CS 6375
Machine Learning, Fall 2024
Telecommunication Churn Prediction
By,
FNU Anuja (AXN220123)

Introduction

In the tough telecom sector, keeping current clients is just as important as attracting new ones. High churn rates raise the cost of acquiring new customers and cause significant revenue losses. In order to help the business allocate resources efficiently and create focused retention tactics, this project intends to provide a predictive machine learning solution that will identify customers who are likely to leave.

Problem Statement

1. Find High-Risk Clients: Identify the clients who are most likely to leave, enabling focused interventions.
2. Recognise Churn Drivers: Examine characteristics and trends that affect customers' decisions to leave, such as poor service, excessive expenses, or low engagement.
3. Create Successful Retention Programs: To increase customer satisfaction and loyalty, adjust marketing tactics, boost service offerings, and respond to consumer concerns.

This project aims to:

1. Use transactional, behavioural, and demographic data about your customers to forecast turnover.
2. Give practical advice for developing customer retention plans.

Objectives

Create a churn prediction machine learning model that is both accurate and efficient.
To obtain important insights, examine the connection between churn and client attributes.

Dataset Overview:
Total Rows: 7043
Total Columns: 22
This dataset consists of customer activity data (features) as well as a churn label indicating whether a customer cancelled their subscription, and will be used to develop predictive models.

Column Details:

1. customerID: A distinct, non-null identifier for every customer.
2. gender: The customer's gender (male or female).
3. SeniorCitizen: Denotes if the client is an elderly person (0 for No, 1 for Yes).
4. Partner: Shows whether or not the client has a partner.
5. Dependents: Selects whether the client has dependents.
6. Tenure: The length of time a client has been with the business.

7. TENURE RANGE: A range-based system for classifying tenure, which typically uses null values.
8. PhoneService: Shows whether or not the client has phone service.
9. MultipleLines: Indicates whether or not the customer has more than one line (Yes/No/No phone service).
10. InternetService: The customer's internet service type (DSL, fibre optic, or no).
11. OnlineSecurity: If the client has internet service, does he or she have online security add-ons?
12. OnlineBackup: Whether or whether the client has internet service and online backup add-ons.
13. DeviceProtection: Whether or whether the client has internet service, device protection add-ons, or both.
14. TechSupport: Whether or not the client has internet service or tech support add-ons.
15. StreamingTV: Shows whether the consumer has internet service or not.
16. StreamingMovies: Shows whether or not the customer has internet service and streams films.
17. Contract: The customer's contract type (month-to-month, one-year, or two-year).
18. PaperlessBilling: Shows whether or not the client has chosen to use paperless billing.
19. The customer's chosen form of payment, such as a mailed cheque or an electronic cheque.
20. MonthlyCharges: The sum that is billed to the client each month.
21. TotalCharges: Total amount charged to the customer (stored as object; might need conversion).
22. Churn: Indicates if the customer has churned (Yes/No).

    Cleaned dataset .csv file has been attached.

Methodology

1. **Data Preprocessing**:
    1. Deal with inconsistent or missing data.
    2. To ensure model readiness, carry out feature engineering and scaling.
    3. Categorical variables should be encoded.
2. **Exploratory Data Analysis (EDA)**:
    1. To comprehend the distribution and significance of features, perform univariate and bivariate analysis.
    2. See churn patterns for different features.
3. **Model Development**:
    1. Use R's Rpart package to create a Decision Tree model, then compare it to other models such as SVMs, Random Forest and Logistic Regression.
    2. Improve performance by optimizing hyperparameters.
4. **Model Evaluation**:
    1. Evaluate the model's accuracy, and precision.
    2. For performance validation, use ROC-AUC curves and confusion matrices.

