Customer Churn Prediction Model

1st Manal Rizwan Qureshi
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore, Pakistan
l200915@lhr.nu.edu.pk

2nd Mariyam Ali
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore, Pakistan
l202156@lhr.nu.edu.pk

3rd Aisha Muhammad Nawaz
Department of Computer Science
National University of Computer and Emerging Sciences
Lahore, Pakistan
l200921@lhr.nu.edu.pk

***Abstract*—In a world full of options, customers find it hard to stay loyal to a single organization. The significance of retaining customers for businesses is therefore very high which is why there has been an increase in the efforts made to retain customers and machine learning techniques have proved to be of great use. Adding to the vast research already done on the topic, this research provides a detailed literature review of past researches and provides result of experiments made using several models that include XGBoost Classifier, Logistic Regression, Random Forest Classifier and Decision Tree to predict whether a customer will churn or not. Results indicate that XGBoost Classifier is the most effective model with an accuracy of 0.80 while also showing a considerable trade-off between precision and recall for both churn and non-churn scenarios. Strong discriminative power is indicated by its ROC curve, which has an AUC of 0.84. Decision Tree model exhibits limits, indicating limited discriminative capabilities, with an accuracy of 0.72 and a lower AUC of 0.64.**

***Keywords*—Churn, Customer, Model, Machine Learning, XGBoost, Logistic Regression, Random Forest Classifier, Decision Tree**

I. Introduction

A. Background

With the growing number of competition in the telecommunication sector, organizations are looking for a variety of ways to reach new and retain old customers. Previously, the focus used to be mainly on attracting new customers but if the rate at which an business attracts new customers is roughly the same as the rate at which customers are leaving the business then there is little advantage of the efforts put and money spent on attracting new customers. This is especially troublesome if the business offers more subscription based services like ones in the telecommunication sector. This is not to say that efforts should not be made to attract new customers but just to clarify that it is equally, if not, more important to ensure that the old customers continue to use and enjoy the services of organizations in all sectors and not just the telecommunication sector.

B. Problem

In order to retain old customers it is necessary to know the factors that effect their decision to churn or not these can include price, location change, preferences change etc. Customer retention strategies can be costly if applied to all the customers , some even say that it can be 10 percent more costly than applying to just some customers. It is therefore more wise to apply such strategies (like discounts vouchers, special deals etc) to people who are more likely to churn but how do we know which customer is more likely to churn and how do we know that using a method less costly? This is where machine learning comes. The aim of this research paper is to dive deeper into the world of machine learning and find the best ways to predict whether a customer is likely to churn based on the data of customers who churned in the past and the ones who did not. This problem is that of classification and a machine learning model shall be built that can predict whether a customer will churn or not based on the features provided.

C. Project Overview

This project involves several phases. It starts with data collection, cleaning of raw data. Then comes the data visualization phase, to see patterns and trend in the data. Next phase is data pre-processing, in which data is normalized to to be prepared for the model. After this, comes feature selection that creates a sparse matrix of huge dimension that are used are training data for machine learning algorithm. Machine Learning Model has been trained and evaluated, and at the end results have been presented and evaluated using metrics that include Accuracy, Loss, Confusion Matrix, Precision, Recall, F1 Score and ROC/AUC.

II. Literature Review

A. Customer Churn Prediction by Hybrid Neural Networks

The research paper [12] looks into the prediction of telecom churn using hybrid neural network models, specifically self-organizing maps (SOM) and back-propagation artificial neural networks (ANN). Two hybrid models, ANN + ANN and SOM + ANN, are introduced, in which unrepresentative data is filtered away using data reduction in the first component. Both hybrid models outperform the single neural network baseline in terms of prediction accuracy and Type I and II error rates, according to experimental results, with the ANN

+ ANN model doing better overall. The study's strengths are found in its thorough assessment employing several testing sets, which provides a sophisticated grasp of the efficiency of the model. Fuzzy testing sets are used to evaluate the robustness of the model even more, and subsets are taken into account to improve the findings' generalizability.There was limited discussion on the interpretability of the models, potentially hindering practical implementation.

B. A Comparison of Machine Learning Techniques for Customer Churn Prediction

This research paper [13] examines the use of different machine learning classifiers for telecom customer churn prediction, such as Back-Propagation Networks (BPN), Support Vector Machines (SVM) with Polynomial (POLY) and Radial Basis Function (RBF) kernels, Decision Trees (DT-C5.0), Naïve Bayes (NB), and Logistic Regression (LR). The AdaBoost.M1 method is used in the study to improve classifier performance. The findings show that DT-C5.0 and BPN with fewer than 20 hidden neurons are the most successful classifiers, with SVM coming in second. Boosting significantly raises the F-measure and accuracy of the classifiers, especially SVM-POLY, it yielded an accuracy of 97%. The study emphasizes how important boosting strategies are for improving prediction accuracy. However, weaknesses include the limited exploration of boosting impact on NB and LR, and the absence of a detailed discussion on the interpretability and scalability aspects of the models in real-world deployment. Overall, the study contributes valuable insights into classifier performance for telecom churn prediction, emphasizing the efficacy of boosting techniques, but further research is warranted for a more comprehensive understanding.Nevertheless, shortcomings include the insufficient investigation of the boosting effect on NB and LR and the lack of a thorough examination of the interpretability and scalability of the models in practical application.

C. A Survey On Data Mining Techniques In Customer Churn Analysis For Telecom Industry

The application of data mining techniques in forecasting customer churn within the telecom business is thoroughly explored in the research piece [1]. The study examines several approaches, including covering algorithms, neural networks, statistical-based techniques, and decision trees, in order to address the difficulties and economic importance of client retention. The paper's strengths are found in its thorough discussion of the CRISP-DM model and its comparative study of various approaches. Nevertheless, shortcomings include a narrow emphasis on RULES family algorithms, an absence of specific results, and an insufficient explanation of boosting effects. Despite these limitations, the paper serves as a valuable resource for understanding data mining applications in customer churn analysis, providing insights into the strengths and weaknesses of each technique.

D. Customer Churn Analysis in Telecom Industry

This research article [5] employs Decision Trees (J-48) and Logistic Regression as data mining approaches to predict churn in the telecom industry. It highlights how difficult it is to deal with high churn rates and how important it is to have reliable prediction models. Using data from Orange, a French telecom firm, the paper proposes a unique paradigm for churn prediction. A thorough examination of the implementation, along with in-depth screenshots and performance assessments, is one of its strengths.The comparison shows that the decision tree performs substantially better in terms of accuracy than logistic regression, with an outstanding accuracy of 99.7%. The lack of a thorough investigation of hybrid classification methods, potential biases in the dataset selection process, and the minimal explanation of the findings' generalizability are among the study's drawbacks.

