TELECOM CUSTOMER CHURN ANALYSIS USING MACHINE LEARNING

Submitted by

NIVEDA DAS M
Roll No: 223040

RAFEEK RAHIM
Roll No: 223041

RAHMA FAHIM K
Roll No: 223042

RAHUL
Roll No: 223043

In partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics of

Kerala University of Digital Sciences, Innovation and
Technology (Digital University Kerala), Thiruvananthapuram 695317

September 2023

CERTIFICATE

This is intended to authenticate that the project report titled “ TELCOM CUSTOMER CHURN ANALYSIS USING MACHINE LEARNING”, submitted by Niveda M Das (Roll No.223040), Rafeek R (Roll No. 223041), Rahma Fahim K (Roll No. 223042) and Rahul(Roll No. 223043), in partial fulfilment of the requirements for the award of Master of Science in Computer Science with a Specialization in Data Analytics, is a true record of the work completed at Kerala University of Digital Sciences, Innovation and Technology under our supervision.

SUPERVISOR	**COURSE COORDINATOR**
Dr. T. K Manoj Kumar Professor DUK	Dr. T. K Manoj Kumar Professor DUK

DECLARATION

This report is largely the result of our own work, unless otherwise stated in the text, and was completed between July 2023 and September 2023, as per Niveda M Das, Rafeek R, Rahma Fahim K, and Rahul, students pursuing Master’s Degree in Computer Science with Specialization in Data Analytics.

NIVEDA DAS M
Roll No: 223040

RAFEEK RAHIM
Roll No: 223041

RAHMA FAHIM K
Roll No: 223042

RAHUL
Roll No: 223043

Place: Trivandrum

Date: 07- September 2023

ACKNOWLEDGEMENT

We thank Dr. T.K Manoj Kumar, Professor, DUK, for giving us the chance to work on a genuine project and for his ongoing guidance and support. For providing these facilities, we would also like to express our sincere thanks to Dr. Saji Gopinath, Vice-Chancellor of Kerala University of Digital Science, Innovation and Technology. We would like to extend our gratitude to our family and friends for their ever-ending support and motivation in making this journey a success.

ABSTRACT

Analysis of Customer churn is a crucial tool for businesses to understand and control customer attrition. Churn means the rate at which customers stop doing business with a company over a given period of time. It is crucial to identify the causes of churn and take steps to prevent it because high churn rates can be detrimental to a business's revenue and growth. In our project, we tried to analyse the churning of telecom customers based on “tenure”. We used Decision tree and Random Forest classifier for the same which are two very important algorithms for classification in Machine Learning.

TABLE OF CONTENT

I. INTRODUCTION

Loss, also known as customer churn, is an important metric and concept in the world of marketing and customer management. It refers to the situation where the customer or customers terminate their relationship with the company or product, usually by banning the subscription, cancelling the service, or not making any further purchases. Losing customers can have significant financial and strategic consequences for a business, as retaining existing customers is often more expensive than acquiring new ones.

Understanding and analysing churn is important for companies of all sizes and industries. By identifying why customers are leaving and developing strategies to reduce churn, organizations can improve customer retention, increase loyalty, and ultimately increase revenue.

CAUSES OF CUSTOMER CHURN

Bad customer experience: Dissatisfaction with a company's products, services, or customer support can lead to customer loyalty. Customers who have had a negative experience are more likely to leave the business and may share their dissatisfaction with others.

Competition: Competitors offering better or better options can encourage customers to switch.

Price Sensitivity: Customers can be very price sensitive and may choose this option if they can find a better price elsewhere, even if the price difference is small.

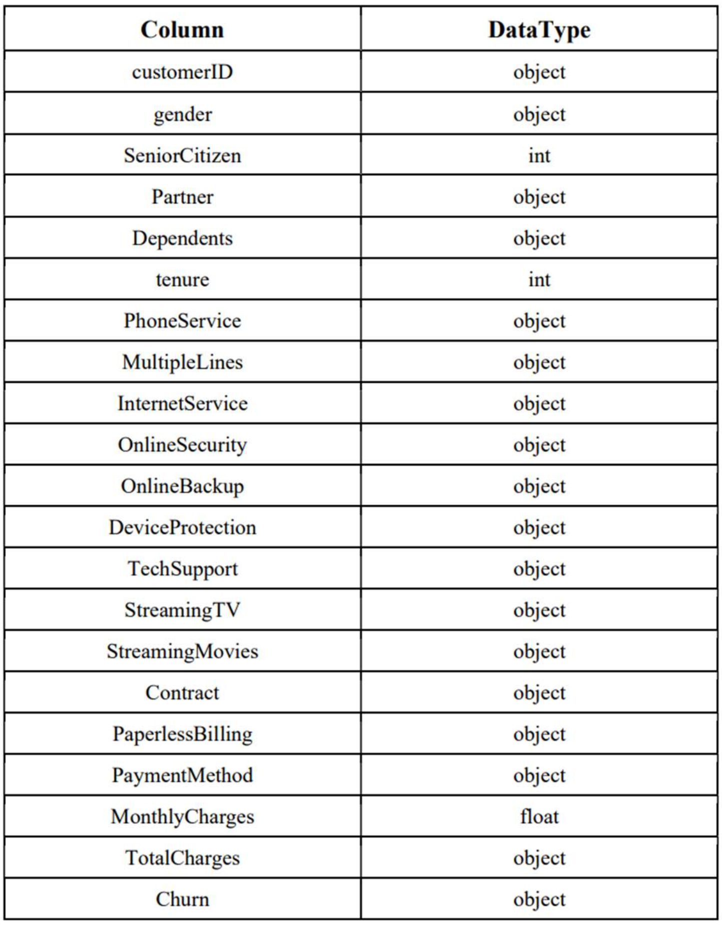
Life Process: Personal or situational factors such as moving to a new location or financial changes can lead to loss of customers.

Disagree: Customers who do not use or participate in products or services may eventually cancel or cancel altogether.

Expectations Unmet: When customers are dissatisfied with product performance, delivery time or service quality, they are more likely to leave.

DATA DISCRIPTION: TELECOM DATASET

We have considered the Telecom data set for the analysis of the Churning of customers which was created by IBM. Each customer is represented by a row of the dataset, and each column contains the customer’s attributes. It contains a total of 21 attributes and 7037 records. (Data types: float64(1), int64(2), object (18))



Column	DataType
customerID	object
gender	object
SeniorCitizen	int
Partner	object
Dependents	object
tenure	int
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	float
TotalCharges	object
Churn	object

customerID - The customerID is an attribute of the telecom dataset that contains ID about the customer. In our analysis we will discard this column as it is of less importance.

