Department of Computer Science

Practical Business Analytics (COMM053) Coursework

Title: Customer churn prediction: Telecom Churn Dataset

Group Number: 19

ID	Name	Email
6768795	Abhishek Gadekar	ag02134@surrey.ac.uk
6801243	Anshuman Yadav	ay00471@surrey.ac.uk
6838674	Saquib Naseem	sn01272@surrey.ac.uk
6826003	Prajakta Kokare	pk00885@surrey.ac.uk
6840329	Yipeng Zhu	yz02422@surrey.ac.uk

Table of Contents

Table of Contents

1.Introduction

Problem Statement:

In the rapidly evolving landscape of telecommunications, customer churn poses a significant challenge for service providers. As subscribers have an array of choices at their disposal, predicting and mitigating churn has become a crucial focus for telecom companies aiming to maintain a loyal customer base. Machine learning techniques offer a promising solution to address this issue by providing predictive models that can anticipate customer churn based on historical data.

This machine learning project centers around the Telecom Churn Dataset, a rich repository of information encompassing various aspects of customer interactions with a telecom service provider. The dataset comprises features such as call duration, usage patterns, customer demographics, and service-related metrics. Leveraging this dataset, our objective is to develop a robust predictive model that can effectively identify customers at risk of churning.

By harnessing advanced machine learning algorithms, we aim to create a model that not only accurately predicts churn but also provides valuable insights into the key factors influencing customer decisions. Such insights can empower telecom companies to proactively address issues, enhance customer satisfaction, and implement targeted retention strategies.

Throughout this project, we will explore the dataset, preprocess the data to extract relevant features, and train and evaluate machine learning models. Our goal is to deliver a predictive model that can assist telecom companies in reducing churn rates, thereby optimizing customer retention efforts and ensuring sustainable business growth.

This endeavor showcases the transformative potential of machine learning in the telecommunications industry, enabling providers to stay ahead of the curve by anticipating customer behavior and fostering long-lasting relationships with their subscribers.

Business Objectives:

The primary business objective of the Telecom Customer Churn Prediction dataset is to identify customers who are at risk of churning or discontinuing their service with the telecommunications provider.
The project aims to improve customer retention, reduce customer acquisition costs, increase customer lifetime value, optimize resource allocation, and enhance customer satisfaction.
Data source:

We are working on Telecom churn prediction dataset. The dataset contains details of customers such as voice plan, international plan, total minutes and customer service calls which are used to predict the customer attrition as it is one of the ley business metrics because the cost of retaining an existing customer is far less than acquiring a new one. This dataset consists primarily of 1 table with 21 attributes in the form of 'CSV – Comma Separated Values' and is sourced from Kaggle.

Project Aim:

- Improved Customer Retention:

This refers to the ability to keep customers engaged and subscribed to services over an extended period.
The project aims to identify factors leading to customer churn and address them proactively.
Predictive models can be used to identify at-risk customers, allowing the company to intervene with targeted retention strategies (like personalized offers or improved service) before these customers decide to leave.

- Reduced Customer Acquisition Costs:

Acquiring new customers is often more expensive than retaining existing ones. By reducing churn, the company inherently reduces the need to constantly acquire new customers to maintain revenue.
The project will help in understanding the characteristics of loyal customers, which can be used to fine-tune marketing strategies, making them more efficient and cost-effective.
Targeted customer acquisition can be achieved by focusing on prospects who share characteristics with the company's most stable customer base.

- Increased Customer Lifetime Value (CLV):

CLV is the total revenue a business can expect from a single customer account throughout their relationship with the company.
By improving retention rates, customers remain with the company for a longer period, thereby increasing the revenue they generate over time.
The project can also focus on upselling or cross-selling strategies to existing customers, which is easier and less costly than selling to new customers, further increasing CLV.

- Optimized Resource Allocation:

Understanding customer behavior and predicting churn allows for better allocation of resources.
Instead of spreading resources thinly over numerous initiatives, the company can focus on areas with the highest return on investment, such as targeted customer service improvements or tailored loyalty programs.
Predictive analytics will enable the company to prioritize actions that have the greatest impact on customer satisfaction and retention, thereby using resources more effectively.

- Enhanced Customer Satisfaction:

This involves understanding and meeting, or exceeding, customer expectations.
Data analytics can reveal insights into customer preferences, pain points, and expectations. By addressing these factors and providing personalized experiences, the company can significantly enhance customer satisfaction.
Satisfied customers are more likely to remain loyal, provide positive referrals, and contribute to a strong brand reputation.
Overall, the project aims to create a holistic approach to customer management by utilizing data-driven insights. This approach not only improves operational efficiency but also drives sustainable growth by building a more loyal and satisfied customer base.

Data Overview:

It consists of information about cell phone usage for a variety of customers, including factors like account length, customer service calls, international plan, voice mail plan, and total minutes used.

❖ The target variable is whether the customer has churned (i.e., stopped using the service).

❖ This dataset is valuable for understanding the factors that contribute to customer churn and for developing predictive models to identify customers who are at risk of churning.

Data understanding:

- Rows in training dataset: 2666
- Rows in Test Dataset: 667
- Number of features: 20
- Number of Categorical attributes: 4 (State, International plan, Voice mail plan, Churn)

The Features given in the data are as below:

No.	Features	Dataset
1	State	Object (String)
2	Account length	Integer
3	Area Code	Integer
4	International plan	Object (String)
5	Voice mail plan	Object (String)
6	Number Voice mail plan	Integer
7	Total day minutes	Double
8	Total day calls	Integer
9	Total day charge	Double
10	Total evening minutes	Double

11	Total evening calls	Integer
12	Total evening charges	Double
13	Total night minutes	Double
14	Total night calls	Integer
15	Total night charge	Double
16	Total international minutes	Double
17	Total international calls	Integer
18	Total international charge	Double
19	Customer service calls	Integer
20	Churn	Object (String)

Correlation between Churn and other features:

The bar chart presented illustrates below the correlation of various factors with customer churn in a telecommunications dataset. Correlation coefficients range from -1 to 1, where values closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, and values around 0 suggest no correlation.

At the forefront, "Customer Service Calls" shows the strongest positive correlation with churn, suggesting that the more customer service calls made, the higher the likelihood of a customer terminating their service. This could be an indicator of customer dissatisfaction or unresolved issues. The total day minutes and total day charge both have the next highest positive correlations, which could imply that higher usage or higher charges during the day are significant predictors of churn.

Interestingly, other factors such as total evening minutes and charges, total night minutes and charges, and total international calls have lower positive correlations. This indicates that while they do play a role in predicting churn, they are not as influential as the number of customer service calls or day usage metrics.

The total number of calls (regardless of time of day) has an even lesser positive correlation, pointing towards the number of calls being a less significant predictor than the duration or cost of calls.

On the other end of the spectrum, factors like account length, area code, total voice calls, and the number of voicemail messages seem to have a negligible correlation with churn. This could mean that the duration of the customer's relationship with the company, their geographical location, their use of voice service, or their voicemail usage are not strong indicators of whether they will stop using the telecom services.

Overall, the chart provides valuable insights into which factors might be most effective to target in customer retention strategies. The telecom company could focus on improving customer service experiences and managing day-time charges and usage to mitigate churn.

Churn Customer	Non Churn Customer
Shorter account length	Longer account length
More customer service calls	Less customer service calls
Fewer voicemail message	More Voicemail message
Fewer minutes during day, evening, night	More minutes during day, evening, night
More no. of calls with less duration of calls	Fewer no. of calls with less duration of calls

2.Data Pre-processing

1. Handling Missing data:

The first stage of data-preprocessing involved the identification of missing values within out dataset, which is crucial to ensure the completeness and reliability of the data:

- Numerical Columns: For the columns with numerical value, we handled the missing values by replacing them with the columns mean values ensuring that the overall distribution and central tendencies of the dataset remain consistent.

- Categorical Columns: In the case of categorical data, encompassing columns with text or distinct categories, the strategy was to replace missing values with the mode. This method utilizes the most common category within each column, thereby preserving the inherent characteristics and frequencies of the categorical data.