E. Customer Churn Prediction:A Survey

The authors examine popular prediction approaches like regression analysis, decision trees, statistical methods, and neural networks in this paper [8]. One of the paper's strengths is its thorough analysis of various prediction techniques and how they are used in the telecom sector. Nevertheless, the article does not specifically address client retention tactics and does not provide a thorough review of the performance of certain models, such as accuracy rates. While the research helps to identify churn trends, it would be beneficial to have more comprehensive evaluation metrics and useful consequences for companies that deal with customer attrition.

F. A Multi-Layer Perceptron Approach for Customer Churn Prediction

This research paper [7] suggests a prediction model to forecast customer attrition in a top Malaysian telecom provider by utilizing a multi-layer perceptron of artificial neural network architecture. The study's findings demonstrate that, with a prediction accuracy of 91.28 percent, the neural network technique beats statistical models in prediction hence may be a good substitute for statistical forecasting methods. The strengths of the paper are that it proposes a novel approach to predict customer churn using a multi-layer perceptron of artificial neural network architecture and this approach outperforms the statistical models in prediction tasks with a prediction accuracy of 91.28 percent. Moreover, The paper provides a detailed comparison of the proposed approach with multiple regression analysis and logistic regression analysis. There are a few weaknesses of this paper that include lack of detailed explanation of the data, neural network architecture and training process used in this study.

G. Customer churning analysis using machine learning algorithms

The research paper [3] offers a thorough examination of the research on customer churn analysis and the application of several machine learning algorithms in this area. The authors have looked at studies on customer churn analysis that were carried out in several industries utilizing machine learning methods like random forests, decision trees, logistic regression, and K-nearest neighbor, among others. The application of deep learning methods to customer churn analysis is also examined in this study. It was concluded that deep learning techniques can model more complicated systems and achieve greater success rates. However, machine learning algorithms are considered to be the closest alternative and may be beneficial in assessing time-related events like customer turnover

because they are a new technology and may not produce steady findings. The paper's strengths include applying different machine learning algorithms to customer churn research, doing a thorough evaluation of the literature in this area and a thorough examination of each algorithm's benefits and drawbacks. The authors have also looked at customer churn analysis studies that have been done with similar methods in a variety of industries. Paper's weaknesses are that the technique employed for the investigation is not thoroughly explained, a thorough review of the study's shortcomings is absent from the publication and it is unclear from the paper's conclusion whether machine learning method is best for customer attrition analysis.

H. A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services

The research paper [2] suggests using neural networks to forecast user attrition in cellular phone service subscriptions, emphasizing that customer churn control has become a critical job for mobile communication operators as the market for cellular network services becomes more competitive. The authors believe in order for businesses to keep their most devoted clients, they must be able to accurately estimate consumer churn. A number of churn prediction models are reviewed in detail in this paper. Additionally, the authors offer a thorough description of the neural network-based strategy and how it is used to forecast customer attrition in cellular network services. Experiments show that the neural network-based method is more than 92 percent accurate at predicting customer attrition. The paper's strengths include a comprehensive literature review of existing churn prediction models, proposal of a neural network based approach to predict customer churn in cellular network devices and the result of experiments performed that show that neural network based approach can predict customer churn with more than 92 percent accuracy. The weaknesses of this paper is that it fails to provide a detailed explanation of the neural network based approach and also does not provide a comparison of proposed approach with existing churn prediction models.

I. Customer churn prediction in telecommunications

The research paper [4] suggests a model that makes use of data mining and machine learning approaches to forecast customer churn in the telecom sector. The churn prediction model was constructed by the authors using decision trees and logistic regression. The SyriaTel telephone firm provided a sizable dataset that was built by transforming massive raw data, which was used to test the model. The SyriaTel system was trained, tested, and evaluated using the dataset, which included all customer information collected over a nine-month period. Four methods were used to test the model: Extreme Gradient Boosting (XGBOOST), Random Forest, Gradient Boosted Machine Tree (GBM), and Decision Tree. The XGBOOST algorithm produced the best outcomes. This churn predicting model utilized this algorithm for classification. The model created in this work creates a novel approach to feature engineering and selection by utilizing machine learning techniques on large data platforms. By extracting Social Network Analysis (SNA) features, the authors have additionally utilized customer social networks in the prediction model. In comparison to the AUC standard, the model's performance increased from 84 to 93.3 percent with the usage of SNA. The utilization of a sizable dataset and the application of machine learning techniques to forecast client attrition are two of the paper's strong points. Paper's weakness is that the technique employed for the investigation is not thoroughly explained and a thorough review of the study's shortcomings is absent from the publication. It is also unclear from the paper's conclusion whether machine learning method is best for customer churn analysis.

J. Comparing to Techniques Used in Customer Churn Analysis

This research paper [10] investigates the application of machine learning techniques to customer churn research, including artificial neural networks (ANN), decision trees, support vector machines (SVM), naive bayes, k-nn, and extreme gradient boosting (XGBoost). The application of deep learning techniques and the Cox proportional hazard model to customer churn analysis is also explored. The authors have looked at customer churn analysis studies that were done with these methods in a variety of industries. The study comes to the conclusion that deep learning techniques can model more complicated systems and achieve greater success rates. However, machine learning algorithms are considered to be the closest alternative and may be beneficial in assessing time-related events like customer turnover because they are a new technology and may not produce steady findings. The estimates of the independent variables influencing the time variable, the rate of life expectancy, and the groups under risk were found to be successfully produced by the Cox regression model. Prediction with different neural network's top logies were experimented. Strengths include applying different machine learning algorithms to customer churn research, a thorough evaluation of the literature in this area, a thorough examination of each algorithm's benefits and drawbacks all were included in the study. The authors have also looked at customer churn analysis studies that have been done with similar methods in a variety of industries. Paper's weaknesses are that the technique employed for the investigation is not thoroughly explained, a thorough review of the study's shortcomings is not mentioned and it is unclear from the paper's conclusion whether machine learning method is best for customer churn analysis.

K. A Proposed Churn Prediction Model

This research paper [11] explores the usage of 3 data mining techniques, namely Decision Tree, Support Vector Machine and Neural Network, to predict customer churn. These techniques classify the customers into Churner and Non-Churner classes. After classification, each class is divided into 3 clusters, using k-means algorithm, in order to decide retention strategies according to each cluster. Both SVM and NN techniques have an accuracy rate of 83.7 percent, while DT has an accuracy rate of 77.9 percent. As a result, SVM and NN are found to be. The strength of this paper is that it does not stop at predicting customer churn, but further divided each class into clusters, so that a business can efficiently develop relevant retention strategies. However, the researchers have employed only one evaluation metric for the classification techniques, which limits the legitimacy of the findings.

