Analysing Telecom Customer Churn

Understanding and Predicting Customer Churn

By Robert Solomon

11/06/2024

Background

Importance of Analyzing Customer Churn

Benefits for Telecom Providers:

Identifying patterns in service quality, pricing, and customer satisfaction

Proactive issue resolution and improved customer retention

Enhanced overall service delivery

Staying competitive by adapting to customer preferences

Overview of Presentation

- Objectives
- Data Management
- Exploratory Data Analysis
- Model Development
- Model Evaluation
- Best Model Selection
- Conclusion

Objectives

1. Analyze 'Customer Churn' and understand the factors associated with it.
2. Develop Churn Prediction Model.
3. Implement Machine Learning Algorithms and select the best method for Churn Prediction.
4. Practical implementation of the project was segmented into 4 Phases **(Data Management, Exploratory Data Analysis, Model Development w/Binary Logistical Regression, Model Evaluation Methods and Comparing w/BLR).**

Column Name	Data Type	Type
customerID	object	Categorical
gender	object	Categorical
SeniorCitizen	int64	Numerical
Partner	object	Categorical
Dependents	object	Categorical
tenure	int64	Numerical
PhoneService	object	Categorical
MultipleLines	object	Categorical
InternetService	object	Categorical
OnlineSecurity	object	Categorical
OnlineBackup	object	Categorical
DeviceProtection	object	Categorical
TechSupport	object	Categorical
StreamingTV	object	Categorical
StreamingMovies	object	Categorical
Contract	object	Categorical
PaperlessBilling	object	Categorical
PaymentMethod	object	Categorical
MonthlyCharges	float64	Numerical
TotalCharges	object	Categorical
Churn	object	Categorical

Data Management

- **Overview of Dataset:**
  - Data source: Customer_Analytics_Telecom_Master.xlsx
- **Key variables:**
  - Tenure, SeniorCitizen
  - Partner, Dependents, etc.
- **Initial Data Cleaning and Preparation:**
  - Handling missing values
  - Data type conversion

Data Management

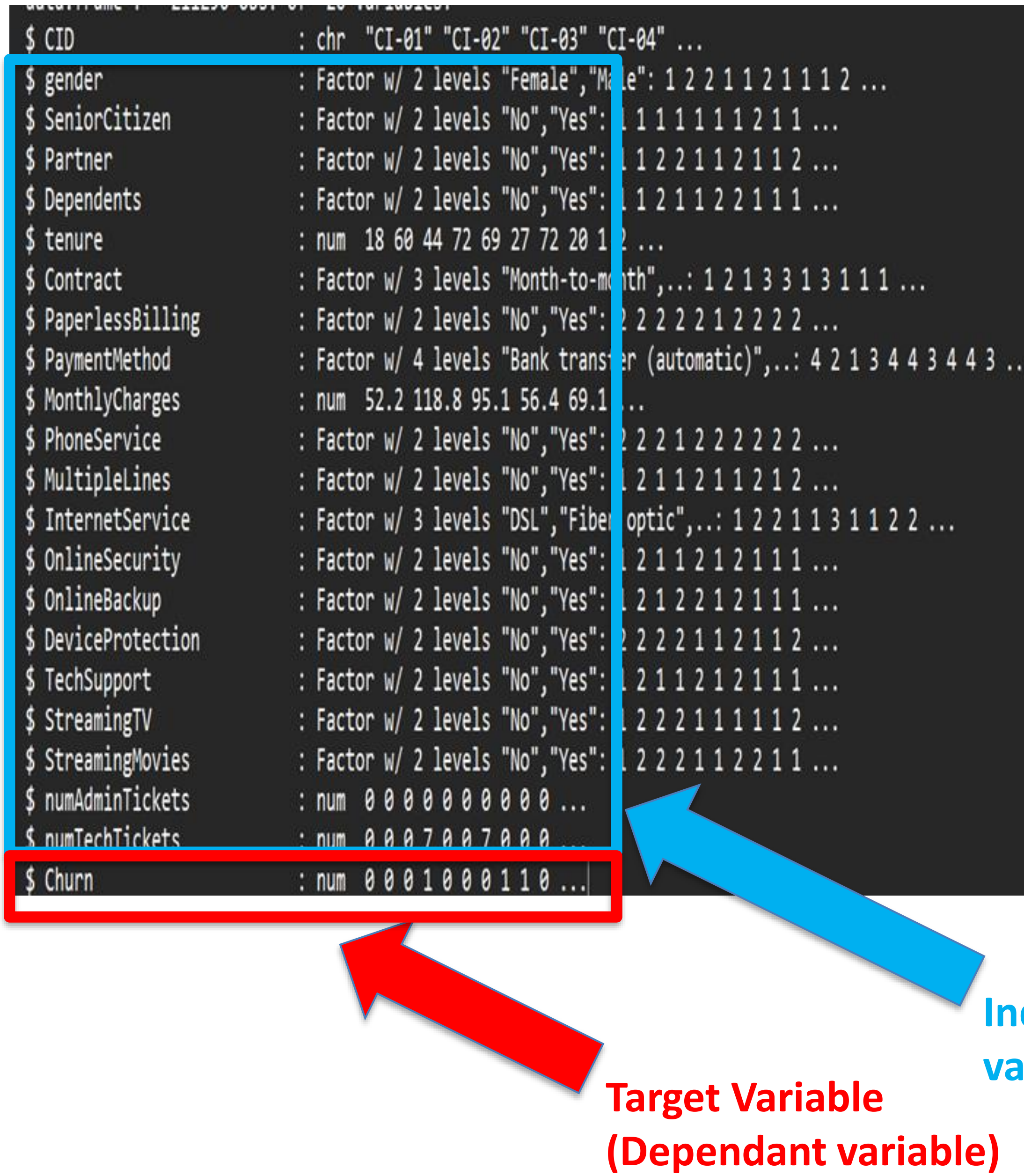
Column Name	Data Type	Type
customerID	object	Categorical
gender	object	Categorical
SeniorCitizen	int64	Numerical
Partner	object	Categorical
Dependents	object	Categorical
tenure	int64	Numerical
PhoneService	object	Categorical
MultipleLines	object	Categorical
InternetService	object	Categorical
OnlineSecurity	object	Categorical
OnlineBackup	object	Categorical
DeviceProtection	object	Categorical
TechSupport	object	Categorical
StreamingTV	object	Categorical
StreamingMovies	object	Categorical
Contract	object	Categorical
PaperlessBilling	object	Categorical
PaymentMethod	object	Categorical
MonthlyCharges	float64	Numerical
TotalCharges	object	Categorical
Churn	object	Categorical

Data Snapshot

- The Column lists the names of the variables in the dataset.
- The Data type indicates the data type for each column, such as object for categorical data, int64 for integer numerical data, and float64 for floating-point numerical data.
- Churn is identified as the dependant variable.