Gender - This contains the gender of the customers, either “Male” or “Female”.

SeniorCitizen – It contains whether the customer is a Senior Citizen or not. It is either 0 or 1

Partner - This column specifies whether the customer has a partner with him/her or not

Dependents - This column specifies whether the customer has dependents such as children, siblings etc.

Tenure - The tenure is a very important attribute for our analysis. It shows the total number of months a customer has stayed with the company.

PhoneService - It specifies whether a customer has Telephone service or not therefore it contains Yes or No values

MultipleLines - It shows whether the customer has Multiple lines or not. It contains 3 values - Yes, No, and No Phone.

InternetService - This feature specifies the type of Internet service the customer has taken whether it is Fibre Optic, DSL, or other.

Online Security- This feature specifies whether the customer has online security or not.

PROBLEM STATEMENT: Our Aim was to analyse the churning of Customers based on the feature tenure.

II. LITERATURE REVIEW

1) The primary objective of the study by Siru Zhu, Xin Hu, Yanfei Yang, and Lanhua Chen was to forecast customer churn using information from a customer database from a Chinese supermarket. This research used a combination model that combines decision trees and neural networks, in contrast to traditional forecasts that rely on a single model. This study took into account transaction trend value, membership card level, age, non-shopping points, and transaction age as predictors. It's noteworthy that the model updated the likelihood of a client departing by taking into account variations in customer confidence levels. The SPSS modeller's C5.0 decision tree was used specifically. The neural network model's average prediction accuracy was determined to be about 96%, while the combined model displayed an excellent prediction accuracy of around 98%.

2) The paper "Churn Prediction: A Comparative Study Using KNN and Decision Trees" by Jamal Alsakran, Ali Rodan, Abdel-Karim Al-Tamimi, and Mohammad A. Hassonah focuses on Customer Relationship Management (CRM) due to the intense competition of various industries globally. Customer retention is the number one priority for most businesses, notably service providers and telecom corporations. This study builds on prior research on predicting customer turnover by employing both a decision tree model and the k-nearest neighbor (KNN) model with k=5. The precision, recall, and F-measure measures are used to assess performance. The dataset used provides data about telecom clients. The almost same AUC (Area Under the Curve) values for both approaches are due to the low specificity of the decision tree (DT), resulting in virtually indistinguishable AUC

values for both strategies. The data show that KNN outperforms decision trees by around 86%, with KNN having a 92% better degree of accuracy.

3) The objective of Ishpreet Kaur and Jasleen Kaur's study, "Customer Churn Analysis and Prediction in Banking Industry Using Machine Learning," is on predicting customer attrition in the financial services sector. Academics offer numerous strategies for this goal, notably decision trees, support vector machines, and logistic regression, among others. Nevertheless, it is worth mentioning that the most widely used algorithms for this task are those based on decision trees, which excel at detecting client attrition.While the computational components of churn prediction have been widely studied in academic studies, the writers present a novel approach aimed at improving churn prediction performance. This strategy is based on the use of a data preparation treatment approach. Pre-processing the dataset eliminates extraneous and redundant information, enabling outstanding performance for the classifiers used in the churn prediction task. This novel data preparation technique shows to be a beneficial upgrade for predicting client attrition in the banking business.

4) Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif ul Islam, and Sung Won Kim investigated machine learning approaches in the telecom industry in order to uncover causes and estimate customer turnover. Their research yielded the creation of a churn prediction model based on the Random Forest (RF) algorithm.During the study's first phase, classification algorithms were used to categorize data from customers who were at danger of attrition. The Random Forest algorithm achieved a remarkable 88.63 percent classification accuracy rate. Considering the critical need for effective retention techniques to reduce churn, the proposed method not only identified data from probable

churners but also divided it into categories using cosine similarity. This segmentation enabled the tailoring of retention offers to certain consumer groups.In addition, the study revealed churn-related elements, offering light on the fundamental causes influencing customer loss. This understanding of the fundamental causes of churn may greatly enhance a company's marketing efforts. It facilitates the deployment of tailored actions, such as providing appropriate incentives to a group of customers who are prone to churn based on similar behavioral patterns, resulting in increased productivity and customer retention rates.The accuracy, precision, recall, f-measure, and Receiving Operating Characteristics (ROC) of the churn prediction model were rigorously tested in the study. The results showed that by integrating the Random Forest technique with k-means clustering, the prediction model was able to attain a better level of precision in identifying churn, hence increasing its usefulness in dealing with client attrition.

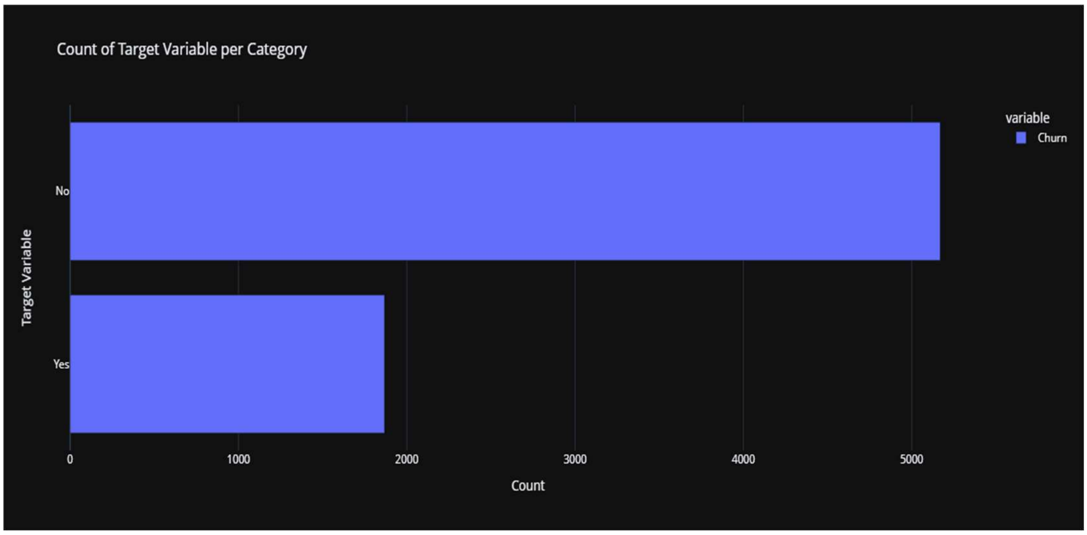
III. EXPLORATORY DATA ANALYSIS

EDA is an important step in the data analysis process, which involves analysing and collecting data for understanding and initial evaluation. EDA helps you understand the underlying structure of your data, identify patterns, identify anomalies, and develop hypotheses. This is usually the first step before doing further testing or building machine learning models. The ultimate aim of EDA is to prepare the data for further analysis or modelling and to generate insights that can inform decision-making. By performing EDA, we received some interesting and useful insights. For doing the same necessary libraries such as Pandas, NumPy etc. should be imported and most important is to read the dataset into a data frame.