L. Model of Customer Churn Prediction on Support Vector Machine

This research paper [14] proposes the use of support vector machine (SVM) with structural risk minimization for customer churn prediction. 4 evaluation metrics have been employed to compare the performance of the proposed SVM against 4 other methods i.e. ANN, Decision Tree, Logistic Regression, and Naïve Bayesian Classifiers. These evaluation metrics are accuracy rate, hit rate, coverage rate, and lift coefficient. Two datasets have been used for model training. Various kernel functions were tested with the constructed SVM, and Radial Basis and Cauchy Kernel Functions were found to show good results with data sets 1 and 2 respectively. For data set 1, SVM outperforms all other methods i.e. ANN, DT, LR, and NBC, in all the evaluation metrics. However for data set 2, ANN has slightly better hit rate and lift coefficient than SVM, which stems from the relatively inferior quality of data set 2. A major strength of this paper is the use of 4 different types of accuracy metrics, and 2 data sets, which express the performance of the model in different views. However, a weakness is that only the factor analysis of the data sets has been presented, leaving the data sets insufficiently explained.

M. Customer Churn Prediction Using Improved One-Class Support Vector Machine

This research paper [15] explores the use of imporved one-class SVM, in order to combat the problem of churn data sets being primarily populated with non-churn data, making it difficult to train the model accurately for both classes. In this one-class SVM, the data points close to enough to the origin are considered members of the negative class (i.e. churn class), and slack variables have been added to the programming problem, in order to allow some error during training. Linear, Polynomial, and Gaussian Kernel functions are used, where the Gaussian KF returns the highest accuracy of 87.15 percent, indicating that the separation plane in this problem is non-linear. The accuracy of the Gaussian SVM, compared with that of the other models (i.e. ANN, Decision Tree, and Naive Bayes), is also the highest. Notable weaknesses of this research item are using only one accuracy metric, and lack of discussion on real-life implications of the model's results.

N. Customer churn analysis in telecommunication sector

This research paper [6] aims to develop a method which can accurately predict the profiles of customers who are likely to churn, so that the telecommunication provider can implement retention strategies beforehand. Logistic Regression and Classification and Regression Trees (CR&T) Decision Tree techniques are used to make the required predictions. In order to have an unbiased result, the data distribution is balanced to have churn and non-churn percentages of 50.03 and 49.07 respectively, instead of the original 25.2 percent and 74.8 percent. The authors aim to predict the affect of each attribute on the likelihood the value of churn being 0 or 1, and hence perform feature selection using Wald test, in order to retain only the most significant features. Logistic Regression gives an accuracy of 70.1 percent, while Decision Tree gives an accuracy of 71.76 percent. Strengths of this research include the data set being adjusted to avoid class imbalance, the authors describing affect of each attribute on the prediction for churn, and a discussion on how this model fits into real-life business scenarios. However, the lack of description of accuracy metrics for the models' evaluation is a major drawback.

O. Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning

The research paper [9] explores 4 tree-based classifier algorithms to predict customer churn in telecommunication industry, using the same data set as our research project. The author has used Python's FeatureTools library to create 724 new features out of the existing 19 features. For feature selection, decision Tree, Random Forest, LightGBM and xGBoost models were used, and xGBoost was found to be the best of them all, shortlisting 11 features. xGBoost, Random Forest, and Decision Tree ML algorithms were implemented, and xGBoost was found to give the best accuracy, 84 percent, on ROC curve. LIME Analysis has been used to explain the predictions of the model. A notable strength of this paper is that it includes a data visualization section, explaining the data set efficiently. In addition, LIME analysis is a reliable method to explain the findings of the ML model. However, the lack of explanation regarding creation of new filters is a limitation.

Reference	Year	Method	Evaluation Metric	Score	Dataset
[12]	2009	ANN and SOM	Accuracy	93.12%	CRM
[13]	2015	SVM-POLY	Accuracy	97%	-
[1]	2014	CRISP-DM	-	-	-
[5]	2015	J-48	Accuracy	99.7%	Orange
[8]	2017	Regression + NN + DT	-	-	-
[7]	2015	MLP NN	-	91.28%	-
[3]	2023	Random Forest	-	81.71%	-
[2]	2011	NN	-	92%	-
[4]	2012	LR,DT	XGBOOST	93.3%	SyriaTel
[10]	2019	XGBoost	-	-	-
[11]	2012	SVM, NN, DT	Accuracy	83.7%	-
[14]	2008	SVM	Accuracy	90.8%	-
[15]	2005	SVM	Accuracy	87.15%	-
[6]	2010	LR, DT	Accuracy	71.76%	-
[9]	2019	xGBoost, RF, DT	AUROC	85%	-

TABLE I. LITERATURE SURVEY

III. DATA COLLECTION AND PROCESSING

The Telco Customer Churn data set was collected from kaggle, and is available in a csv file format. There are a total of 21 columns and 7043 rows in the data set, each row representing information about a customer. The Churn attribute acts as the label, indicating whether or not the customer churned. The attributes in the data set are listed in table [2], along with their data types.

The data set can be divided into 3 segments:

1) *Demographic information*: gender, SeniorCitizen, Partner, Dependents
2) *Customer account information*: tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges
3) *Subscribed services*: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

Attribute	Data type
customerID	object
gender	object
SeniorCitizen	Int64
Partner	object
Dependents	object
tenure	Int64
PhoneService	object
MultipleLines	object
InternetService	object
OnlineSecurity	object
OnlineBackup	object
DeviceProtection	object
TechSupport	object
StreamingTV	object
StreamingMovies	object
Contract	object
PaperlessBilling	object
PaymentMethod	object
MonthlyCharges	Float64
TotalCharges	object
Churn	object

TABLE II. ATTRIBUTES OF THE DATASET

A. Data Wrangling

The are no null values in the data set. Similarly, no duplicates or outliers exist in the data set either. The customerID column contains the unique identifiers of each user, i.e. 7043 unique values which will not be contributing to the machine learning process. Hence, this column was dropped.

B. Data Encoding

As mentioned in table [2], 18 of the columns are in available in object data type. After removing customerID column, 17 object-type columns are left. Fig [1] shows unique values of all these columns, except the TotalCharges column, which has 7000+ unique values.

```
Gender:              ['Female' 'Male']
Partner:             ['Yes' 'No']
Dependents:          ['No' 'Yes']
PhoneService:        ['No' 'Yes']
MultipleLines:       ['No phone service' 'No' 'Yes']
InternetService:     ['DSL' 'Fiber optic' 'No']
OnlineSecurity:      ['No' 'Yes' 'No internet service']
OnlineBackup:        ['Yes' 'No' 'No internet service']
DeviceProtection:    ['No' 'Yes' 'No internet service']
TechSupport:         ['No' 'Yes' 'No internet service']
StreamingTV:         ['No' 'Yes' 'No internet service']
StreamingMovies:     ['No' 'Yes' 'No internet service']
Contract:            ['Month-to-month' 'One year' 'Two year']
PaperlessBilling:    ['Yes' 'No']
PaymentMethod:       ['Electronic check' 'Mailed check' 'Bank transfer (automatic)'
 'Credit card (automatic)']
Churn:               ['No' 'Yes']
```

Fig. 1. Unique values of the object-type attributes

As the columns gender, Partner, Dependents, PhoneService, PaperlessBilling, and Churn had only two unique values, they have been encoded using binary encoding. The columns MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies have been encoded using label encoding. The columns Contract and PaymentMethod have been one-hot encoded. As a result of this one-hot encoding, 5 new columns have been added to the dataset, making the total number of columns 25. Lastly, the TotalCharges column has been converted to numeric datatype. As a result of this conversion, 11 of the rows of the TotalCharges column were left null, so the null rows were dropped.

