of Essex

Department of Mathematical Sciences

MA981 DISSERTATION

Enhancing Customer Retention Strategies through Churn Prediction in Telecommunication Industry

Omkar Parab
2201535

Supervisor: **Dr Jianya Lu**

August 25, 2023
Colchester

Contents

List of Figures

List of Tables

CHAPTER 1

Abstract

Telecommunications services like the internet and phones are crucial in today's related surroundings. In this industry, predicting consumer behaviour has taken on more importance. Customers leaving a business have an effect on revenue as well as the ability of the business to learn customer trends and preferences. My goal is to solve this problem by creating a smart model that will predict customer turnover and help companies in keeping their clients. The main goal of this research is to develop such a machine learning model for predicting client churn. The project begins by visualising the data to find trends and important factors that influence customer churn. Confirmation of these results is achieved through hypothesis testing, including chi-square and ANOVA tests. Since categorical variables make up the majority of the data, they are converted to numerical form for compatibility with other algorithms. Sampling strategies are used to handle the data imbalance, when one class is dominating and the other is a minority. For churn prediction, five models are used. The Random Forest classifier performs best, earning a 90% accuracy rate and a 95% recall score with the selected sampling technique.

CHAPTER 2

Introduction

2.1 Motivation

In this modern world of telecommunication which includes phone and internet services, the ability to predict the customers behaviour has gotten significant importance. With the customer leaving the company it not only effects the revenue of the company it also prevents the company to know valuable information of customer preferences and trends. To solve this problem, we want to create a smart model that can tell us when customers might leave. This can help companies take action to keep customers happy and not lose them. This project is all about making that smart machine learn model which predicts the customer churn.

2.2 Objective

The forecasting power can allow companies to know the reasons about why customers are leaving the company so that they can try to bring up some strategies which can eliminate the reasons and keep customers happy. By understanding the pattern in the data the model can predict the which customer is going to leave the company and we can also get to know the reasons behind it. When the algorithms predict whether customer will churn or not accurately it helps the company in focusing toward the improvement of the services, building good relationship with the customer and ultimately stop as

many customer churn. In recent years there are huge technology advancement because of this prediction models such as machine learning algorithms are becoming viable option for doing task like there. There are various kinds of machine learning algorithms which will be benefinical for this kind of task. The performance of the models can only be considered if they are accurately and precisely predicting the customer churn. Customers and the business could profit from a model that precisely detects and predicts when customers might churn. Customers and the business create a give-and-take relationship as a result, which benefits both parties. Predictive analytics provides a solid foundation for this peaceful and effective partnership. This dissertation explores the interaction between data analysis, prediction models, and smart customer strategies to go deeply into the details of anticipating when customers would depart.

In this project, I'll start by looking closely at the data using different types of graphs like bar plot for comparing the categorical columns, density plot for continuous variables, and boxplot for comparing the categorical and continuous variable. These graphs will help me understand the different parts of the data, like how certain factors relate to whether customers leave or stay with the company. I'll figure out which factors seem to have the most impact on customer decisions. These graphs will help me understand the different parts of the data, like how certain factors relate to whether customers leave or stay with the company. I'll figure out which factors seem to have the most impact on customer decisions. To make sure my findings are strong, I'll also use statistical tests to double-check. Since my data has more of one type of customer decision than the other, it's called "imbalanced." , a dataset is said to be imbalance when one class is in majority as compared to other class. To handle this situation, I have used 3 different sampling technique like under, over and smote sampling. To predict the Customer churn on the normal data, and the data which is formed by three sampling technique I will use 4 different machine learning models and 1 deep learning models which is Logistic regression, Decision tree, Random forest, XGBoost and ANN classifiers. These models will learn from the data patterns and help to make predictions. My aim will be to predict as many as customer churn means my model should able to predict true positives accurately and reduce as many false negative. I won't just rely on one measure like accuracy to know if my models are good. I'll also use precision, accuracy, and a

special score called F1 score. To make things clear, I'll use a graph that shows how well my model is doing when it comes to precision and recall This way, I can be sure my models are doing a great job even with the tricky imbalance in my data.

CHAPTER 3

Data Description

Dataset Link :- https://www.kaggle.com/datasets/blastchar/telco-customer-churn

This dataset contains crucial information related to customer behavior and attributes within a telecom company. The dataset revolves around customers who have recently decided to either stay with the company or discontinue their services, which is indicated by the "Churn" column. The dataset also comprises a variety of other features that shed light on the services customers have subscribed to and their demographic characteristics.

Features:

- **Customer ID:** A unique identifier assigned to each customer.
- **Gender:** Denotes the gender of the customer, indicating whether they are male or female.
- **Senior Citizen:** Indicates if the customer is a senior citizen, represented as 0 for no and 1 for yes.
- **Partner:** Indicates whether the customer has a partner, represented as "Yes" or "No".
- **Dependents:** Indicates whether the customer has dependents, represented as "Yes" or "No".

- **Tenure:** The duration for which a customer has been using the services of the company.
- **Phone Service:** Indicates whether the customer has subscribed to phone service, represented as "Yes" or "No".
- **Multiple Lines:** Specifies if the customer has multiple lines, represented as "Yes" for multiple lines, "No" for a single line, and "No phone service" for no phone subscription.
- **Internet Service:** The type of internet service subscribed by the customer, categorized as DSL, Fiber Optic, or No Internet Service.
- **Online Security:** Specifies if the customer has online security service, represented as "Yes", "No", or "No Internet Service".
- **Online Backup:** Indicates if the customer has online backup service, represented similarly to online security.
- **Device Protection:** Indicates if the customer has device protection service, following the same pattern as online security.
- **Tech Support:** Specifies whether the customer has tech support service, represented similarly to online security.
- **Streaming TV/Streaming Movies:** Denotes whether the customer has subscribed to streaming TV and/or streaming movies, represented as "Yes", "No", or "No Internet Service".
- **Contract:** The type of contract the customer has, categorized as "Month-to-month", "One year", or "Two year".
- **Paperless Billing:** Indicates if the customer has paperless billing, represented as "Yes" or "No".
- **Payment Method:** Specifies the method of payment chosen by the customer, including options like "Electronic check", "Mailed check", "Bank transfer (automatic)", and "Credit card (automatic)".

- **Monthly Charges:** The amount charged to the customer on a monthly basis.
- **Total Charges:** The total charges incurred by the customer.
- **Churn:** Denotes whether the customer churned or not, with "Yes" indicating churn and "No" indicating no churn.

CHAPTER 4

Literature Review

In order to deal with the issue of loss of customers in the telecom company, presents a new set of features. Dial types, account information, billing information, Henley segmentation, aggregated call details, payment information, complaint information, and service information are some of these features [1]. The models abilities to predict events is improved by this new set of features. These seven models Logistic Regression, Linear Classifier, Decision Trees, Multi-Layer Perceptron, Support Vector Machines, and Data Mining by Evolutionary Learning (DMEL) were used when carrying out this task [2]. While logistic regression will perform well if you're looking for churn possibility, decision trees and SVMs were useful for predicting true false churn rates. Future work can be done on solving the imbalance classification problem, and the size of the input variables should also be decreased, according to [3]

This study [4] used a more advance machine learning methodology that is made of many models. In order to forecast churn in the telecom industry. Random forest, AdaBoost, and Gradient Boosting have been compared with traditional algorithms like Naive Bayes, Support Vector Machine (SVM), and decision trees [5]. The idea that it is useful to identify clients who are very likely to leave a firm or its services early on is what the author's research is based on. Clients can increase profitability by doing this and retaining clients. The results show that Random Forest outperforms other techniques thanks to its high accuracy (91.66%), low error rate, higher sensitivity, and lower specificity.

[6] research explains the need to address the data imbalance problem in which there are more customers who didn't churn than customers who did. Algorithms can begin to generate incorrect predictions when classifying the majority class and become biased for the minority class. The author uses a variety of sampling techniques, including under-sampling, over-sampling, boosting, and random forest to resolve this problem [7]. The superior evaluation criteria that are more appropriate for imbalance data are AUC and lift. Under-sampling can increase accuracy especially when the model is evaluated using the AUC metric. Weighted random forests outperform traditional models [8], which is advantageous for churn prediction. According to the results weighted random forests and under sampling are particularly helpful techniques for dealing with data imbalance issues.

This paper [9] suggests a method for predicting which customers could transfer operators by utilising artificial neural networks (ANN). The model draws on datasets from telecom companies for several factors, including demographic information, billing details, and usage trends. The ANN-based technique predicts churn for Pakistan's telecom business with 79 percent accuracy. The ANN results highlight churn factors, allowing measures to be taken to address the causes of client churn. The study supports previous research and shows that ANN outperforms other classification methods. The research also assesses the impact of each variable on telecom customer churn, providing insights into the elements that contribute to churn. The backpropagation technique trains the model for churn prediction.

The class imbalance problem (CIP), also referred to as the issue of an uneven number of samples in a dataset, is discussed in relation to the prediction of customer departures. The authors [10] review research on class disparities and approaches to dealing with the CIP. According to this definition, classifying is the process of using a dataset to train a classifier to properly identify unknown classes of unseen objects. If the samples in the dataset [11] are not distributed equally, the classification process could produce incorrect outcomes. The authors then go into a number of oversampling strategies that can be applied to deal with the CIP, including SMOTE, ADASYN, and MTDF. They stress the value in evaluating the effectiveness of prediction models using metrics like

balance accuracy, imbalance ratio [12], area under the curve (AUC), and McNemar's statistical test. The results indicate that a technique called "MTDF" that uses rules created by a "Genetic algorithm" is the most effective.

Businesses must be able to predict when customers might stop their interactions with them in order to take steps to keep them happy. Similar to computer systems, neural networks are able to identify patterns and make predictions. Utilising multiple technologies in combination to improve predicts is another smart idea [13]. In order to forecast customer loss, this study compares two strategies for combining neural networks. The two neural network types they use are "ANN" and "SOM." The first combination model filters out irrelevant data, while the second model uses the filtered data to provide predictions. They [14] tested these models using a number of test datasets and found that for predicting client loss the combined models outperform using just one type of neural network. They also find that one combination performs better to the other.

CHAPTER 5

Methodology

5.1 Prepossessing

5.1.1 Duplicate values

Checking for duplicate values in the dataset is one of the primary and vital stages in the pre-processing phase. Due to the possibility of negative impact of duplicate values on the overall data quality, this step is of the utmost significance. If duplicates cannot be identified and eliminated, the dataset's quality may be distorted, producing incorrect outcomes and incorrect insights during further analysis. The performance of models can be impacted using duplicate data. A machine learning model is more likely to deliver bad results when trained on such information since it absorbs patterns from repeated instances. The model could have overfitting when it starts identifying noise rather than true patterns, limiting its ability to generalise to new, unknown data. I carefully performed the duplicate check during pre-processing for my specific dataset. Fortunately, I found that there were no duplicate values in my data. This successful conclusion is essential because it provides a solid basis for accurate analysis and reliable model training. I am better able to give my models high-quality input and get more precise and insightful results by making sure there are no duplicate variables.