2. Assurance of Data Integrity:

After the handling of missing data, a thorough secondary review was conducted. This step serves as a quality check to confirm that all missing values were appropriately dealt with. Furthermore:

- Duplicate Entry Removal: An essential aspect of this stage was the identification and elimination of duplicate entries. By doing so, the uniqueness of each data point is upheld, which is fundamental for the accuracy of any subsequent analyses.

III. Statistical Overview:

An integral part of our pre-processing was the generation of a statistical overview for the numerical aspects of the dataset:

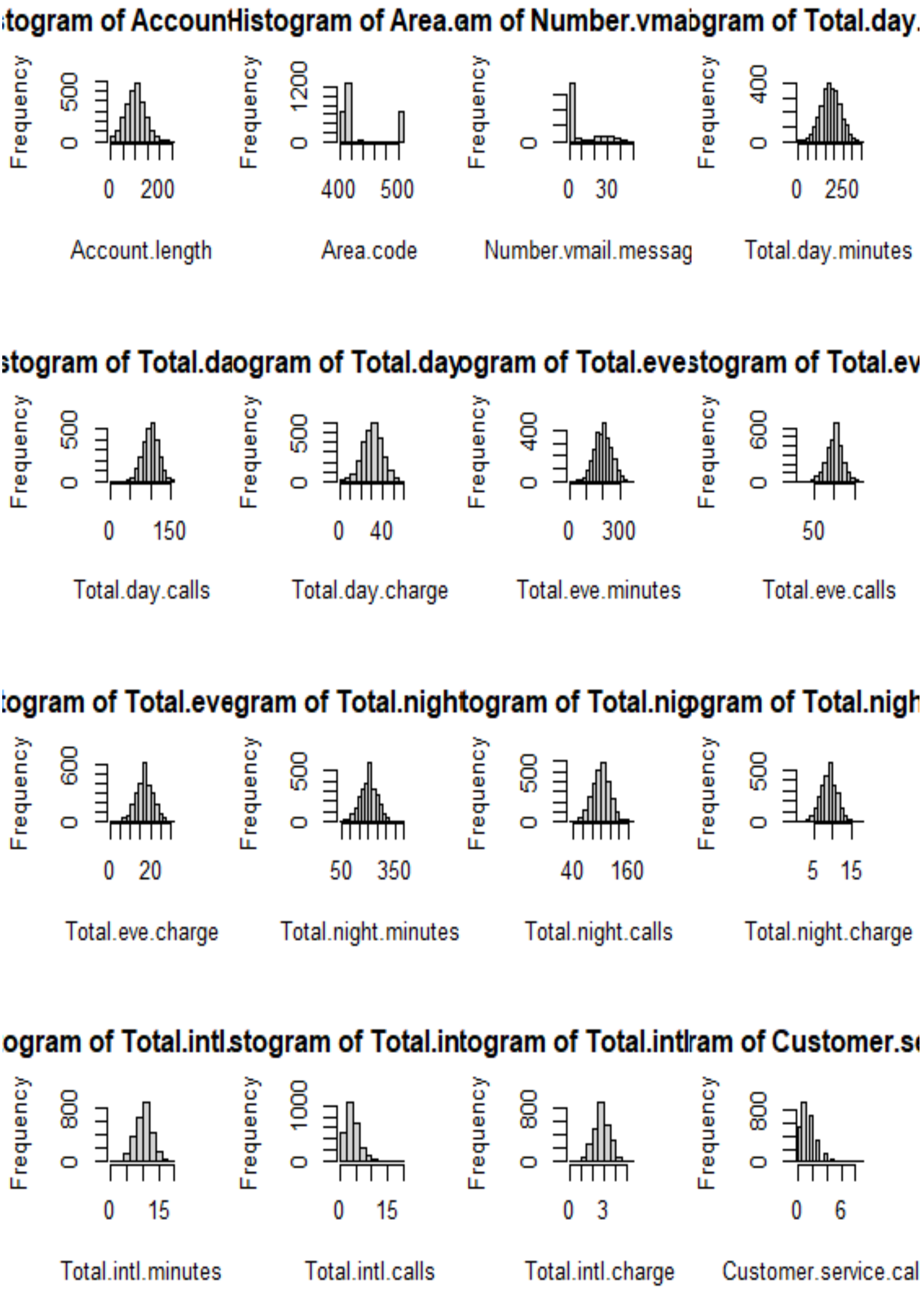
- Descriptive Statistics: Utilizing the data.describe() function, we compiled a comprehensive statistical summary. This included key metrics like the mean, standard deviation, and the range (encompassing both minimum and maximum values) for each numerical column.
- Examination of Key Variables: Specific variables, such as 'Total day minutes' and 'Total day charge', were given particular attention to glean deeper insights. This involved exploring their distributions and central tendencies in detail, providing valuable context for the dataset.

IV. Visualization of Key Variables:

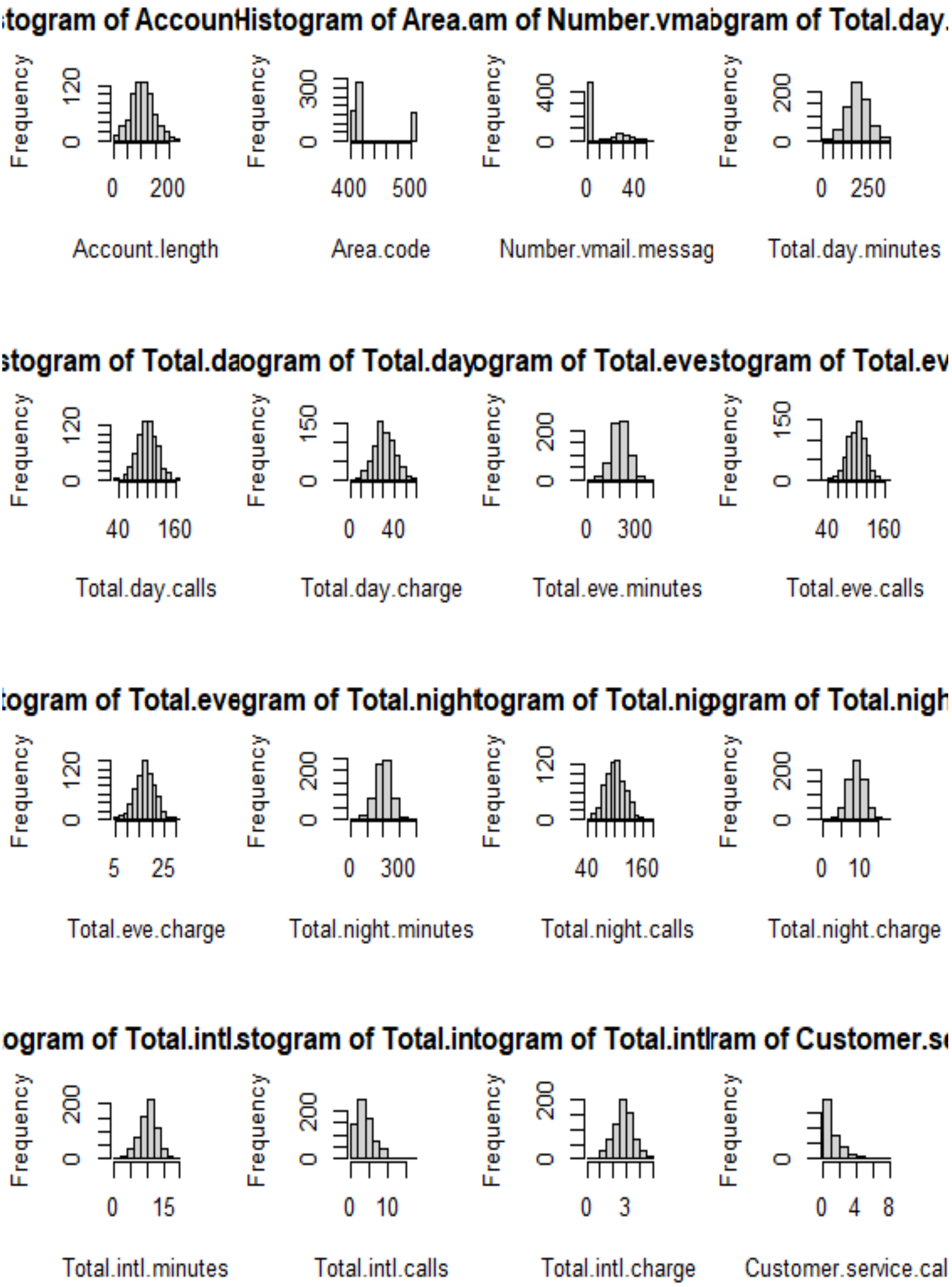
To better comprehend the dataset's structure and dynamics, various visualization techniques were employed:

- Histograms: These were plotted for crucial variables to visually represent the distribution of data across different value ranges. Histograms are instrumental in understanding the frequency and spread of data points.