IV. Data Exploration and Transformation

A. Data Visualization

The distribution of the data in Churn column is visualized in Fig. [2], showing that the majority of the customers in the data set did not churn. A bias towards non-churn can be observed.

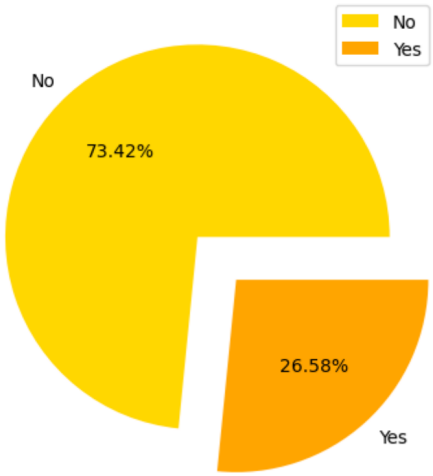

Fig. 2. Churn column

Fig [3] - [21] visualize the relationship between each attribute in the data set and Churn. The correlation between all the attributes has been summarized in the heatmap in Fig [22].

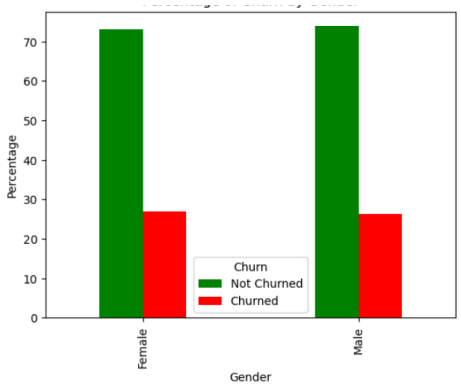

Fig. 3. Percentage of Churn by Gender

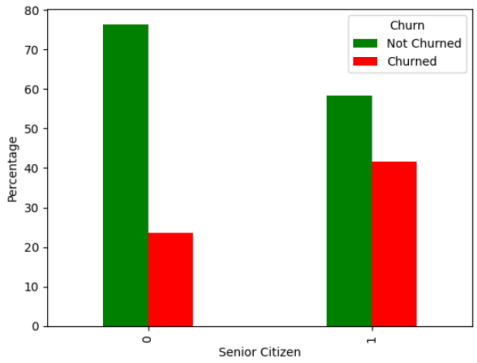

Fig. 4. Percentage of Churn by Senior Citizen

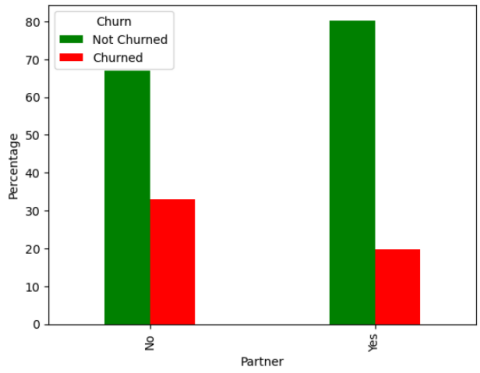

Fig. 5. Percentage of Churn by Partner

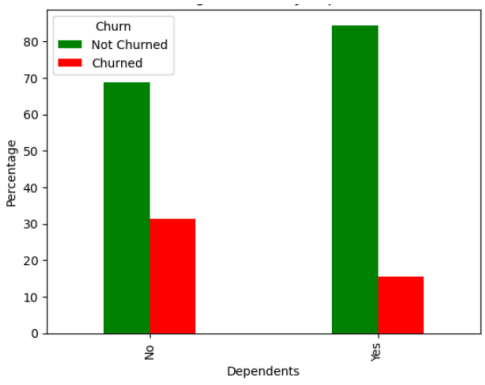

Fig. 6. Percentage of Churn by Dependents

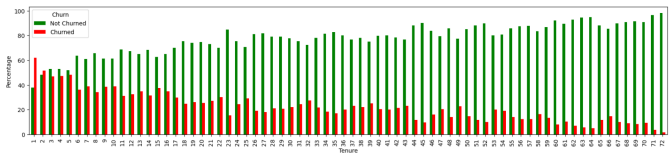

Fig. 7. Percentage of Churn by Tenure

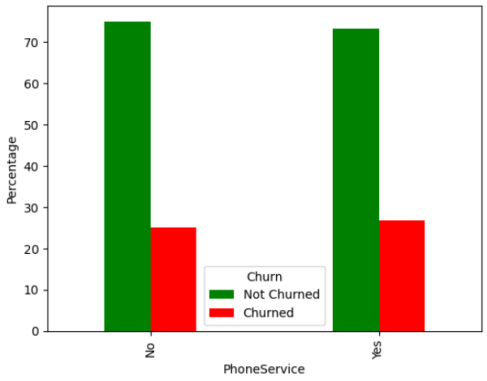

Fig. 8. Percentage of Churn by PhoneService

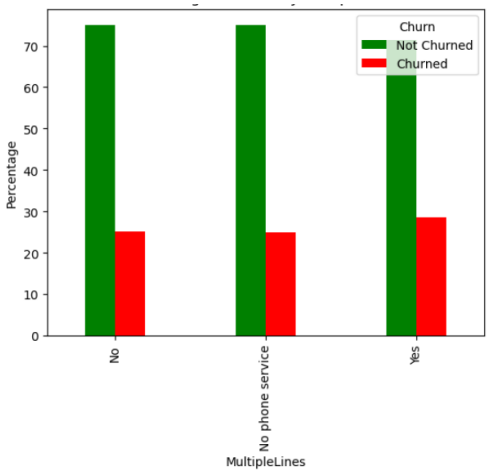

Fig. 9. Percentage of Churn by MultipleLines

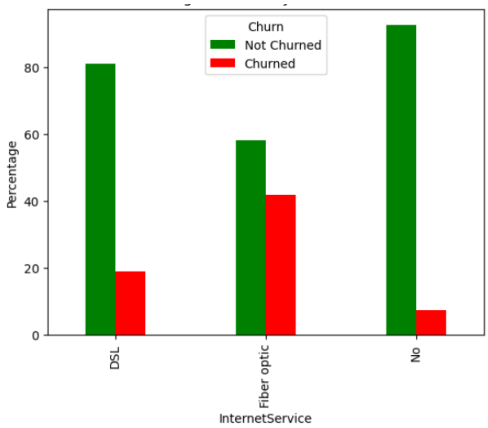

Fig. 10. Percentage of Churn by InternetService

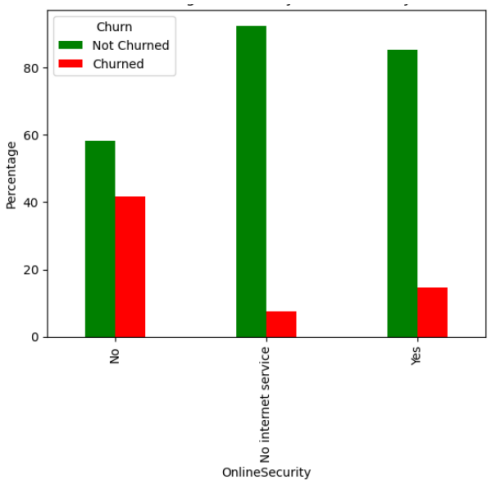

Fig. 11. Percentage of Churn by OnlineSecurity

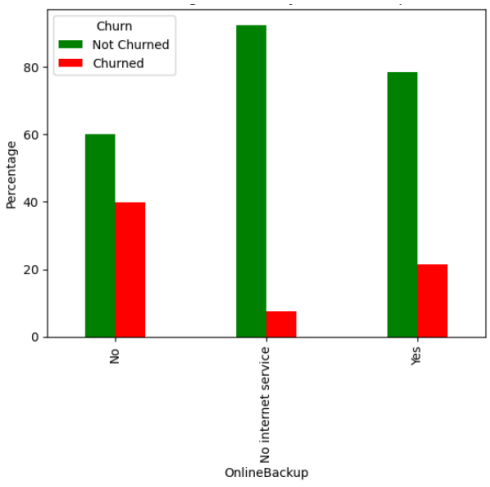

Fig. 12. Percentage of Churn by OnlineBackup

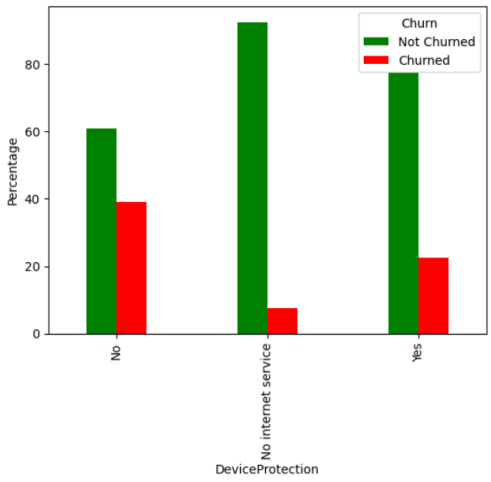

Fig. 13. Percentage of Churn by DeviceProtection

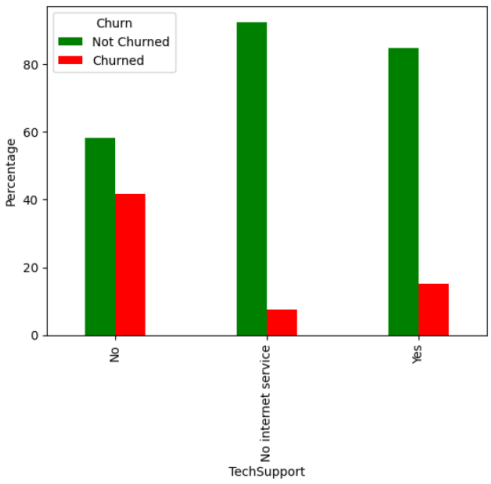

Fig. 14. Percentage of Churn by TechSupport

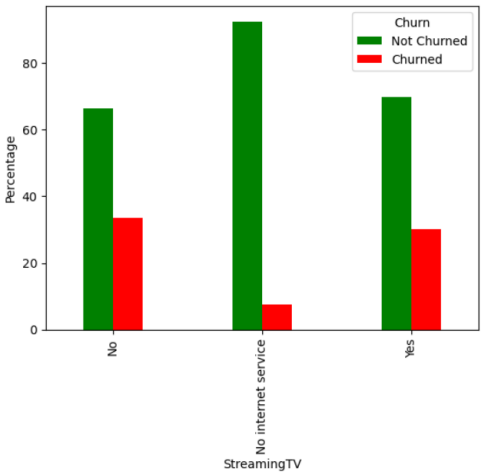

Fig. 15. Percentage of Churn by StreamingTV

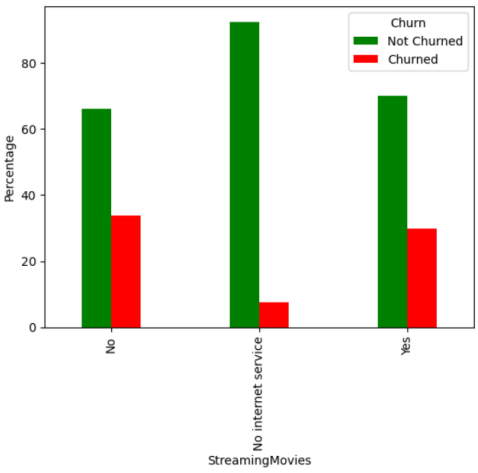

Fig. 16. Percentage of Churn by StreamignMovies

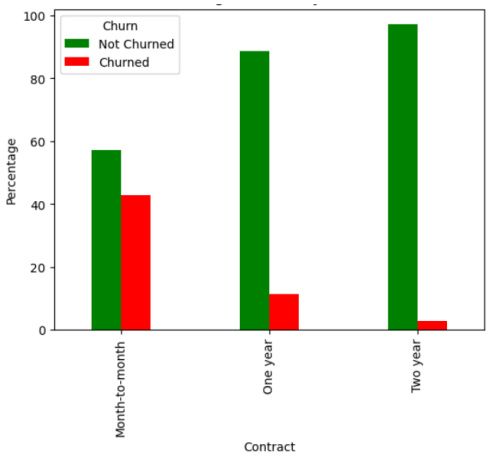

Fig. 17. Percentage of Churn by Contract

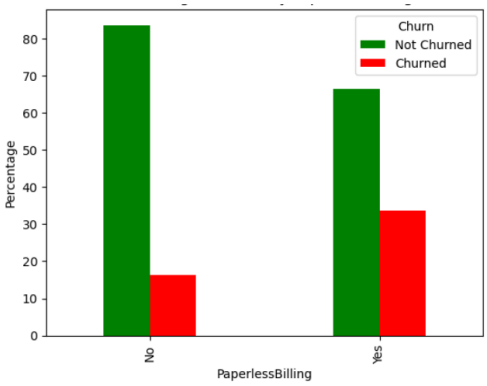

Fig. 18. Percentage of Churn by PaperlessBilling

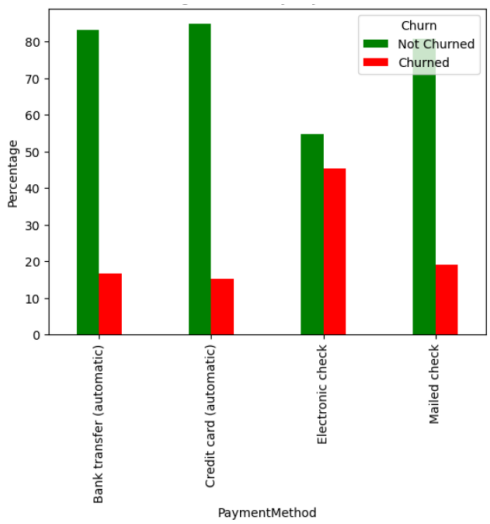

Fig. 19. Percentage of Churn by PaymentMethod

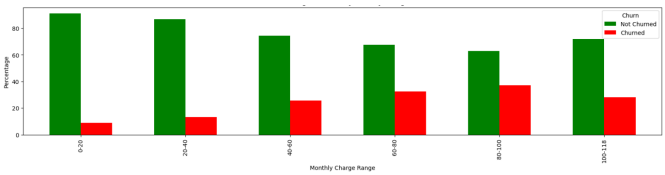

Fig. 20. Percentage of Churn by MonthlyCharges

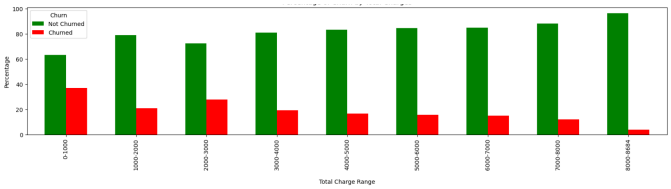

Fig. 21. Percentage of Churn by TotalCharges

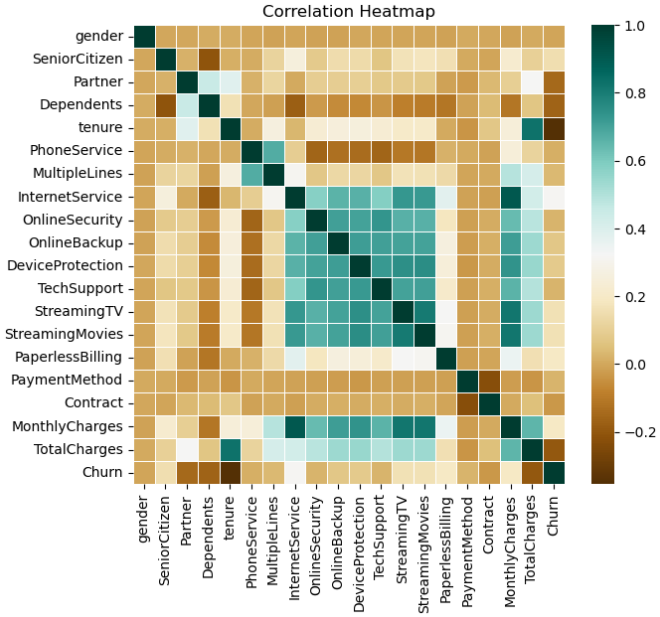

Fig. 22. Correlation between all the attributes of the data set

B. Transforming the Data through Normalization