5.1.2 Missing values

I carefully examine the dataset for the presence of null values in the following stage of the pre-processing. Prior to starting model training on the dataset, this step acquires the most significance. Null values can contribute a wide range of errors, weakening the dataset's durability and making any models trained on them to do incorrect analyses. Null values have a negative impact on several dataset attributes. Their impact on statistical analyses is particularly significant because the inclusion of null values can distort measurements like mean, median, and standard deviation. The effects also apply to data visualisation. When discrete data is represented by bar graphs, null values unintentionally appear as a separate category. This misrepresentation can cause incorrect interpretations because null values can accidentally be given too much weight, which affects the visual representation.

When it comes to data scaling, a requirement for algorithms like the K-Nearest Neighbours (KNN) classifier, the challenges created by null values are particularly obvious. The scaling process can be greatly impacted by null values, which affects the model's precision and potency. The fact that many machine learning algorithms are unable Strategies are required for dealing with null values. A simple strategy involves replacing missing values with the mean or median of the column for datasets where the proportion of missing values is low (5% to 10%). More advanced imputation approaches are used when the number of missing values is greater than 10. The KNN imputer, which uses nearby data points to anticipate missing values, and the multiple imputation technique, which creates numerous imputed datasets that include the uncertainty of imputation, are noteworthy options.

By reviewing my dataset, I got to know the absence of null values. This effective outcome improves the accuracy of the data can lead to an in-depth and exact modelling method. By carefully addressing the null value issue, I can say that my model will perform well on my data.

5.1.3 Feature Engineering

It's essential to do this crucial steps before we can apply machine learning models to our data. Since the majority of algorithms use mathematical calculations to find patterns in

the data, they need numerical data for their calculations. Categorical data, which represents non-numeric categories and presents difficulties for mathematical computations, unfortunately causes algorithms difficulties. Directly trying to include categorical data can result in errors and affect the performance of the model. Nominal and ordinal data are the two main forms of categorical data. Categories without a natural order, such as different fruits like apples, oranges, bananas, and mangoes, are included in nominal data. These categories are distinct from one another and are not arranged in any logical sequence. However, ordinal data has a defined ranking or order, as seen by terms like "bad," "good," and "best." There is a different ranking among these categories, with "best" ranking higher than "good." According to the type of categorical data, we use a different strategy to address it. One-hot encoding is a common technique for nominal data. This method converts each category into a binary column that indicates whether it is present in a data point or not. During the process, a category like "No" might turn into a new column with the same name. This method makes sure that algorithms don't infer relationships between categories that aren't necessarily related.

5.1.4 One-Hot encoding

It is a method for turning category information into numerical information. It transforms categorical data into a binary matrix format, making it easier for machine learning algorithms to use it for training as most of them require numerical input. The category in the column is changed to a new binary column, where 1 denotes the presence of a value in the column and 0 denotes its absence. One-hot encoding allows the method to efficiently analyse categorical data by preventing it from assuming any ordinal relationship between categories to ensure that the model treats each category fairly and prevents the introduction of unexpected patterns.

5.1.5 Ordinal Encoding

Ordinal encoding is a technique specifically designed for handling ordinal categorical data, where the categories have a meaningful order or hierarchy. Instead of converting categories into separate binary columns, ordinal encoding assigns numerical values to the categories based on their position in the order.

One column in my dataset which is 'MultipleLines' had three categorical values each: "No", "Yes", and "No phone service." I changed "No phone service" to "No" to make the analysis easier to follow. Additionally, many columns just like ' MultipleLine' have 3 categories which is 'Yes', 'No' and 'No Internet Service' so in this i converted the 'No Internet services' to 'No' same as we did for previous variable. Many columns included binary input like "Yes" and "No," among others. I changed "No" to 0 and "Yes" to 1 to convert them to numerical as most of the algorithms works well on numerical data. In addition, I had a column called "Gender" with the values "Male" and "Female," which I similarly changed to "0" and "1" for analysis. After dealing with binary categorical data, I used the pandaâs library's 'get_dummies' method to deal with nominal data. This method creates binary columns for each category, which makes simpler to process algorithms. The first column generated during encoding was eliminated nevertheless, to avoid problems with multicollinearity because it can cause the dataset to become redundant. The algorithms' ability to recognise patterns and relationships is enhanced by the conversion of category data into a numerical data, eventually leading to more accurate and informative conclusions."

5.2 Exploratory Data Analysis

5.2.1 Data Visualization

Payment Method Analysis and Customer Churn.

As shown in Fig. 5.1 The goal of this plot is to determine the preferred payment methods by customers and whether any specific techniques contribute to customer churn, or when consumers stop using the services. We look at four methods of payment which is bank transfer, credit card, traditional mailed cheque and modern electronic cheque. The graph 5.1 shows how these techniques affect customer behaviour. Customers prefer to use credit card payments since they are simple to use and convenient. Bank transfers received positive feedback as it shows that many users are comfortable in this method of payment. The graph points out a big issue with the electronic cheque option, as over 1000 clients have chosen it and then stop using the services. This suggests that there

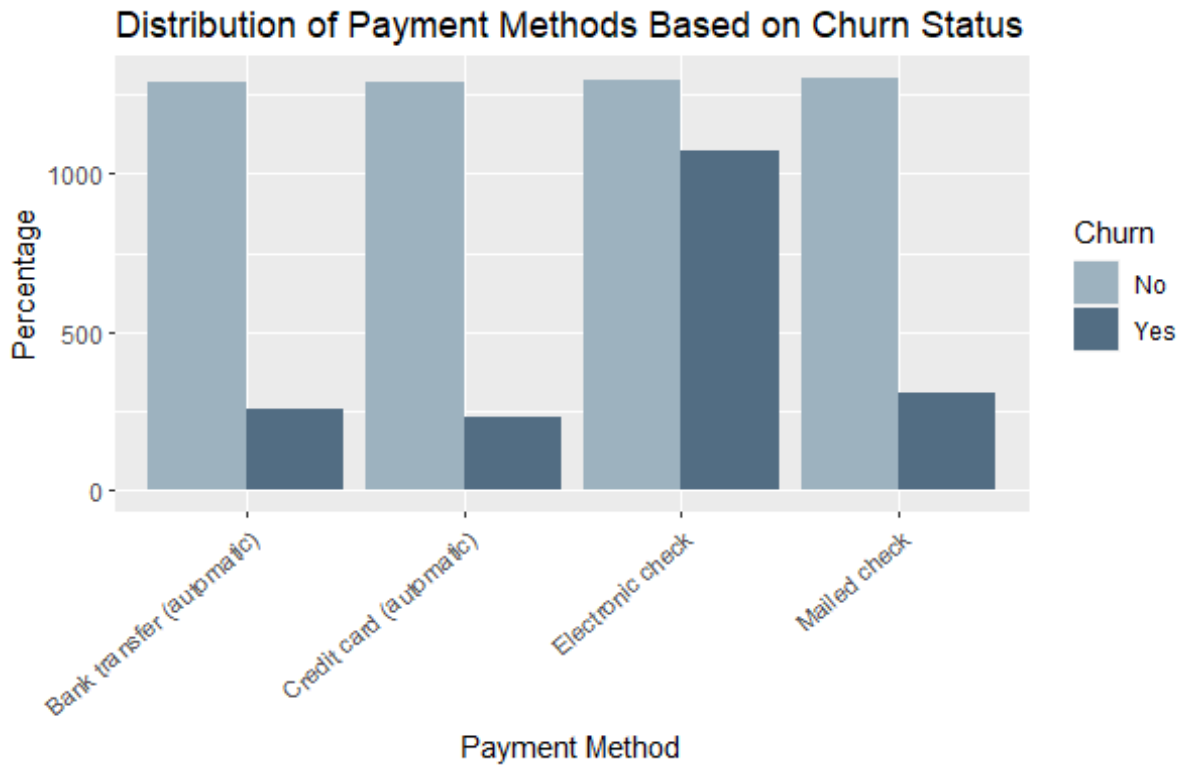

Figure 5.1: Comparision between payment methods and churn.

may be technical problems or a dissatisfaction with electronic checks. Mailing checks is the most use payment method. Customers who choose this method have lower churn rates which indicates that customer are highly satisfied. The graph conclude that to meet customer expectations and lower the churn rate the electronic cheque service may need to be improved. To promote a better customer experience, the business should concentrate on improving this payment option.

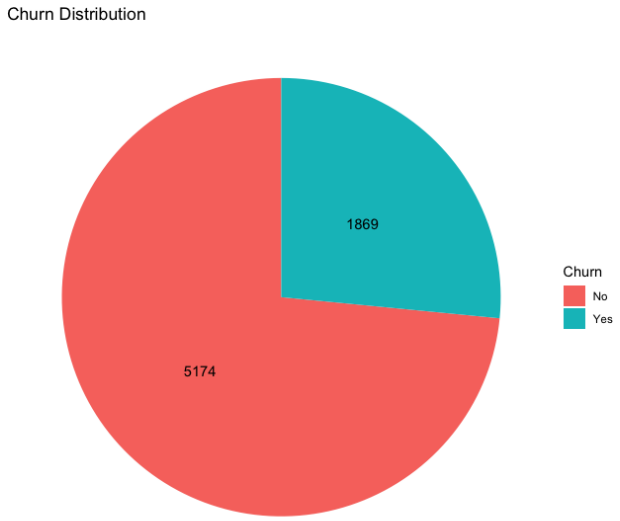

Figure 5.2: Churn.

Churn class distribution

We generated a pie plot Fig. 5.2 and found that the "No" class which represents customers who didn't churn is present in larger quantity than the "Yesâ class. When we try to use this data to train models on this it will be bias for majority class. Due to the greater number of cases in one group, the model may focus more on that group and perform poorly when predicting the smaller group.