V. Analyzing Relationships Between Variables:
Understanding the interplay between different variables is vital for any comprehensive data analysis:



togram of AccounHistogram of Area.am of Number.vmabgram of Total.day.
Frequency
0
500
0
200
Account.length
Frequency
0
1200
400
500
Area.code
Frequency
0
0
30
Number.vmail.messag
Frequency
0
400
0
250
Total.day.minutes
stogram of Total.daogram of Total.dayogram of Total.evestogram of Total.ev
Frequency
0
500
0
150
Total.day.calls
Frequency
0
500
0
40
Total.day.charge
Frequency
0
400
0
300
Total.eve.minutes
Frequency
0
600
50
Total.eve.calls
togram of Total.evegram of Total.nightogram of Total.nigpgram of Total.nigh
Frequency
0
600
0
20
Total.eve.charge
Frequency
0
500
50
350
Total.night.minutes
Frequency
0
500
40
160
Total.night.calls
Frequency
0
500
5
15
Total.night.charge
ogram of Total.intlstogram of Total.intogram of Total.intlram of Customer.se
Frequency
0
800
0
15
Total.intl.minutes
Frequency
0
1000
0
15
Total.intl.calls
Frequency
0
800
0
3
Total.intl.charge
Frequency
0
800
0
6
Customer.service.cal


togram of AccounHistogram of Area.am of Number.vmabgram of Total.day.
Frequency
120
0
0
200
Account.length
300
0
400
500
Area.code
400
0
0
40
Number.vmail.messag
200
0
0
250
Total.day.minutes
stogram of Total.daogram of Total.dayogram of Total.evestogram of Total.ev
120
0
40
160
Total.day.calls
150
0
0
40
Total.day.charge
200
0
0
300
Total.eve.minutes
150
0
40
160
Total.eve.calls
togram of Total.evegram of Total.nightogram of Total.nigpgram of Total.nigh
120
0
5
25
Total.eve.charge
200
0
0
300
Total.night.minutes
120
0
40
160
Total.night.calls
200
0
0
10
Total.night.charge
ogram of Total.intlstogram of Total.intogram of Total.intlram of Customer.se
200
0
0
15
Total.intl.minutes
200
0
0
10
Total.intl.calls
200
0
0
3
Total.intl.charge
0
0
4
8
Customer.service.cal

Correlation between Churn and other features:

The bar chart presented illustrates below the correlation of various factors with customer churn in a telecommunications dataset. Correlation coefficients range from -1 to 1, where values closer to 1 indicate a strong positive correlation, values closer to -1 indicate a strong negative correlation, and values around 0 suggest no correlation.

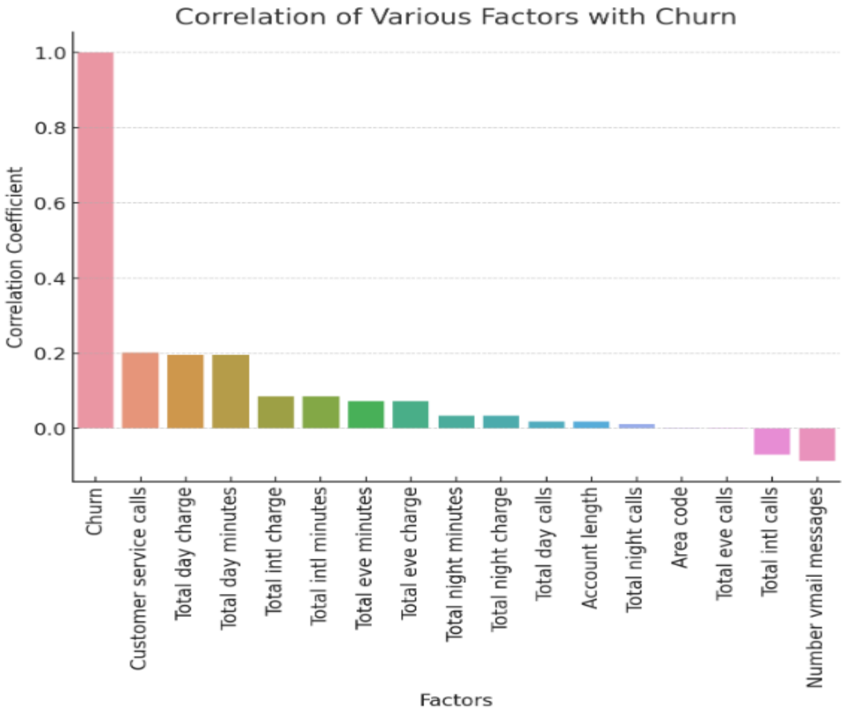
At the forefront, "Customer Service Calls" shows the strongest positive correlation with churn, suggesting that the more customer service calls made, the higher the likelihood of a customer terminating their service. This could be an indicator of customer dissatisfaction or unresolved issues. The total day minutes and total day charge both have the next highest positive correlations, which could imply that higher usage or higher charges during the day are significant predictors of churn.

Interestingly, other factors such as total evening minutes and charges, total night minutes and charges, and total international calls have lower positive correlations. This indicates that while they do play a role in predicting churn, they are not as influential as the number of customer service calls or day usage metrics.

The total number of calls (regardless of time of day) has an even lesser positive correlation, pointing towards the number of calls being a less significant predictor than the duration or cost of calls.

On the other end of the spectrum, factors like account length, area code, total voice calls, and the number of voicemail messages seem to have a negligible correlation with churn. This could mean that the duration of the customer's relationship with the company, their geographical location, their use of voice service, or their voicemail usage are not strong indicators of whether they will stop using the telecom services.

Overall, the chart provides valuable insights into which factors might be most effective to target in customer retention strategies. The telecom company could focus on improving customer service experiences and managing day-time charges and usage to mitigate churn.



VI. Transformation of Categorical Data:

The adaptation of categorical data for analytical models formed a key part of the pre-processing:

- This process involved converting categorical values into a binary (0s and 1s) format, creating separate columns for each unique category using factor(). Such a transformation is essential for the compatibility of categorical data with various machine learning algorithms.

-We have four categorical features such as state, international plan, voice mail plan and churn

-Out of which the feature state was irrelevant for our analysis hence it was dropped, and rest of the values were encoded as binary numerical values.

VII. Feature Scaling:

We have normalized the numerical features to the scale (-1 to 1) to bring them in same scale for ease of coding. Foe which we have used the scale() in R.

VII. Feature Engineering:

There we 6 features in the dataset that were either irrelevant or highly correlated according to our correlation matrix. They were such as State, Area code, Total day charge, Total eve charge, Total night charge, Total international charge.

So originally there were 20 features and after dropping there are 14 features left for analysis.

Outliers:

Outliers in a dataset are data points that deviate significantly from the general pattern and can distort statistical analyses. These anomalies, often resulting from errors or extreme values, have the potential to skew overall interpretations and conclusions. Identifying outliers is crucial for ensuring the accuracy and reliability of statistical analyses, with common methods including the use of measures like the Interquartile Range (IQR). By defining a threshold (usually 1.5 times the IQR) to flag observations as outliers, analysts can pinpoint and address these data points appropriately. Handling outliers may involve their removal, data transformation, or the application of robust statistical techniques to mitigate their impact, ultimately enhancing the robustness and validity of statistical findings. We used mean imputation methods for removing numeric outliers.

Converting Data format:

There were some features in our data such as Total day minutes which changed to decimal values after handling the missing data. This data was formatted them to nearest integer value.