As the object-type columns, except TotalCharges, have been binary, label, or one-hot encoded, their values exists in the range of 0-1 or 0-2. However, the ranges of values for tenure, MonthlyCharges, and TotalCharges columns are very wide. As a result, Normalization has been applied to transform the data in these columns. The choice of using Normalization arises from the fact that the data is not normally distributed.

C. Feature Selection

After performing encoding, there are 25 columns in the data set, which must be narrowed down to avoid the Curse of Dimensionality. The research work [9], which is using the same data set as ours, has used FeatureTools library to create 724 new features, and then used xGBoost model to select the top features. We have selected the same features in our work. These selected features are:

1) Contract-Month-to-month
2) Contract-One year
3) Contract-Two year
4) TechSupport
5) OnlineSecurity
6) tenure
7) MonthlyCharges
8) OnlineBackup
9) InternetService
10) PaperlessBilling
11) SeniorCitizen
12) tenure MOD MonthlyCharges

V. Data Modeling

In this phase, different models were explored and their accuracies were evaluated by fine-tuning different hyperparameters. The models utilized are given below:

A. XGBoost Classifier

XGBoost functions by use of an ensemble learning method called gradient boosting, in which a sequence of weak learners (usually decision trees) are trained one after the other to fix the errors of the previous ones.To avoid overfitting, it uses a regularized objective function with a loss term and a complexity term. Because of its proficiency in creating ensemble models of decision trees, XGBoost is well-suited to capture the complicated patterns that exist across different features in the customer churn prediction problem.

B. Logistic Regression

Second classifier used in this project was Logistic regression. It is a classification algorithm.The model is particularly well-suited for scenarios where the relationship between the features and the binary outcome can be expressed in terms of probabilities. This model was chosen because customer churn prediction is a binary classification problem. Logistic Regression is easier to implement, computationally efficient and uses regularization techniques which curb the problem of overfitting.

C. Random Forest Classifier