REMOVAL OF CustomerID and gender

The CustomerID and gender were 2 features or columns of the telco dataset that are just ids given to customers and the gender column consisted for values male or female but they were not much useful or necessary for our analysis so we removed these two columns.

CHURNED CUSTOMERS VS NON-CHURNED CUSTOMERS



The above shown visualization is a Horizontal bar chart created using plotly library of python for making it interactive to visualize the count of churn variable. This helped us in understanding the distribution of churn across the dataset. We were able to obtain the following insights from this

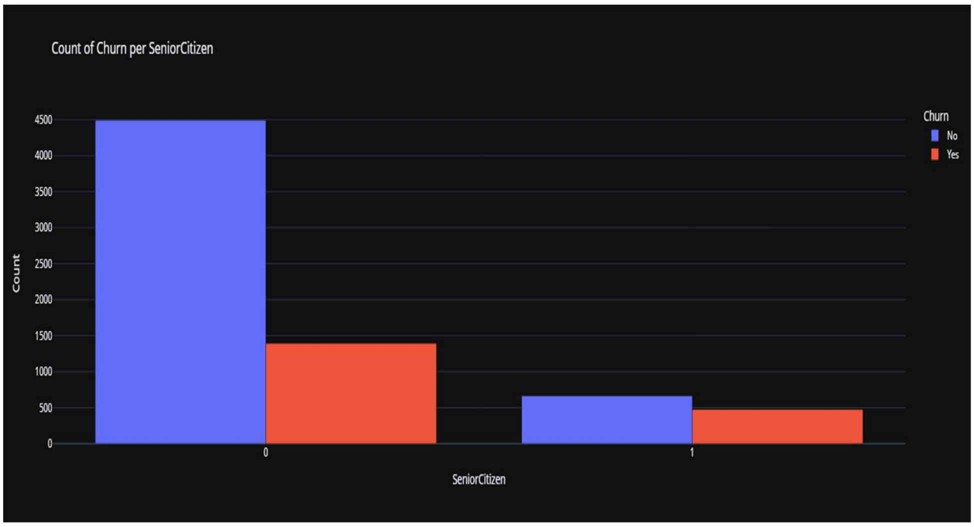
1. Among 7037 customers 1868 are Churned customers
2. 26.5% of the customers are churned ones
3. 73% remains unchurned

MISSING VALUE TREATMENT

In the initial stage of Exploratory Data Analysis, we couldn't find any missing values in the data so we decided to go deep into it. We noticed that the Total Charges column was not in numeric form, so we converted it to numeric. The following insight was obtained

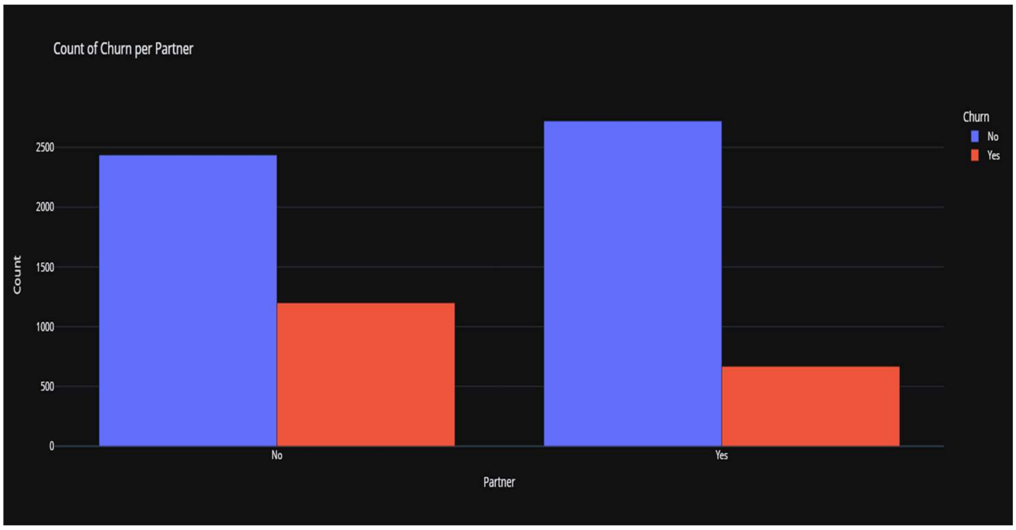
1. There were 11 missing values in the "Total Charges" column.
2. Since the percentage of missing values when compared to the whole data was less ie 0.15% it was safe to ignore them from further processing

COUNT OF CHURN PER SENIOR CITIZEN



The above given bar graph which was created using the plotly library of python shows that there are more non-Senior citizens than Senior citizens but the churn rate in Senior citizens are very high.