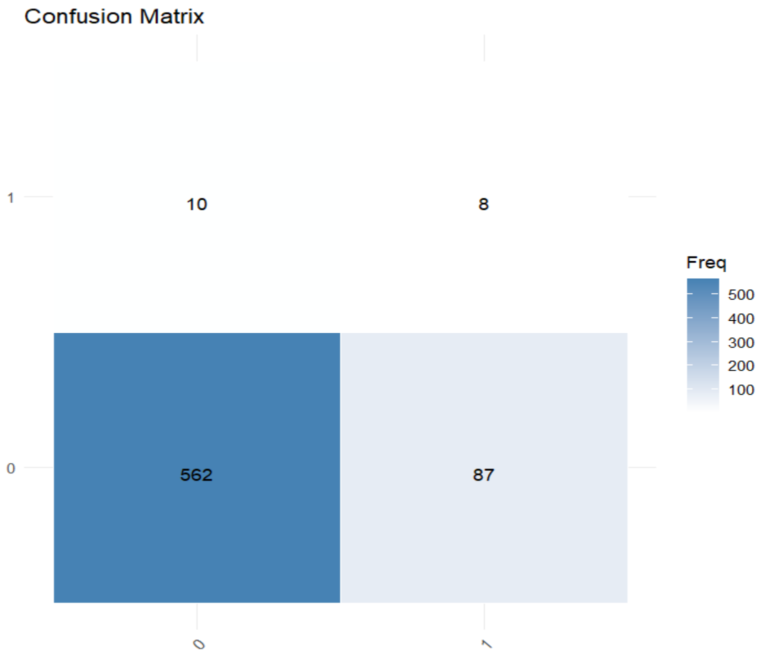
Balancing:

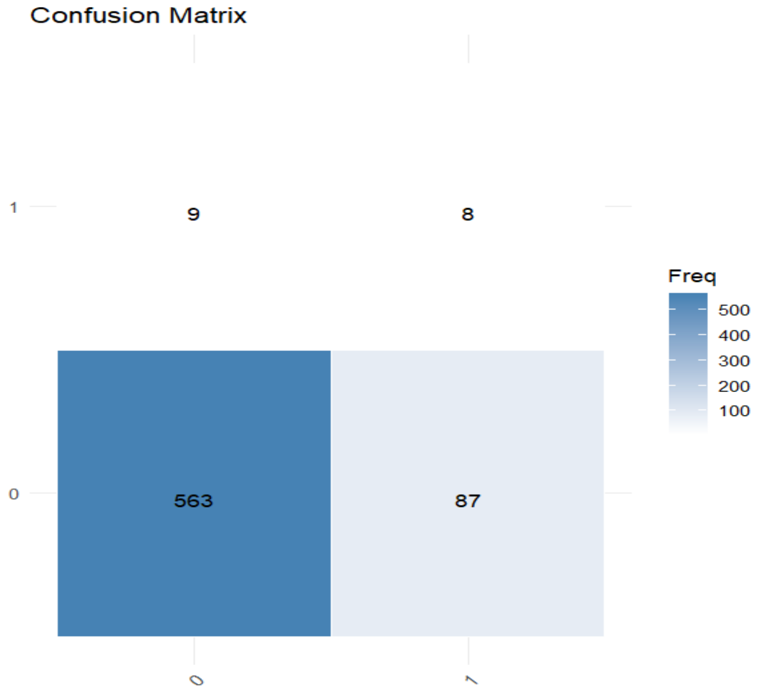
Under-sampling is a technique employed in imbalanced classification scenarios to address the disproportionate representation of different classes in a dataset, particularly prevalent in tasks such as fraud detection or medical diagnosis. This method involves reducing the size of the majority class by randomly removing instances from that class until a more balanced distribution is achieved between the minority and majority classes. By equalizing the class sizes, under-sampling aims to prevent the model from being biased toward the majority class and ensures that the algorithm considers both classes equally during training. While under-sampling helps in achieving better class balance, it comes with the trade-off of potentially discarding valuable information from the majority class. Careful consideration of the dataset characteristics and the specific goals of the analysis is necessary to determine the most suitable approach for addressing class imbalance, which may include a combination of under-sampling, over-sampling, or more advanced techniques such as synthetic data generation.

3.Data Modeling:

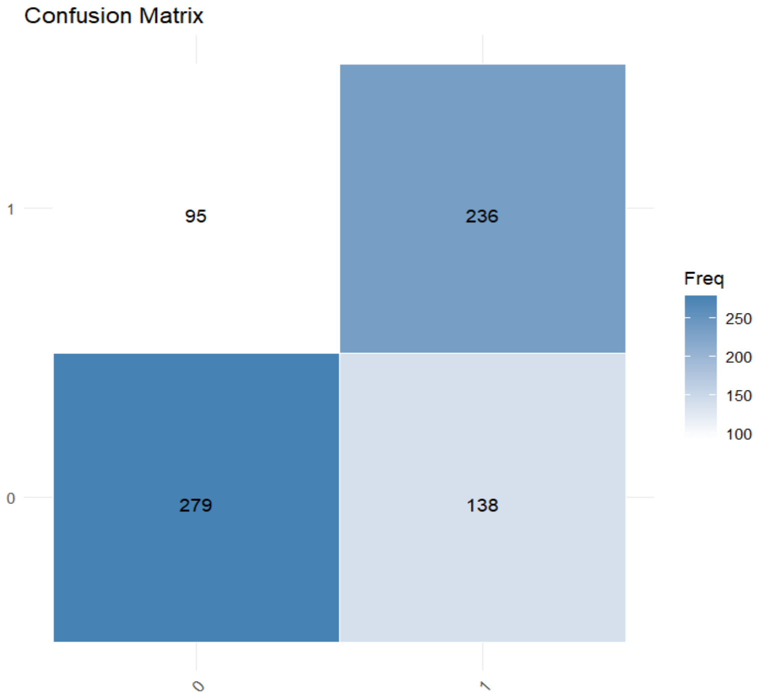
Logistic regression:

A logistic regression model was applied to predict telecom churn, displaying high initial accuracy (85.75%) that dipped slightly after removing correlation (85.45%) and more substantially post-adjustment of balancing (68.85%). Evaluated using a confusion matrix and ROC curve, the model, while potent, suggests a need for additional metrics for thorough analysis and raises concerns about potential overfitting. Recommendations for future applications include further optimization and validation to ensure robust and generalizable performance across various datasets in the telecom sector.