Data Management

Detailed snapshot of Dataset structure


```
$ CID              : chr  "CI-01" "CI-02" "CI-03" "CI-04" ...
$ gender           : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 2 1 1 1 2 ...
$ SeniorCitizen    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
$ Partner          : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 1 2 1 1 2 ...
$ Dependents       : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 2 1 1 1 ...
$ tenure           : num  18 60 44 72 69 27 72 20 12 ...
$ Contract         : Factor w/ 3 levels "Month-to-month",..: 1 2 1 3 3 1 3 1 1 1 ...
$ PaperlessBilling : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 2 2 ...
$ PaymentMethod    : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 2 1 3 4 4 3 4 4 3 ...
$ MonthlyCharges   : num  52.2 118.8 95.1 56.4 69.1 ...
$ PhoneService     : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 2 2 ...
$ MultipleLines    : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 2 1 2 ...
$ InternetService  : Factor w/ 3 levels "DSL","Fiber optic",..: 1 2 2 1 1 3 1 1 2 2 ...
$ OnlineSecurity   : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 2 1 1 1 ...
$ OnlineBackup     : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 1 1 1 ...
$ DeviceProtection : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 2 1 1 2 ...
$ TechSupport      : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 2 1 1 1 ...
$ StreamingTV      : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 1 1 2 ...
$ StreamingMovies  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 2 2 1 1 ...
$ numAdminTickets  : num  0 0 0 0 0 0 0 0 0 0 ...
$ numTechTickets   : num  0 0 0 7 0 0 7 0 0 0 ...
$ Churn            : num  0 0 0 1 0 0 0 1 1 0 ...
```


- Dataset contains various features related to customer information and service usage, including both numerical and categorical data.
- 22 variables present in the master dataset.
- ‘CID’ variable is a unique identifier and a dummy variable which was eventually dropped from future analysis.
- Variable **<u>Churn</u>** was identified as the dependant (target) variable, while the rest were treated as independent variables. The target variable ‘Churn’ indicating whether a customer has churned, which is crucial for our predictive modeling tasks.

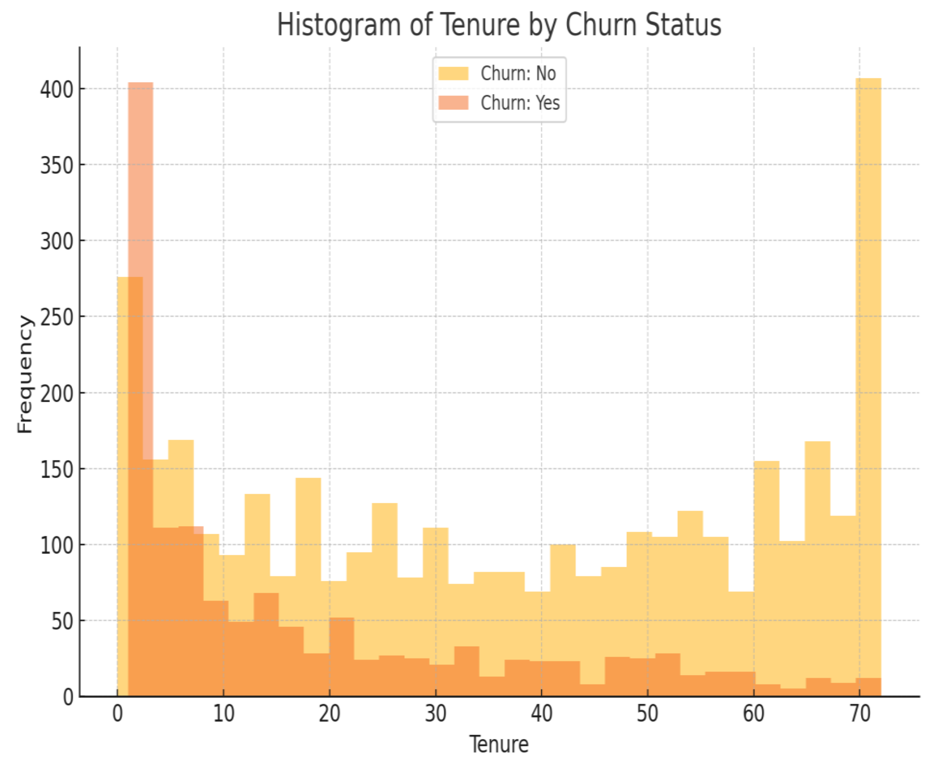
Exploratory Data Analysis (EDA)

Key Factors Influencing Churn:

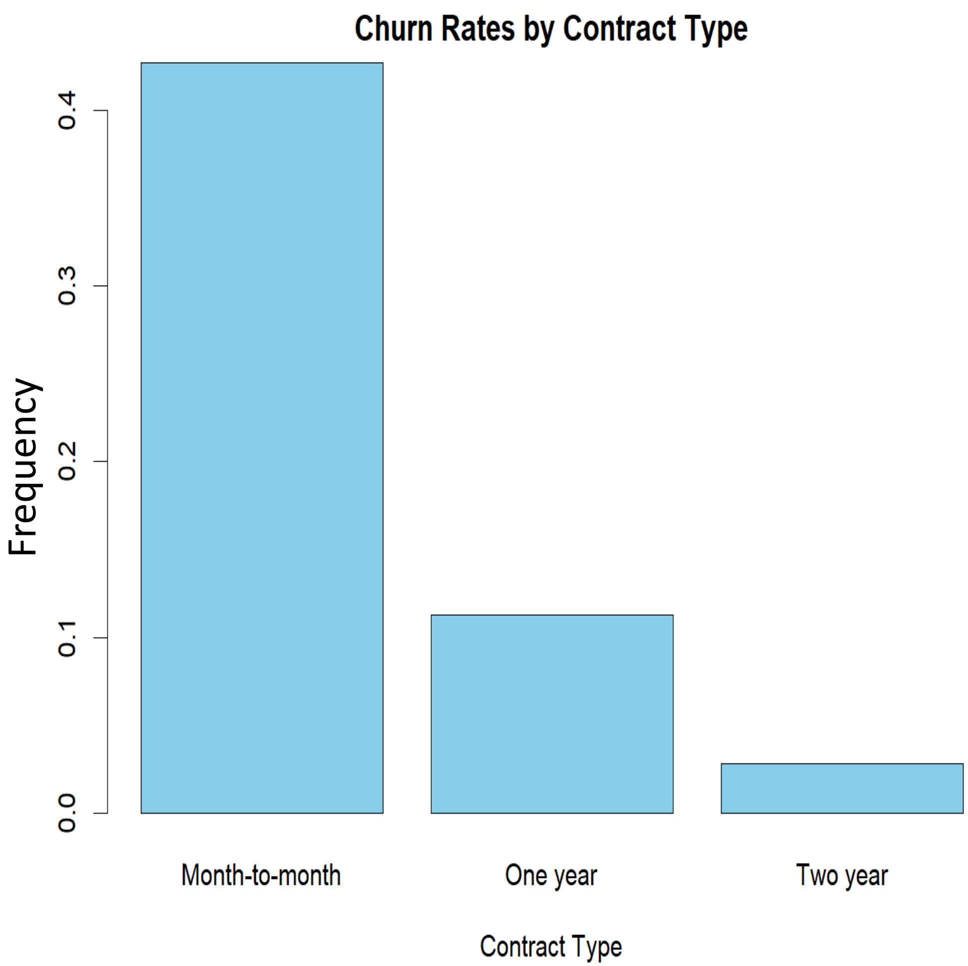
- Distribution of tenure by Churn status

Visualizations:

- Histogram of tenure with Churn



Exploratory Data Analysis (EDA)



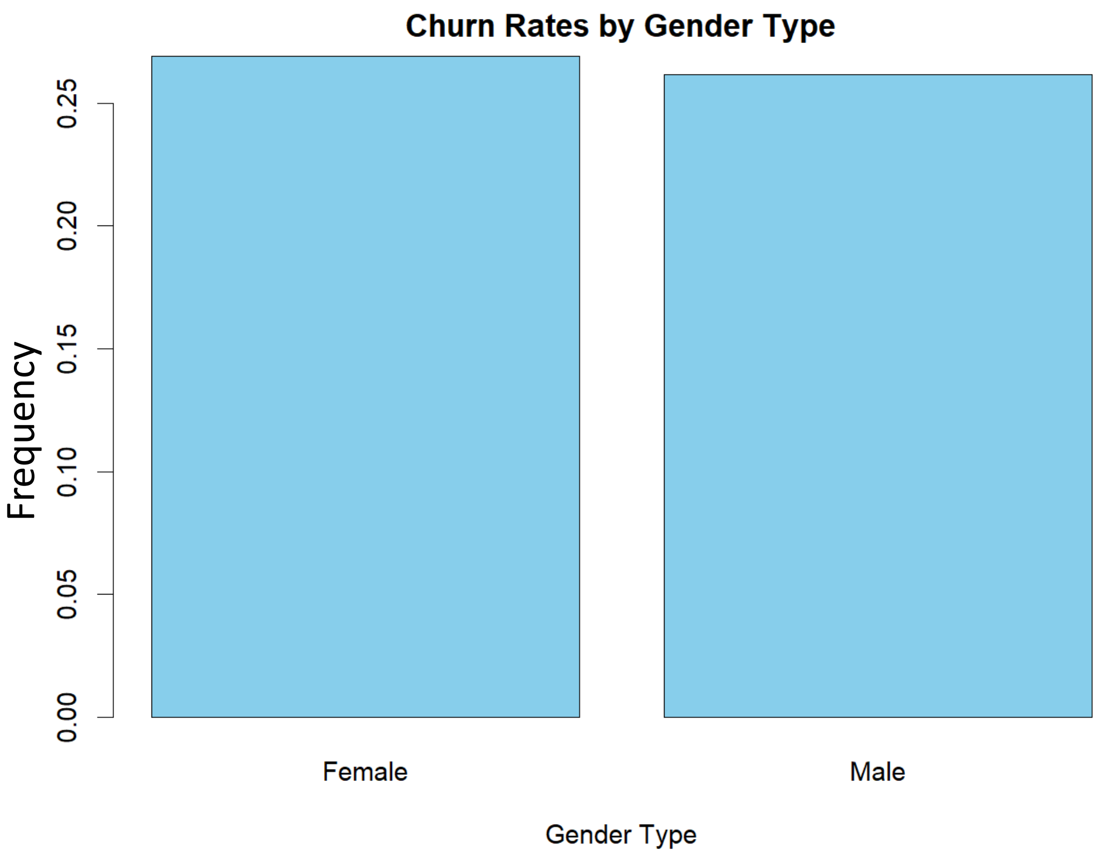
Key Factors Influencing Churn:

- Distribution of Contract by Churn status

Visualizations:

- Bar chats of Contract Type with Churn

Exploratory Data Analysis (EDA)



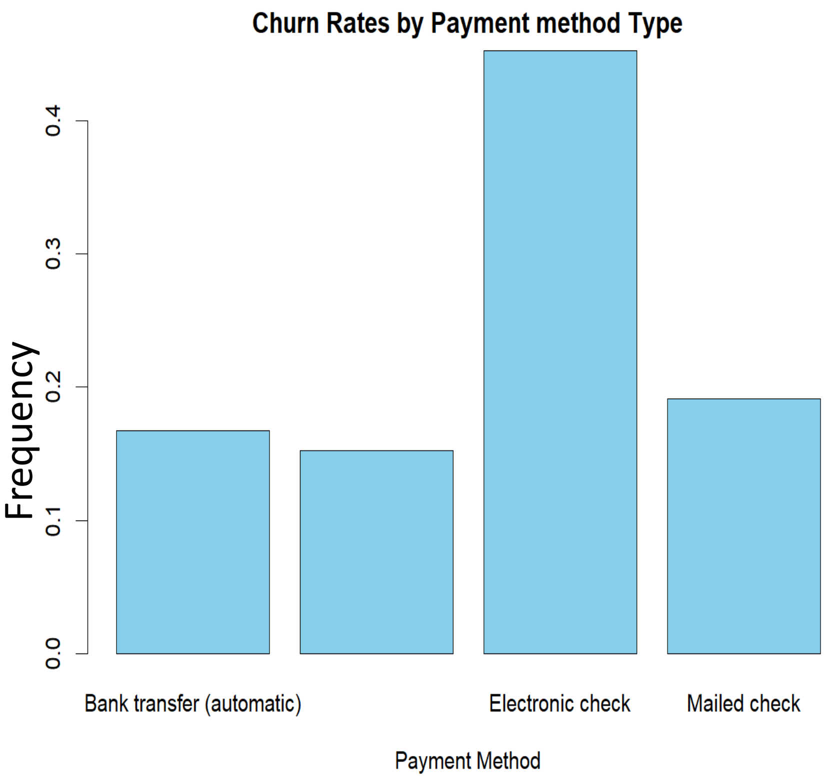
Key Factors Influencing Churn:

- Distribution of Gender by Churn status

Visualizations:

- Bar chats of Gender with Churn

Exploratory Data Analysis (EDA)



Key Factors Influencing Churn:

- Distribution of Payment method by Churn status

Visualizations:

- Bar chats of Payment method with Churn

Table of Coefficients for BLR


```
# BLR Model Summary
```

summary() generates a detailed description of the model.

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	1.727979	0.225309	7.669	1.73e-14	***
genderMale	-0.10287	0.015581	-6.603	4.04e-11	***
SeniorCitizenYes	0.274804	0.021264	12.924	< 2e-16	***
PartnerYes	-0.068832	0.01921	-3.583	0.000339	***
DependentsYes	-0.072561	0.022031	-3.294	0.000989	***
tenure	-0.082224	0.000796	-103.3	< 2e-16	***
ContractOne year	-0.853694	0.03198	-26.694	< 2e-16	***
ContractTwo year	-2.392295	0.060708	-39.407	< 2e-16	***
PaperlessBillingYes	0.334508	0.017424	19.198	< 2e-16	***
PaymentMethodCredit card (automatic)	-0.209419	0.029178	-7.177	7.11e-13	***
PaymentMethodElectronic check	0.13631	0.02409	5.658	1.53e-08	***
PaymentMethodMailed check	-0.249314	0.027278	-9.14	< 2e-16	***
MonthlyCharges	-0.039183	0.00819	-4.784	1.72e-06	***
PhoneServiceYes	0.28021	0.166762	1.68	0.092899	.
MultipleLinesYes	0.503598	0.045002	11.191	< 2e-16	***

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 195862 on 169031 degrees of freedom

Residual deviance: 102454 on 169007 degrees of freedom

AIC: 102504

Number of Fisher Scoring iterations: 7

Interpretation :

- *Accepts null hypothesis that the following variables are significant and have a p-value <0.05*
- *All variables except **"StreamingMoviesYes"** are significant (p-values < 0.05 or better)*
- ***"StreamingMoviesYes"** is not significant (p-value: > 0.05 i.e. 0.0992678)*

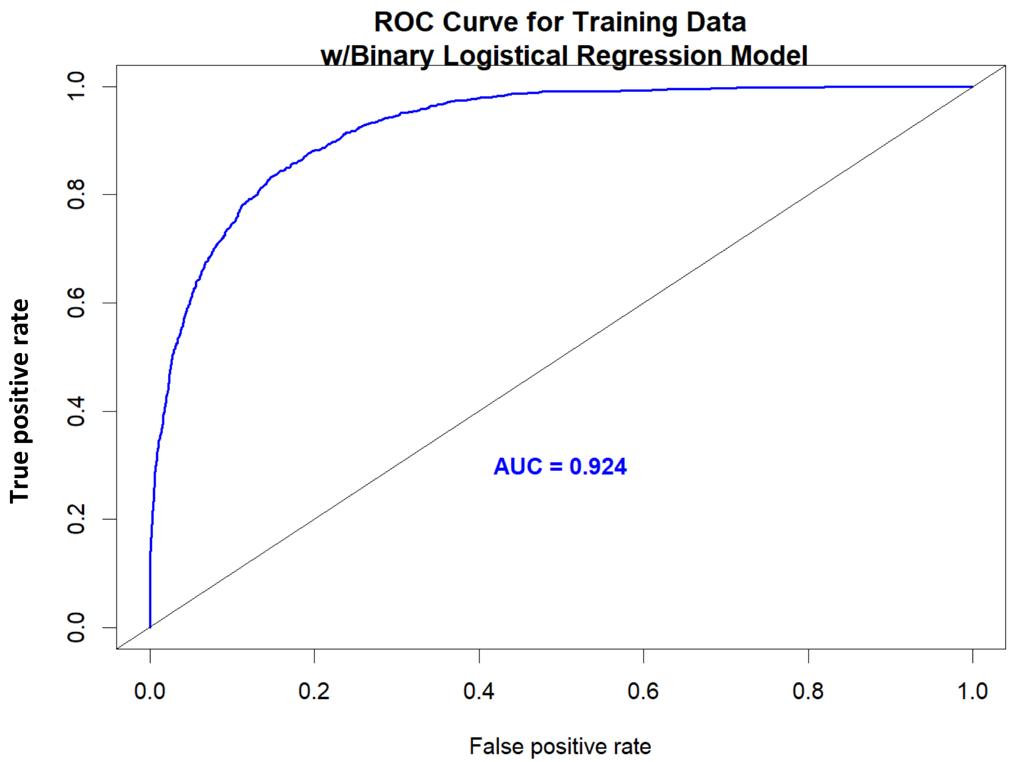
Model Development

- **ML Models Used:**
  - Logistic Regression
  - Decision Tree (DT)
  - Naïve Bayes
  - Random Forest

- **Feature Selection and Engineering:**
  - Conversion of categorical variables to factors
  - Splitting data into training and test sets

Model Development

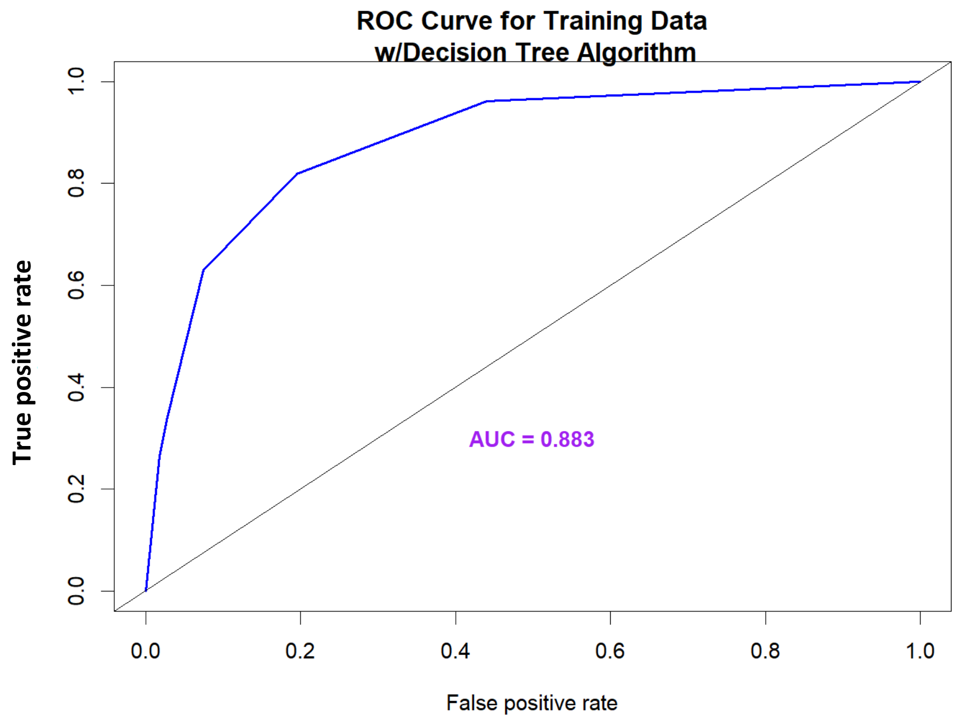
Binary Logistic Regression



- Classifier is relatively effective in distinguishing between the positive and negative classes.
- The curve then moves towards the top-right corner (1,1), but not as steeply as a perfect model would.
- The AUC value **(92.4%)** indicates excellent performance of the BLR model in distinguishing between the classes. This also means that the model has a high true positive rate and a relatively low false positive rate across different threshold level.

Model Development

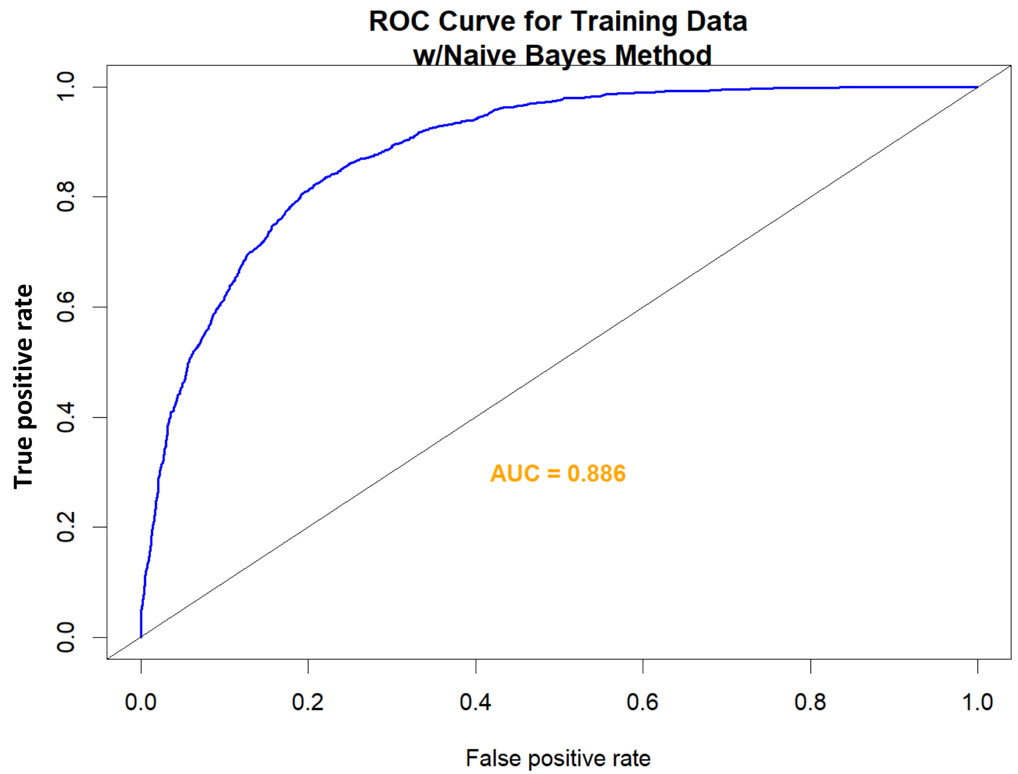
Decision Tree Method



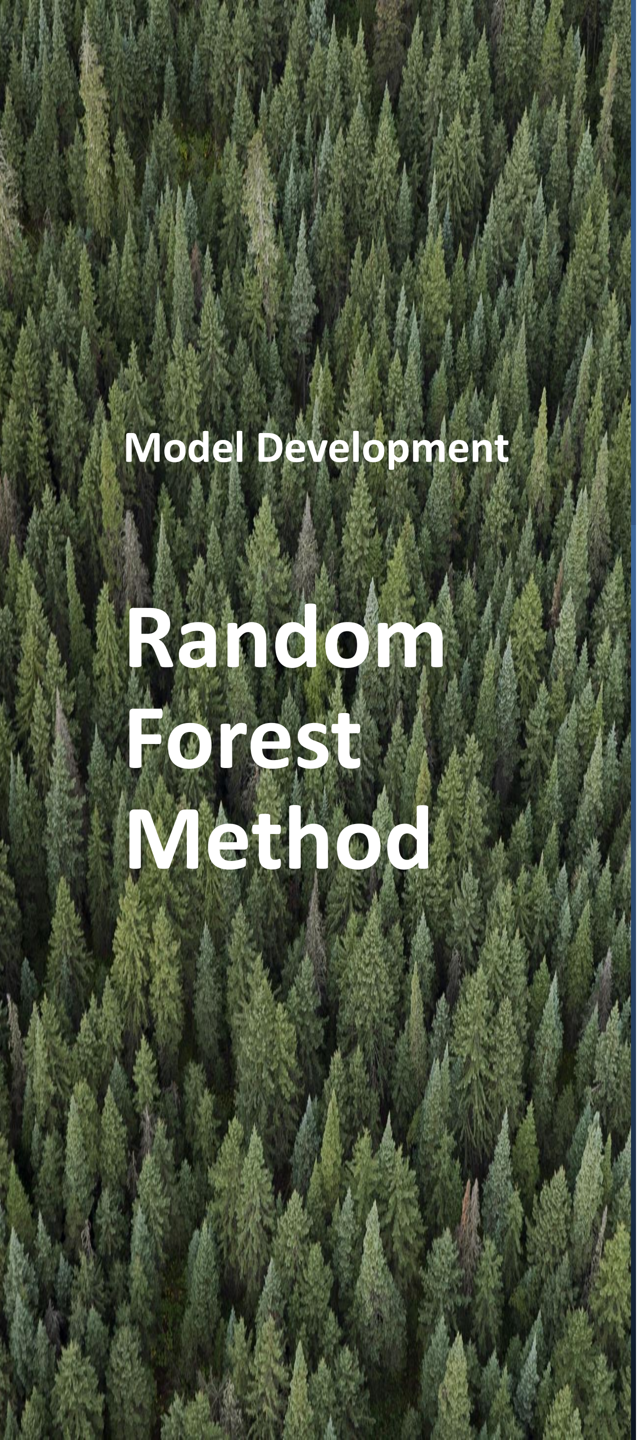
- Classifier is relatively effective in distinguishing between the positive and negative classes.
- The curve then moves towards the top-right corner (1,1), but not as steeply as a perfect model would.
- The AUC value of **88.3%** indicates good performance of the Decision Tree algorithm in distinguishing between the classes.
- This also suggest that the model has a high true positive rate and a relatively low false positive rate across different threshold levels.

Model Development

Naïve Bayes Classifier

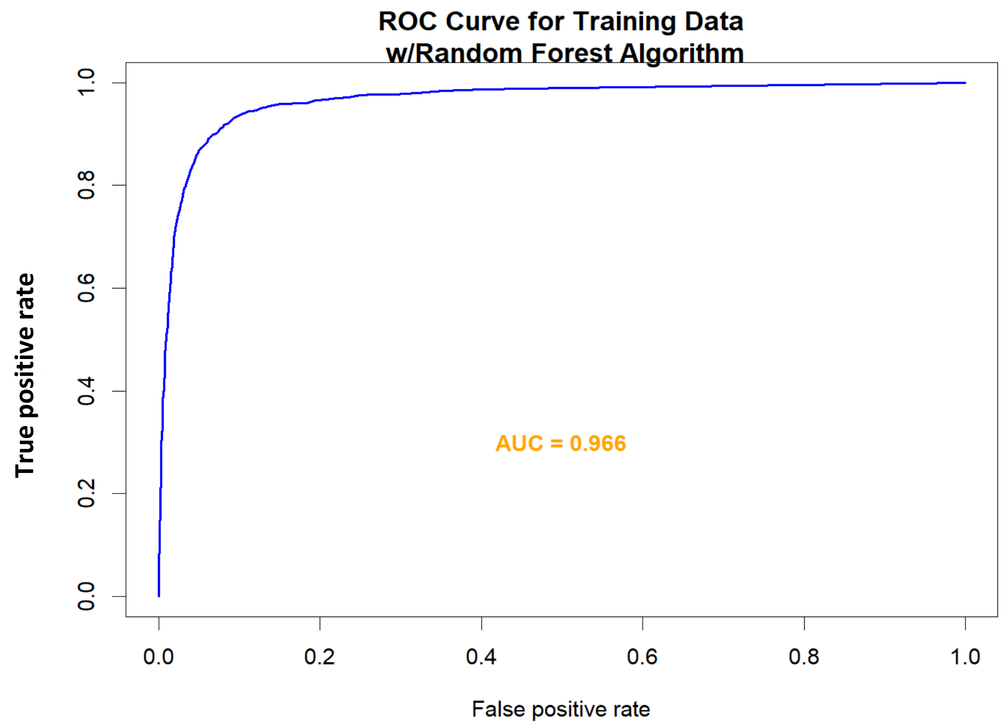


- Classifier is relatively effective in distinguishing between the positive and negative classes.
- The curve then moves towards the top-right corner (1,1), but not as steeply as a perfect model would.
- The AUC value of **88.6%** indicates good performance of the Naïve Bayes Method in distinguishing between the classes. This also means that the model has a high true positive rate and a relatively low false positive rate across different threshold levels.



Model Development

Random Forest Method

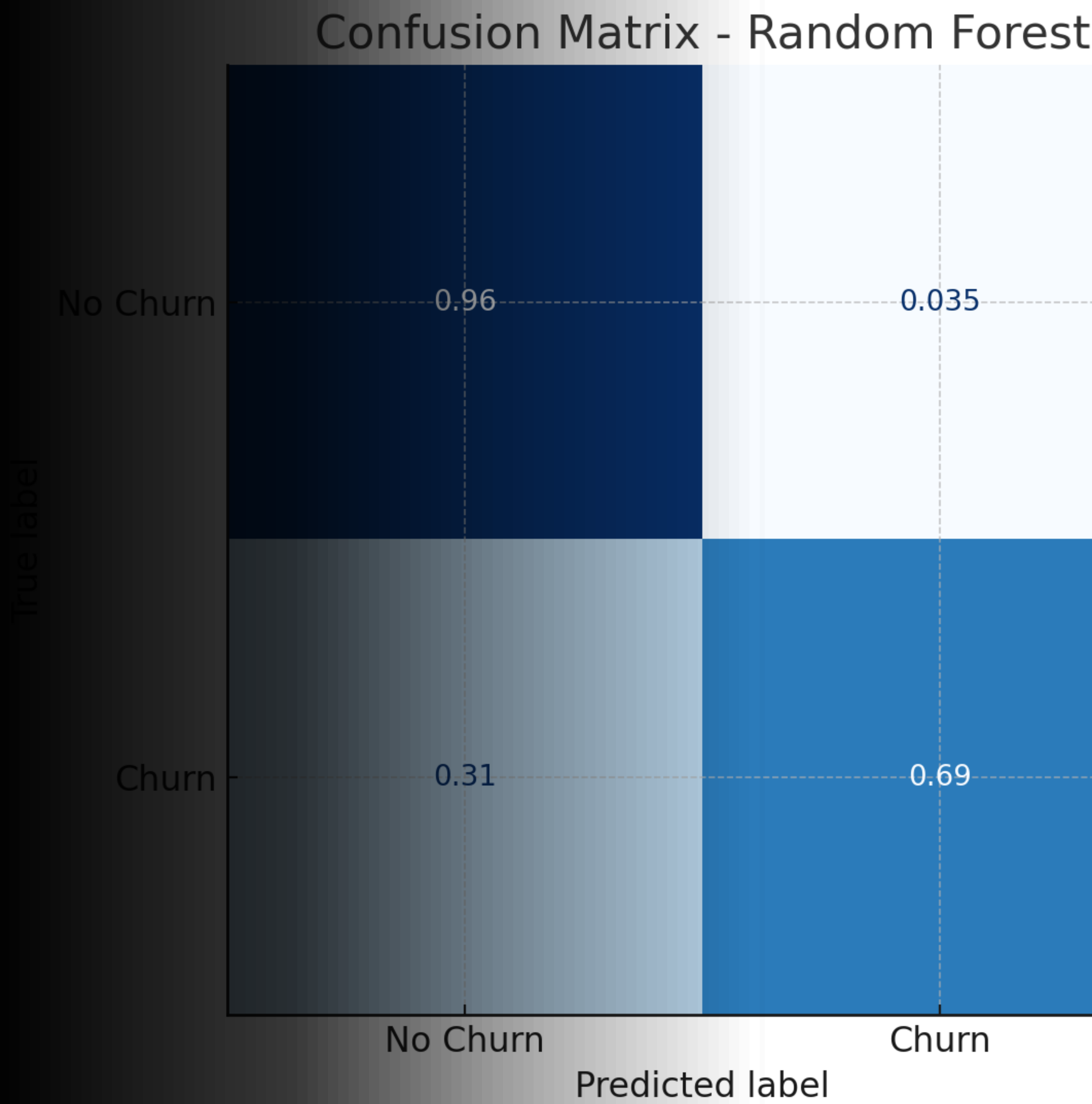


- Classifier is very effective in distinguishing between the positive and negative classes, especially at lower false positive rates.
- The model's performance with an AUC of **96.6%** indicates a significantly better than random guessing, which would be represented by the diagonal line.

Model Evaluation

- Metrics for Evaluation:
  - Accuracy, Precision, Recall, F1 Score

- Comparison of Model Performance:
  - Confusion Matrix for each model



Binary Logistic Regression / Confusion Matrix

Model Evaluation

Binary Logistic Regression / Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 113224  12699
         1  10825  32284

               Accuracy : 0.8608
                 95% CI : (0.8592, 0.8625)
    No Information Rate : 0.7339
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6389

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9127
            Specificity : 0.7177
         Pos Pred Value : 0.8992
         Neg Pred Value : 0.7489
             Prevalence : 0.7339
         Detection Rate : 0.6698
   Detection Prevalence : 0.7450
      Balanced Accuracy : 0.8152

       'Positive' Class : 0
```

- Overall accuracy of the model is 86.08%, indicating that the model correctly predicts the outcome of in 86.08% of the cases.
- Kappa statistic accounts for agreement occurring by chance. A Kappa of 0.6389 suggests moderate agreement beyond chance.
- Sensitivity is 91.27%, meaning the model correctly identifies 91.27% of the actual positive cases.
- The Specificity is 71.77%, indicating the model correctly identifies 71.77% of the actual negative cases.

Accuracy Formula:
(TP+TN)/(TP+TN+FP+FN)

Specificity Formula:
TN/(TN+FP)

Sensitivity Formula:
TP/(TP+FN)

Decision Tree / Confusion Matrix

Decision Tree / Confusion Matrix

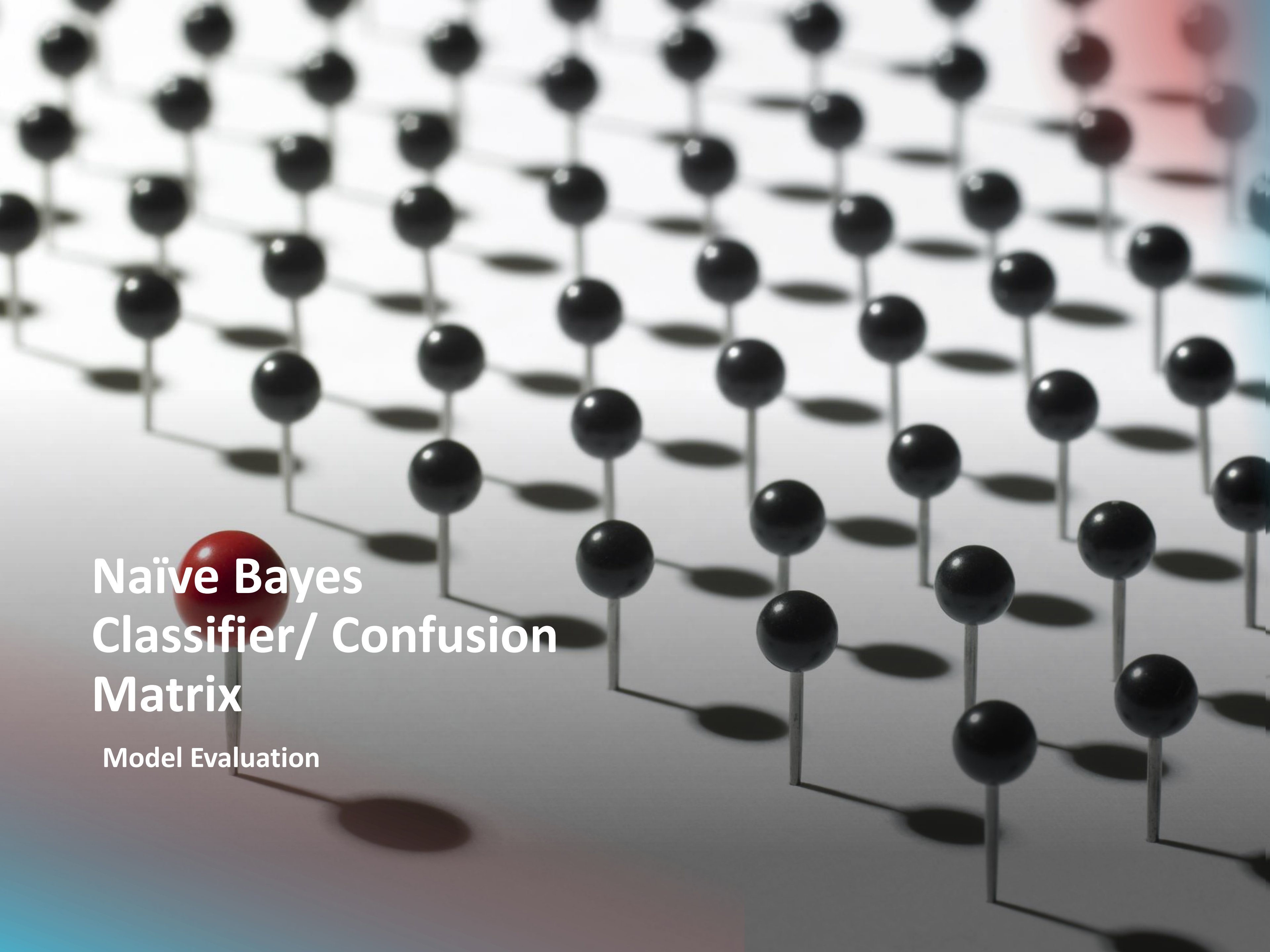
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 99795  8137
         1 24254 36846

               Accuracy : 0.8084
                 95% CI : (0.8065, 0.8102)
    No Information Rate : 0.7339
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5597

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8045
            Specificity : 0.8191
         Pos Pred Value : 0.9246
         Neg Pred Value : 0.6030
             Prevalence : 0.7339
         Detection Rate : 0.5904
   Detection Prevalence : 0.6385
      Balanced Accuracy : 0.8118

       'Positive' Class : 0
```

- Overall accuracy of the model is 80.84%, indicating that the model correctly predicts the outcome of in 80.84% of the cases.
- Kappa statistic accounts for agreement occurring by chance. A Kappa of 0.5597 suggests moderate agreement beyond chance.
- Sensitivity is 80.45%, meaning the model correctly identifies 80.45% of the actual positive cases.
- The Specificity is 81.91%, indicating the model correctly identifies 81.91% of the actual negative cases.

Accuracy Formula:
(TP+TN)/(TP+TN+FP+FN)

Specificity Formula:
TN/(TN+FP)

Sensitivity Formula:
TP/(TP+FN)



Naïve Bayes Classifier/ Confusion Matrix

Model Evaluation

Naïve Bayes / Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 97553  7814
         1 26496 37169

               Accuracy : 0.797
                 95% CI : (0.7951, 0.7989)
    No Information Rate : 0.7339
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5411

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7864
            Specificity : 0.8263
         Pos Pred Value : 0.9258
         Neg Pred Value : 0.5838
             Prevalence : 0.7339
         Detection Rate : 0.5771
   Detection Prevalence : 0.6234
      Balanced Accuracy : 0.8063

       'Positive' Class : 0
```

Overall accuracy of the model is 79.7%, indicating that the model correctly predicts the outcome of in 79.7% of the cases.

Kappa statistic accounts for agreement occurring by chance. A Kappa of 54.11% suggests moderate agreement beyond chance.

Sensitivity is 78.64%, meaning the model correctly identifies 78.64% of the actual positive cases.

The Specificity is 82.63%, indicating the model correctly identifies 82.63% of the actual negative cases.

Accuracy Formula:

$$(TP+TN)/(TP+TN+FP+FN)$$

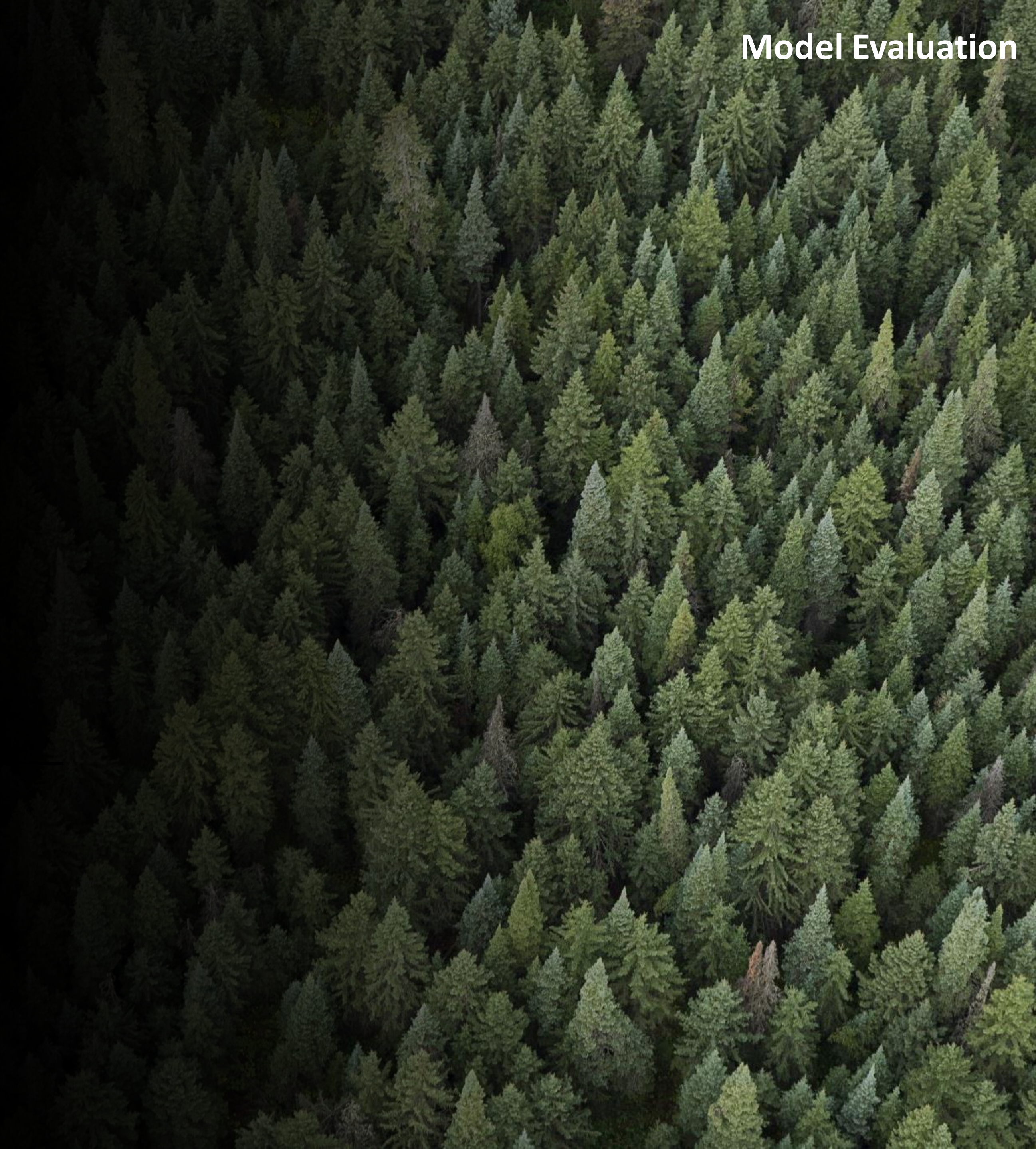
Specificity Formula:

$$TN/(TN+FP)$$

Sensitivity Formula:

$$TP/(TP+FN)$$

Random Forest / Confusion Matrix



Random Forest Method/Confusion Matrix

Model Development

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 112433   3030
         1  11616  41953

               Accuracy : 0.9134
                 95% CI : (0.912, 0.9147)
    No Information Rate : 0.7339
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7909

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9064
            Specificity : 0.9326
         Pos Pred Value : 0.9738
         Neg Pred Value : 0.7832
             Prevalence : 0.7339
         Detection Rate : 0.6652
   Detection Prevalence : 0.6831
      Balanced Accuracy : 0.9195

       'Positive' Class : 0
```

- Accuracy of the model is 91.34%, indicating that the model correctly predicts the outcome of in 91.34% of the cases.
- Kappa statistic accounts for agreement occurring by chance. A Kappa of 79.09% suggests substantial agreement beyond chance.
- Sensitivity is 90.64%, meaning the model correctly identifies 90.64% of the actual positive cases.
- The Specificity is 93.26%, indicating the model correctly identifies 93.26% of the actual negative cases.

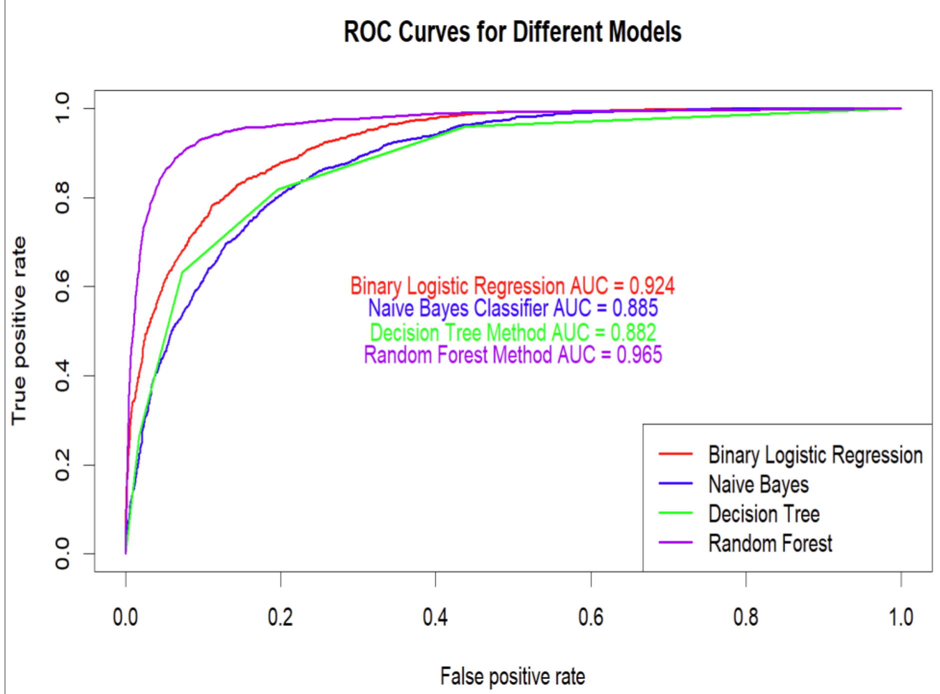
Accuracy Formula:
(TP+TN)/(TP+TN+FP+FN)

Specificity Formula:
TN/(TN+FP)

Sensitivity Formula:
TP/(TP+FN)

Best Model Selection

- Detailed Analysis of Best Performing Model:
  - Summary of model performances using resamples
  - Confusion Matrix and ROC Curve



- The Random Forest model has the highest AUC value, indicating the best performance among the four models.
- The ROC curve for the Random Forest model is closest to the top-left corner, showing high true positive rates and low false positive rates across different thresholds
- The Binary Logistic Regression model also performs very well, with a high AUC value of 0.924. The curve is slightly below the Random Forest curve but still indicates strong discriminatory power.
- The Decision Tree model has the lowest AUC value of 0.882 among the four models

Conclusion

Summary of Findings:

- **Key factors influencing churn:**
  - Tenure
  - Gender
  - Contract
  - Payment Method
- **Initial Data Cleaning and Preparation:**
  - Handling missing values
  - Data type conversion
- **Best model for predicting churn**
  - Random Forest Classifier (96%)

Future Work and Potential Improvements:

- Further parameter tuning of models
- Incorporation of additional data sources

THANK YOU!