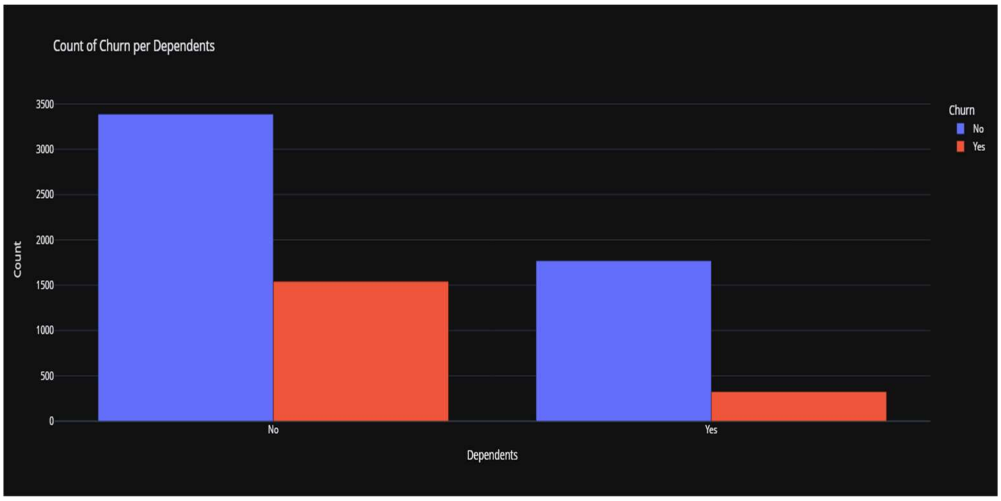
COUNT OF CHURN PER PARTNER



The insight obtained from this visualization is that

1. Churner Ratio is high when the customer is single or he/she does not have a partner.

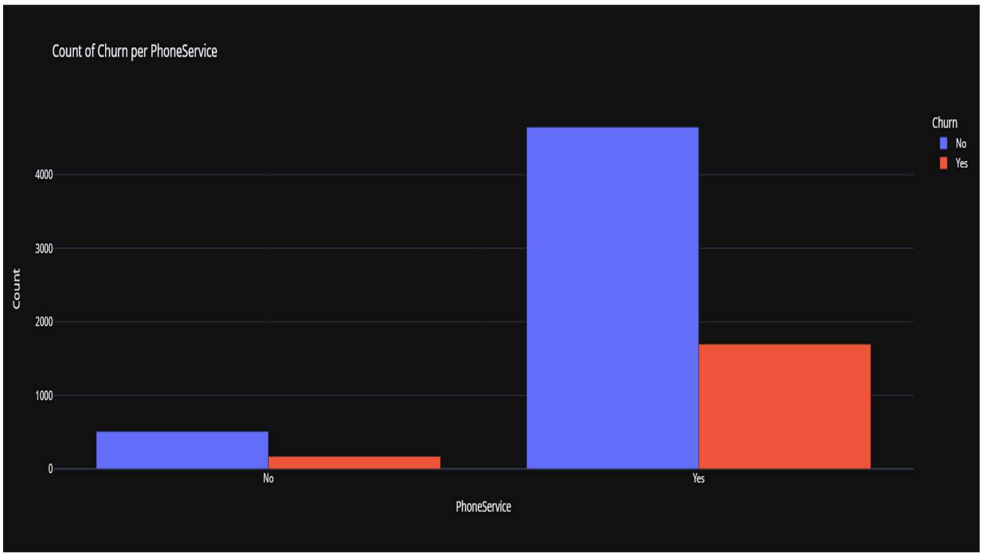
COUNT OF CHURN PER DEPENDENT



The insight obtained from this visualization is that

1. The customers with dependents have a low churn rate.
2. The partner churn rate and Dependent are negatively correlated.

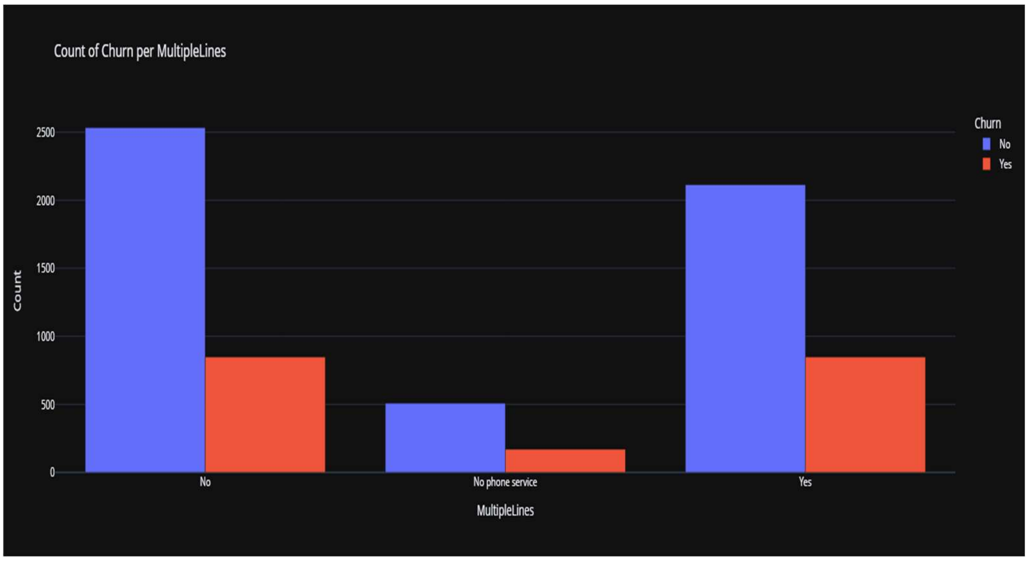
COUNT OF CHURN PER PHONE SERVICES



The insight obtained from this visualization is that

1.People with Phone Services are more likely to churn.