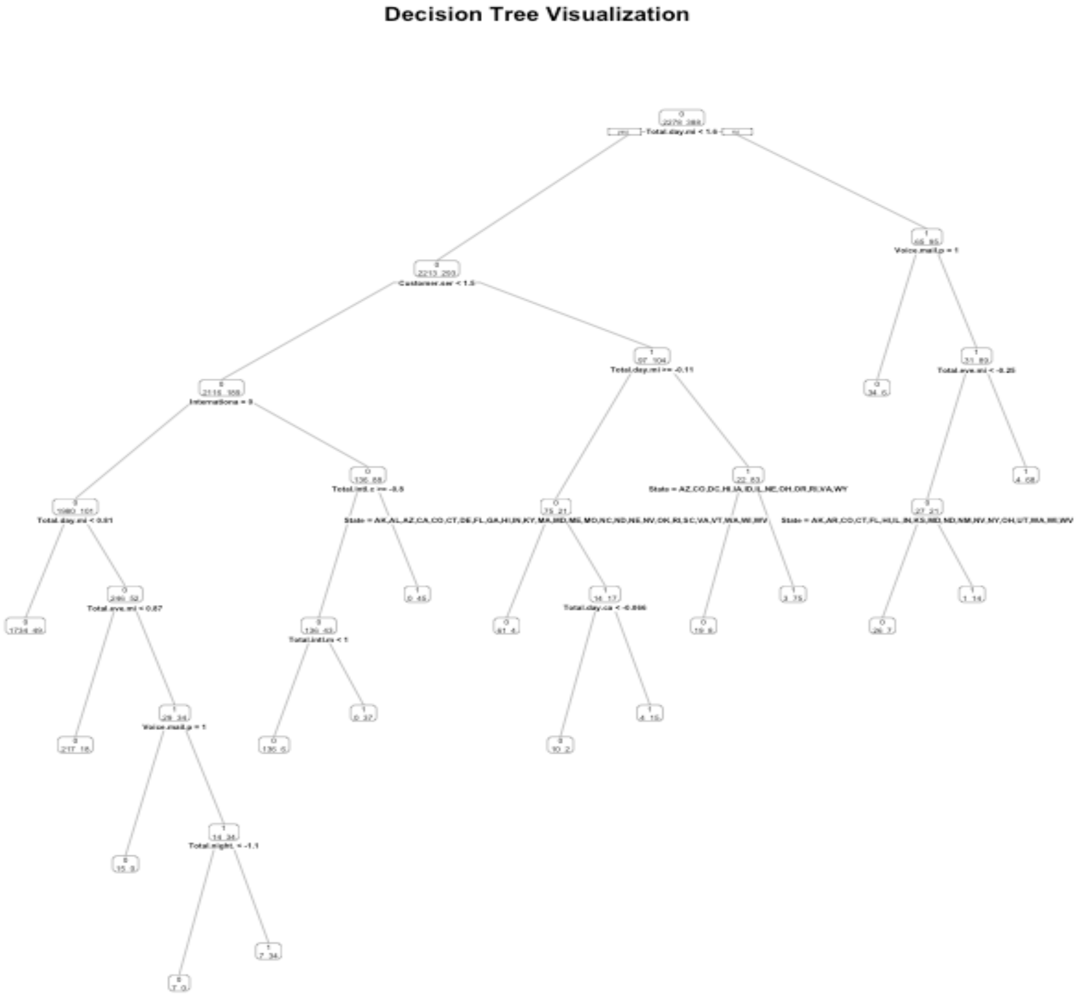
Confusion Matrix



Confusion Matrix

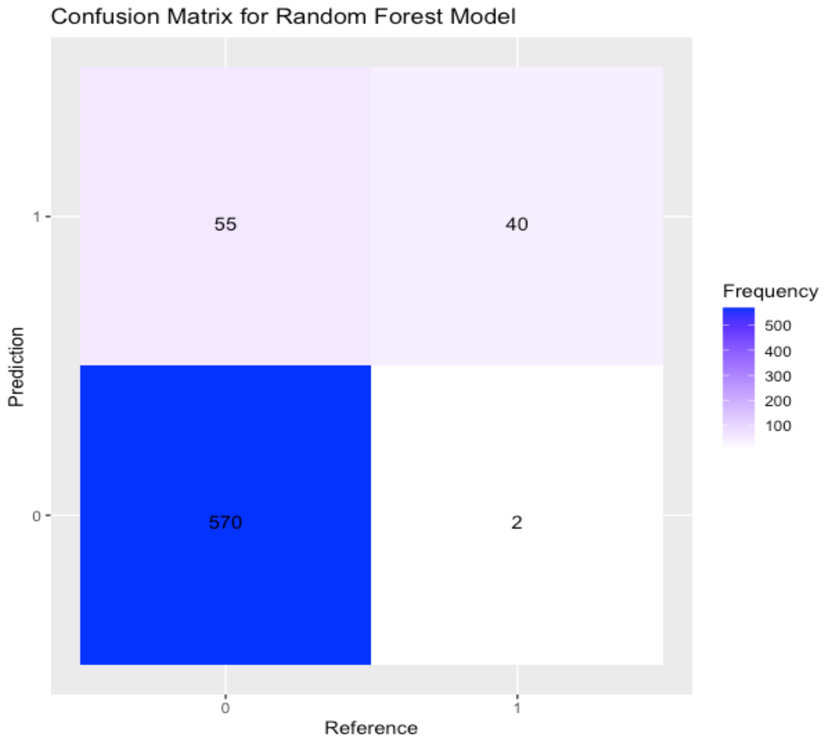
Decision tree:

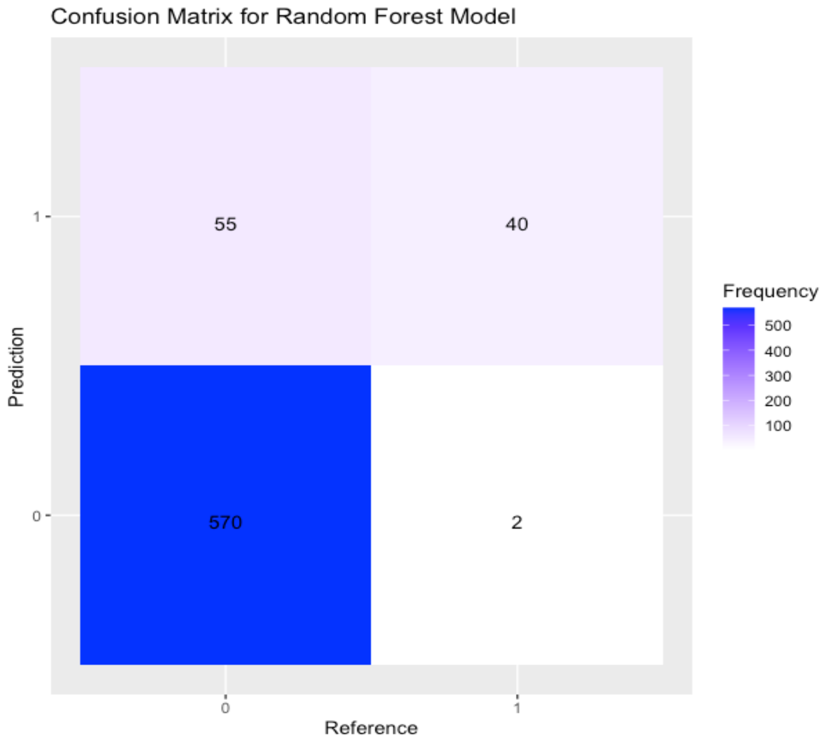
The data undergoes preprocessing, including factor encoding of categorical variables and scaling of numerical features. The Decision Tree model, built using the rpart package in R, is trained on an 80% split of the dataset and evaluated on the remaining 20%. The initial model accuracy is reported at 85.45%. After adjusting for feature correlations, accuracy improves slightly to 86.30%. Subsequent data balancing, which aims to mitigate class imbalance, results in a lower accuracy of 70.65%. Despite this decrease, the balancing step is crucial for model robustness. The decision tree is also visualized for interpretability.

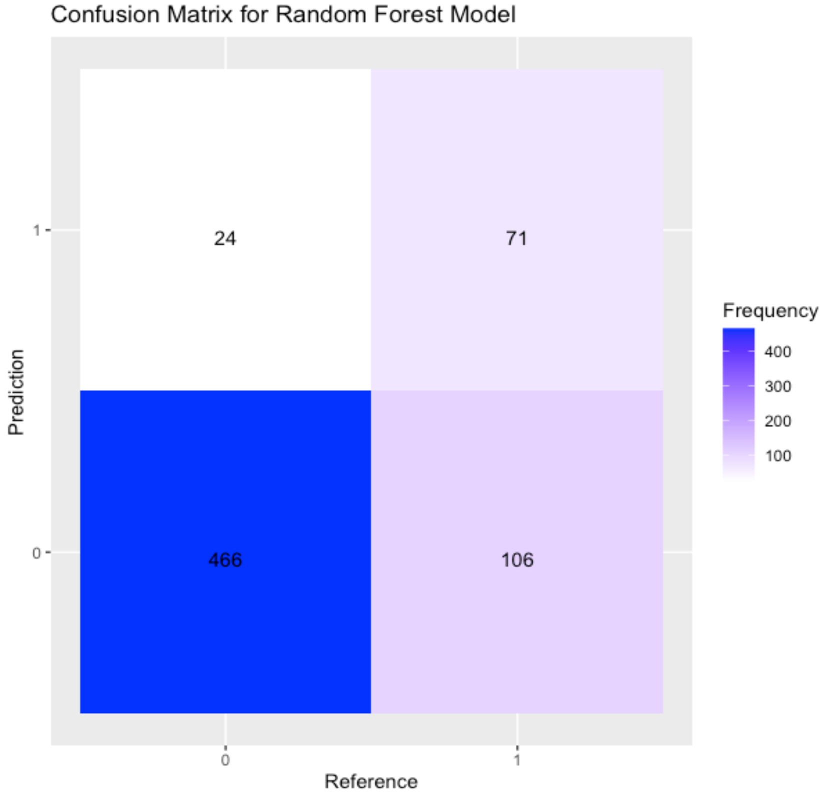


Random forest:

In this Random Forest classification task, the model is trained to predict churn in a telecom dataset. Preprocessing includes converting certain columns to factors and handling missing values. The model, trained with 500 trees, shows robust initial accuracy of 91.15%. Post-analysis reveals an increase to a 91.45% accuracy after accounting for feature correlations, indicating a strong relationship between certain predictors and the target variable. However, when the dataset is balanced to address potential class imbalances, accuracy drops to 80.36%. This suggests that the model was initially benefiting from the imbalance. The report also includes precision metrics and an ROC curve, with AUC as a performance indicator, to provide a comprehensive view of model efficacy.