Evaluation Criteria

A confusion matrix is calculated for each of the above-mentioned algorithms. The confusion matrix has Actual values vs Predicted values. The matrix can be divided into four parts:

1. True Positive – Predicted value is positive, and the actual value is also positive. This is called sensitivity.

2. True Negative – Predicted value is negative, and the actual value is also negative. This is called specificity.

3. False Positive – Predicted value is positive, but the actual value is negative. This is a Type I error.

4. False Negative – Predicted value is negative, but the actual value is positive. This is a Type II error.

With the help of confusion matrix, all the evaluation criteria can be computed,

1. Recall = True positive / (True positive + False negative)

2. Precision = True positive / (True positive + False positive)

3. F1 = (2 * Recall * Precision)/(Recall + Precision)

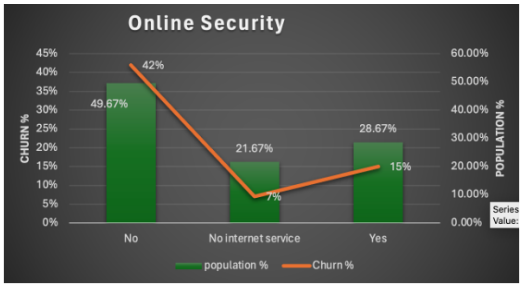
4. Accuracy = (True Positive + True Negative) / Total.

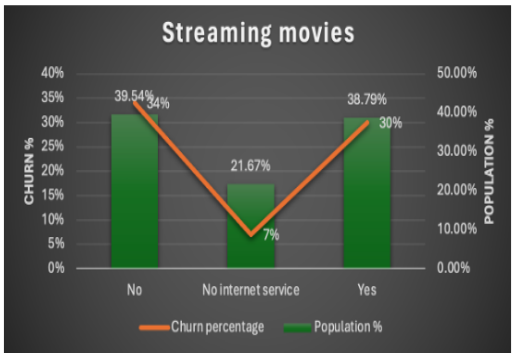
Bivariate analysis:

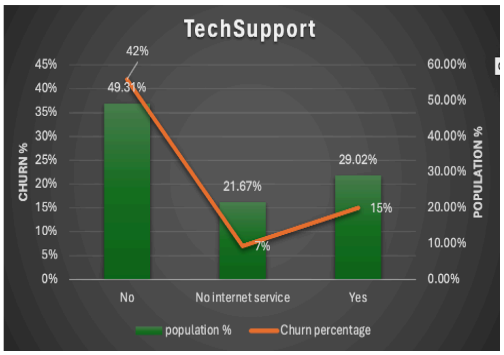
Bivariate analysis is one of the statistical analysis where two variables are observed. One variable here is dependent while the other is independent. These variables are usually denoted by X and Y. So, here we analyse the changes occured between the two variables and to what extent.

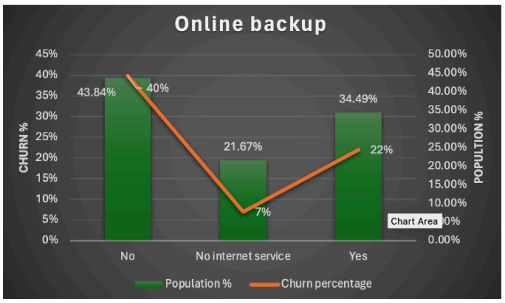
Below are the snapshots of bivariate analysis :

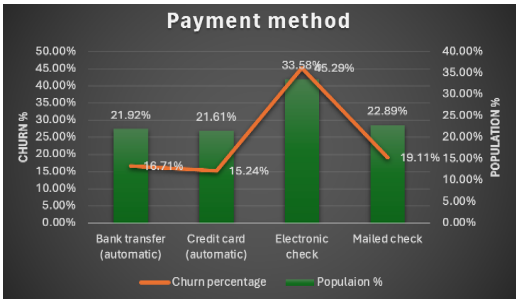
Here we are finding a relationship between population and churn rate.