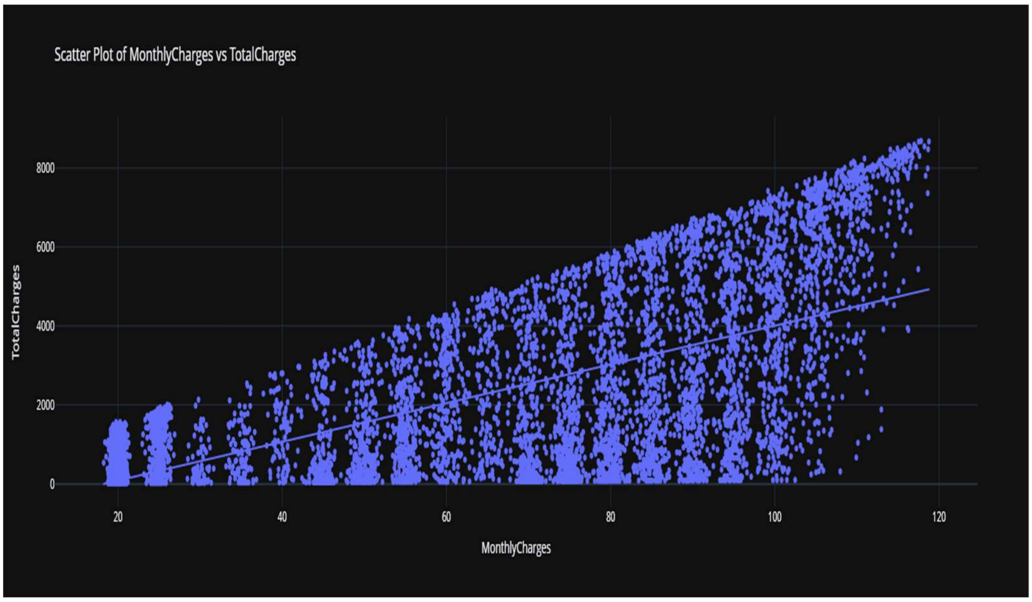
COUNT OF CHURN PER MULTIPLELINES



The insight obtained from this visualization is that

1. Customers that use payment mode as electronic checks are highest churners.

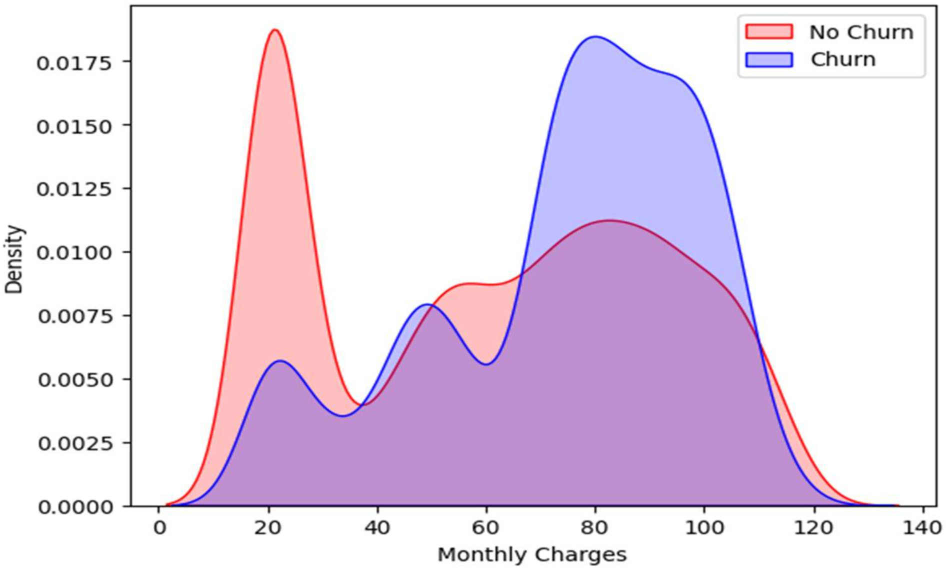
MONTHLY CHARGES VS TOTAL CHARGES



The above-given visualization is a Scatter plot, for finding the relationship between MonthlyCharges and TotalCharges. We were able to obtain the following insights,

1. Total Charges increase as Monthly Charges increase.

CHURN BY MONTHLY AND TOTAL CHARGES

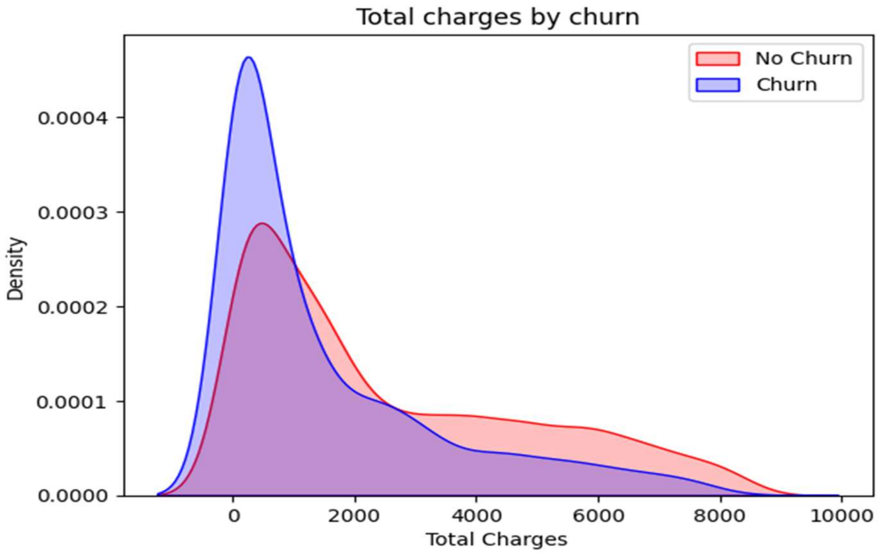


The above given visualization uses Seaborn to create a Kernel Density Estimate (KDE) plot to visualize the distribution of monthly charges for two different groups: customers who churned (Churn == 1) and customers who did not churn (Churn == 0).

The following insights were obtained

1. When monthly charges are high the churn rate is high.

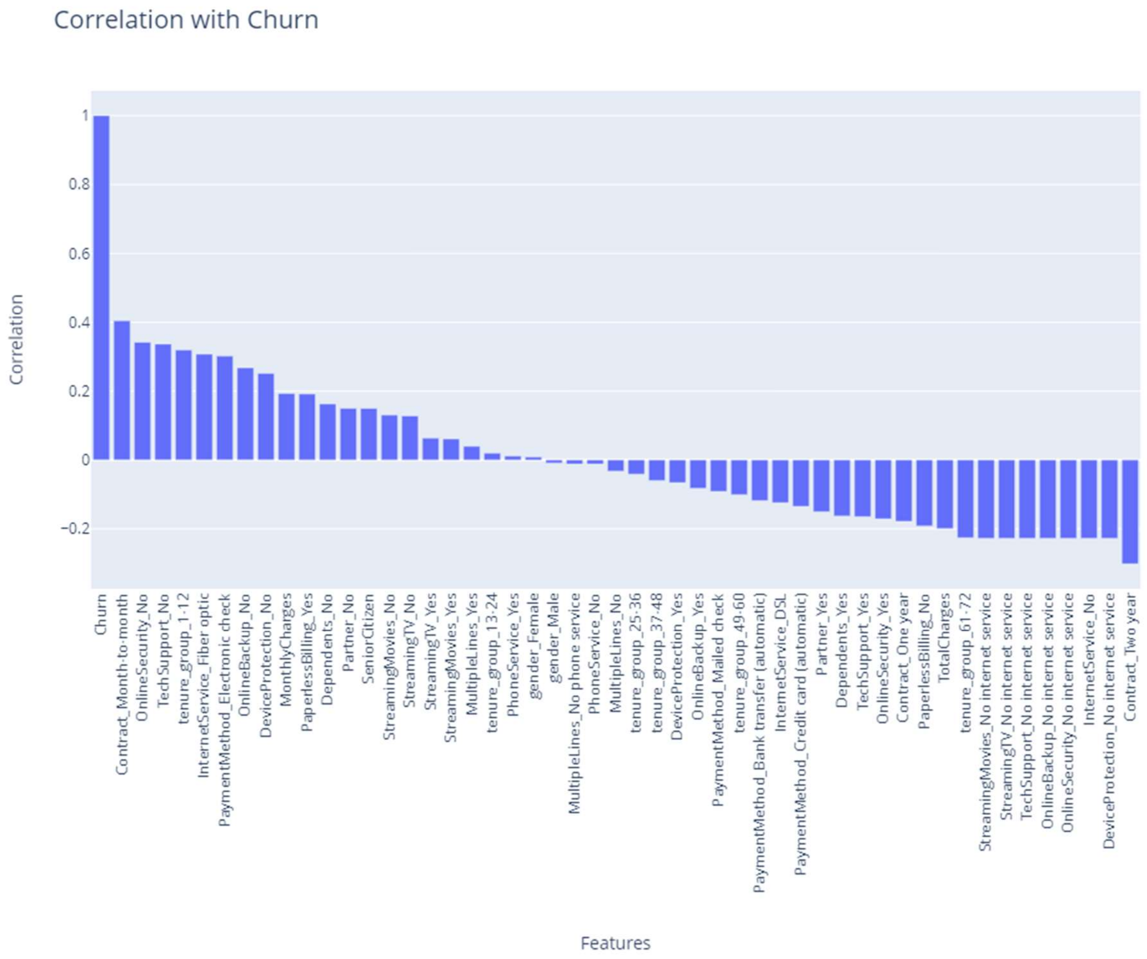
TOTAL CHARGES BY CHURN



The following insights were obtained from the above KDE graph

1. Higher churn at Lower TotalCharges
2. However, when we consider the insights from the combination of 3 parameters that are Tenure, TotalCharges and MonthlyCharges then the picture is a bit clear. Higher monthly charges at lower tenure results in lower total charges.
3. Therefore these 3 factors which are higher monthly charges, lower tenure and lower total charges are linked to higher Churn.

BUILDING CORRELATION FOR ALL PREDICTORS WITH CHURN



The above-shown visualization is created using Plotly to create a bar chart that visualizes the correlation between different features (variables) and the target variable 'Churn' . From this we were able to infer the following insights.

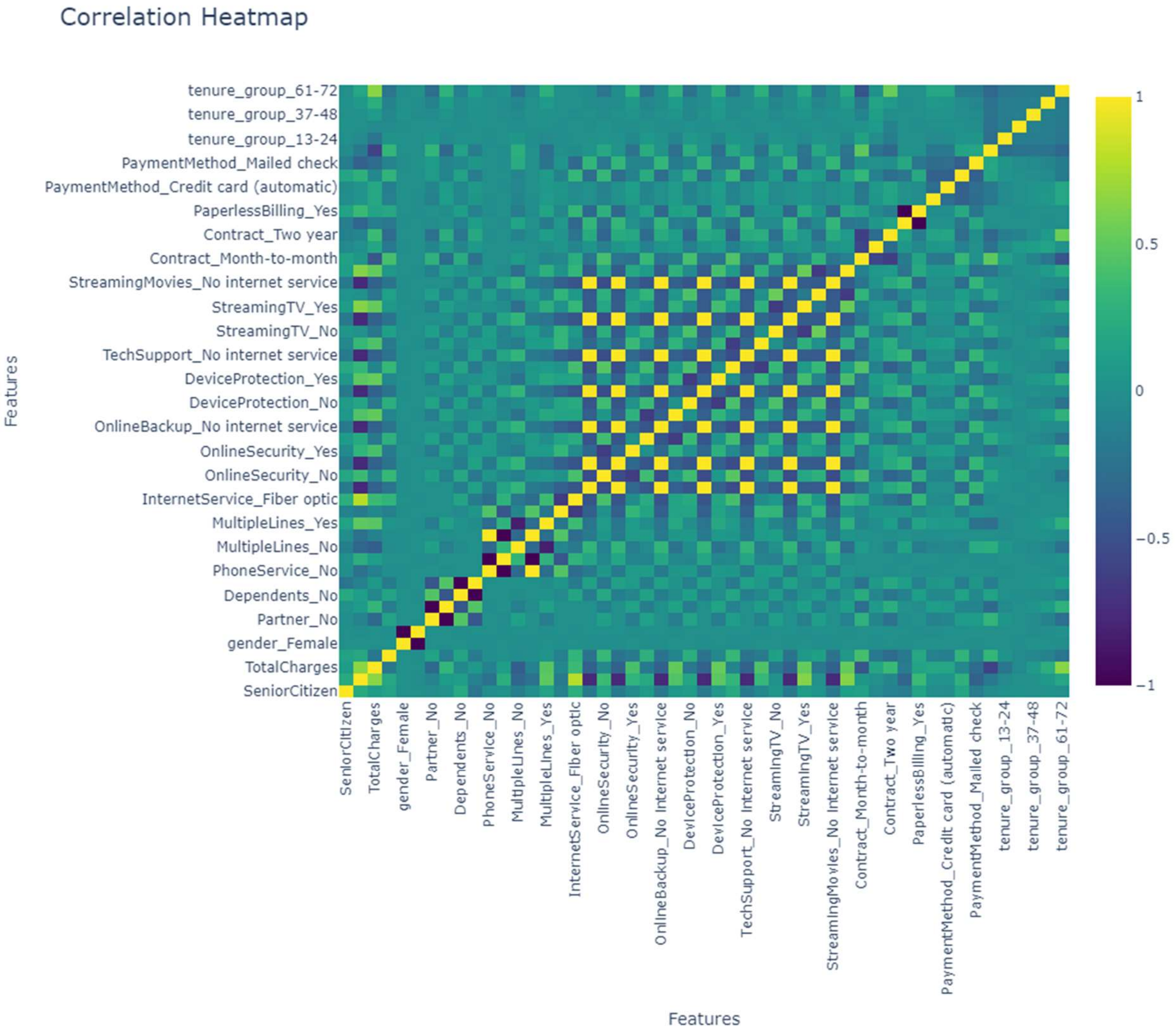
1.Higher churn is seen in the cases of Month-to-month contracts, With no online security, First year of subscription, and Fiber Optics internet.

2.Lower churn is seen in the cases of long-term contracts, Subscriptions without internet service, and customers who have been engaged for more than 5 years.

3.Factors such as Availability of Phone service and multiple connections has no impact on Churn.

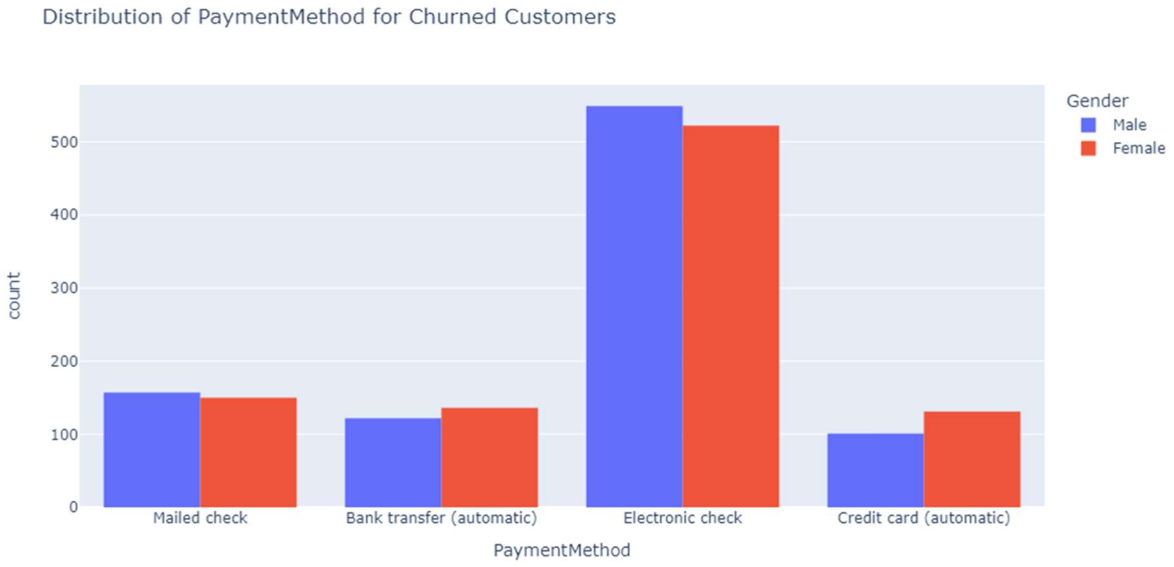
This all are evident from the heatmap below

Heatmap

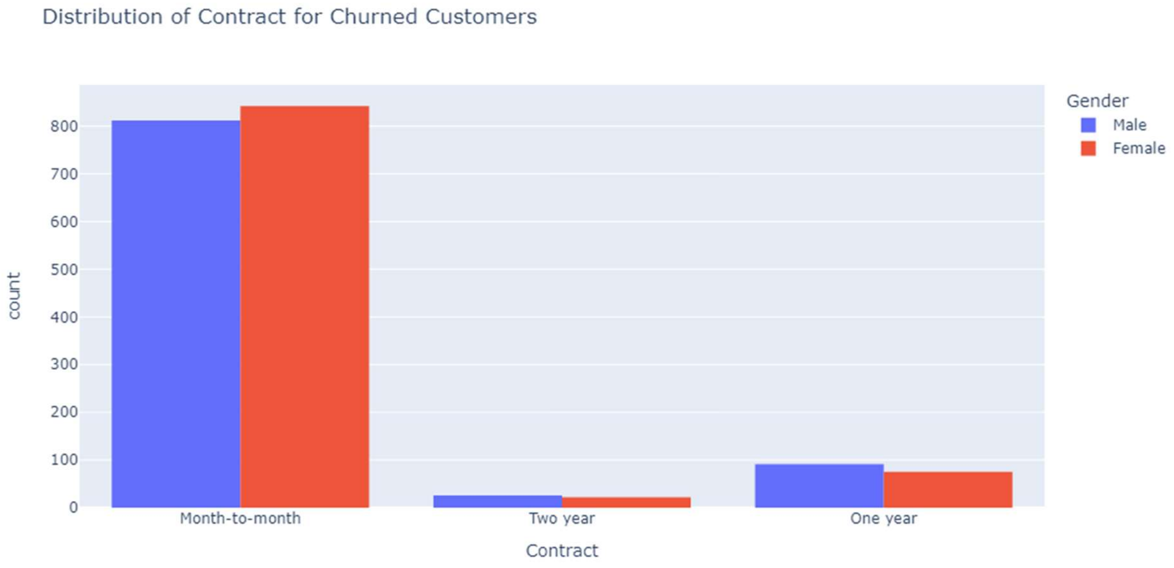


HISTOGRAMS

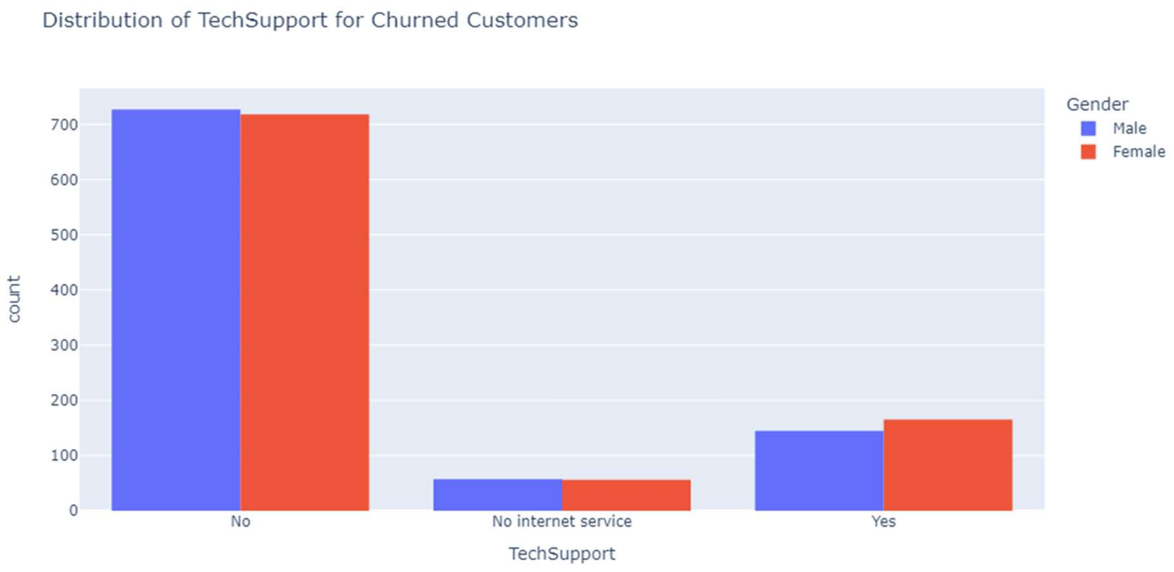
DISTRIBUTION OF PAYMENT METHODS OF CHURNED CUSTOMERS



DISTRIBUTION OF CONTRACT FOR CHURNED CUSTOMERS



DISTRIBUTION OF TECH SUPPORT FOR CHURNED CUSTOMERS



IMPORTANT INSIGHTS

1. Customers with Electronic cheque medium are highest churners.
2. Contract type - Monthly customers have more chances to churn because there is no contract terms as they are free-to-go customers.
3. Customers that have no online security and with no tech support are high churners.
4. Customers that are not senior citizens are high churners.

IV. FINDINGS AND DISCUSSION

Model Performance

Decision Tree Classifier

The following results were produced by the decision tree classifier:

- A 78% accuracy rate
- A Class 1 precision score of 60% for recognizing consumers who have left the company
- A recall rate of 43% for identifying clients who had left (Class 1)
- F1-score of 50% for Class 1 (churned customers)

These outcomes show that the Decision Tree Classifier's overall accuracy was good. However, as evidenced by its considerably lower recall score for Class 1, it showed certain difficulties in accurately identifying clients at danger of leaving.

Using a decision tree classifier and SMOTE

When class imbalance was addressed using SMOTE (Synthetic Minority Over-sampling Technique), the Decision Tree Classifier demonstrated considerable improvements:

- A stunning 93% accuracy rate
- A precision score of 91% for consumers who have left (Class 1)
- An outstanding 96% recall rate for correctly recognizing consumers who have left (Class 1)
- An F1-score of 94% for clients who have left (Class 1)

The model's accuracy and capacity to recognize clients who have abandoned their accounts have significantly improved as a result of the inclusion of SMOTE. This change makes it a better model for anticipating customer turnover in the telecom industry.

Random Forest Classifier

These were the outcomes of the Random Forest Classifier:

- 80 percent accuracy
- A 66% accuracy score (Class 1) for recognizing clients who have left the company
- Recall rates of 45% for Class 1 correctly identifying clients who have left
- An F1-score of 54% for clients who have left (Class 1)

Although it struggled to accurately identify clients at danger of leaving, as evidenced by its significantly lower recall score for Class 1, the Random Forest Classifier achieved commendable overall accuracy.

Random Forest Classifier (with SMOTE)

Significant improvements were obtained by using SMOTE to balance the dataset:

- A remarkable 94% accuracy rate
- A precision score of 93% for consumers who have left (Class 1)
- A Class 1 recall rate of 95% for correctly recognising clients who have left the company
- An F1-score of 94% for clients who have left (Class 1)

SMOTE was used, and as a result, the model's accuracy and capacity to identify churned consumers were significantly improved, making it the most effective model of those considered.

Principal Component Analysis (PCA)

The results were as follows after dimensionality was reduced using Principal Component Analysis (PCA):

- A 73% accuracy percentage

- An accuracy rating of 74% for classifying clients who have left (Class 1)
- An 81% recall rate for identifying clients who had left (Class 1)
- An F1-score of 77% for clients who have left (Class 1)

The lack of superior outcomes with PCA suggests that dimensionality reduction may not be helpful in this particular case. It is critical to think about the features of the dataset and how dimensionality reduction methods will affect them.

Discussion

The results of this experiment on predicting customer attrition in the telecom industry offer the following notable conclusions:

SMOTE for Handling Unbalanced Data: SMOTE's implementation significantly enhanced the performance of both the Decision Tree and Random Forest Classifiers. By oversampling the minority class, SMOTE successfully reduced the problem of unbalanced data, resulting in increased accuracy and better recall for consumers who had left the company.

In terms of accuracy, precision, and recall, Random Forest outperforms Decision Tree: Overall, the Random Forest Classifier showed greater performance to the Decision Tree Classifier. This emphasises how effective ensemble approaches are for predictive modelling.

Impact of Dimensionality Reduction: Principal Component Analysis (PCA) did not produce better findings, indicating that lowering the dimensionality of the data would not be helpful for this particular issue. It is essential to thoroughly assess the nature of the dataset and the effects of dimensionality reduction methods.

Business Implications: For telecom firms, the high recall obtained with the Random Forest Classifier (with SMOTE) is quite valuable. As a result, more customers who are at risk of leaving can be correctly recognised, allowing for more effective retention strategies.

Based on the data, it can be concluded that the Random Forest Classifier with SMOTE is the best model for predicting customer turnover in the telecom sector. To improve prediction performance, though, more research and adjustment may be required. When developing a churn prediction system in a real-world telecom context, it is crucial to take into account the cost-benefit trade-offs related to various models.

V. CONCLUSION

We applied machine learning approaches in our project, which was aimed at anticipating telecom customer churn, to address the crucial problem of client retention. The Random Forest Classifier, strengthened by SMOTE to balance class distribution, won out as the most effective model, reaching a noteworthy 94% accuracy and strong recall performance for consumers who had left. This methodology gives telecom companies a strong tool for proactively locating and retaining consumers who are at risk of leaving.

In addition, we discovered that Principal Component Analysis (PCA) only produced modest improvements, highlighting the significance of careful feature selection and a thorough comprehension of the dataset. These insights highlight the crucial role that data-centric tactics play in reducing churn, increasing customer satisfaction, and boosting financial viability.

Our study offers telecom businesses a useful road map for reducing customer churn, boosting loyalty, and ultimately boosting financial success. This investigation marks a significant step forward in the fight against customer churn in the telecom sector by providing practical answers that may be applied to other sectors dealing with related problems.

VI. REFERENCES

1.Churn Prediction: A Comparative Study Using KNN and Decision Trees by Mohammad A. Hassonah; Ali Rodan; Abdel-Karim Al-Tamimi; Jamal Alsakran

2. Telecom churn prediction and used techniques, datasets and performance measures by Hemlata Jain, Ajay Khunteta & Sumit Srivastava

3. Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network by: Xin Hu; Yanfei Yang; Lanhua Chen; Siru Zhu

4. Churn Prediction in Telecommunication using Logistic Regression and Logit Boost by Hemlata Jain, Ajay Khunteta, Sumit Srivastava

5. Systematic Review on Churn Prediction Systems in Telecommunications by Gireen Naidu, Tranos Zuva & Elias Mmbongeni Sibanda

6. Telecom Customer Churn Prediction by Mehul Bhargava, Shruti Singh, Jaya Sharma & D. Franklin Vinod