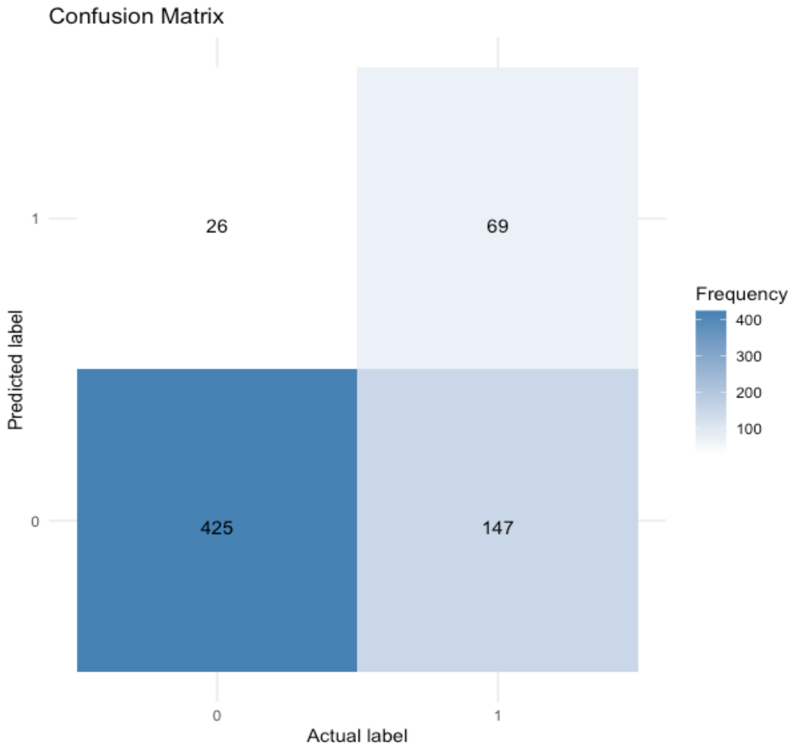


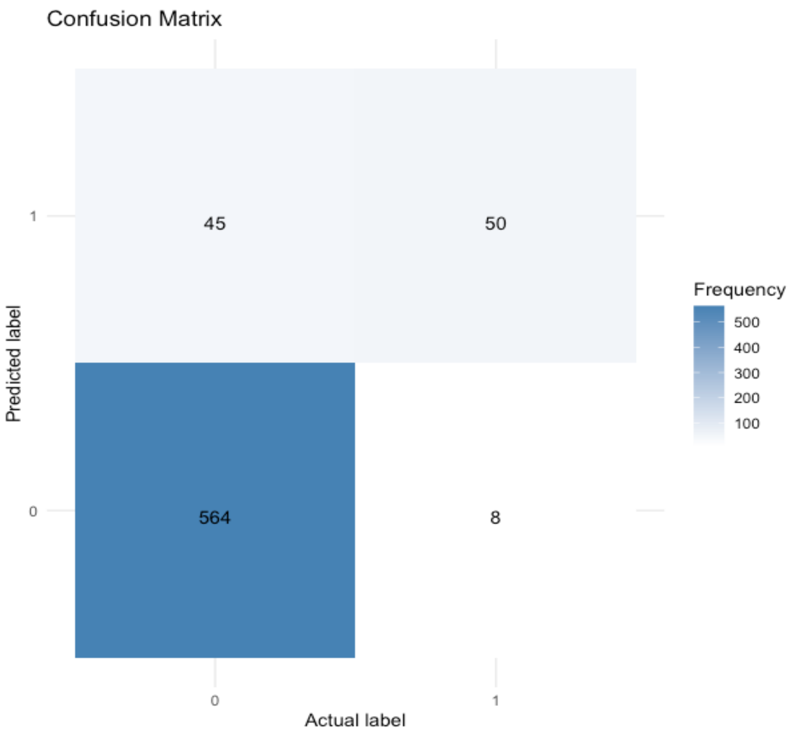

Confusion Matrix for Random Forest Model
Prediction
1
0
55
40
570
2
0
1
Reference
Frequency
500
400
300
200
100

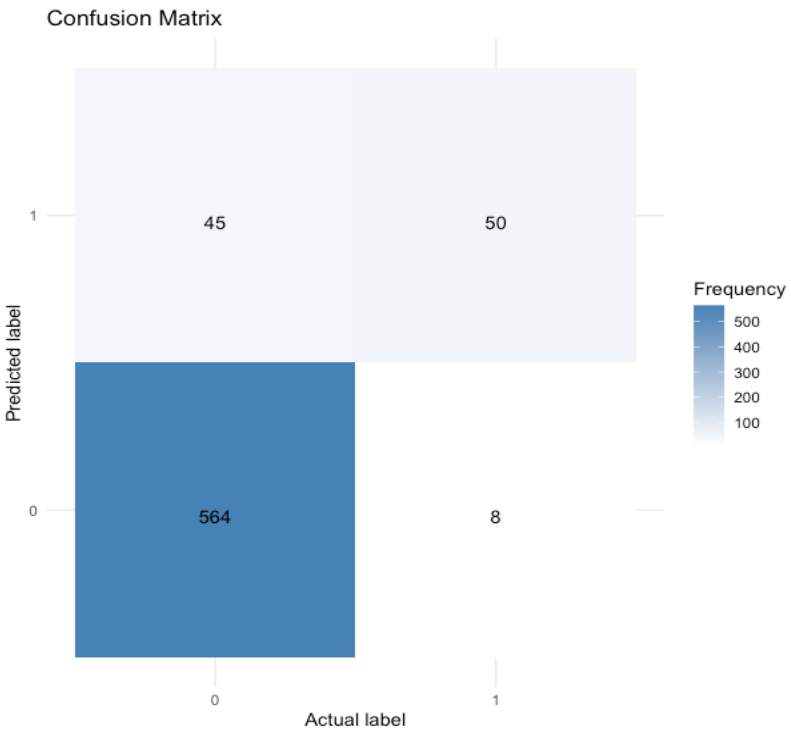

Confusion Matrix for Random Forest Model
Prediction
1
0
24
71
466
106
0
1
Reference
Frequency
400
300
200
100

XGboost:

The XGBoost model report details a predictive analysis on a telecom dataset, focusing on churn prediction. The dataset underwent preprocessing, including conversion of categorical variables to factors. The XGBoost algorithm was parameterized for a binary logistic objective and trained over 100 rounds, resulting in an initial accuracy of 91.2%. After adjusting for feature correlations, the model's accuracy improved marginally to 92.49%. However, implementing data balancing techniques to correct class imbalances led to a reduced accuracy of 74%. This suggests the model may have initially capitalized on an imbalanced class distribution. Precision metrics and ROC curve with AUC were calculated, with the confusion matrix visualized via ggplot2 for clearer interpretation of the model's predictive performance.