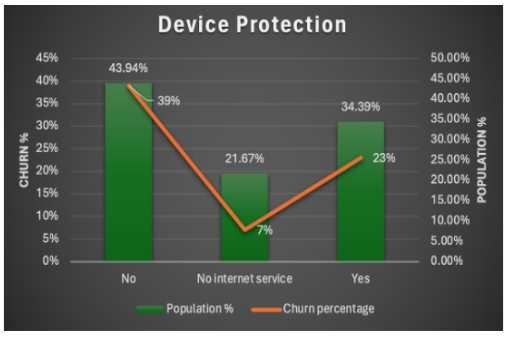


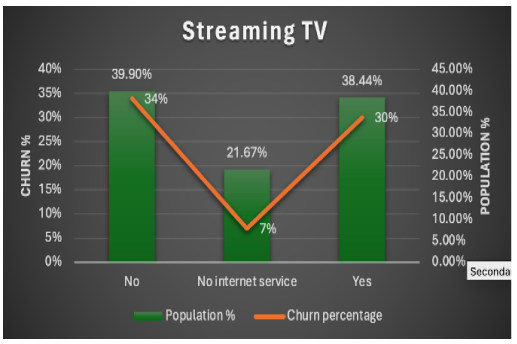




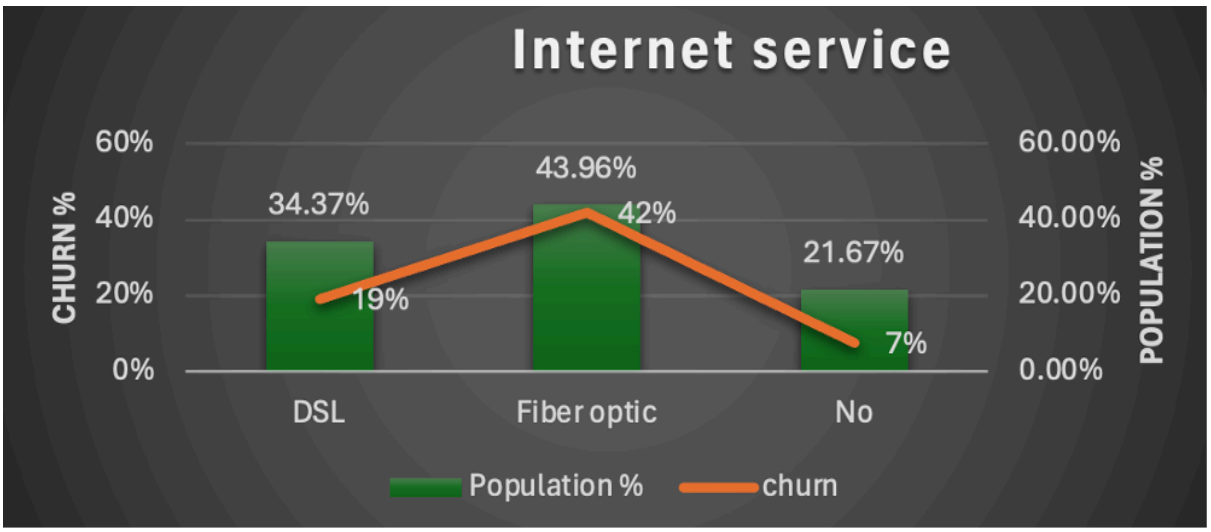








Bivariate analysis example:



1. The population percentage of consumers that use DSL as their internet service is 34.37%.
   Churn Rate: A comparatively low 19% of DSL consumers have experienced churn.
2. Fiber Optic: Population Percentage: Fiber Optic is the most widely used internet service, with 43.96% of consumers using it. Churn Rate: The largest percentage of any category, 42% of Fiber Optic consumers have churned. This implies that this group is greatly impacted by discontent or rivalry.
3. Lack of Internet Service: Population Percentage: 21.67% of clients lack internet access. Because there may be fewer competing options for customers without internet services, the churn rate of 7% is the lowest of the three groups.

Experiments:

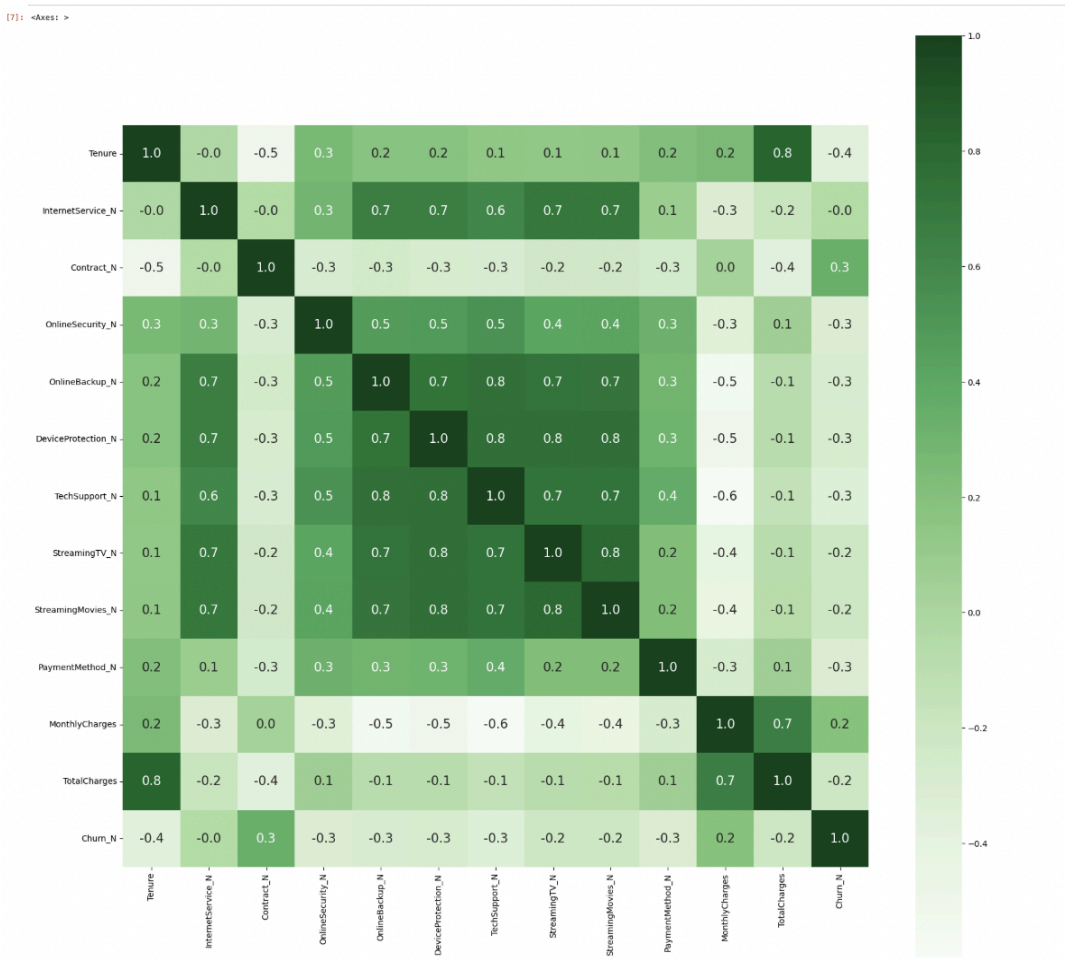
```
]: data.head()
   data.isnull().sum()

]: Tenure                0
   InternetService_N     0
   Contract_N            0
   OnlineSecurity_N      0
   OnlineBackup_N        0
   DeviceProtection_N    0
   TechSupport_N         0
   StreamingTV_N         0
   StreamingMovies_N     0
   PaymentMethod_N       0
   MonthlyCharges        0
   TotalCharges         11
   Churn_N               0
   dtype: int64
```