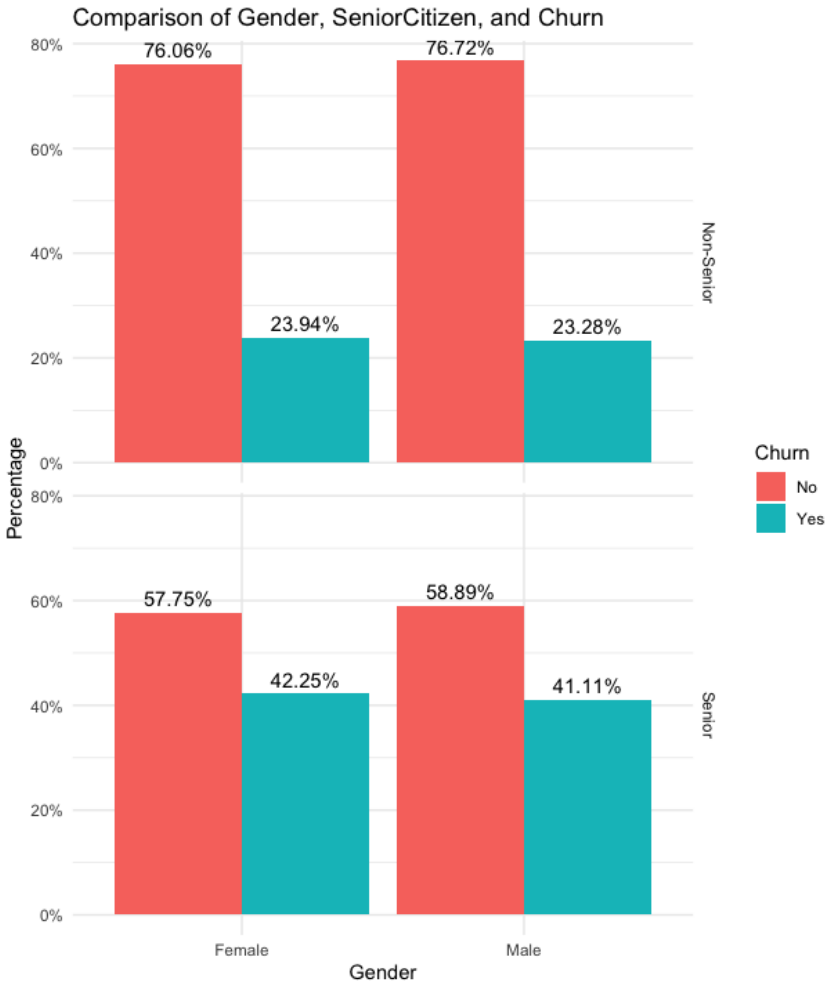

Figure 5.3: Comparison between senior citizen, gender and churn.

Gender and Senior Citizen Analysis in Relation to Churn

In the given figure 5.3, I've looked for the relationship between gender, senior citizen status, and churn. What I observe is that there's a higher rate of churn among senior citizens especially among women as compared to men. The churn rate is lower among non-senior citizens and this could imply that the services provided by the company might be more challenging for senior citizens to use, while non-senior citizens find it easier to adapt to. The company should consider simplifying their services and making them more user-friendly, particularly for senior citizens. In this way they can ensure a smoother experience and better retention among this demographic.

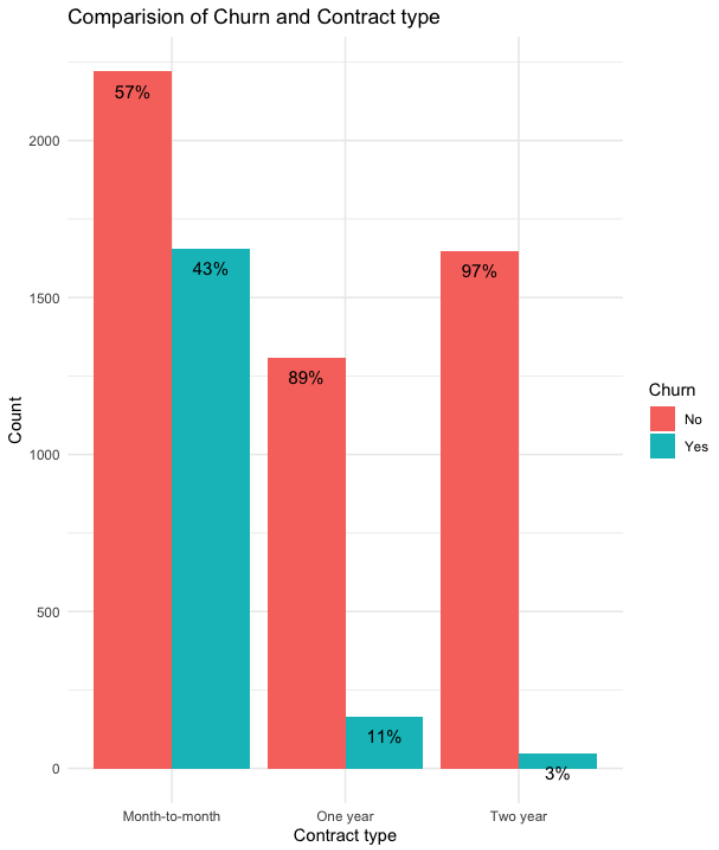

Figure 5.4: Comparison between contract type and churn

Analysis of Contract Types in Relation to Churn.

In this plot 5.4 I made comparison of contract type and churn, I wanted to see which contract type is letting people leave the service. There are three contract categories which is 'Month-to-month', 'One year', and 'Two year'. We can see from the graph that most of the customers who stayed with the company are in the 'Month-to-month' contract category but many of the customers who left the company are also in the same 'Month-to-month' category. For the 'One year' and 'Two year' contract more customers prefer to stay rather than leaving this service. This could mean that customers who choose longer contracts tend to stick around more. So, it seems like customers with 'Month-to-month' contracts might be more likely to leave, while those with longer contracts are more likely to stay. This gives the company a hint that they might want to focus on keeping 'Month-to-month' customers happy and finding ways to encourage them to stay.

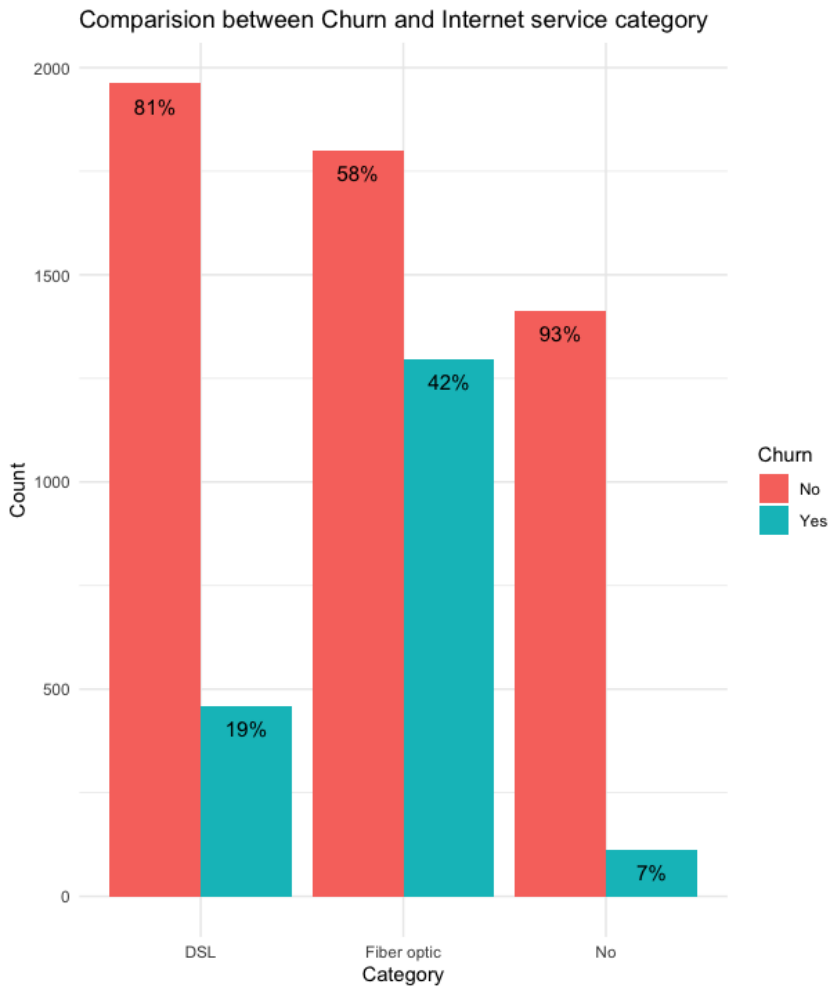

Figure 5.5: comparison between internet service and churn.

Comparison of Internet Service Types and Churn Rates

This graphs shows the comparison between different internet service and Churn. There are three main types of internet services which is 'DSL', 'Fiber optic', and 'No Internet service'. While looking at âDSLâ we can observe that it consists of 81 percent who hasnât churn the company while 19% have churn the company. In Fiber optic category around 58% of them stayed loyal but a significant 42% opted to discontinue the service, indicating a considerable churn rate in this category. In No Internet Service category customers without any Internet service showed a high retention rates with approximately 93% remaining with the company. This graph 5.5 suggest thatâs the company should more on their Fiber optic service as company must be unsatisfied with the services provided by company.

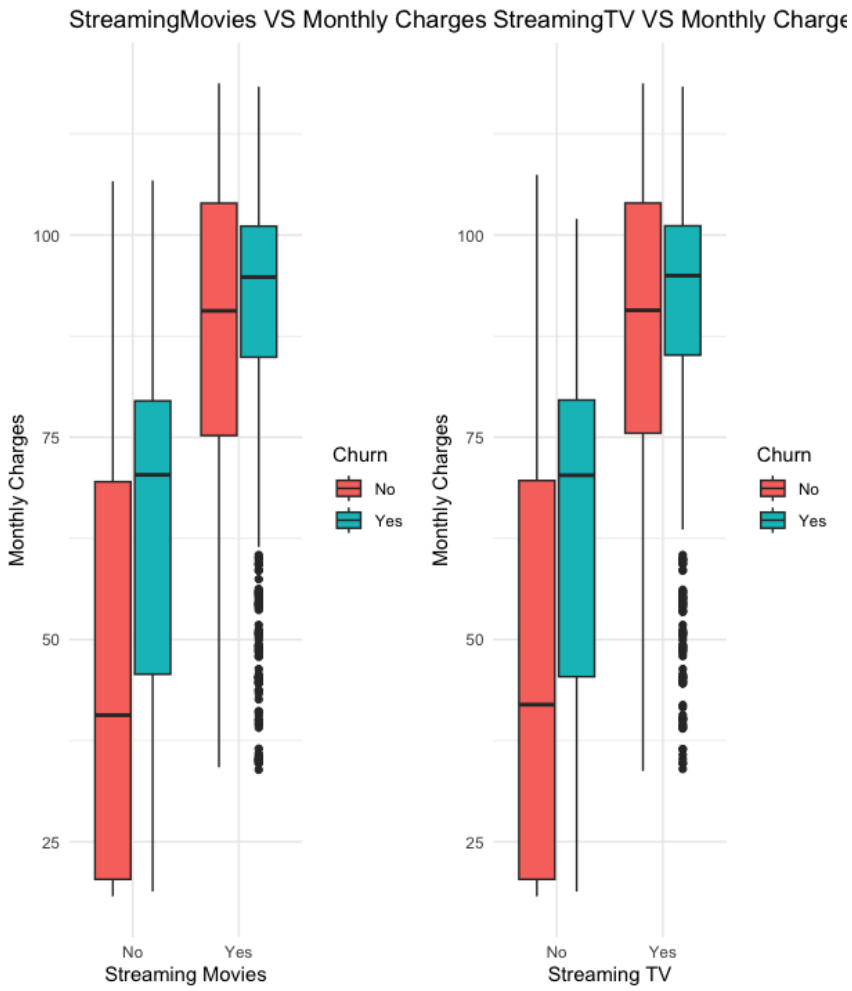

Figure 5.6: comparison between streaming movies, streaming tv and churn.

Analysis of Streaming Services and Monthly Charges in Relation to Churn

In the boxplot 5.6 we can see a significant number of customers who haven't subscribed to streaming movies and TV services, as large proportion of these customers are leaving the company. This suggests that customers who haven't opted for these services might be dissatisfied and more likely to leave. When we the graph for comparison of monthly charges,we observe that when the monthly charges exceed 60 there's a trend of customers leaving the services. This indicates that higher pricing for these services could be a reason for customer churn. The analysis suggests that the company should consider improving the quality of its streaming services to retain more customers and adjusting the pricing strategy to make the services more affordable could help prevent customer attrition.

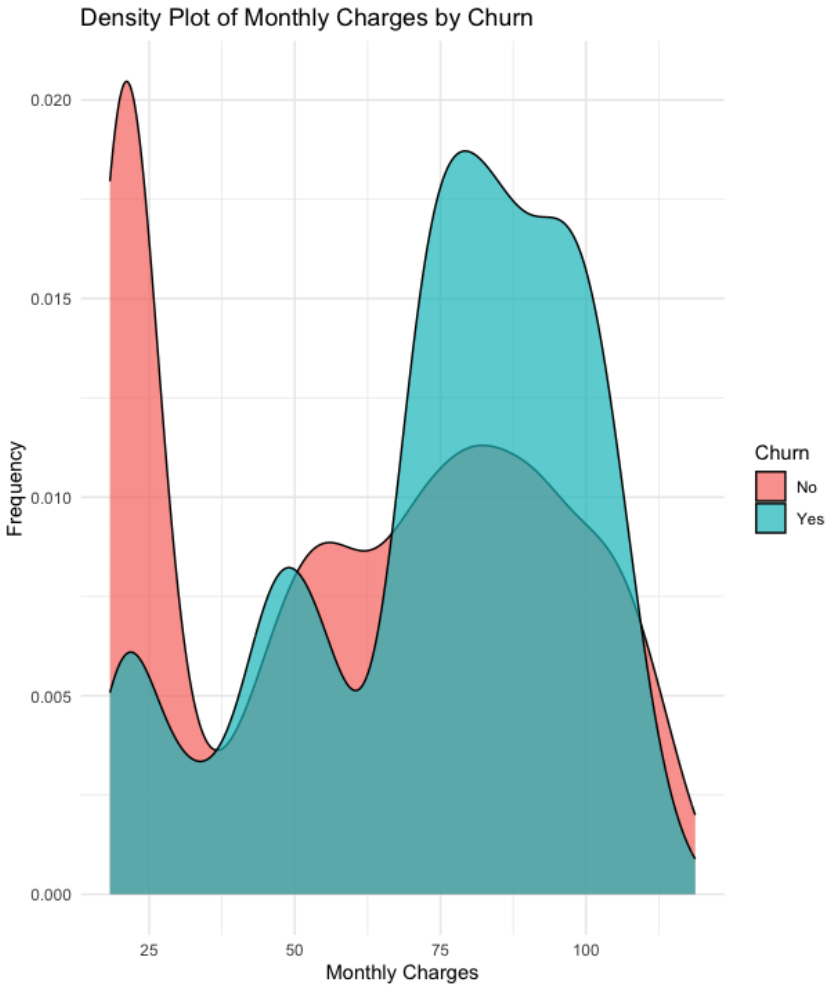

Figure 5.7: Comparison between monthly charges and churn.

Affordability's Impact on Customer Churn: A Comparative Analysis of Monthly Charges

In this graph 5.7 I've compared the Monthly Charge with customer Churn. The findings reveal that when services are offered at affordable rates customers are more likely to remain with the company. As we saw in our previous observation, where we noticed that customers tend to discontinue services like TV and movies when the prices become high. This suggests that affordability plays a significant role in customer decision-making. The company should carefully review its pricing strategy. By offering services at competitive and reasonable prices the company can improve customer satisfaction and retention. This strategic adjustment has the potential to keep more customers and increase the company's position in the market.

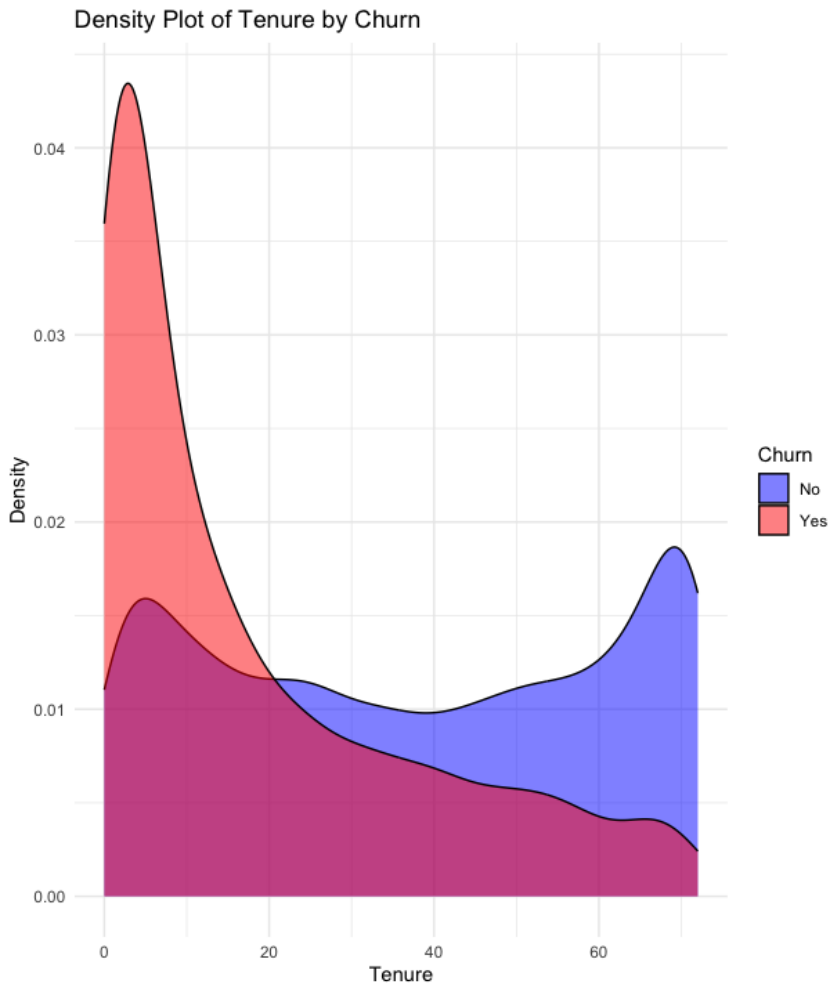

Figure 5.8: Comparison between tenure and churn.

Tenure's Influence on Churn: Unveiling the Role of Customer Loyalty and Trust

Upon analysing this graph 5.8 a clear pattern emerges that the customers with shorter tenure periods are more likely to leave the company. This trend suggests that these customers might lack confidence in the company's ability to continuously provide satisfactory service over time. Individuals who have been with the company for a longer duration shows higher satisfaction levels with the services they've received. This observation highlights the significance of building trust and delivering quality service from the early stages of a customer's relationship with the company. Ensuring positive experiences during the initial stages of tenure can play a pivotal role in keeping customers and reducing churn rates

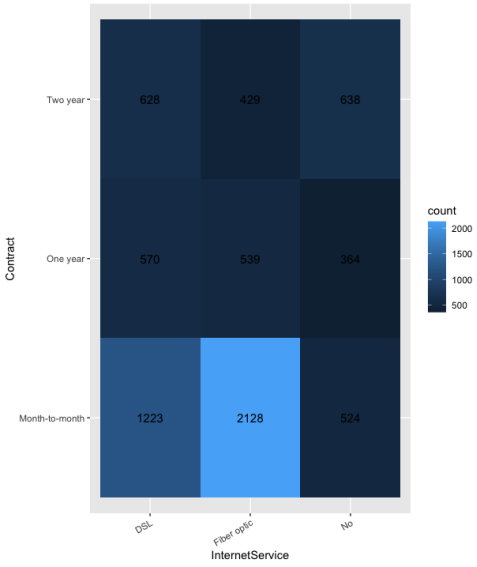

Figure 5.9: Heatmap.

Decoding Churn Trends: The Intersection of Internet Service, Contracts, and Affordability

After analyzing this graph we can observe a significant trend that the majority of customers are choosing for internet services from the company, with a preference for fiber optic service. A substantial number of customers are choosing to pay on a monthly basis, rather than committing to one or two-year contracts. As we previously analyze the internet service and contract type versus churn rates. We got to know customers who opt for fiber optic service or month-to-month contracts are more likely to leave the company. These two categories fiber optic and month-to-month contracts, also have a higher customer count. The potential reasons for this churn trend could be more. The affordability of fiber optic service might be an issue which is causing customers to discontinue the service. Additionally, the service quality might not match with price they are paying, leading to dissatisfaction among customers. The company should consider a two-pronged approach which is improving the quality and value proposition of fiber optic services could entice customers to stay and evaluating the pricing structure and exploring options for making the service more affordable might also contribute to retaining customers who are currently churning due to cost concerns. This approach of addressing both affordability and quality can potentially help reduce churn rates and increase overall customer satisfaction.

5.3 Hypothesis test

5.3.1 Chi square test

Considering the overall occurrences of each category, the Chi-square test analyses the arrangement of observations to help us decide whether particular combinations of categories appear more frequently than what would happen at random. This test looks for relationships between different variables. The Chi-square test [15] is used when the categories are separate and not a part of a continuous scale, as compared to using a correlation coefficient, which is appropriate for continuous data.

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.1}$$

Does age, particularly whether a customer is a senior citizen or not, have an impact on the likelihood of customers churning from the company's services?

Pearson's Chi-squared test with Yates' continuity correction
Data: df$SeniorCitizen and df$Churn
X-squared = 159.43, df = 1, p-value < 2.2e-16

Based on the chi-square test results we are going understand whether age plays a important role in customers deciding to leave the company's services. This research question was driven by our initial visualization that showed that senior citizens might encounter difficulties with the services provided by the company. This was driven by the belief that the service might not be well-suited or understood by senior citizens. The test results show that there is a significant association between age and customer churn. The calculated X-squared value is 159.43 with 1 degree of freedom and the p-value is extremely small (less than 2.2e-16). The test results provide strong evidence that there is a connection between being a senior citizen and the likelihood of leaving the company's services. This suggests that there might be challenges for senior citizens in using the services which is leading to a high chance of the customer deciding to leave. It's important for the company to consider this finding and potentially adjust their services to the needs and preferences of senior citizens. By improving these challenges the company could work towards improving customer retention among this particular

demographic.

Does the choice of payment method significantly impact customer churn in service-based companies?

Pearson's Chi-squared test
Data: df$PaymentMethod and df$Churn
X-squared = 648.14, df = 3, p-value < 2.2e-16

As in above visualization we saw plot with Churn and payment method, the visual revealed that a significant number of customers who left the company were using the online check payment system. This strongly suggested that this particular payment method might be a main reason for driving customers to discontinue the company's services. To validate this observation, I conducted a chi-square test. The test results showed that there is indeed a meaningful connection between the payment method and customer churn. The calculated X-squared value was 648.14 with 3 degrees of freedom, and the p-value was extremely low (less than 2.2e-16). These results confirm that the choice of payment method significantly influences customer churn. This outcome underscores the need for the company to address the issues with the online check payment method. By improving this method or offering alternative payment options, the company can work towards reducing customer churn and enhancing overall customer satisfaction.

Are customers with higher monthly expenses more or less likely to churn compared to those with lower expenses?

Variable	df	Sum Sq	Mean Sq	F value	Pr(>F)
Churn	1	238374	238374	273.5	$< 2 \times 10^{-16}$ ***
Residuals	7041	6137530	872		

Table 5.1: ANOVA Table for Churn Analysis

A study using an ANOVA test was carried out in the context of understanding the potential relationship between monthly fees and customer churn. The objective was to determine whether changes in monthly fees have a major effect on the chance that

consumers will leave the business. An ANOVA test was performed with the categorical variable "Churn" as the focal point as montly expense was continuous and churn was categorical. The calculated p-value is significantly less than the standard significance level of 0.05 (2e-16). As a result, it appears that the null hypothesis, which states that there is no meaningful connection between churn status and monthly charges, can be rejected. Consequently, there is strong statistical support for the claim that monthly fees do affect client churn.

5.4 Logistic Regression

By looking at the name of this algorithm [16] it sounds like it is related to regression problem, while itâs a powerful tool for solving classification problems. When we have binary class classification problem it helps us to detail the features belongs to which class. Its highly used in various industries due to its effectiveness and simplicity. It's a go-to method for classification tasks and is particularly useful when the classes are linearly separable, meaning they can be separated by a straight line. This model can also handle the multiclass problem where there is more than two class presents in the target variable to classify.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}}$$

$P(Y = 1|X)$ is the probability of the dependent variable Y being 1 given the values of the independ

e is the base of the natural logarithm.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients that determine the relationship between the independent variabl

$X_1, X_2, \ldots, X_p$ are the independent variables.

5.5 Decision tree

Decision tree algorithm can be used for both classification and regression task. It is programmed to do partition of the data continuously into subset of data using the

features in the data. Nodes are used to show the partition of the trees based on various decision paths. The decision tree grows by asking questions about the input feature and based on the answer it decides the outcome. The final decision or outcome is represented by the leaf node. This algorithm can be useful for predicting both type of data like numerical and categorical, as they can handle various types of attributes. Decision tree is easy to explain someone as there is no mathematics involve in the process. The process of constructing a decision tree involves selecting the most informative features at each node to split the data into subsets that are as homogeneous as possible with respect to the target variable. Various algorithms exist for constructing decision trees, including the widely used ID3, C4.5, CART, and Random Forest algorithms [16].

5.6 Ensemble Methods

In machine learning ensemble technique is most powerful method to increase the accuracy and robustness of the algorithms. It combines the power of multiple models to make prediction about the data rather than relying on one single model. It's like having a group of experts with diverse viewpoints collaborating to arrive at a collective decision. By using different technique or subset of the original data multiple models are trained to make a precise prediction of the output variable. The individual models are called weak learners. By combining their predictions, ensemble methods often outperform individual models by capturing different aspects of the data and reducing the risk of overfitting [16].

5.6.1 Bagging Technique

This technique involves training multiple models on different kind of data of the original data and with the help of combining power of all algorithms we use them for making prediction. The prediction that is made by all the models is then get selected by a majority vote (classification) or by averaging (regression) to reach the final decision [16].

Random Forest

As we distribute different subsets of the original data to the decision tree models, the distinctive aspect lies between them. This distribution of subsets of data ensures the uniqueness of each decision tree as well as their independence from previous decisions. Random forest uses multiple decision trees to collectively classify and process the information. Random Forest introduces upper boundaries based on two important variables to control adaptation error and achieve balanced performance. The approach aims to achieve a balance between each classifier's power and the degree of correlation between them, which is related to the raw margin of predictions [16].

5.6.2 Boosting

Consider Boosting [16] as a group of students who try to improve on each try. Boosting helps learners in improving by learning from their mistakes, unlike bagging, where we depend on randomness. Think about managing a band of musicians. They speak about what went wrong and how to make the next performance better after each performance. This is how Boosting's students interact with one another to enhance one other's accuracy. We begin with a large amount of data and split it into three sections. Let's call the first student, d1, and have them study with the first section. The second part is our challenge to d1. It makes some errors and some good decisions, and d2 the second learner learns from these mistakes. The third part now connects to d1 and d2. When d1 and d2 are at odds, those situations act as practise for d3. When it's time for the test, the three students, D1, D2, and D3, work together. If d1 and d2 agree, we use their example. If they disagree, however, we respect d3's opinion. It's like having a panel of three experts who deliberate and reach their agreement.

XGBoost

The powerful gradient boosting library XGBoost [17] was created to train machine learning models quickly and effectively. Using the advantages of several weak models to produce a stronger prediction, it works as an ensemble learning strategy. The word "XGBoost" stands for "Extreme Gradient Boosting," and it has become quite well known for its ability to handle large datasets and produce remarkable performance in a variety

of machine learning tasks, including classification and regression.

The ability of XGBoost to handle missing data values is an important characteristic. Due to this property, the method may operate on real-world datasets with missing data without the need for costly preprocessing. Additionally, XGBoost has built-in parallel processing capabilities, which enables it to handle sizable datasets while keeping appropriate training speeds.

5.6.3 Artificial Neural Network

In our brains, decisions are made using natural neural networks made up of tiny units called neurons. These neurons consist of parts like dendrites that receive information, a cell body that processes it, and an axon that sends out the decision as an electrical signal through synapses.

Similarly, in artificial neural networks, which mimic the brain's functioning, inputs like X1, X2... Xn are received by each neuron. These inputs are added up with certain weights and a bias, and then an activation function is applied for decision-making. The output is based on this process across the network of neurons.

$$O_k = F\left(\sum_{i=1}^{n} W_i X_i + b_j\right) \tag{5.3}$$

A neural network consists of layers, including input, hidden, and output layers. During training, errors between the actual and predicted outputs are continually reduced by adjusting the weights of inputs.

$$E_i = A - \hat{A} \quad (2)$$

where E_i is the Error between actual and predicted outputs

$A =$ Actual Output

$\hat{A} =$ Predicted/Modeled Output

There are different activation functions that help neural networks perform tasks. Sigmoid is one that outputs values between 0 and 1, but it has limitations. Another function called Hyperbolic Tangent (tanh) is better for optimization. Rectified Linear Units (ReLU) is another popular function, especially for hidden layers, and it overcomes some limitations of others. There's also Leaky ReLU, a modified version of ReLU that addresses certain issues during training.

Relu Function

For neural networks [18], the ReLU (Rectified Linear Activation) function functions like a smart calculator. It operates in an easy way, if the number you provide it is positive, it keeps that value. But it returns 0 if the number is negative. The network can learn more quickly and avoid many difficult maths problems because to this intelligence. Using the same "rule" for positive values makes ReLU incredibly quick to calculate, and it helps the network keep simple and ordered. Additionally, ReLU can say "zero" to some numbers, which is excellent for simplifying the network.

$$f(x) = \max(0, x) \tag{5.4}$$

Here, $f(x)$ is the output of the ReLU function for the input x. If x is a positive number, the function returns x itself. If x is negative, the function outputs zero. This makes the ReLU function a piecewise linear function that introduces non-linearity in the neural network while being computationally efficient.

Sigmoid Function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5.5}$$

Equation illustrates how the sigmoid activation function [19] converts input values from the negative infinity to positive infinity range into a limited range between 0 and 1. Figure 1 shows how it presents an even derivative and adds nonlinearity to the neural network. Since the sigmoid's output range is restricted to [0, 1], the output of each unit is limited within this range, which causes the output of each unit to be compressed, the gradient decreases, especially in deeper networks. The network optimisation can be difficult because of this diminishing gradient, and this difficulty increases beyond a certain point.

Binary cross entropy

The formula for binary cross entropy, also known as log loss, is as follows:

$$H(y, \hat{y}) = -\left(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right) \tag{5.6}$$

Where:

$H(y, \hat{y})$ represents the cross entropy between the true binary label y and the predicted probability $\hat{y}$.

y is the true binary label, which can be either 0 or 1.

$\hat{y}$ is the predicted probability of the positive class (class 1) provided by the model.

A popular loss function [20] in machine learning, especially for tasks involving binary classification, is known as binary cross entropy, sometimes known as binary log loss or binary cross-entropy loss. Its goal is to calculate the difference between a dataset's actual binary labels and the expected probability distribution.

The projected probabilities for each class are compared to the actual class outputs, which are either 0 or 1, in the setting of binary cross entropy. The function calculates a score that penalises the probabilities according to how far they deviate from the predicted values. Basically, this assesses how closely or far the probabilities are predicted from the actual values.

Adam Optimizer

The Adam optimizer [21], frequently referred to as the Adaptive Moment Estimation optimizer, is a well-liked deep learning method for improving neural network training. It is an enhancement to the stochastic gradient descent (SGD) technique and is used for precisely adjusting a neural network's weights during training.

The term "Adam" is a term for "adaptive moment estimation," highlighting its unique capacity to dynamically adjust the learning rate for each weight inside the network. The Adam optimizer generates individual learning rates for each weight by taking into account historical gradients and their squared values, compared to the fixed learning rate used by SGD during training. Due to this flexibility, weight updates during training are more efficient, improving learning and convergence.

5.7 Evaluation metrics

5.7.1 Accurcay

The ratio of a model's correct predictions to its total number of predictions is calculated by the accuracy metric, which is frequently used in machine learning. It quantifies the

performance of a model by calculating the percentage of instances that were properly classified out of all the instances it examined [22].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.7}$$

5.7.2 Recall (True Positive Rate or Sensitivity)

Recall [22] predicts the percentage of cases that are actually positive and that the algorithm accurately classifies as positive. It evaluates the model's capacity for avoiding missing positive cases. A high sensitivity means the model is successful in capturing the majority of the positive cases

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.8}$$

5.7.3 Precision

Precision [22] measures the percentage of cases that the algorithm classifies as positive and are actually positive. It assesses the model's ability to correctly identify positive cases without falsely labeling negative cases as positive. A high precision indicates that the model has a low rate of false positives and is reliable in identifying true positive cases.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.9}$$

5.7.4 F1-score

The F1 Score takes into account both false positives (misclassifications of negative instances as positive) and false negatives (misclassifications of positive instances as negative). It balances the trade-off between precision (accuracy of positive predictions) and recall (sensitivity to true positive instances). A greater value of the F1 Score denotes better classification performance, and it ranges from 0 to 1. It's particularly useful when trying to find a balance between correctly recognising both good and negative examples [22].

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.10}$$

5.7.5 k-Fold Cross-Validation

Cross-validation [22] is like evaluating students' performance by giving them various sets of questions to respond to. We give them several tests with different questions instead of to simply one. It allows us to evaluate how well they actually understand the subject. Cross-validation in machine learning allows us to test a model's performance using several subsets of the data. To make sure the model is truly effective and not just lucky with one particular set of data it is like presenting it with a variety of challenges.

5.7.6 Precision-recall trade-off

This visual representation is used to evaluate the trade-off between accuracy and recall for various threshold settings in binary classification problems. Recall indicates the ability of the classifier to identify any relevant instances, whereas precision reflects the accuracy of positive predictions. While the x-axis represents recall, the y-axis represents accuracy. The graph [22], where recall indicates the ability of the classifier to identify every relevant occurrence and precision reflects the accuracy of positive predictions, is especially helpful when the dataset is unbalanced. Depending on the precision and recall levels you want for your classifier, the curve could help you in selecting a suitable threshold.

$$\text{Precision-Recall Trade-off} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.11}$$

5.8 Sampling techniques

To solve the problem of an unbalanced data distribution, sampling techniques are used in the context of data analysis and machine learning. It can result in biassed model performance when one class (in this case, positive examples) in the dataset is greatly neglected when compared to another class (in this case, negative instances). To ensure that the model can effectively learn from both positive and negative cases, sampling methods try to balance the distribution of classes in the dataset. These methods involve adding new data points, changing those that already exist, or changing how training data is used.

5.8.1 Under-sampling

With this strategy [23], we reduce the size of the majority class to that of the minority class by eliminating some instances of the majority class. Machine learning is particularly effective at extracting patterns from data when the data is almost balanced. By doing this, we guarantee that the dataset is free of bias.

5.8.2 Over-sampling

It is a method through which we make the minority class more prevalent until it is about the same size as the majority class. This is done by duplicate the minority class instance to make sure the model doesn't miss any crucial data for prediction. The technique [23] is helpful for dealing with skewed data because SVM and ANN find it difficult to handle it.

5.8.3 SMOTE

It adds instances from the minority class until the minority and majority classes are equal. Its stands for Synthetic Minority Oversampling Technique [24]. Instead of duplicating the minority class values to increase them, data augmentation is used to produce new synthetic data with plausible values that are near to the minority class's "feature space."

CHAPTER 6

Results

Accurate predictive modelling is difficult to achieve because of the dataset's inherent imbalance between the minority class (Class 1) and the majority class (Class 0). Many types of sampling methods, including under-sampling, over-sampling, and SMOTE (Synthetic Minority Over-sampling Technique), have been used to deal with this problem successfully. To determine how well five different models perform in fixing the class imbalance, their performance was evaluated as well under each of these sampling techniques.

My objective is to identify which customers will leave, thus I need a model that can identify as many true positives as possible and predict less false positive. This will help the business identify the reasons why customers are leaving so they can take action to prevent it and increase customer retention. Since we want to know how well the model predicts future positive values and also when the data set is imbalanced, traditional metrics like accuracy might not accurately represent the performance of the models. Therefore, an in-depth evaluation of the models' capacity to accurately classify instances of the minority class(true positive) using recall, precision, and F1-score has been performed.

The precision-recall graph has been used in addition to these measures to illustrate the trade-off between precision and recall, corresponding with the complexity of imbalanced classification problems. This graph tells us how much balance between precision and recall the model is having as we want the model to predict more true positive (recall) value and less less false positive value (precision). So the balance

between two metrics is necessary and this plot help us to understand whether model is giving balance result between precision and recall. Along with sampling technique generated data I also evaluated the models' performance against a Normal data which were having imbalance data to see is there different in the performance of the model between imbalance data and balance data. Both conventional algorithms like logistic regression and decision trees as well as advanced approaches like ensemble methods like random forests and XGBoost classifiers are included in the models' toolkit. Moreover, an Artificial Neural Network (ANN) design is used to take advantage of deep learning's powers. This various sampling techniques will allows us to see how each approach affects the classification performance of the models.

6.1 Performance Analysis of Models under Different Sampling Technique

6.1.1 Normal data

I found important details about how different models work when I looked at the original dataset in its original form.The Table 6.1 shows the perfomance of five different models on normal data. A common technique, logistic regression, showed an acceptable accuracy rating of about 81.5%. But since the dataset contains more of one kind of object than the other, I suggest using caution when relying only on this number. This can lead to inaccurate accuracy. I must look at other important indicators in order to truly understand how effectively the models are performing. Recall is an essential topic to understand because it evaluates how well models can find instances of the positive class, or less common category.

When using Logistic Regression its Recall score shows that it correctly identified about 53.8% of the true positive instances. Precision is another important factor to consider because it sheds light on how well the models predict instances that belong to the positive class. To be more specific, its Precision in the context of Logistic Regression was approximately 70%, suggesting that it correctly predicted true positive instances. I included the F1 Score, a metric that strikes a balance between Recall and Precision,

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.815472	0.538874	0.695502	0.607251
Random Forest	0.798439	0.479893	0.665428	0.557632
XGBoost	0.7956	0.506702	0.645051	0.567568
Decision Tree	0.71824	0.474531	0.468254	0.471372
ANN	0.749468	0.109920	0.661290	0.188506

Table 6.1: Performance Metrics of Different Models on Imbalance data

for an in-depth view. The F1 Score demonstrated an overall effectiveness of 60.7% in the context of logistic regression. I first observed the excellent accuracy rates of about 79.8% and 79.5% when I investigated into advanced techniques like Random Forest and XGBoost. On further examination, it became clear that these models had some issues correctly identifying cases of the minority group. For instance, XGBoost captured 50.4% of true positive cases, compared to Random Forest's 47.7%. Random Forest showed a higher precision rate than XGBoost at 66.5%, in terms of precision. I looked at their performance more thoroughly with the F1 Scores, which offer an in-depth analysis by considering both Recall and Precision.

These showed that XGBoost had an F1 Score of 56.7% compared to Random Forest's 55.5%. The Decision Tree model was used for analysis among the different techniques I examined. In terms of accuracy, it was about 72%. However, its Recall score was determined to be about 47.4% and its Precision, indicating its accuracy in predicting the rarer category, was around 4686% when it came to detecting occurrences of the rarer category. The F1 Score was 47.9%. These numbers highlighted the advantages and disadvantages of the Decision Tree model in the dataset situation. Compared to Logistic Regression and Random Forest, it has comparatively lesser precision and recall. In my final investigation, I used an Artificial Neural Network (ANN) to study deep learning. Its Precision score of 66% showed that it could reasonably forecast the rarer category. The fact that its Recall was lower, at about 10%, indicates that it had trouble accurately identifying a significant number of these situations. The final F1 Score was 1885%. This revealed how difficult it is to produce a balanced performance on datasets like these.

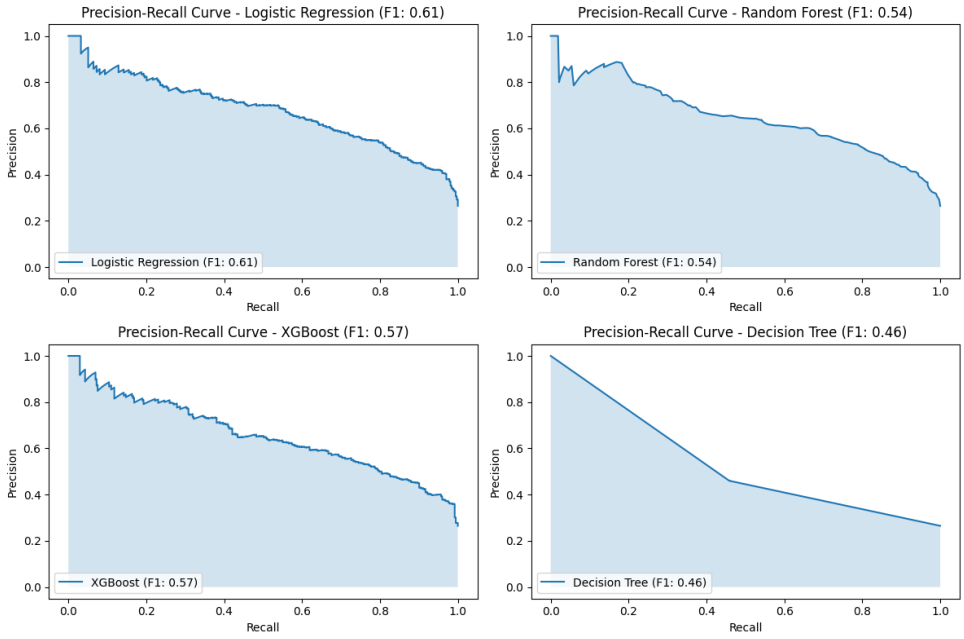

Figure 6.1: Machine learning models precision recall graph for Imbalance data

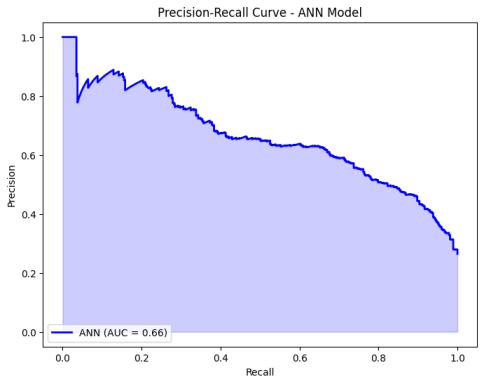

Figure 6.2: ANN model performance

I made an interesting discovery while examining the performance of different models, as depicted in Figure 6.1 and Figure 6.2, along with their corresponding precision-recall graphs. When we aim for more accurate positive predictions and fewer incorrect positive predictions, the logistic regression model stands out with high precision and recall values. Its graph aligns closely with the upper-right corner, which is a positive sign. Next, the random forest and XGBoost models strike a good balance between recall and precision. However, the decision trees and artificial neural networks struggle in correctly predicting positive cases and minimizing false positive cases. These two models exhibit the weakest balance between precision and recall. The artificial neural network, particularly stands out as it has the lowest recall value, as seen in the graph.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.770053	0.785908	0.75718	0.771277
Random Forest	0.75	0.769648	0.735751	0.752318
XGBoost	0.760695	0.777778	0.747396	0.762284
Decision Tree	0.713904	0.710027	0.710027	0.710027
ANN	0.573529	0.986450	0.536873	0.695320

Table 6.2: Performance Metrics of Different Models on undersampling data

6.1.2 Under-sampling

I evaluated the models again after using the under-sampling technique to fix the class imbalance. I compared these fresh findings with my earlier findings as you can you see in table 6.2. The accuracy for Logistic Regression decreased a little to about 77.1%. Recall, however, saw a remarkable change; it rose significantly to around 78.0%. This shows that the model became more adept at identifying the rarer category. However, the precision, which measures how precise the model is in classifying something as being rarer, remained at 76%. The model is balanced in how it handles the dataset's imbalance, as shown by the final F1 Score, which considers both recall and precision, which came out at 77.1%. Similar events happened with the Random Forest and XGBoost models. Their relative levels of accuracy dropped somewhat to roughly 75% and 76%. However, recall improved, which meant that they were better able to identify instances of the rarer categoryâroughly 76.9% for Random Forest and 77% for XGBoost.

The precision also got increased to roughly 73.5% and 74.7%. These modifications led to F1 Scores for Random Forest and XGBoost of roughly 75.2% and 76.2%, respectively. The model using a decision tree improved the most. After under sampling, its accuracy remained same to roughly 71.3%, which was expected. However, the recall improved, reaching about 71%, indicating that it became more successful at identifying examples of the more True positive. The F1 Score came in at 71.0%, while the precision got changed to approximately 71%. The Artificial Neural Network's (ANN) precision reduce somewhat to 53.6%. Recall also increased significantly, reaching around 98.8%, indicating that it became more successful at locating examples of the uncommon class. An F1 Score of roughly 69.5% which is higherâwas achieved as a result of these adjustments to recall and precision. The improvements resulting from the modifications of the data for the

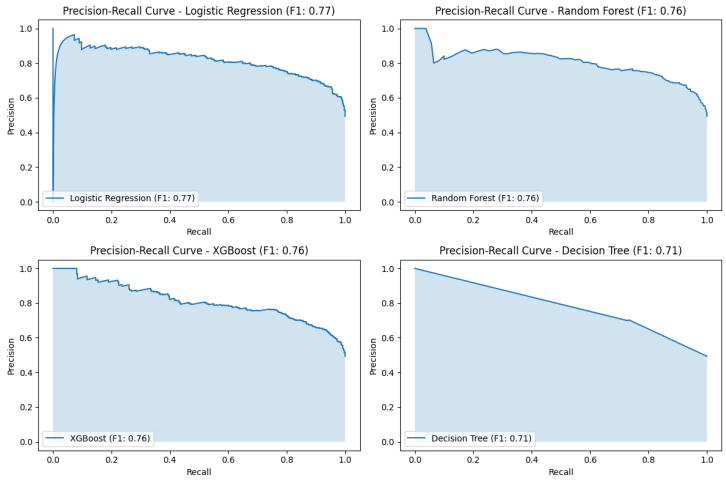

Figure 6.3: Machine learning models precision recall graph for under-sample data

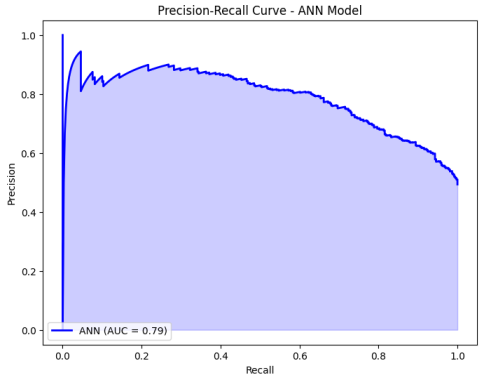

Figure 6.4: ANN model performance in under-sampling

precision-recall graphs are quite evident, as shown in figures 6.3 and 6.4. These graphs effectively showcase how under-sampling techniques influenced the performance of the models.

One noteworthy observation is the considerable improvement in the precision-recall graph of the Artificial Neural Network (ANN). This indicates that the ANN achieved a more balanced performance, achieving good precision and recall scores. Among the models, the Logistic Regression displayed strong performance in terms of precision, recall, and F1 Score. This was supported by the graph, which illustrated its capability to predict more true positives and fewer false positives. Similar to the ANN, the graph for the decision tree also displayed noticeable enhancement when compared to the imbalanced data. Additionally, both Random Forest and XGBoost models demonstrated improvements in their respective precision-recall graphs, further solidifying their performance enhancements.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.783092	0.797903	0.77933	0.788507
Random Forest	0.9	0.951382	0.864818	0.906037
XGBoost	0.8657	0.932316	0.825316	0.87556
Decision Tree	0.876329	0.952336	0.829046	0.886424
ANN	0.757488	0.745472	0.768928	0.757018

Table 6.3: Performance Metrics of Different Models on over-sampling data

6.1.3 Over-sampling

I saw major improvements in the models' performance, especially when handling the challenge involving class imbalance, by employing the over-sampling strategy. Important models like the Decision Tree, Logistic Regression, Random Forest, XGBoost are showing obvious gains in many metrics as it is mention in table 6.3. The accuracy of Logistic Regression improved to around 78.4%, and Random Forest performed even better, with an amazing 90% accuracy. Notably, Random Forest is much better at locating events of the rarer category, as shown by a rise in recall score to 95.1%. The level of accuracy for XGBoost and the Decision Tree increased as well, reaching 86.5% and 95%, respectively.

The ability of the Decision Tree, XGBoost, and Logistic Regression to find examples of the more uncommon class all increased. Their F1 Scores increased because of the increased precision and recall. Although the ANN model's recall dropped to 74.8%, its precision rose to 76.6%. It suggests a rise in its ability to reduce false positive predictions of the rarer class. This increased recall and precision led to a significantly higher F1 Score, demonstrating the value of over-sampling.

In figures 6.5 and 6.6, the precision-recall tradeoff for five different models using the over-sampling technique is depicted. These graphs help us understand how well the models balance precision and recall. Among them, Random Forest consistently showcases the best results between precision and recall. This aligns with our earlier evaluation metrics where Random Forest achieved the highest scores. The graph confirmed its superior performance. On the other hand, ANN's poor results are seen in its

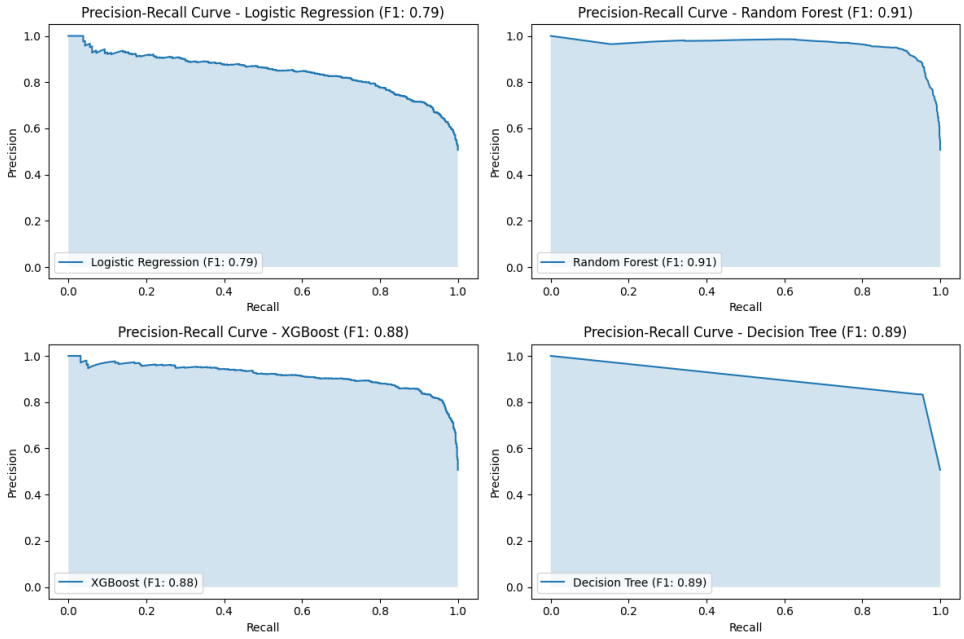

Figure 6.5: Machine learning models precision recall graph for over-sample data

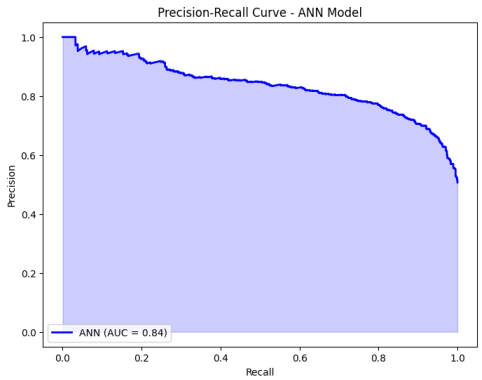

Figure 6.6: ANN model performance in over-sampling

graph, indicating lower precision and recall. Logistic Regression's performance slightly declined, evident from its lower precision and recall scores on the graph. Decision Tree and XGBoost, though not as optimal as Random Forest, still demonstrate a commendable trade-off between precision and recall, making them the second-best performing models.

Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.839614	0.861773	0.828598	0.84486
Random Forest	0.847826	0.854147	0.846881	0.850498
XGBoost	0.853623	0.861773	0.851224	0.856466
Decision Tree	0.803865	0.797903	0.811833	0.804808
ANN	0.758454	0.593899	0.893831	0.713631

Table 6.4: Performance Metrics of Different Models on smote data

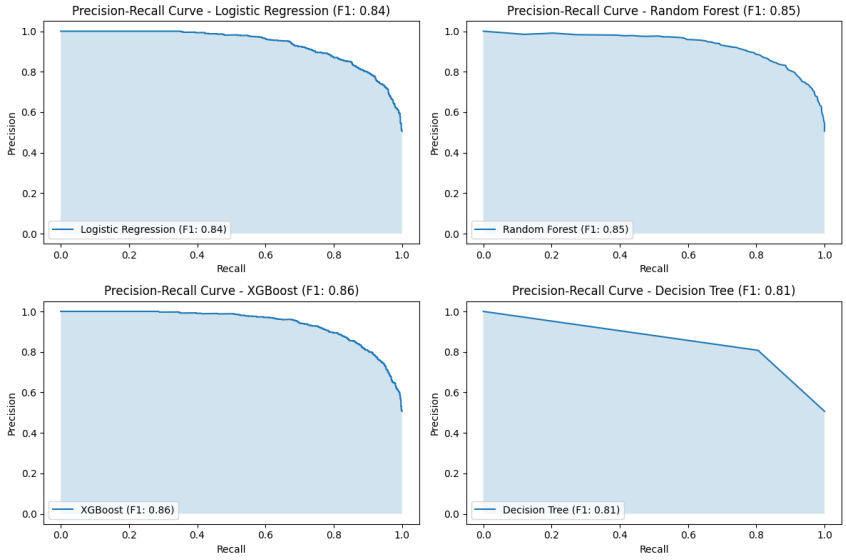

Figure 6.7: Machine learning models precision recall graph for smote-sample data

6.1.4 Smote-sampling

When we compared the results of using SMOTE-sampling to over-sampling, as shown in Table 6.4, we noticed a decrease in the performance of most models. For instance, Random Forest, which initially had the highest accuracy and recall scores, saw a decline to 84% accuracy and 85% recall. Similarly, XGBoost and Decision Tree experienced slight drops in performance, with recall rates of 86% and 79%, and precision rates of 85% and 81%, respectively. The performance of ANN worsened significantly, struggling to predict true positives, as indicated by its 59% recall score. Despite achieving a good precision score, ANN's ability to capture positive cases was compromised. On the other hand, Logistic Regression displayed good improvements when compared to over-sampling. It demonstrated improve recall, precision, accuracy, and F1 scores.

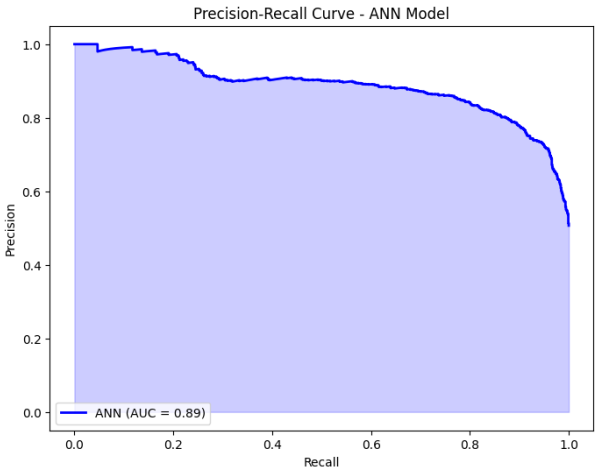

Figure 6.8: ANN model performance in smote

As observed in the figures 6.7 and 6.8, both Random Forest and Logistic Regression models exhibit a favorable trade-off between recall and precision. Particularly noteworthy is the improvement seen in the logistic regression graph when compared to the over-sampling technique. In this technique XGBoost and ANN also demonstrate a respectable balance between precision and recall, indicating their effectiveness in capturing positive cases while minimizing false positives. However, the decision tree model's graph displays poor performance, implying its limitations in accurately identifying true positive instances.

6.1.5 Cross validation

Random Forest outperformed the other models I examined, especially when I balanced the data via oversampling. I applied cross-validation to ensure the accuracy of my results. This means I split the data into five different parts and tested the model on each one. Surprisingly, the accuracy was around 89% regardless of the part I used. The average of these scores was also close to 89%. This consistent performance across different parts of the data confirms that my model is reliable and works well. My cross-validation study proved that, independent of the particular data used, the Random Forest model consistently performs with excellent accuracy. This tells that the model's capacity to predict churn or not churn is good.

$$\text{Fold 1:} \quad \text{Accuracy} = 0.9010$$

$$\text{Fold 2:} \quad \text{Accuracy} = 0.8957$$

$$\text{Fold 3:} \quad \text{Accuracy} = 0.8792$$

$$\text{Fold 4:} \quad \text{Accuracy} = 0.8922$$

$$\text{Fold 5:} \quad \text{Accuracy} = 0.8980$$

$$\text{Average Accuracy} = 0.8932$$

6.1.6 Comparative Analysis

Since unbalanced data was the problem at hand while I was working on this kind of task, I learned how models function with various types of data. Even though I was getting decent accuracy scores while I kept my data unbalanced, there was no apparent difference in my model's performance. Because of this, I should now pay more attention to how my model is performing when making rare case predictions (true positives). Then, when I checked additional metrics like precision, recall, and f1 score, I discovered that these models were doing well on the majority class, which is 0 but were very poor on predicting 1 as the instances of 1 were quite rare.

However, the performance of every model greatly increases when I employ different sampling techniques, with smote and over-sampling technique doing significantly better than other techniques, Logistic regression, random forest, XGBoost, and ANN were among the models that performed well. Along with their accuracy, other measures were also improved. I was receiving an accuracy score of 90% in random forest for the over-sampling technique, with precision and recall of 86% and 95%, respectively. This was the greatest result I got in random forest when compared to other techniques, and it is the top score among other techniques. As my aim is to create models that reliably predicted True positive as I want to focus on which customers may churn the firm, having a strong recall will be extremely helpful, and for this random forest in over-sampling approach is producing very nice results. Logistic regression does well in the smote method as well, with a recall score of 87%. The random forest's performance was supported by the precision-recall trade-off graph, which showed that it was effective in predicting most of the instances since the graph was in the top right corner, which is

positive because it indicates that this model is working well.

CHAPTER 7

Conclusions

As I worked on this project, I looked through a number of research papers on customer churn in the telecom industry and discovered a number of solutions. I started out by observing the data, which included both continuous and categorical variables. I discovered the connections between various variables and the churn rate using techniques like bar plots, box plots, density plots, pie charts, and correlation plots. By comparing categorical characteristics with the objective variable "churn," I was able to identify important factors that the company could change to reduce churn rates. I noticed that senior adults had trouble using complicated features and services, which showed the need for user-friendly options. Additionally, customers who chose fiber-optic internet service were more likely to leave, and the monthly payment schedule presented difficulties. I discovered through the analysis of continuous variables that lowering the price of streaming TV and films could be useful in keeping clients. Shorter tenure was associated with greater churn rates, and the density plot showed high monthly charges as a primary churn cause. I used hypothesis tests like the chi-square and ANOVA tests to confirm significant associations in my visual observations.

The problem of processing categorical data for machine learning modelsâwhich by default require numerical inputs was also issue I tackled. I looked into modern methods to address this problem because the data was imbalanced and had fewer instances of churn. I used a variety of sampling techniques, drawing on the literature, to give the machine learning models a variety of datasets. My main objective was to reliably anticipate genuine positives and minimise false negatives, allowing the organisation to

identify future customer departures and use successful customer retention actions. I evaluated precision, recall, and F1 score because I was aware of the disadvantages of depending just on accuracy. To account for the imbalance in the data I also confirmed these measures with precision-recall plots.

The outcomes showed that Random Forest performed well under the over-sampling approach, reaching an astounding recall score of 95%. To prevent false positives, it's crucial to find a balance between recall and accuracy. The precision-recall graph showed this model to have remarkable precision and recall, demonstrating its superior performance. The experience also showed that various approaches might work with various models.

The results of this initiative will have a big impact on the telecom sector and beyond. This analysis provides insightful information by thoroughly analysing the patterns of client churn, which can be extremely beneficial to telecoms organisations. These businesses can be guided in developing plans to address these problems by learning what causes customers to quit, such as excessive costs and poor service. With this information, they may modify their offerings to better suit client needs, lowering churn and increasing customer satisfaction. Additionally, the methods used in this study, such as improved sampling and predictive modelling, can serve as a template for other businesses dealing with similar difficulties in handling imbalanced data and predicting consumer behaviour. The lessons learnt here may help in better planning and decision-making across a variety of industries, leading to better tactics that prioritise the needs of the customer.

7.1 Future Scope

In this project, dealing with significantly imbalance data was difficult because most customers stayed with the business. I used numerous sample strategies to address this. In the future, even more modern methods might be taken into account. Future methods to manage imbalanced data more effectively include ones like ADASYN, Borderline-SMOTE, and MSMOTE. These methods may be helpful in building more balanced datasets for training models to improve prediction accuracy.

Bibliography

[1] Customer Churn Prediction for Broadband Internet Services. https://link.springer.com/chapter/10.1007/978-3-642-03730-6$_1$9.

[2] Bingquan Huang, Brian Buckley, and T-M Kechadi. Multi-objective feature selection by using nsga-ii for customer churn prediction in telecommunications. *Expert Systems with Applications*, 37(5):3638–3646, 2010.

[3] https://www.sciencedirect.com/science/article/pii/S0957417411011353. Customer churn prediction in telecommunications.

[4] https://ieeexplore.ieee.org/abstract/document/8365230. A comparative study of customer churn prediction in telecom industry using ensemble based classifiers.

[5] Abinash Mishra and U. Srinivasulu Reddy. A novel approach for churn prediction using deep learning. pages 1–4, 2017.

[6] https://www.sciencedirect.com/science/article/pii/S0957417408002121. Handling class imbalance in customer churn prediction.

[7] Kristof Coussement, Dries F Benoit, and Dirk Van den Poel. Improved marketing decision making in a customer churn prediction context using generalized additive models. *Expert systems with Applications*, 37(3):2132–2143, 2010.

[8] Dirk Van den Poel and Bart Lariviere. Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, 157(1):196–217, 2004.

[9] Yasser Khan, Shahryar Shafiq, Abid Naeem, Sheeraz Ahmed, Nadeem Safwan, and Sabir Hussain. Customers churn prediction using artificial neural networks (ann) in

telecom industry. *International journal of advanced computer science and applications*, 10(9), 2019.

[10] https://ieeexplore.ieee.org/abstract/document/7707454. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study.

[11] Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, and Sajid Anwar. Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, 94:290–301, 2019.

[12] Adnan Amin, Faisal Rahim, Imtiaz Ali, Changez Khan, and Sajid Anwar. A comparison of two oversampling techniques (smote vs mtdf) for handling class imbalance problem: A case study of customer churn prediction. In *New Contributions in Information Systems and Technologies: Volume 1*, pages 215–225. Springer, 2015.

[13] Chih-Fong Tsai and Yu-Hsin Lu. Data mining techniques in customer churn prediction. *Recent Patents on Computer Science*, 3(1):28–32, 2010.

[14] https://www.sciencedirect.com/science/article/pii/S0957417409004758. Customer churn prediction by hybrid neural networks.

[15] https://www.simplypsychology.org/chi square.html. Chi-square test.

[16] Saudi Arabia PRACTICAL MACHINE LEARNING FOR DATA ANALYSIS USING PYTHON AbdulhAmit SubASi Professor of Information Systems at Effat University, Jeddah. Practical machine learning for data analysis using python.

[17] XGBoost. https://www.geeksforgeeks.org/xgboost/.

[18] https://iq.opengenus.org/relu activation/. Relu activation function.

[19] https://ieeexplore.ieee.org/abstract/document/9108717. A review of activation function for artificial neural network.

[20] https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary classification/. Binary cross entropy.

[21] https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/h-adam-optimizer formula. Adam optimizer.

[22] Python book. MÃ¼ller, andreas c, and sarah guido. introduction to machine learning with python: A guide for data scientists. sebastopol: Oâreilly media, incorporated, 2016. print.

[23] https://www.numpyninja.com/post/under sampling. Sampling technique.

[24] https://practicaldatascience.co.uk/machine-learning/how-to-use-smote-for-imbalanced classification. Smote technique.