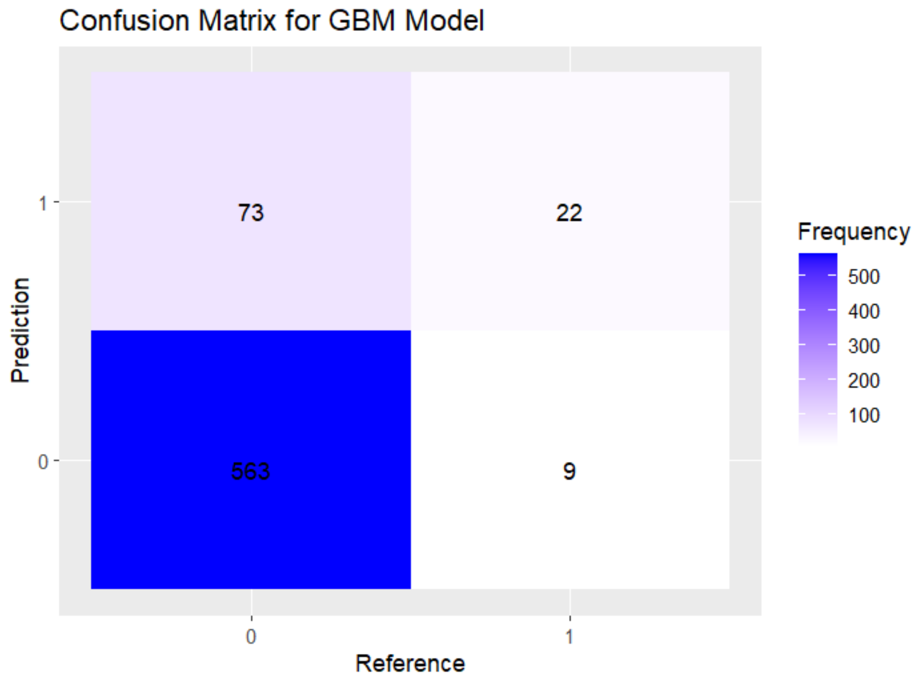


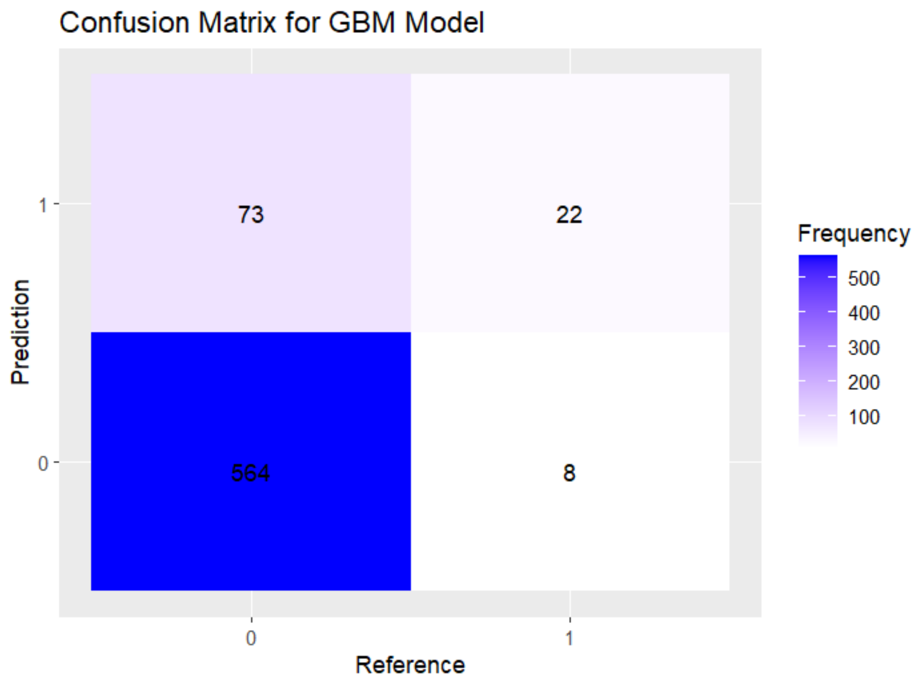

Confusion Matrix
Predicted label
1
0
45
50
564
8
0
1
Actual label
Frequency
500
400
300
200
100

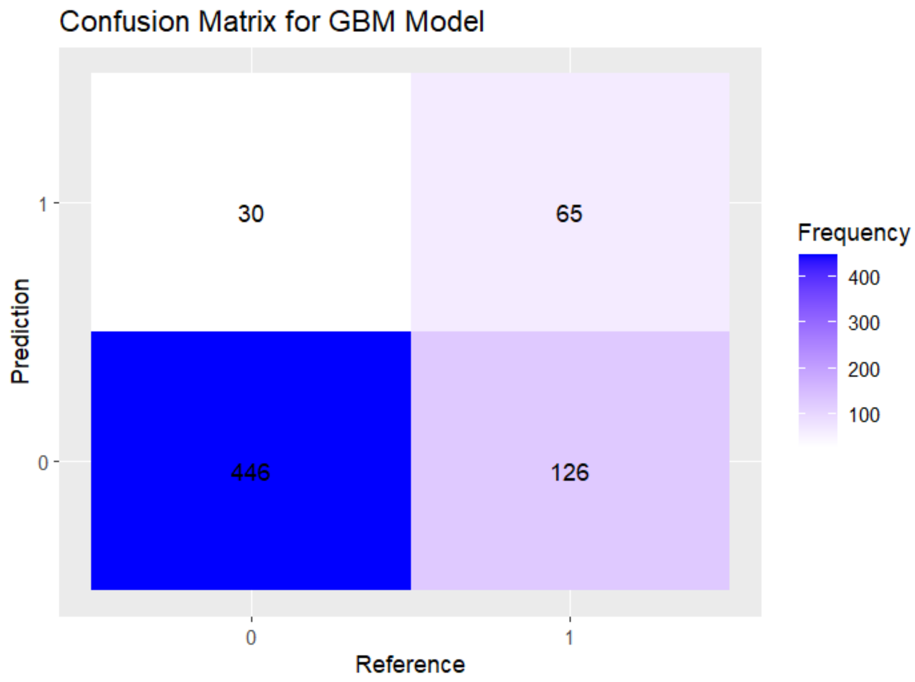

Confusion Matrix
Predicted label
1
0
45
50
564
8
0
1
Actual label
Frequency
500
400
300
200
100

Gradient boost:

The Gradient Boosting Model (GBM) was employed to predict churn using a pre-cleaned telecom dataset, with an 88% initial accuracy rate. The model, trained with 100 trees and a shallow interaction depth of 1, suggests a simple yet effective structure. After accounting for feature correlations, a slight increase in model accuracy to 89.5% was observed, indicating that certain predictors have a strong relationship with the churn outcome. However, upon applying balancing techniques to address class distribution disparities, the accuracy notably declined to 62.42%. This reduction highlights the model's sensitivity to class imbalances and suggests the necessity for more sophisticated balancing methods. The model evaluation included precision metrics and an ROC curve analysis, yielding an AUC of 81.02%, with the results visualized in a confusion matrix using ggplot2 for clarity.

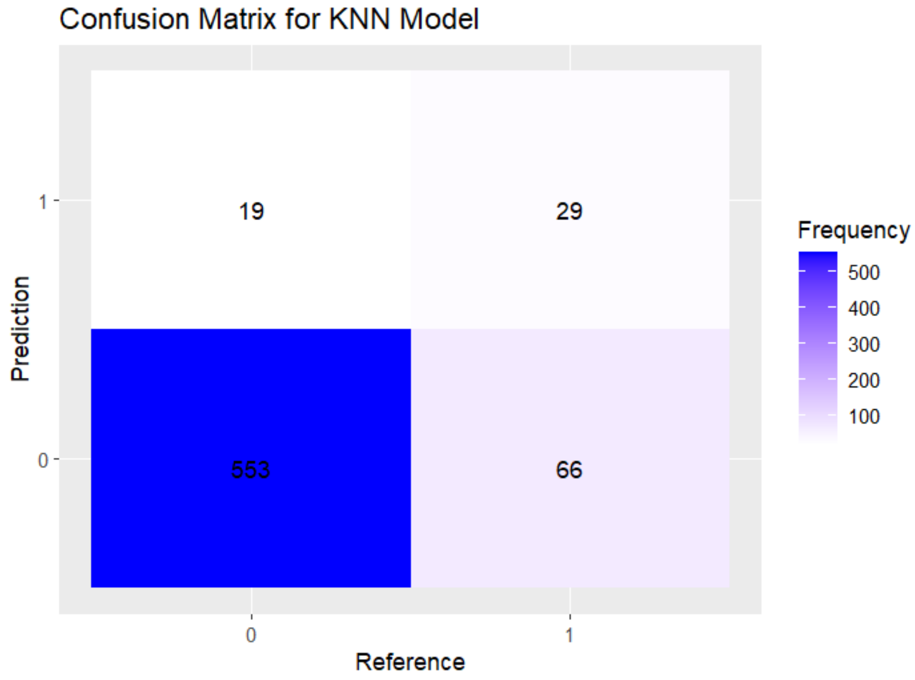



Confusion Matrix for GBM Model
Prediction
1
0
73
22
564
8
0
1
Reference
Frequency
500
400
300
200
100