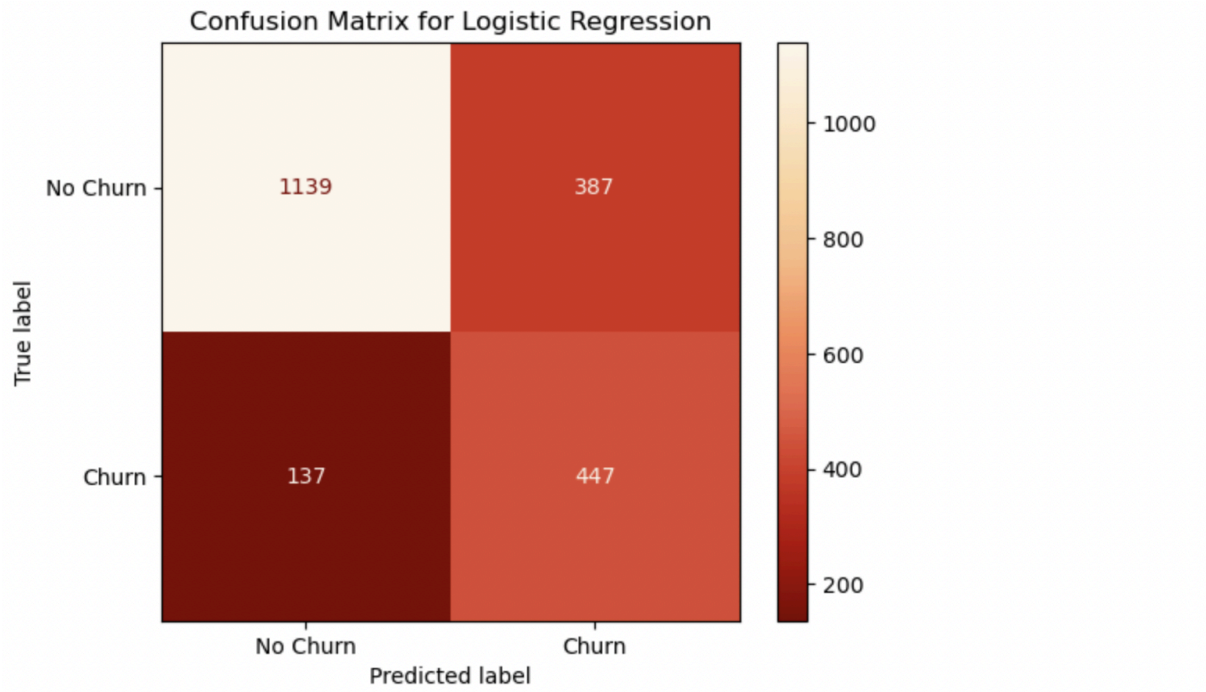
- This code checks for missing values (NaNs) in each column of the dataset and provides the total count of missing values per column.

data.isnull(): Creates a boolean DataFrame where True indicates missing values, and False indicates non-missing values.
.sum(): Adds up the True values (count of missing entries) for each column.

- **data = data.dropna(subset=["TotalCharges"])**
  Removes rows in the DataFrame where the column TotalCharges has missing values (NaN).
  corr = data.corr() - Computes the pairwise correlation matrix of all numeric columns in the dataset.
  Corr.shape - Returns the dimensions of the correlation matrix (e.g., number of rows and columns).
  plt.figure(figsize=(20,20)) - Creates a figure object for the plot, setting its size to 20x20
  Inches.
  sns.heatmap(...)- Plots a heatmap of the correlation matrix, visualizing the relationships between variables.



- Here is the confusion matrix of Linear regression model:



1. True Negatives (TN) (Top Left: 1139)

The model correctly predicted No Churn when the actual label was No Churn. These are correct rejections.

2. False Positives (FP) (Top Right: 387)

The model predicted Churn when the actual label was No Churn. These are Type I errors, indicating the cost of incorrectly targeting non-churning customers.

3. False Negatives (FN) (Bottom Left: 137)

The model predicted No Churn when the actual label was Churn. These are Type II errors, indicating the cost of missing actual churners.

4. True Positives (TP) (Bottom Right: 447)

The model correctly predicted Churn when the actual label was Churn. These are correct detections.

Accuracy: Proportion of correctly classified instances.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

= 447+1139/447+1139+387+137

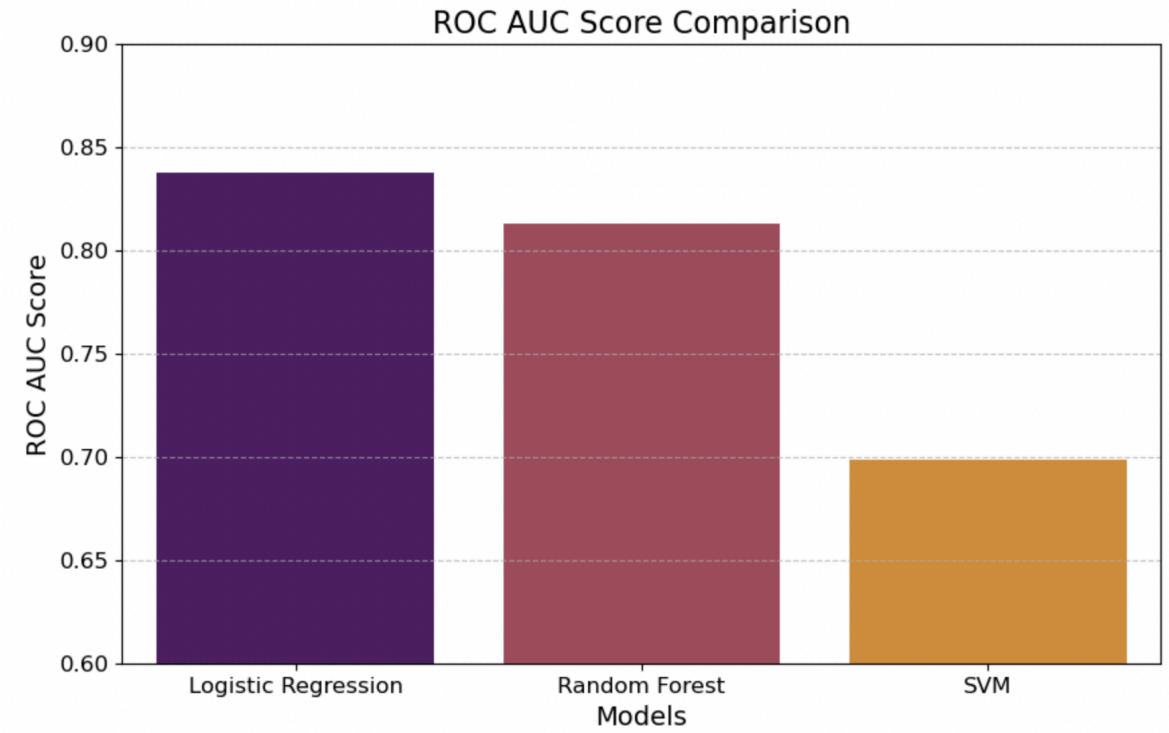
≈78.86%

Precision (for Churn): Focuses on how many predicted churns were true churns.

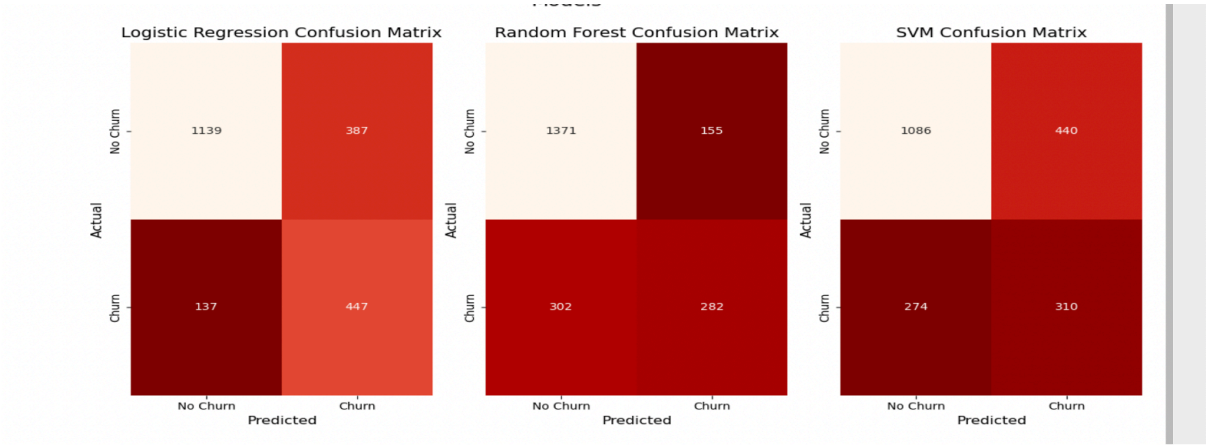
Precision = TP/(TP+FP)

= 447/(447+387)

≈53.58%



Three models Logistic Regression, Random Forest, and SVM that are used to forecast customer attrition in the provided dataset are compared using a bar plot. The model's capacity to differentiate between churn and non-churn clients is measured by the ROC AUC score. A higher number indicates better performance. Logistic Regression achieves the highest ROC AUC score, around **0.83** indicates strong predictive performance in distinguishing between churn and no-churn.



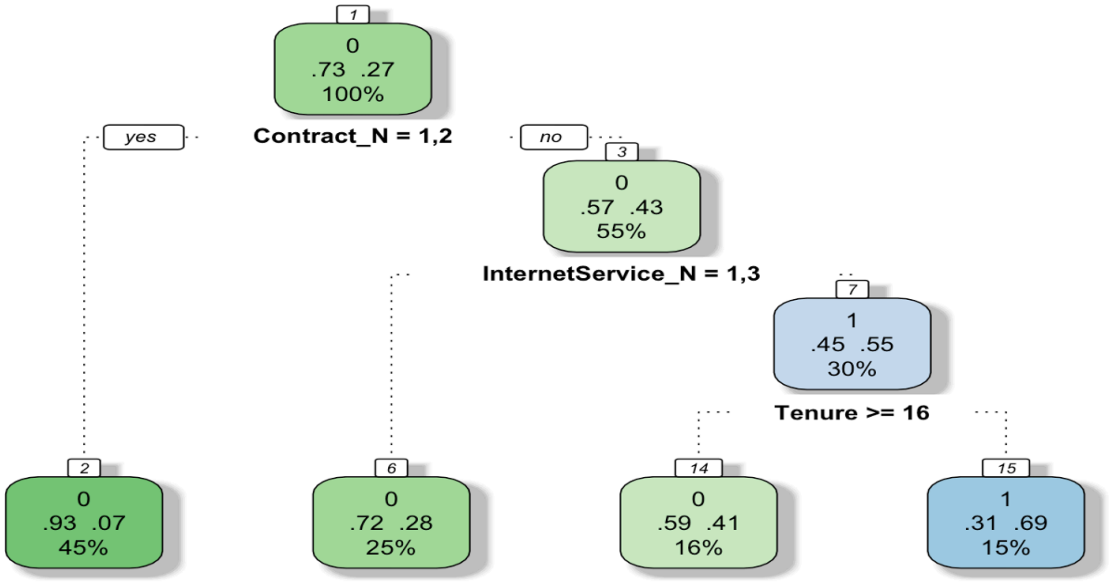
This displays the confusion matrices for the three machine learning models used to forecast customer attrition: SVM, Random Forest, and Logistic Regression.

- With 387 false positives and 137 false negatives, logistic regression forecasts 447 true positives and 1139 true negatives.

- With 155 false positives and 302 false negatives, Random Forest outperforms with 1371 true negatives and 282 genuine positives.
- With 440 false positives and 274 false negatives, SVM predicts 1086 true negatives and 310 real positives.

Among all these accuracy of logistic regression is best with 78.86%

RPART MODEL:



Rattle 2024-Dec-08 20:17:13 anujabarin

Derived variable Contract_N:
1 denotes Contract for a Year
2 denotes a two-year contract.
3 denotes a monthly contract
Derived variable Internet Service:
1= DSL
2= No Internet Service
3= Fiber Optics
Interpretation of the model
Population percentage is 15%, and churn rate is 69% for samples with Contract_N=3 [monthly contract] & Internet Service = 3, i.e. Fibre Optics & Tenure less than 16.
The population with yearly and two-year contracts, Contracts N 1 and 2, have the lowest churn rate. In this case, the demographic data shows a 45% churn rate of 7%.

Conclusion

Machine learning was effectively used in the Telco Churn Prediction project to forecast customer attrition and pinpoint the main causes of it. With a high ROC-AUC score (~0.83), Logistic Regression was the most successful model, followed by Random Forest. It showed that while value-added services like online security greatly decreased the chance of customer loss, clients with shorter tenure, monthly contracts, or electronic cheque payments were more likely to churn. The telecom company can use these results to create focused retention efforts, like encouraging long-term agreements, combining services, and providing targeted outreach to high-risk clients. The initiative offers a strong basis for data-driven decision-making to increase customer loyalty and lower attrition rates, even several drawbacks.

What I learned from the project

I gained knowledge about data preparation and selection, feature vector identification, and their effects on the target variable. Using measures like ROC-AUC and confusion matrices, I evaluated the performance of several machine learning models, including SVM, Random Forest, and Logistic Regression. I also learnt the benefits of handling missing values, cleaning data, and using heatmaps and other tools to visualise data relationships. Through this project, I improved my ability to understand model results and apply them to practical issues like forecasting loss of clients and suggesting workable solutions.