# MACHINE LEARNING PROJECT – PHASE 3

DATA SCIENCE – PART TIME

NICHOLAS KIRUI

# OUTLINE

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Model fitting**
5. **Model prediction**
6. **Model evaluation**
7. **Conclusion and Recommendations**

# BUSINESS UNDERSTANDING

- **Stakeholder**
- Syriatel has been leading the Syrian mobile telecommunication market since 2000. The company has successfully established its reputation by focusing on customer satisfaction and social responsibility. Syriatel believes that its first responsibility is to offer its customers, a wide array of high quality products and services which meet their needs and make their life easier at reasonable prices.
- In this project, the stakeholder is the Board of Directors and senior management of the SyriaTel Televommunications company.
- The Board od Directors and senior management are interested in knowing if they can predict in advance the customers who are unhappy with their services and therefore likely to stop using the telecommunication services offered by the company in the near future.

# BUSINESS PROBLEM

- The business problem in this project is being able to predict which customer is likely to leave the company in the next few months.
- Predicting which customer is likely to leave will enable the company put in place measures to retain the customers and therefore an increased likelihood of retaining the current company revenue.

# DATA UNDERSTANDING

- The data for the project was obtained from the Kaggle website: https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset
- This dataset consists of a single csv file: bigml.csv
  - Dataset has 21 variables: 'state', 'account length', 'area code', 'phone number', `international plan', 'voice mail plan', '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', 'churn'
  - The data has 3333 data instances, 0 to 3332

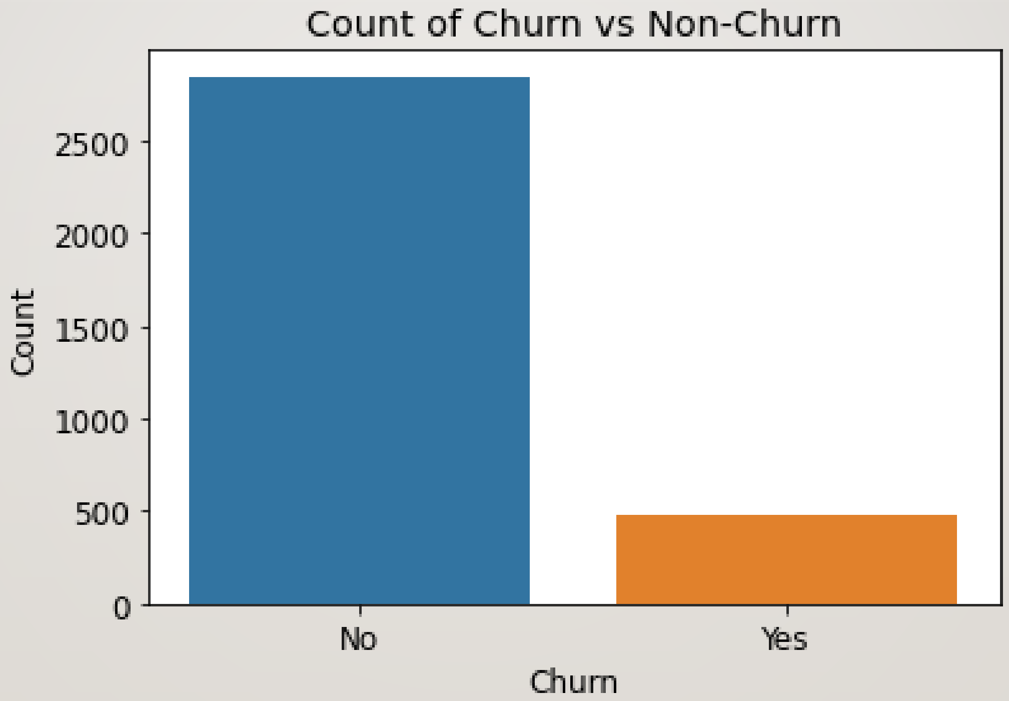
# DATA PREPARATION, ANALYSIS & MODELING

- Data was read into python using pandas
- Data was manipulated and cleaned in pandas dataframe
- Data analysis was conducted.
- Data visualization was done using matplotlib and seaborn with bar plots and line plots
- Feature engineering was done with MinMax scaler

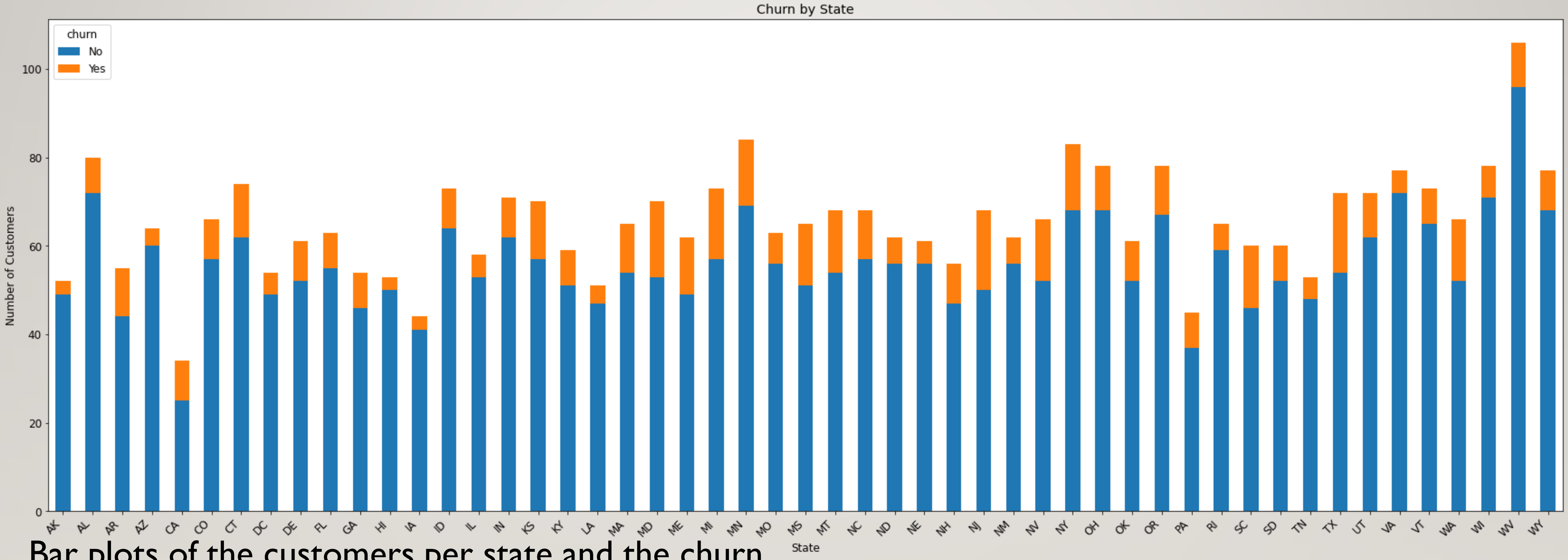
# DATA PREPARATION, ANALYSIS & MODELING

- Predictions were made using two models:
  - Primary model – Logistic regression
  - Secondary model – Decision Trees
- Evaluation of the models was done:
  - Accuracy scores
  - Classification matrices

# RESULTS 1 – DESCRIPTIVE ANALYSIS

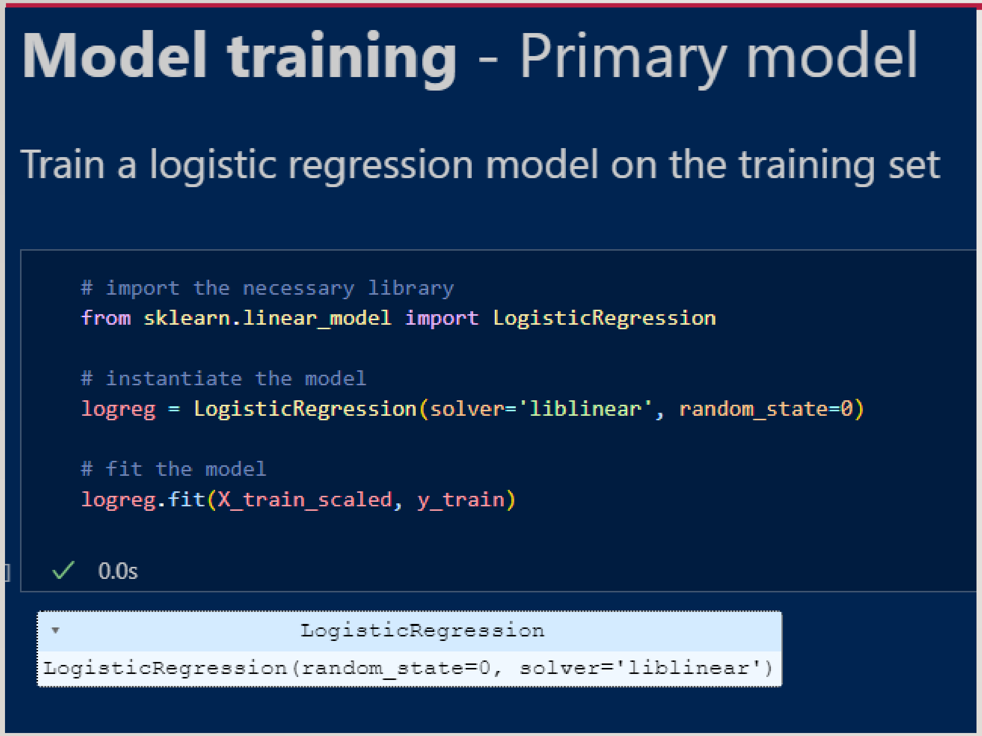


# RESULTS 2 – DESCRIPTIVE ANALYSIS



Bar plots of the customers per state and the churn

# RESULTS 3 – TRAINING THE MODEL



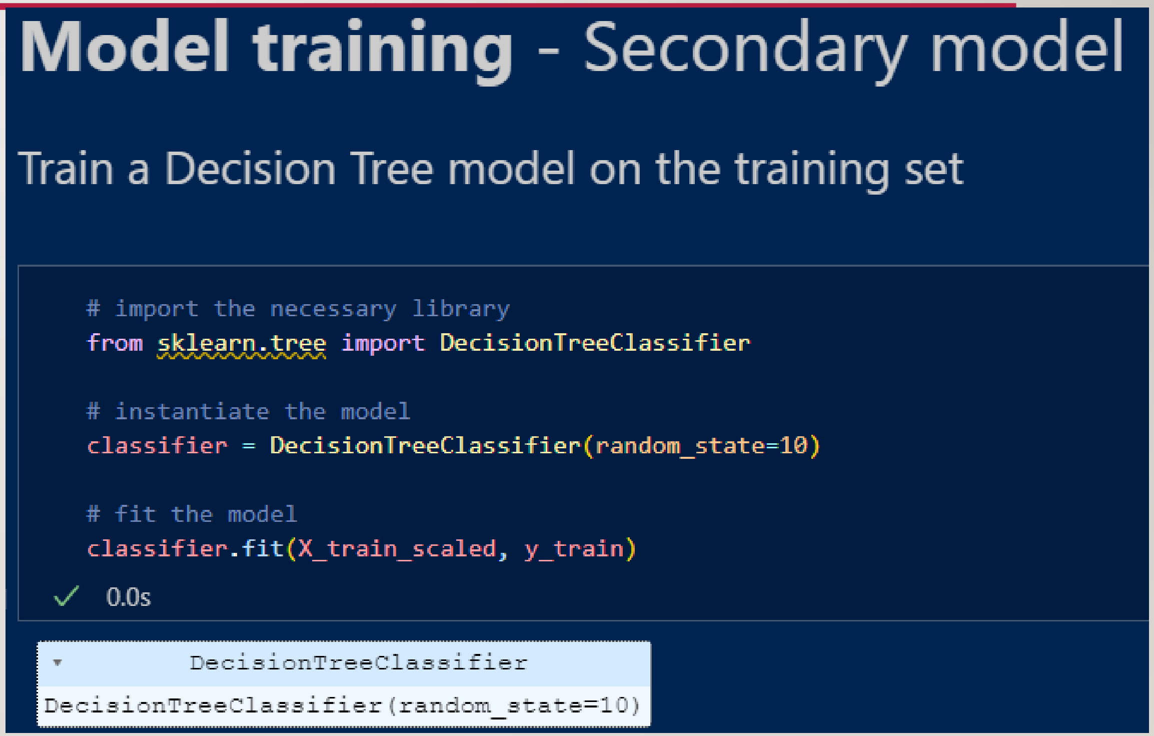
## Model training - Primary model

Train a logistic regression model on the training set

```
# import the necessary library
from sklearn.linear_model import LogisticRegression

# instantiate the model
logreg = LogisticRegression(solver='liblinear', random_state=0)

# fit the model
logreg.fit(X_train_scaled, y_train)
```

✓ 0.0s

```
LogisticRegression
LogisticRegression(random_state=0, solver='liblinear')
```



## Model training - Secondary model

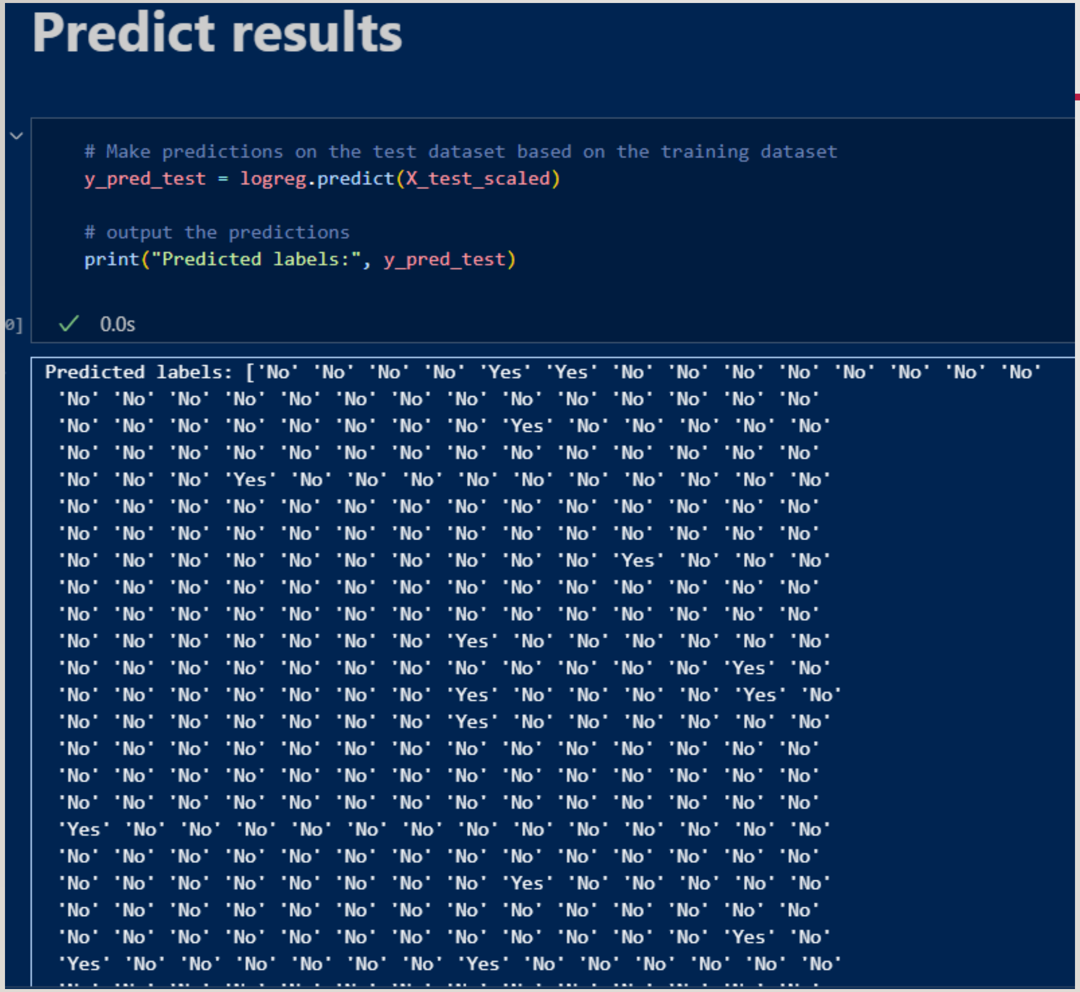
Train a Decision Tree model on the training set

```
# import the necessary library
from sklearn.tree import DecisionTreeClassifier

# instantiate the model
classifier = DecisionTreeClassifier(random_state=10)

# fit the model
classifier.fit(X_train_scaled, y_train)
```

✓ 0.0s

```
DecisionTreeClassifier
DecisionTreeClassifier(random_state=10)
```

# RESULTS 4 – PREDICTION OF RESULTS



## Predict results

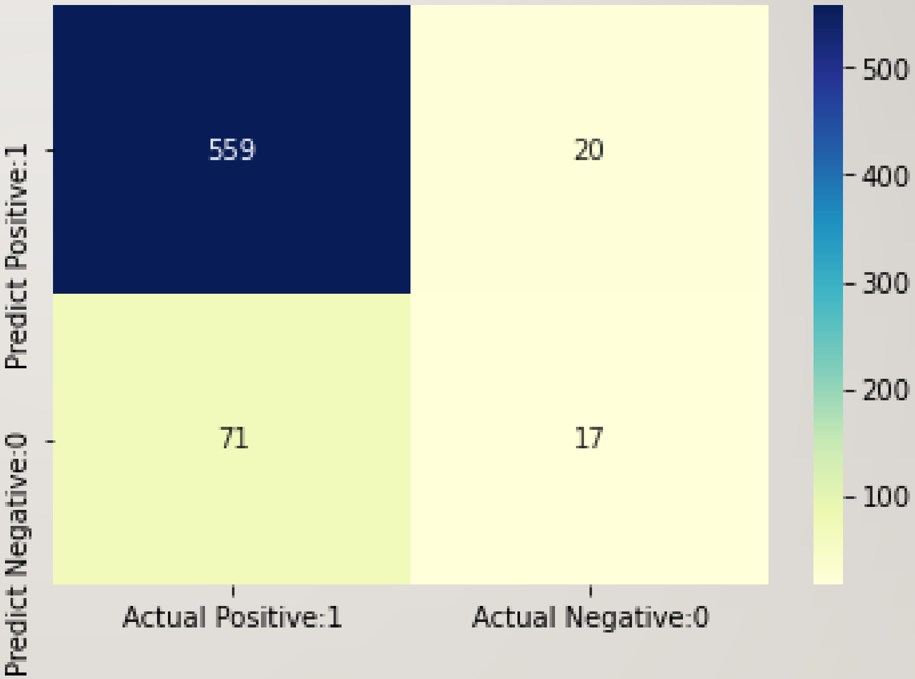
```
# Make predictions on the test dataset based on the training dataset
y_pred_test = logreg.predict(X_test_scaled)

# output the predictions
print("Predicted labels:", y_pred_test)
```

✓ 0.0s

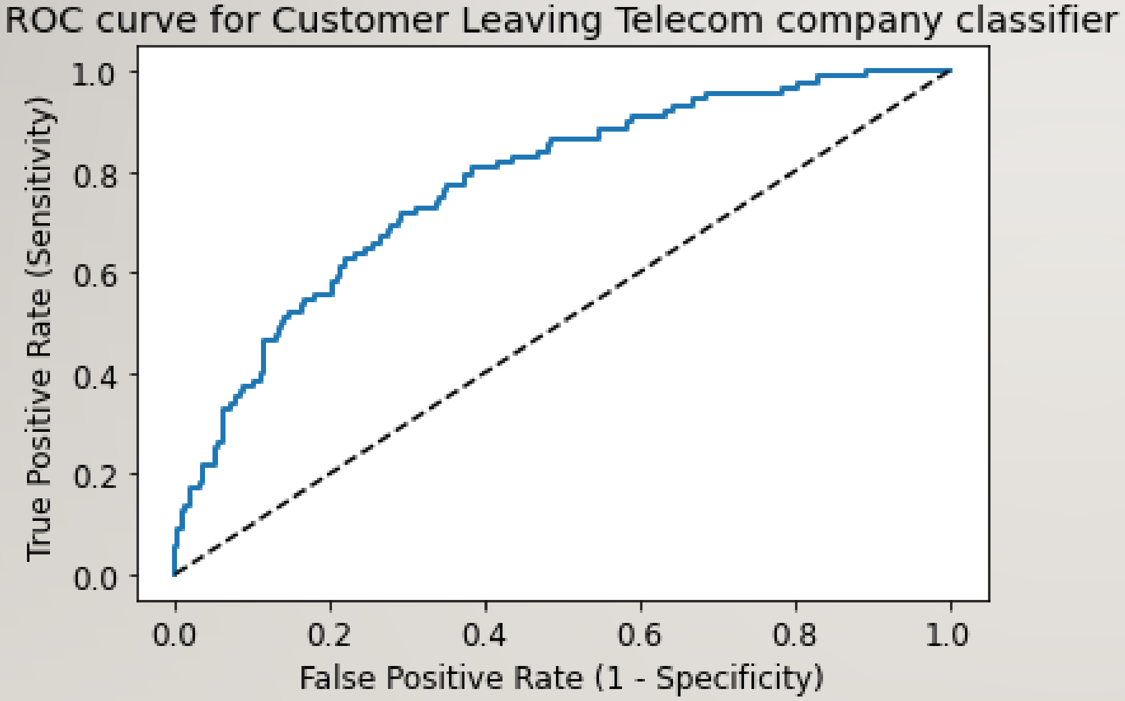
```
Predicted labels: ['No' 'No' 'No' 'No' 'Yes' 'Yes' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'Yes' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'Yes' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No'
 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No'
 'Yes' 'No' 'No' 'No' 'No' 'No' 'No' 'Yes' 'No' 'No' 'No' 'No' 'No' 'No'
```

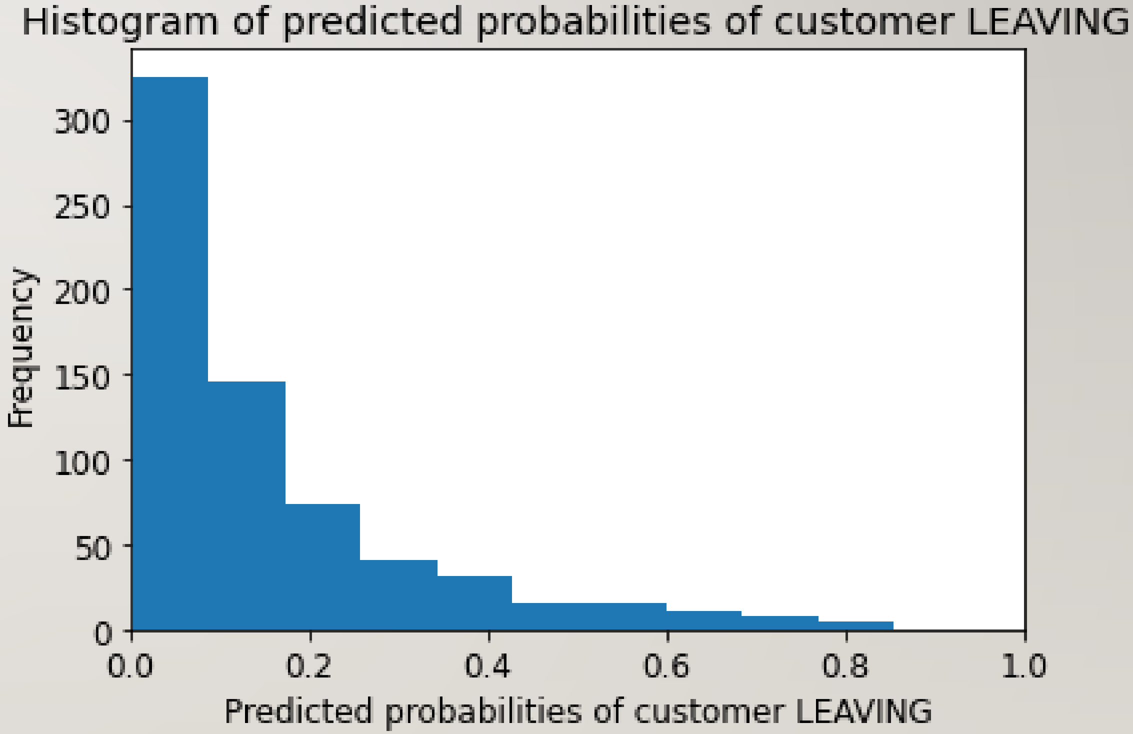
# EVALUATION OF MODEL 1

- Training set score: 0.8728
- Test set score: 0.8636



# EVALUATION OF MODEL 2





# RESULTS AND CONCLUSION

1. The logistic regression model accuracy score is **0.8636**. The model does a good job in predicting whether or not the customer will leave the telecom company.
2. Small number of observations predict that the customer will leave the telecom company. Majority of observations predict that the customer will not leave the telecom company.
3. The model shows no signs of overfitting.
4. Increasing the value of C results in lower test set accuracy but also a slightly increased training set accuracy. Therefore we conclude that a more complex model will not improve the performance of the prediction.
5. ROC-AUC original model score is found to be **0.7731**. The average ROC-AUC cross-validation score is **0.862**. Therefore, we conclude that cross-validation results in a major performance improvement.
6. Our original model test accuracy is 0.8636 while GridSearch CV accuracy is 0.8531. The GridSearch CV does not improve the performance for this particular model.

# RECOMMENDATION

- The SyriaTel Communication company can use the prediction model to determine which customers will leave the company.
- The prediction will be correct 86.4 % of the time
- The company can therefore put in place measures to mitigate the customers from leaving the company and therefore not lead to loss of revenue based on the prediction model