A group of decision trees in this classifier are trained using various subsets of the training set and features. Every tree ”votes” on the result during prediction, with the final prediction being decided by the majority of votes. By using this method, overfitting is reduced and the model's ability to generalize to new data is enhanced. Because of its resilience in managing complex patterns and non-linear interactions, it is a good option for customer churn prediction.

D. Decision Tree

The Decision Tree classifier is a flexible machine learning technique that builds a prediction tree by segmenting the feature space according to the most informative attributes. Decision Trees are a good option for predicting customer churn because of their interpretability and capacity to identify important factors that impact churn.

VI. Presentation and Results

The results of the different models evaluated are described below:

A. XGBoost Classifier

The classifier was trained with a test size of 0.3 and a random state of 3. The hyperparameters were fine-tuned as well. The learning rate was set to 0.01 , the maximum depth of the decision trees was set to 7 and the parameter n-estimators which indicates the number of boosting rounds was set to 200. It yielded an accuracy of 0.80.

1) Confusion Matrix: The confusion matrix provided in Fig. [23] indicates that there were 1451 True Negatives , 127 False Positives , 297 False Negatives and 235 True Positives.

```
Confusion Matrix:
[[1451  127]
 [ 297  235]]
```

Fig. 23. Confusion Matrix of XGBoost Classifier

2) Classification Report: The classification report indicates that the classifier performed well in predicting customers who were likely to stay (Class 0) with high precision (0.83) and recall (0.92). However, there was room for improvement in predicting customer churn (Class 1) with lower precision (0.65) and recall (0.44).

```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.92      0.87      1578
           1       0.65      0.44      0.53       532

    accuracy                           0.80      2110
   macro avg       0.74      0.68      0.70      2110
weighted avg       0.78      0.80      0.79      2110
```

Fig. 24. Classfication Report of XGBoost Classifier

3) ROC Curve: A strong model performance was indicated by a ROC curve with an AUC of 0.84. According to the AUC value, the model correctly assigned a randomly selected churn instance a better ranking than a randomly selected non-churn instance.