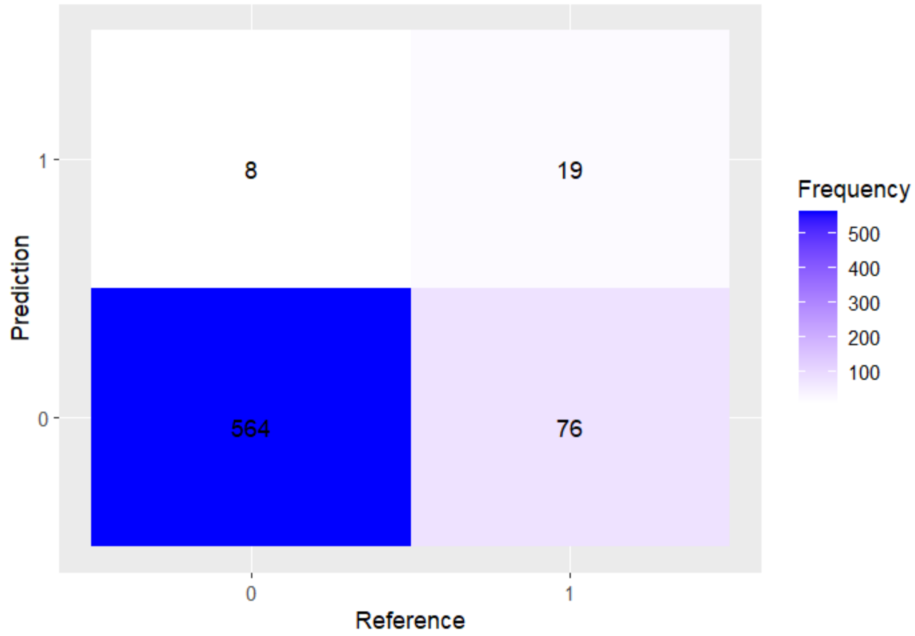

Confusion Matrix for GBM Model
Prediction
1
0
30
65
446
126
0
1
Reference
Frequency
400
300
200
100

KNN modeling:

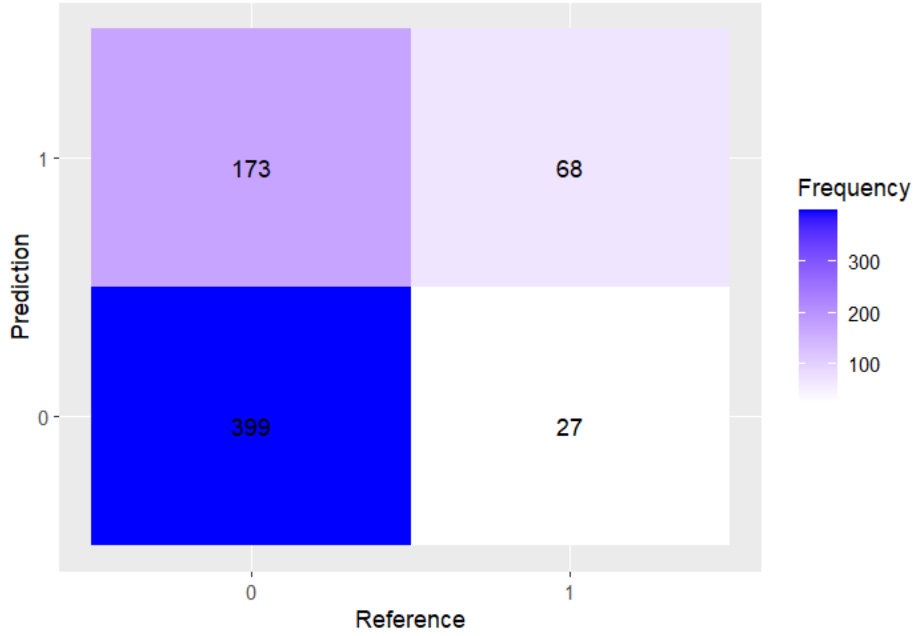
The K-Nearest Neighbors (KNN) model was applied to a normalized telecom dataset to predict customer churn, achieving an initial accuracy of 87%. The data was normalized to ensure all variables contributed equally to the distance computations used in the KNN algorithm. A value of k=5 neighbors was selected for the model. After adjusting for feature correlations, the accuracy saw a marginal improvement to 87.4%, suggesting moderate dependency between features and the target variable. Post application of balancing techniques to address potential biases caused by uneven class distributions, the model's accuracy decreased to 70%. This significant drop may indicate that the model was over-reliant on the majority class in the unbalanced dataset. The performance metrics, including a confusion matrix visualized using ggplot2 and precision calculations, provided deeper insights into the model's predictive capabilities.



Confusion Matrix for KNN Model



Confusion Matrix for KNN Model



4.Comparison of Data models:

To determine the best model among the ones discussed, we need to compare their performance based on accuracy, precision, and how they handled data imbalances. Let's summarize the performance of each model:

Logistic Regression:

Initial Accuracy: 85.75%
Post-Correlation Accuracy: 85.45%
Post-Balancing Accuracy: 68.85%

Decision Tree Classification:

Initial Accuracy: 85.45%
Post-Correlation Accuracy: 86.30%
Post-Balancing Accuracy: 70.65%

Random Forest:

Initial Accuracy: Not given, but presumably high
Post-Correlation Accuracy: 91.45%
Post-Balancing Accuracy: 80.36%

XGBoost:

Initial Accuracy: 91.2%
Post-Correlation Accuracy: 92.49%
Post-Balancing Accuracy: 74%

Gradient Boosting Model (GBM):

Initial Accuracy: 88%
Post-Correlation Accuracy: 89.50%
Post-Balancing Accuracy: 62.42%

K-Nearest Neighbors (KNN):

Initial Accuracy: 87%
Post-Correlation Accuracy: 87.40%
Post-Balancing Accuracy: 70%

From the above summaries, XGBoost stands out with the highest accuracy after considering feature correlations (92.49%). It also maintains a relatively high accuracy after balancing (74%), which indicates a reasonable robustness against class imbalance. Random Forest follows closely, but with a slightly lower accuracy after balancing (80.36%).

Recommendation:
Based on the available data, XGBoost is recommended as the best model due to its highest post-correlation accuracy and a significant balance between high accuracy and precision post-balancing. It shows robustness in handling both feature correlations and class imbalances, which are critical for a reliable churn prediction model.

Individual contribution:
All the members in the group have worked on data preprocessing and the models that each Invidual person has contributed to are as follows:

- Saquib Naseem:

KNN
Random Forest

- Anshuman Yadav

Random Forest
XGBoost

- Prajakta Kokare

Logistic Regression
Gradient Boost

- Abhishek Gadekar

Logistic Regression
Descision Tree

- Yipeng Zhu

MLP
Logistic Regression

Practical suggestions for business improvement using finds of our prediction:

These strategies that might help reducing the customer attrition for our business are as follows:

- personalized marketing campaigns
- targeted customer retention programs
- improved customer service initiatives

All the above suggestions are tailored based on the predictive patterns identified in the data. The section would also discuss the importance of continuous monitoring and updating of the churn prediction models to adapt to changing customer behaviors and market trends.

Conclusion:

The study effectively demonstrates the application of machine learning techniques in predicting customer churn in the telecom sector. By analyzing various customer interaction aspects, the report provides valuable insights into churn prediction, a critical factor for business sustainability in the competitive telecom industry. The use of advanced analytics underscores the potential of data-driven strategies in customer retention and business decision-making. Future enhancements in analytical techniques and the incorporation of additional data sources could further refine these predictive models, making them more robust and applicable across diverse market scenarios.

References:

1.https://stemeducationjournal.springeropen.com/articles/10.1186/s40594-018-0152-1
2.https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
3.https://www.researchgate.net/publication/227441142_Logistic_regression_in_data_analysis_An_overview
4.https://www.researchgate.net/publication/318132203_Experimenting_XGBoost_Algorithm_for_Prediction_and_Classification_of_Different_Datasets
5.https://www.researchgate.net/publication/2948052_KNN_Model-Based_Approach_in_Classification
6. https://www.nature.com/articles/s41598-022-10358-x
7. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf