


# **PREDICTION OF SYRIATEL COMPANY CUSTOMER CHURN**



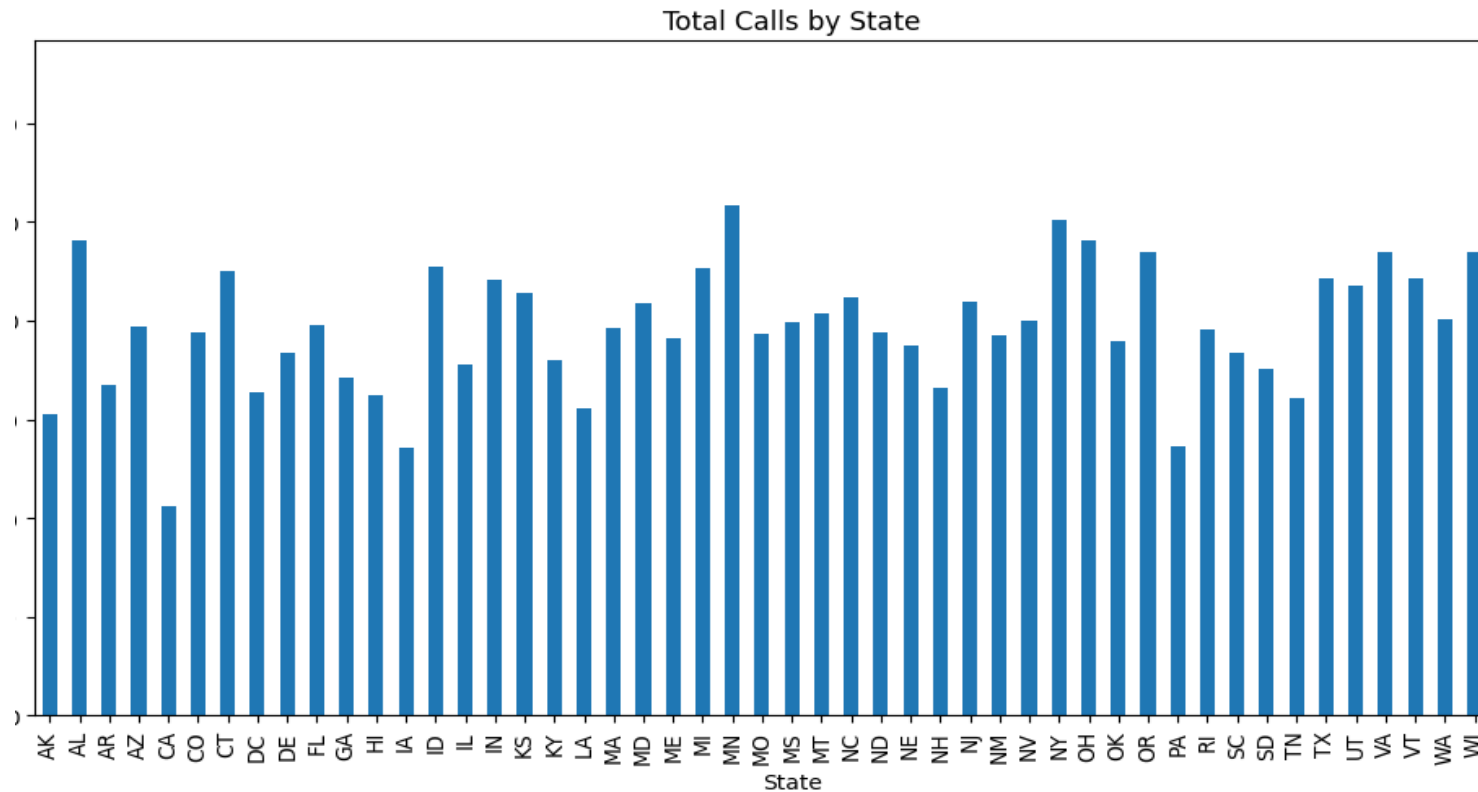
# Business Understanding

Due to increased competition in the telecommunication industry, there has been concerns on predicting the customer churn in order to retain customers. In order to maintain competitive parity by SyriaTel company, customer retention is a key to the business. Since retaining customers will be cheaper than acquiring new ones it is therefore paramount to deduce ways of making sure customers are not lost by the SyriaTel company. By analyzing various data points and using predictive analytics, SyriaTel company will identify patterns and indicators that help them anticipate customer behavior and take proactive measures to reduce churn rates. Therefore, SyriaTel is seeking to models that will predict whether customers are likely to churn or not hence taking a proactive measures to retain them. Data on demographics like location and usage patterns like calls, charge etc will be used to analyze and build predictive machine learning models, hence SeriaTel must continuously adapt its models to reflect changing customer behavior and market dynamics to ensure its retention strategies remain effective.

# Objectives

- Create machine learning models that can predict customer churn by using data to analyze customer features.
  - Comparing the build machine learning models and determine the most accurate model in prediction.
  - The analysis aims to identify the specific features that have a significant impact on the customer churn rate in SyriaTel, provide valuable recommendations based on the findings hence help to mitigate churn rates in the company and improve customer retention.
- 

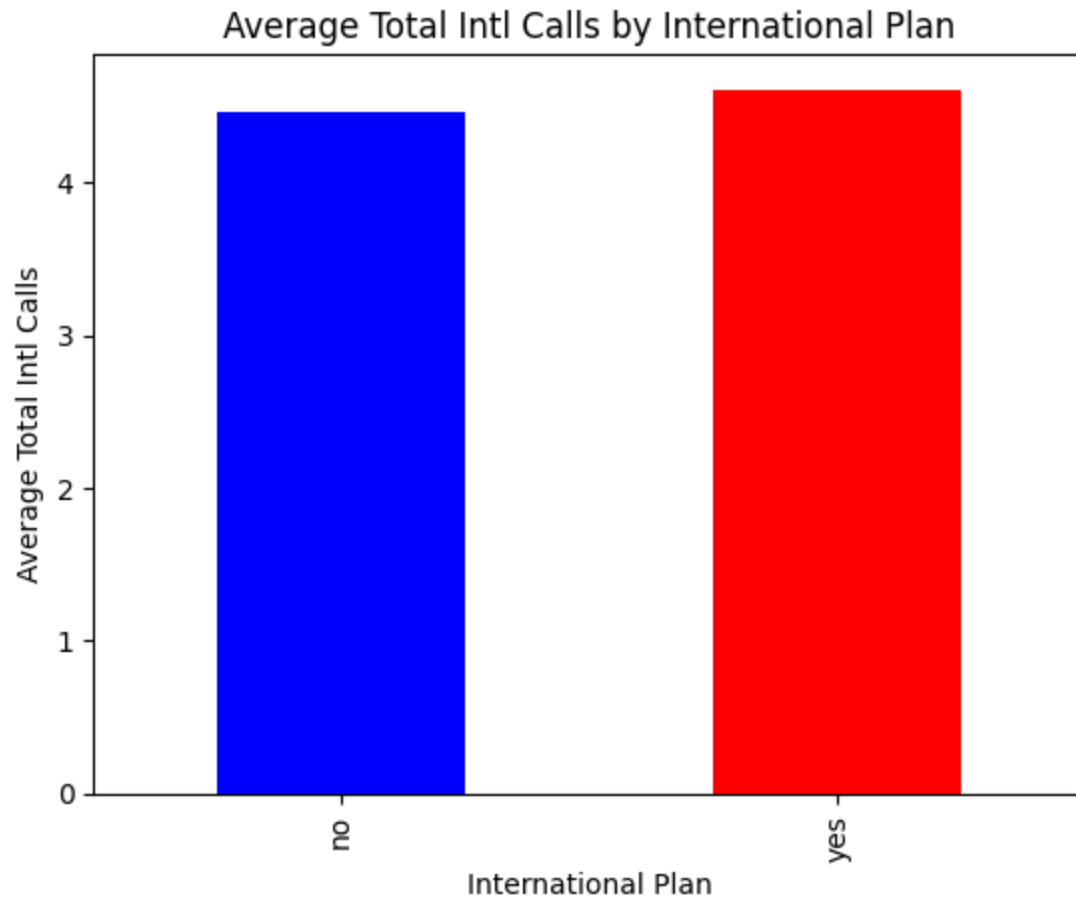
# DATA EXPLORATION



## Total calls per state:

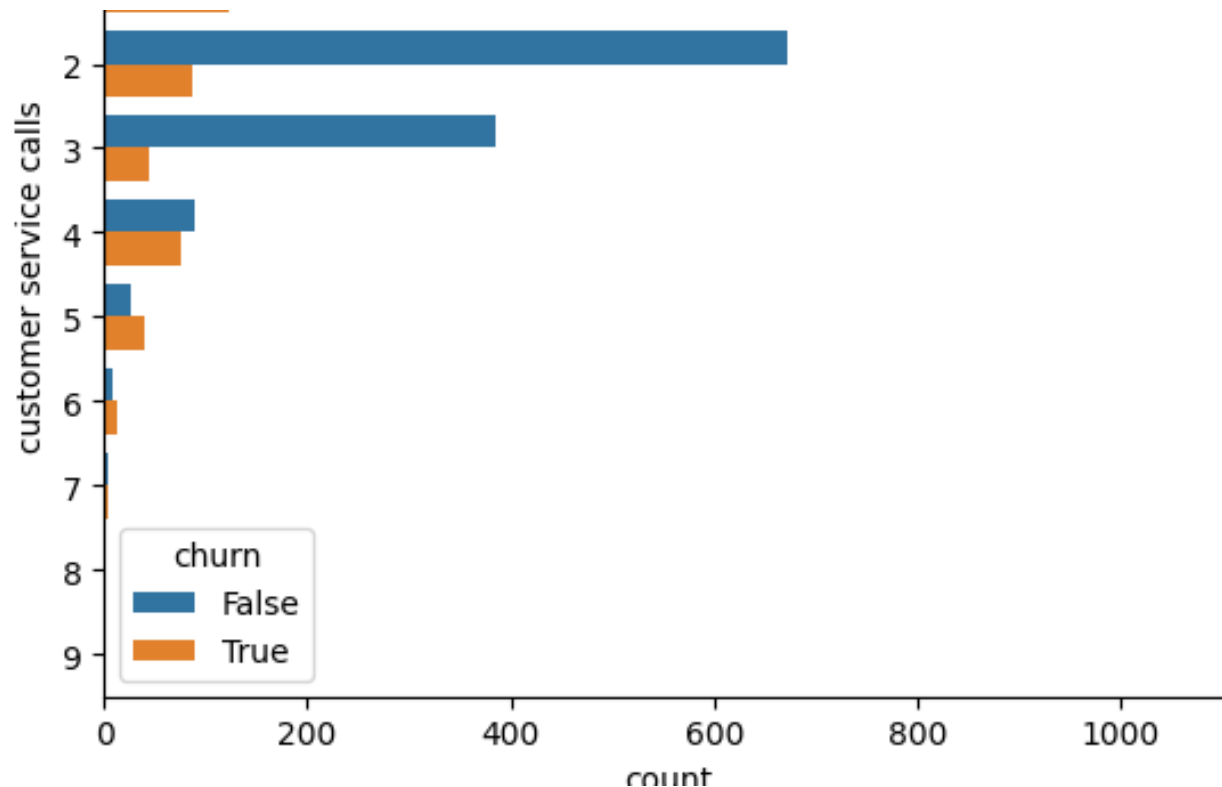
The chart indicates that West (WV) state has the highest number of total calls while (CA) has the least number of total calls

## Visualizing for the international calls Made



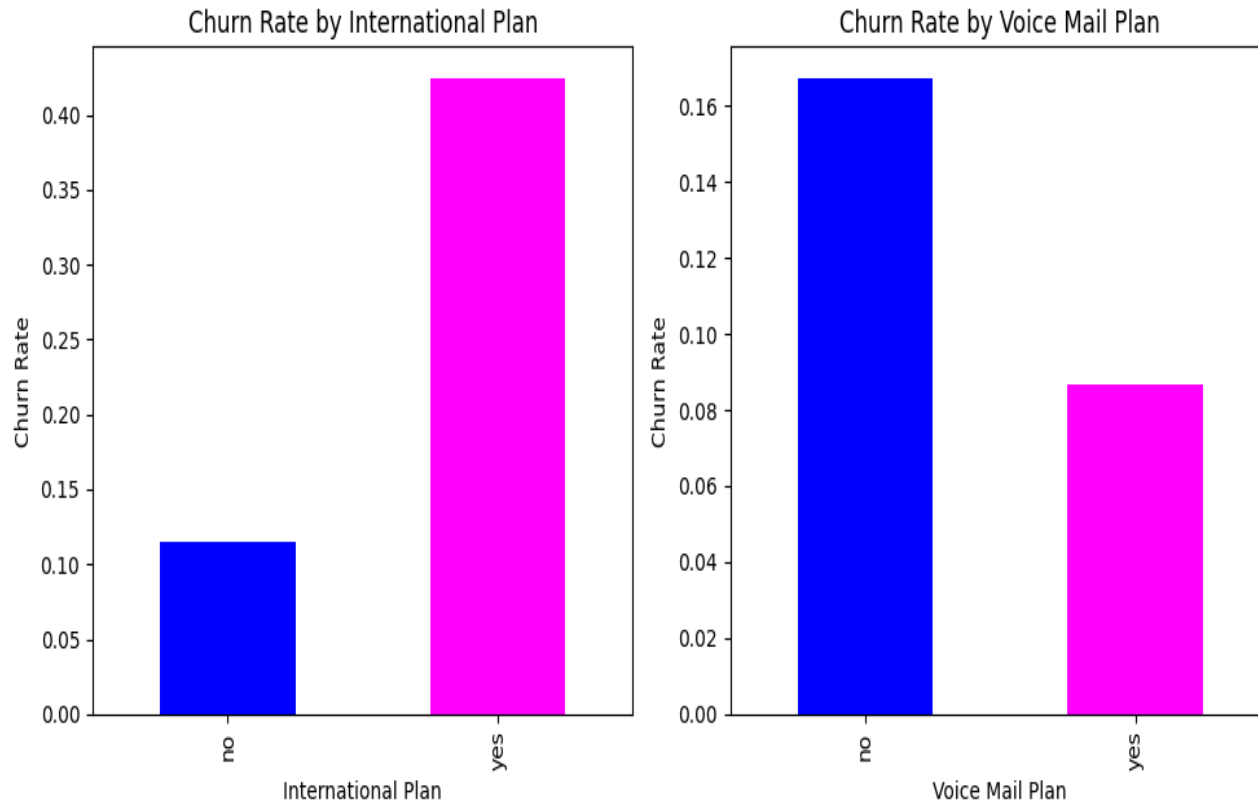
The bar graph indicates that there is little variation among people making international calls concerning their choice of an international plan. On average, the number of international calls remains nearly consistent, regardless of whether individuals opt for an international plan or not.

visualizing the relationship between the number of calls to the call center and loyalty



The chart provided illustrates the connection between the frequency of calls to the call center and customer loyalty. The data on the chart reveals a significant correlation between the number of calls and loyalty. Specifically, it suggests that a substantial portion of individuals who engage in these calls are loyal to Syriatel, resulting in a lower likelihood of them switching to another service provider.

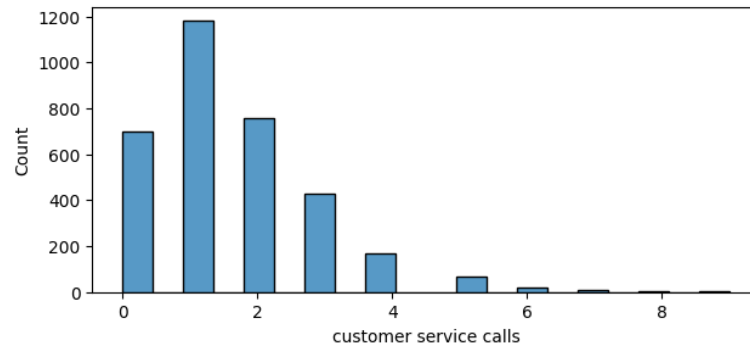
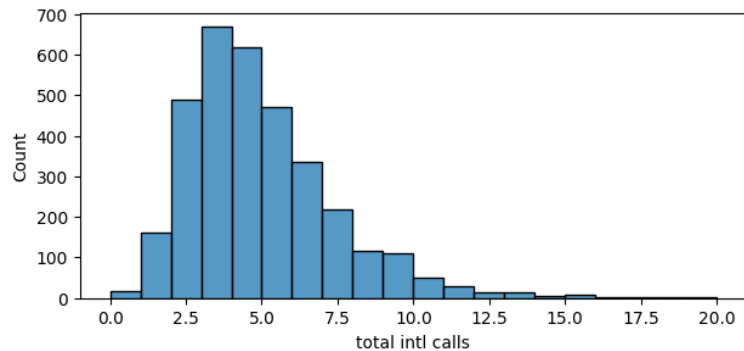
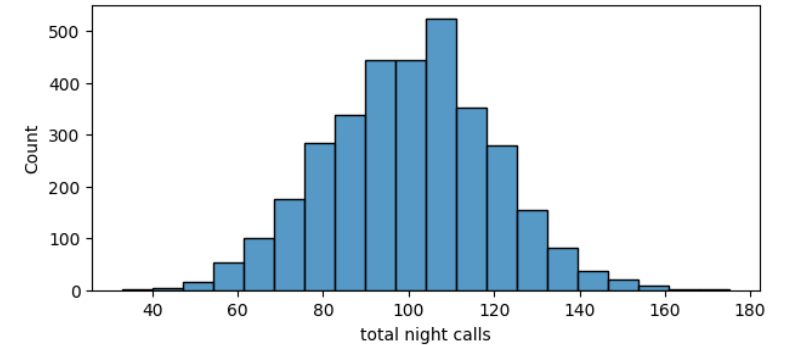
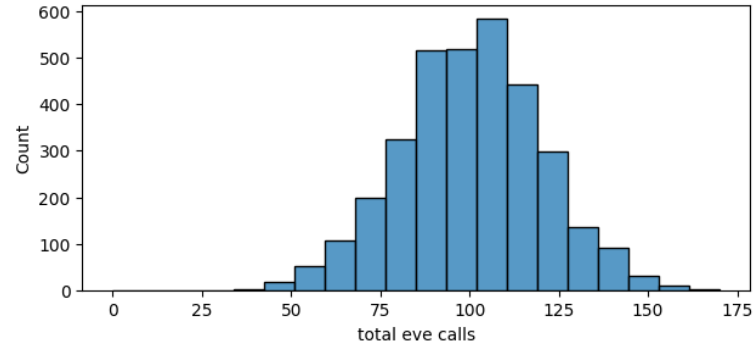
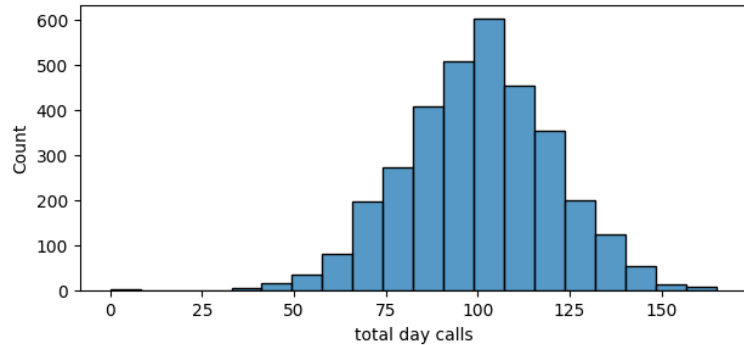
## visualization of churn rate given the international plan and the voice mail plan



- The graphs below indicates that Voice Mail plan subscribers exhibit higher customer loyalty, primarily attributed to their lower churn rate. Conversely, customers who have subscribed to the international plan display a higher churn rate, indicating a greater probability of switching from Syriatel.
- These findings suggest that customers express greater satisfaction with the Voice Mail plan compared to the international plan.

## visualization of calls distribution

- All calls exhibit a normal distribution, with the exception of customer service calls. Total international calls display a slight right skew, although it still maintains a relatively normal distribution.
- Customer service calls, exhibit multiple peaks indicating the presence of several modes within the population. This observation is logical since customer service calls are discrete integers and not continuous float numbers





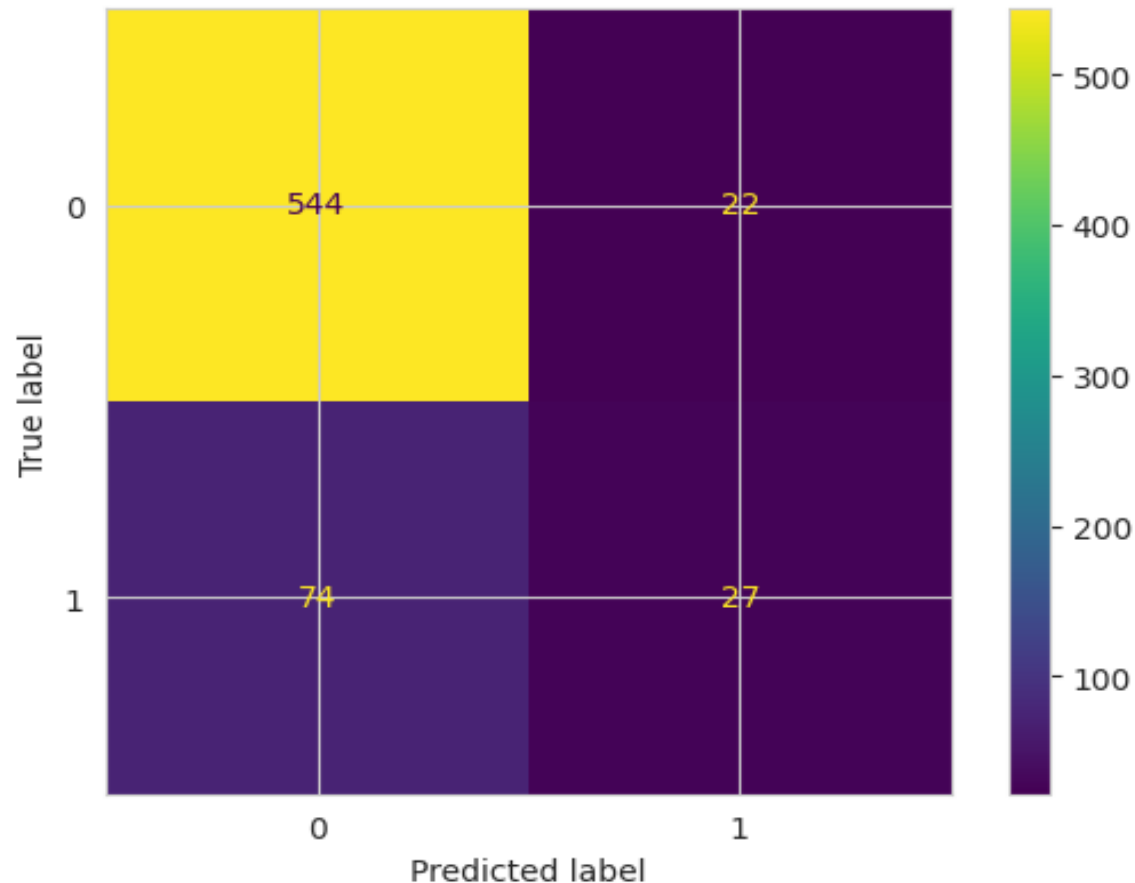
heatmap to check the correlation existing between variables in order to select the best features for modelling

Correlation Matrix

messages	1.00	0.00	-0.01	0.00	0.02	-0.01	0.02	0.01	0.01	0.01	0.00	0.01	0.00	-0.01
minutes	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01
day calls	-0.01	0.01	1.00	0.01	-0.02	0.01	-0.02	0.02	-0.02	0.02	0.02	0.00	0.02	-0.02
day charge	0.00	1.00	0.01	1.00	0.01	0.02	0.01	0.00	0.02	0.00	-0.01	0.01	-0.01	-0.01
minutes	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01
eve calls	-0.01	0.02	0.01	0.02	-0.01	1.00	-0.01	-0.00	0.01	-0.00	0.01	0.02	0.01	0.00
day charge	0.02	0.01	-0.02	0.01	1.00	-0.01	1.00	-0.01	0.01	-0.01	-0.01	0.00	-0.01	-0.01
minutes	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01
night calls	0.01	0.02	-0.02	0.02	0.01	0.01	0.01	0.01	1.00	0.01	-0.01	0.00	-0.01	-0.01
night charge	0.01	0.00	0.02	0.00	-0.01	-0.00	-0.01	1.00	0.01	1.00	-0.02	-0.01	-0.02	-0.01
minutes	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01
intl calls	0.01	0.01	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	-0.01	0.03	1.00	0.03	-0.02
intl charge	0.00	-0.01	0.02	-0.01	-0.01	0.01	-0.01	-0.02	-0.01	-0.02	1.00	0.03	1.00	-0.01
customer service calls	-0.01	-0.01	-0.02	-0.01	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	1.00
number vmail messages														
total day minutes														
total day calls														
total day charge														
total eve minutes														
total eve calls														
total eve charge														
total night minutes														
total night calls														
total night charge														
total intl minutes														
total intl calls														
total intl charge														
customer service calls														

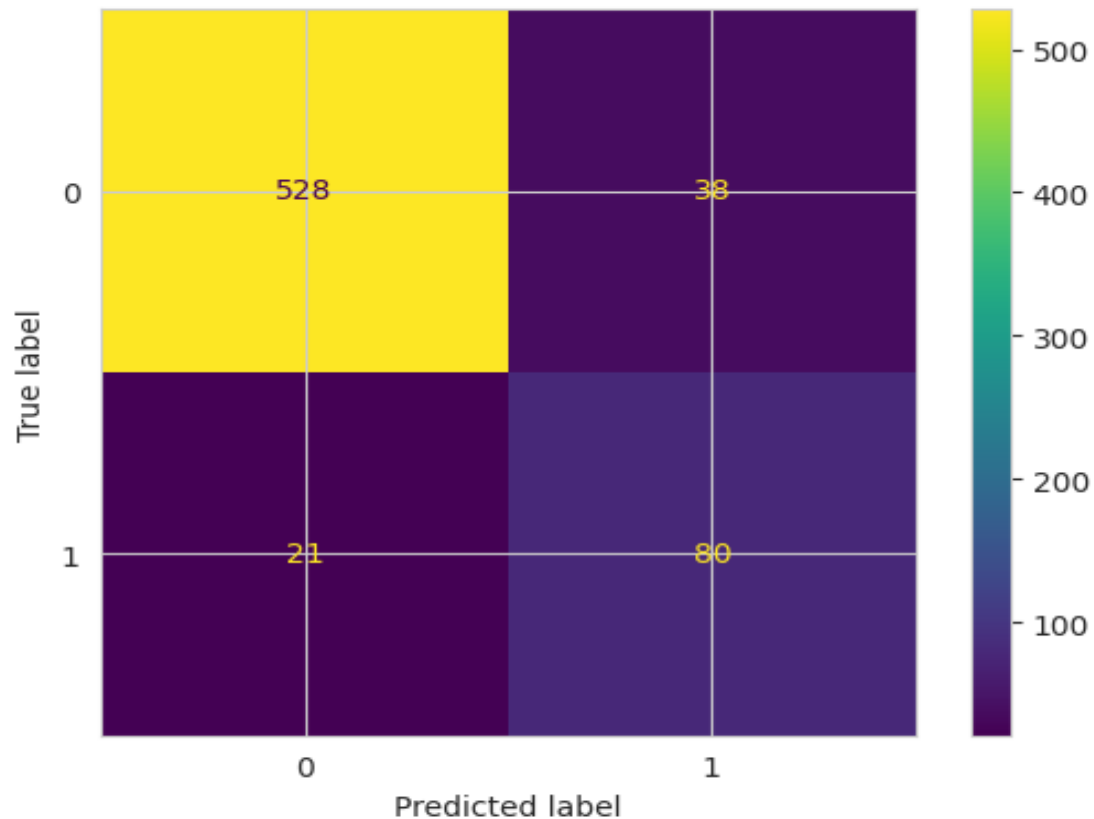
From the above heat-map, some variables exhibit a correlation coefficient of 1, indicating a perfect correlation with other variables, thereby making them redundant for certain analyses. Highly correlated variables potentially cause instability thus producing unreliable estimates for the model parameters. Therefore we remove the highly correlated variables

## MODEL 1: BASELINE LOGISTIC REGRESSION



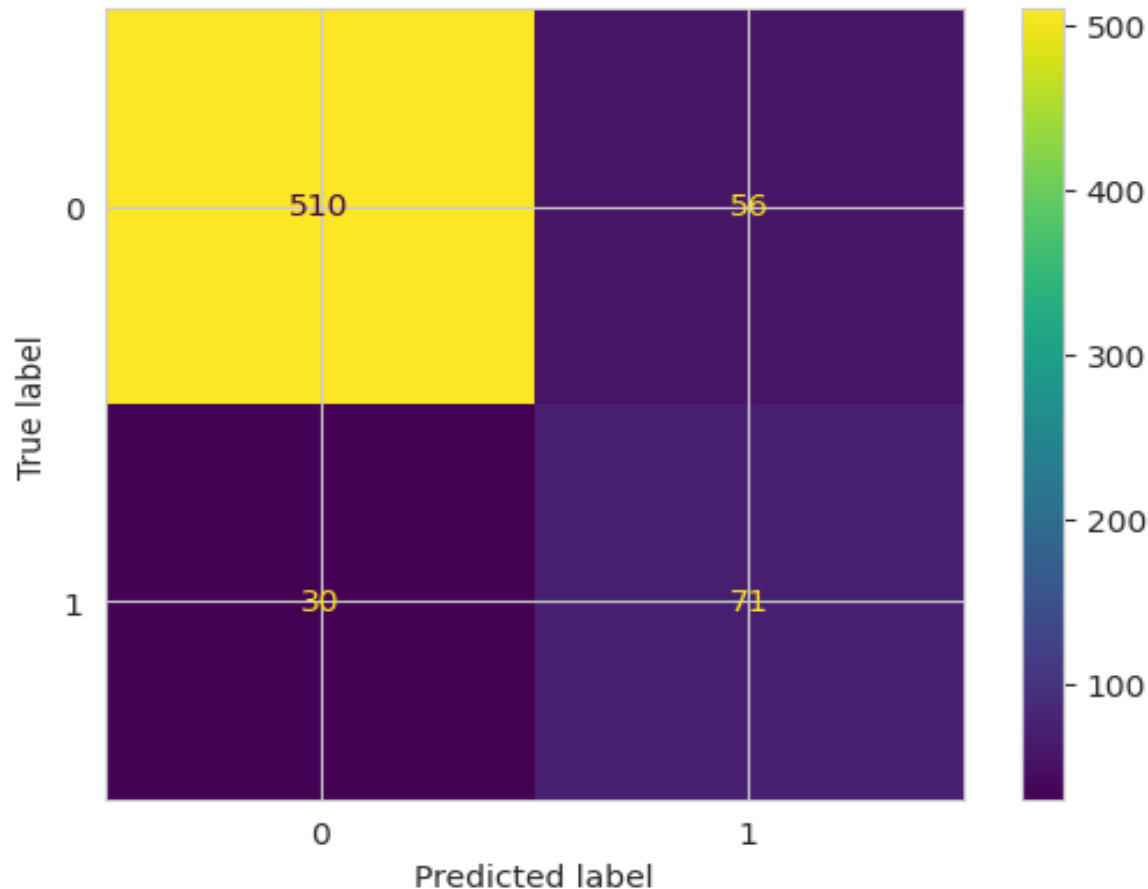
- The model has a training accuracy of approximately 89.8% and a testing accuracy of around 85.6%.
- There was an overfit and hence cross-validation with 5 folds was done
- This improved testing data accuracy to 84.8% and training data accuracy dropped to 89.7%

## MODEL 2: Decision Trees classifier model



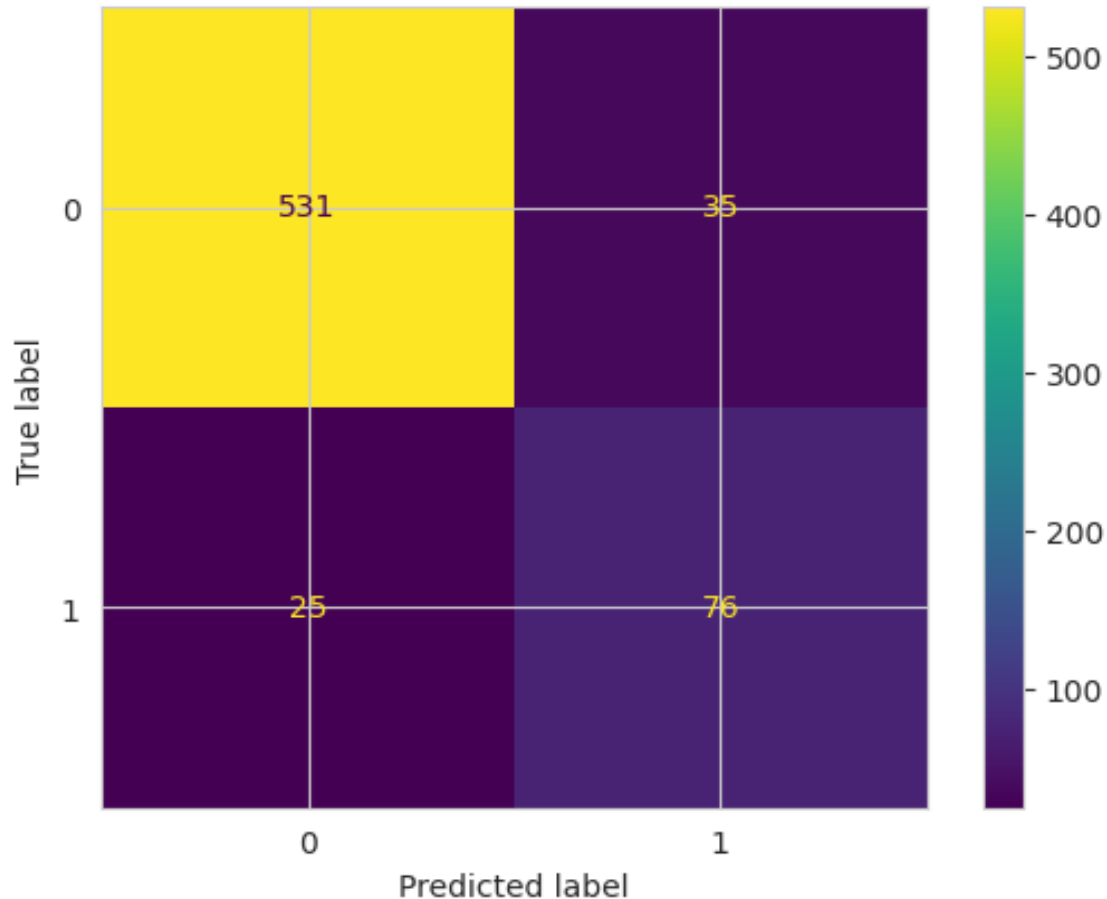
- The model has a training accuracy of approximately 88.6% and a testing accuracy of around 99.1% on testing data for predicting customer churn

## MODEL 3. Random Forest Model



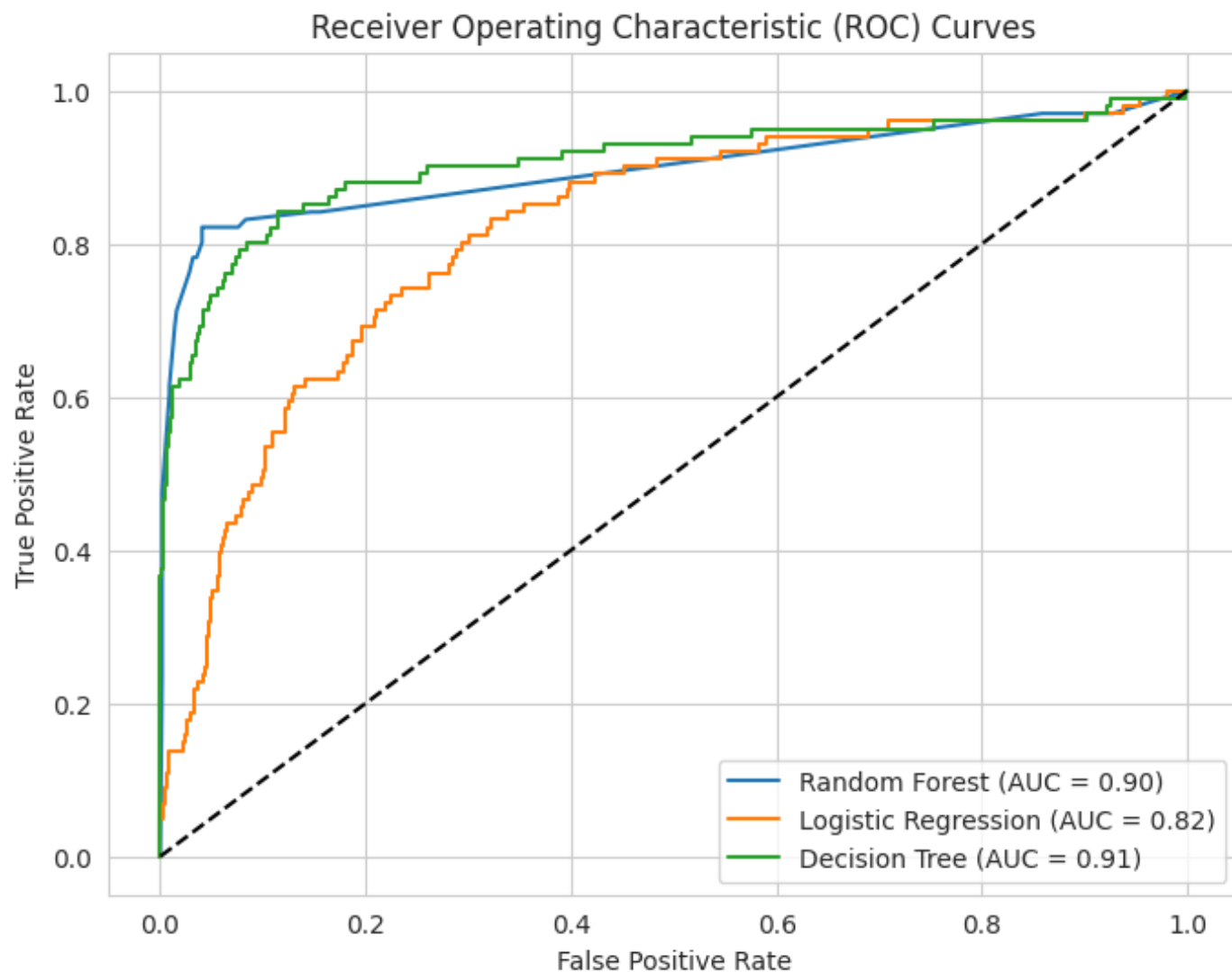
Random Forest classifier has an accuracy of approximately 86.8% on the training data and 87.1% on the testing data. This indicates a good performance in distinguishing between the positive and negative classes.

# Hyperparameter Tuning



## Random Forest

The random forest model improved on its performance after hyperparameter tuning, the accuracy on testing data is 91.0%, which indicates that the model correctly predicted the class labels for the test data with an accuracy of approximately 93.0% on the training data, hence perfect for predicting customer churn.



## Using ROC curve to check the best model

- **Decision Tree has an AUC of 0.91:** This indicates that the Random Forest model has good discriminatory power and is able to distinguish between the positive and negative classes effectively.
- **Random Forest has an AUC of 0.90:** The Random Forest model also performs well but slightly lower than decision tree classifier. It has a good ability to classify the two classes correctly, but it may have slightly higher false positive and false negative rates compared to decision tree.
- **Logistic Regression has an AUC of 0.82:** The Logistic Regression model performs the lowest among the three models in terms of discrimination. It may have a relatively higher false positive and false negative rates compared to the other models, resulting in a smaller area under the ROC curve.
- Decision Tree has the largest area under the ROC curve followed by random forest, indicating the former has the best overall discriminatory power and performs better in distinguishing between the positive and negative classes.

# Evaluation

Logistic regression performed poorly having training and testing rates of 89.8% and 85.6% respectively. Prediction of training data dropped to 89.7% while the testing data dropped to 84.8%

Decision tree classifier and Random forest Models gave better accuracy. The decision tree training and testing accuracy is 88.6% and 91.1% respectively, while random forest training accuracy is 86.8% on training data and testing accuracy of 87.1%.

Decision tree model had the best prediction in average followed by random forest model . In order to improve our prediction accuracy, we hyper-tuned the models using grid-search.

Following the hyperparameter tuning, decision tree model results improved to 91.7% accuracy on training data and 93.8% on testing data. Decision tree results improved to 93.0% on training data and 91.0% on testing data.

This shows that there was a little overfit in random forest model hence preferring decision trees. in terms of precision, recall and F1 score tuned decision tree performed better than random forest classifier.



# Conclusion

- Therefore, from the above data analysis, Seriatel company will be able to achieve alot by using the Decision Tree Classifier model which includes:
- Accurate Customer Churn Prediction:
- Cost Savings:
- Customer Retention:
- Business Strategy and Decision-Making:



# Recommendations

Continuous Model Optimization

Benchmarking

Use the insights from the model to shape the company's long-term strategic direction  
Customer Feedback Analysis

Segmentation

# Next Steps

1

Regular Model  
Reevaluation

2

Customer  
Retention  
Strategies

3

Training and  
Awareness

4

Formulate long-  
term business  
strategies that  
align with  
customer needs  
and market