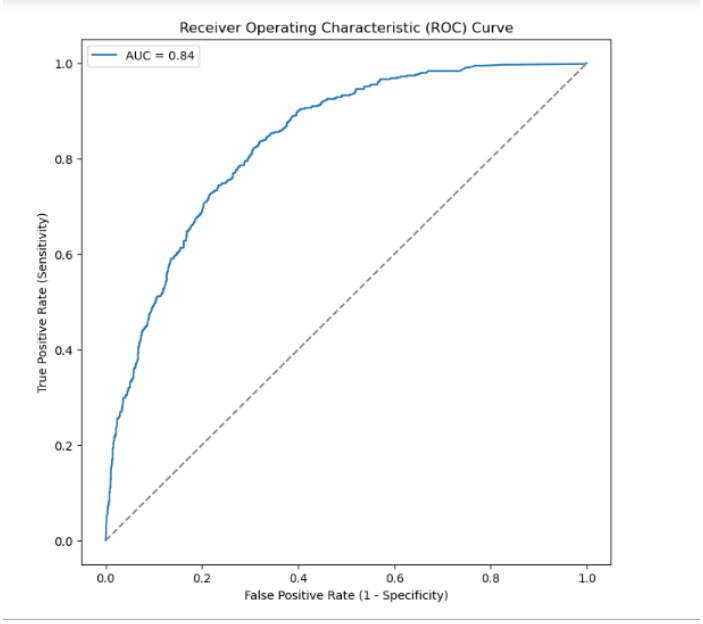

Fig. 25. ROC Curve of XGBoost Classifier

B. Logistic Regression

The model was trained with a test size of 0.3 and a random state of 42. L2 regularization was used to mitigate overfitting. It had an accuracy of 0.79.

1) Confusion Matrix: The confusion matrix, provided in Fig. [26], indicates that there were 1397 True Negatives , 150 False Positives , 286 False Negatives and 277 True Positives.

```
Confusion Matrix:
[[1397  150]
 [ 286  277]]
```

Fig. 26. Confusion Matrix of Logistic Regression Model

2) Classification Report: A model with noteworthy recall (0.90) and precision (0.83) for consumers likely to stay (Class 0) is shown in the classification report. On the other hand, the precision (0.65) and recall (0.49) for churn (Class 1) prediction needed improvement.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      1547
           1       0.65      0.49      0.56       563

    accuracy                           0.79      2110
   macro avg       0.74      0.70      0.71      2110
weighted avg       0.78      0.79      0.78      2110
```

Fig. 27. Classfication Report of Logistic Regression Model

3) ROC Curve: A ROC curve with an Area Under the Curve (AUC) of 0.83 suggested a robust performance in distinguishing between positive and negative instances.

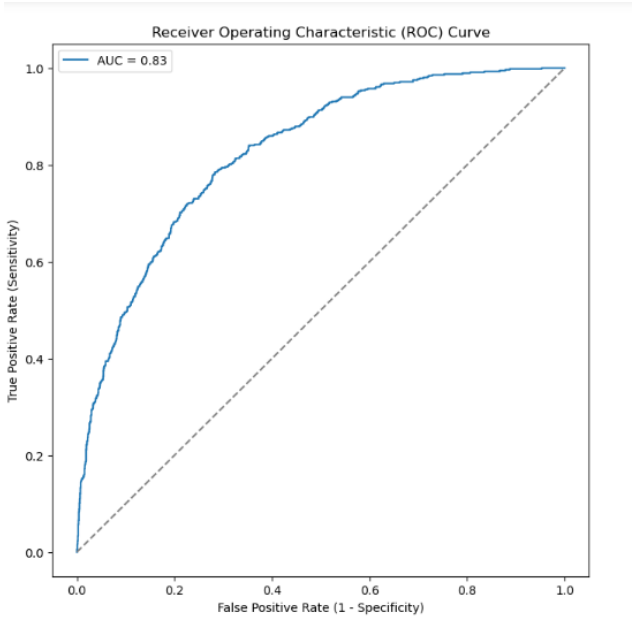

Fig. 28. ROC Curve of Logistic Regression

C. Random Forest Classifier

The classifier was trained with a test size of 0.4 and a random state of 67. The hyperparameter n-estimators which determines the number of decision trees in the ensemble was set to 100. It had an accuracy of 0.79.

1) Confusion Matrix: The confusion matrix, provided in Fig. [29], indicates that there were 1842 True Negatives , 227 False Positives , 371 False Negatives and 373 True Positives.

```
Confusion Matrix:
[[1842  227]
 [ 371  373]]
```

Fig. 29. Confusion Matrix of Random Forest Classifier

2) Classification Report: The model exhibited a precision of 0.83 for predicting non-churn instances (Class 0), indicating strong accuracy. However, there was room for improvement in recall (0.89), suggesting potential missed non-churn cases. For churn predictions (Class 1), precision was at 0.62, denoting accuracy, while recall was 0.50, indicating a moderate ability to identify actual churn.

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	2069
1	0.62	0.50	0.56	744
accuracy			0.79	2813
macro avg	0.73	0.70	0.71	2813
weighted avg	0.78	0.79	0.78	2813

Fig. 30. Classfication Report of Random Forest Classifier

3) ROC Curve: While an AUC of 0.81 in the ROC curve signified a satisfactory performance in distinguishing between positive and negative instances, it suggested that there was room for improvement.

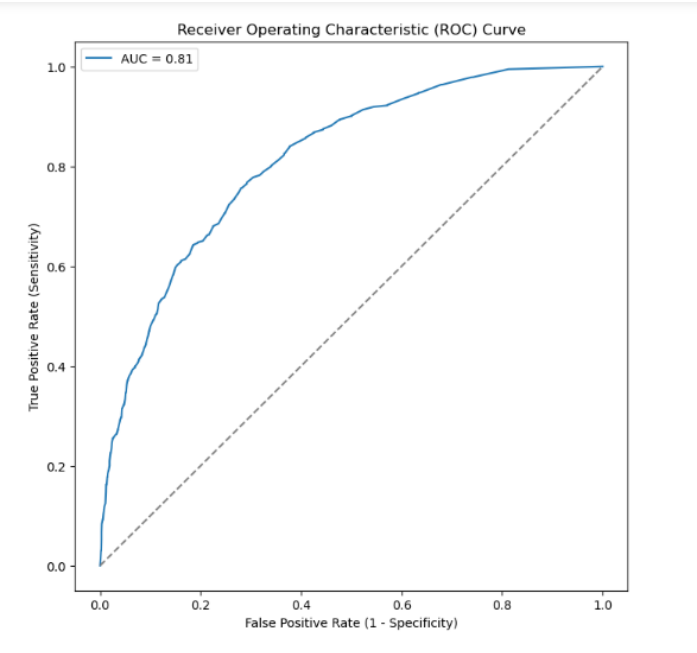

Fig. 31. ROC Curve of Random Forest Classifier

D. Decision Tree

The classifier was trained with a test size of 0.3 and a random state of 42. It yielded an accuracy of 0.72.

1) Confusion Matrix: The confusion matrix, provided in Fig. [32], indicates that there were 1242 True Negatives , 305 False Positives , 296 False Negatives and 267 True Positives.

```
Confusion Matrix:
[[1242  305]
 [ 296  267]]
```

Fig. 32. Confusion Matrix of Decision Tree

2) Classification Report: The classification report indicated that although the model's precision for non-churn cases (0.81) was rather good, its recall (0.47), suggested that the model might miss a large percentage of true non-churn situations. In addition, the model's precision and recall for churn instance prediction were both 0.47, indicating a decent level of accuracy but also a difficulty in accurately identifying real churn situations.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.80	0.81	1547
1	0.47	0.47	0.47	563
accuracy			0.72	2110
macro avg	0.64	0.64	0.64	2110
weighted avg	0.72	0.72	0.72	2110

Fig. 33. Classification Report of Decision Tree

3) ROC Curve: Restricted discriminative power was suggested by a ROC curve with an AUC of 0.64. According to the AUC, the model's performance in differentiating between churn and non-churn cases was subpar.

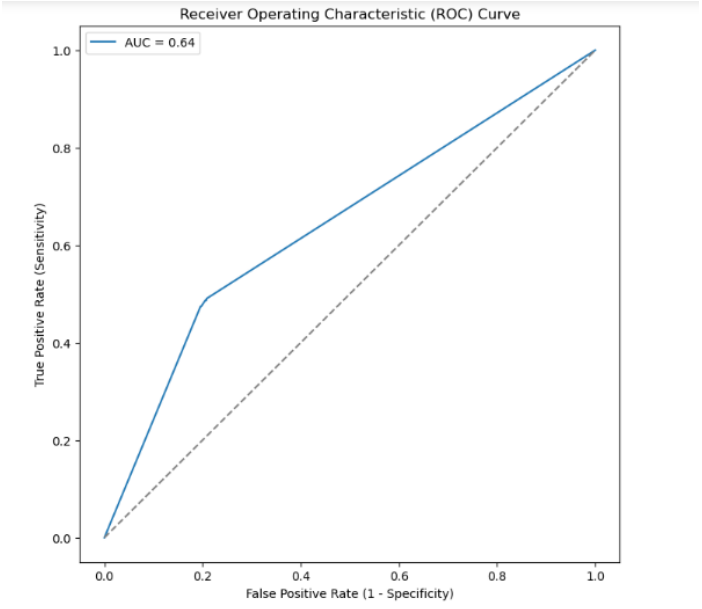

Fig. 34. ROC Curve of Decision Tree

E. Comparison of Results

The XGBoost Classifier is the most effective model among those that were assessed for customer churn prediction; it

achieved a high accuracy of 0.80 and showed a considerable trade-off between precision and recall for both churn and non-churn scenarios. Strong discriminative power is indicated by its ROC curve, which has an AUC of 0.84. On the other hand, the Decision Tree model exhibits limits, indicating limited discriminative capabilities, with an accuracy of 0.72 and a lower AUC of 0.64. All things considered, the XGBoost Classifier turns out to be the most successful model, providing a viable method for precise customer churn prediction.

VII. Conclusion and Future Work

In conclusion, this study has navigated the complexities of predicting customer churn through the use of thorough data processing, encoding strategies, exploratory analysis, normalization procedures, and model evaluations. With an outstanding accuracy of 0.80 and a well-balanced precision and recall for both churn and non-churn instances, the XGBoost Classifier stood out among the other models. The model is well-positioned for accurate customer churn prediction due to its great discriminative strength, which is demonstrated by its high AUC of 0.84 in the ROC curve. Nevertheless, the investigation of substitute models, such as the Decision Tree, revealed certain constraints, indicating possible avenues for enhancement. In order to improve prediction capabilities, future work will concentrate on optimizing hyperparameters, investigating other feature engineering techniques, and integrating ensemble models.

References

[1] Amal M Almana, Mehmet Sabih Aksoy, and Rasheed Alzahrani. "A survey on data mining techniques in customer churn analysis for telecom industry". In: *International Journal of Engineering Research and Applications* 4.5 (2014), pp. 165–171.

[2] Dr. Prabin Kumar Panigrahi Anuj Sharma. "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services". In: *International Journal of Computer Applications* 27.11 (2011), pp. 0975–8887.

[3] B.R. Kavitha B. Prabadevi R. Shalini. "Customer churning analysis using machine learning algorithms". In: *International Journal of Intelligent Networks* 4.14 (2023), pp. 145–154.

[4] Brian Buckley Bingquan Huang Mohand Tahar Kechadi. "Customer churn prediction in telecommunications". In: *Expert Systems with Applications* 39.1 (2012), pp. 1414–1425. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2011.08.024. URL: https://www.sciencedirect.com/science/article/pii/S0957417411011353.

[5] Kiran Dahiya and Surbhi Bhatia. "Customer churn analysis in telecom industry". In: *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*. 2015, pp. 1–6. DOI: 10.1109/ICRITO.2015.7359318.

[6] Umman Şimşek Gürsoy. "Customer churn analysis in telecommunication sector". In: *İstanbul Üniversitesi İşletme Fakültesi Dergisi* 39.1 (2010), pp. 35–49.

[7] Mohammad Ridwan Ismail et al. "A Multi-Layer Perceptron Approach for Customer Churn Prediction". In: *International Journal of Multimedia and Ubiquitous Engineering* 10.7 (2015), pp. 213–222.

[8] Chinnu P Johny and Paul P Mathai. "Customer churn prediction: A survey". In: *International Journal of Advanced Research in Computer Science* 8.5 (2017), pp. 2178–2181.

[9] Kriti. "Customer Churn: A Study of Factors Affecting Customer Churn using Machine Learning". Iowa State University, 2019.

[10] Usame O. Osmanoglu Ozer Celik. "Comparing to Techniques Used in Customer Churn Analysis". In: *Journal of Multidisciplinary Developments* 4.1 (2019), pp. 30–38. ISSN: 2564-6095.

[11] Essam Shaaban et al. "A proposed churn prediction model". In: *International Journal of Engineering Research and Applications* 2.4 (2012), pp. 693–697.

[12] Chih-Fong Tsai and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks". In: *Expert Systems with Applications* 36.10 (2009), pp. 12547–12553. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2009.05.032. URL: https://www.sciencedirect.com/science/article/pii/S0957417409004758.

[13] T. Vafeiadis et al. "A comparison of machine learning techniques for customer churn prediction". In: *Simulation Modelling Practice and Theory* 55 (2015), pp. 1–9. ISSN: 1569-190X. DOI: https://doi.org/10.1016/j.simpat.2015.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1569190X15000386.

[14] Guo-en Xia and Wei-dong Jin. "Model of customer churn prediction on support vector machine". In: *Systems Engineering-Theory & Practice* 28.1 (2008), pp. 71–77.

[15] Yu Zhao et al. "Customer churn prediction using improved one-class support vector machine". In: *Advanced Data Mining and Applications: First International Conference, ADMA 2005, Wuhan, China, July 22-24, 2005. Proceedings 1*. Springer. 2005, pp. 300–306.