Machine Learning Based Telco Customer CHURN Prediction

Capstone Project Report

Name : Chathura Peiris

Course : Machine Learning Foundations

Date : 21st Nov 2021

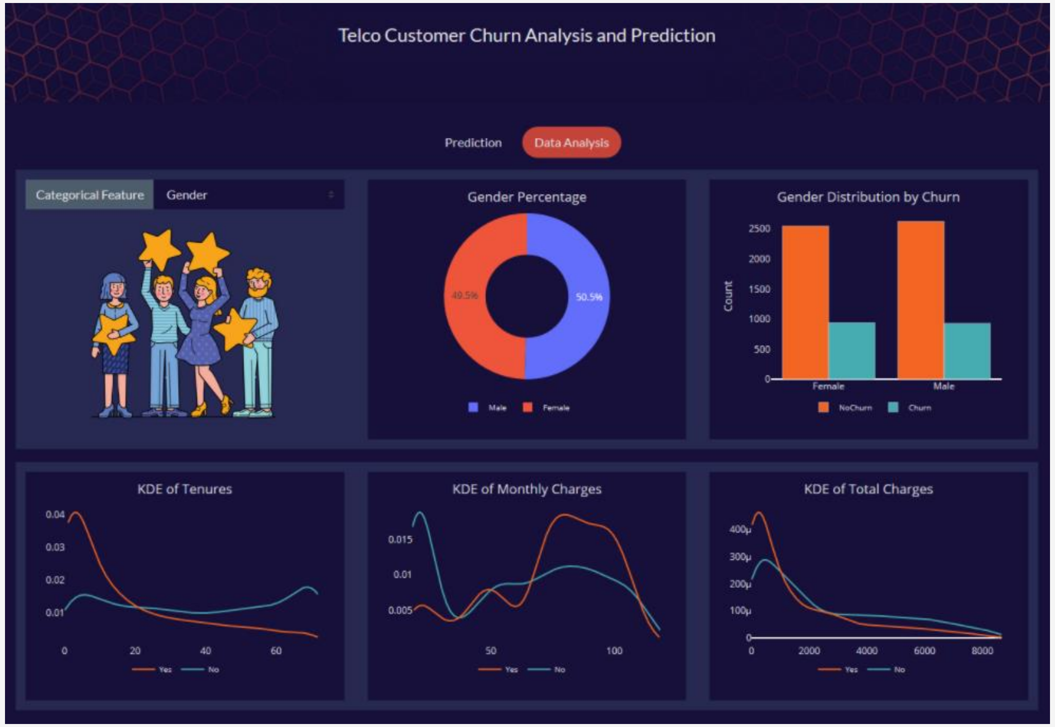
1. INTRODUCTION

- Customer churn is a major challenge and one of the most important concerns for telecommunication service providers.

- Although there are many reasons for customer churn, some of the major reasons includes overall service dissatisfaction, high subscription chargers including monthly bill shock and better alternatives.

- Primarily telcos try to retain their customers than acquiring new ones as it proved to be much costlier. Hence predicting churn in the telecom industry is very important.

- Scope of this capstone project is to build a Machine Learning based Telco CHURN Prediction model to handle above problem.



Categorical Feature: Gender

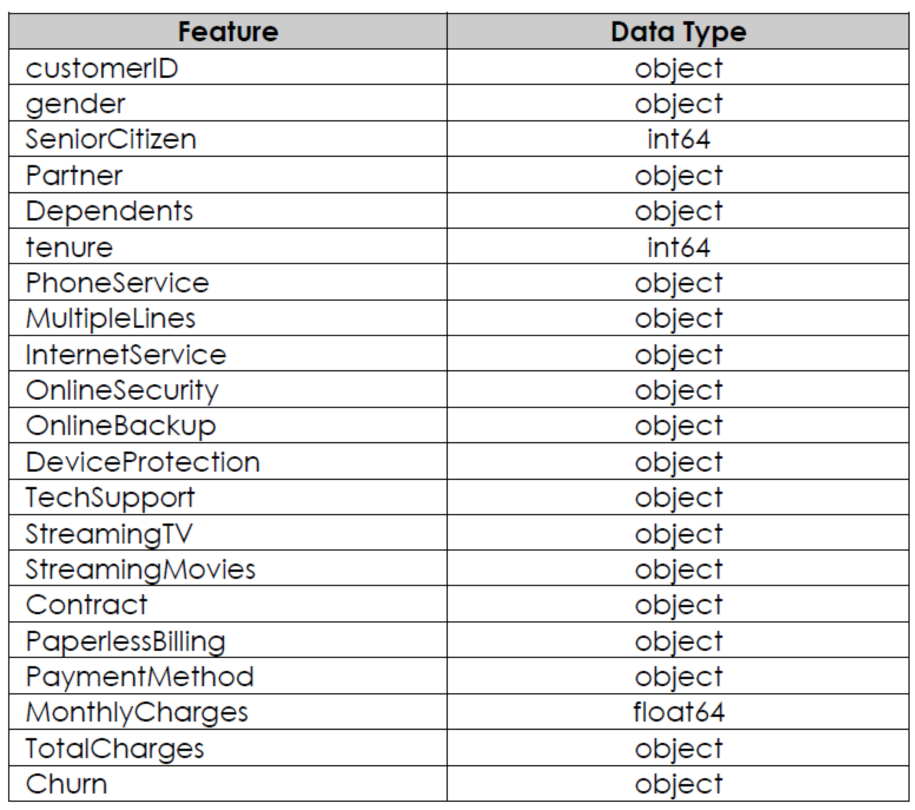
2. DATA : Summary

Data Source : https://www.kaggle.com/blastchar/telco-customer-churn

Data Shape : 7043 Rows & 21 Columns

Data Description :

- Customers who left within the last month : Churn
- Value Added Services(VAS) that each customer has subscribed : Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV and Movies
- Customer account information – Tenure, Contract, Payment Method, Paperless Billing, Monthly Charges and Total Charges
- Customer demographic information : Customer ID, Gender, Age range, Partners and Dependents

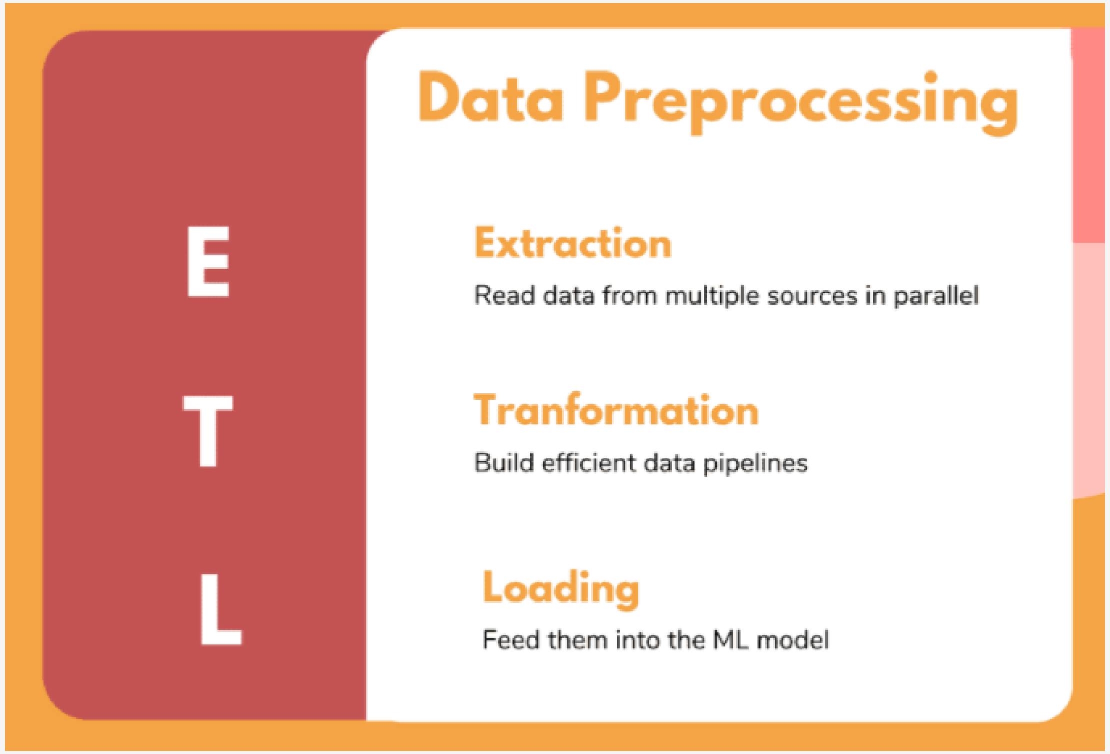


Feature	Data Type
customerID	object
gender	object
SeniorCitizen	int64
Partner	object
Dependents	object
tenure	int64
PhoneService	object
MultipleLines	object
InternetService	object
OnlineSecurity	object
OnlineBackup	object
DeviceProtection	object
TechSupport	object
StreamingTV	object
StreamingMovies	object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

2. DATA : Data Pre-Processing Pipeline

During data pre-processing following transformations were made.

- Change the data type of 'TotalCharges' from 'Object' to 'float'.
- Analyse missing values : 11 records from ‘TotalCharges’ was dropped.
- Drop non value adding features : ‘customerID’ dropped.
- Add new features : ‘TotVAS’ was added by summing up all subscribed services.
- Normalize unique values : ‘No internet service’ & ‘No phone service’ was replaced with ‘No’
- Apply label encoding : Applied for respective features with 2 unique values.
- Apply One-Hot encoding : Applied for respective features with 3 or more unique values.
- Feature re-scaling : Min-Max Scaling was applied for ‘tenure’, ‘MonthlyCharges’ & ‘TotalCharges’



Data Preprocessing

E T L

Extraction

Read data from multiple sources in parallel

Tranformation

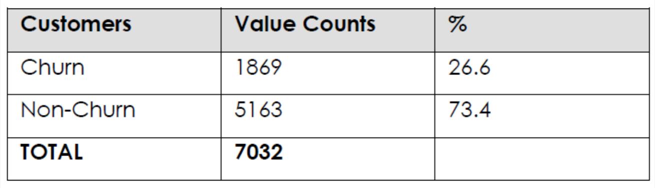
Build efficient data pipelines

Loading

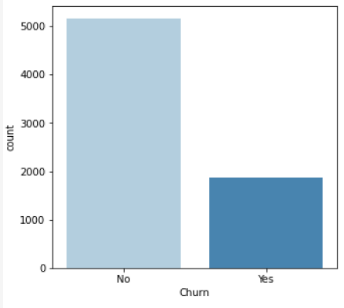
Feed them into the ML model

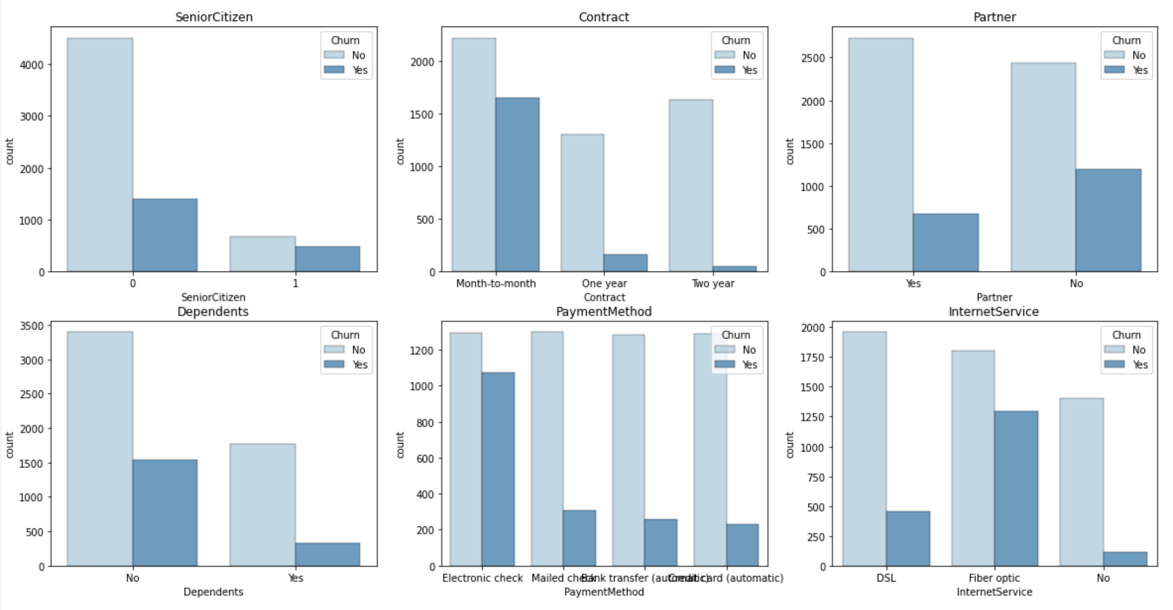
2. DATA : Exploratory Data Analysis (EDA)

- Analysis shows class imbalance of data between Churn & Non-Churners.

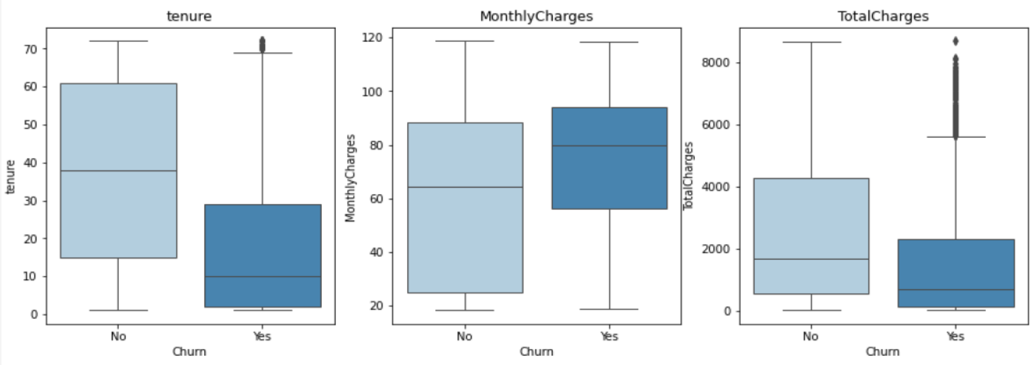


Customers	**Value Counts**	**%**
Churn	1869	26.6
Non-Churn	5163	73.4
TOTAL	**7032**	



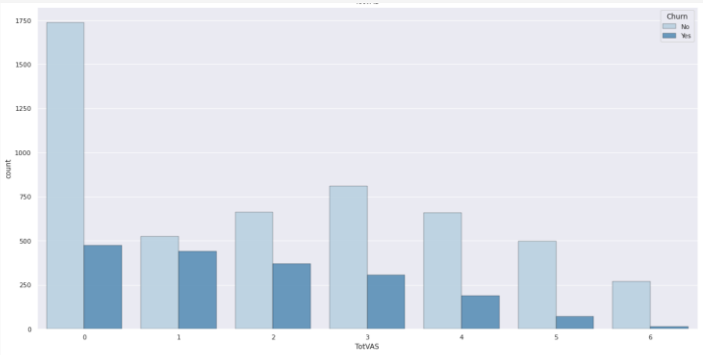


- Customers who churned had much lower tenure with a median of 10 months & much lower inter quartile range (IQR) as compared to non-churners. (median of 38 months)
- Customers who churned had higher monthly charges with a median of 80 USD and much lower IQR compared to that of non-churners (median of 65 USD).
- TotalCharges are the result of tenure and MonthlyCharges, which are more insightful on an individual basis.



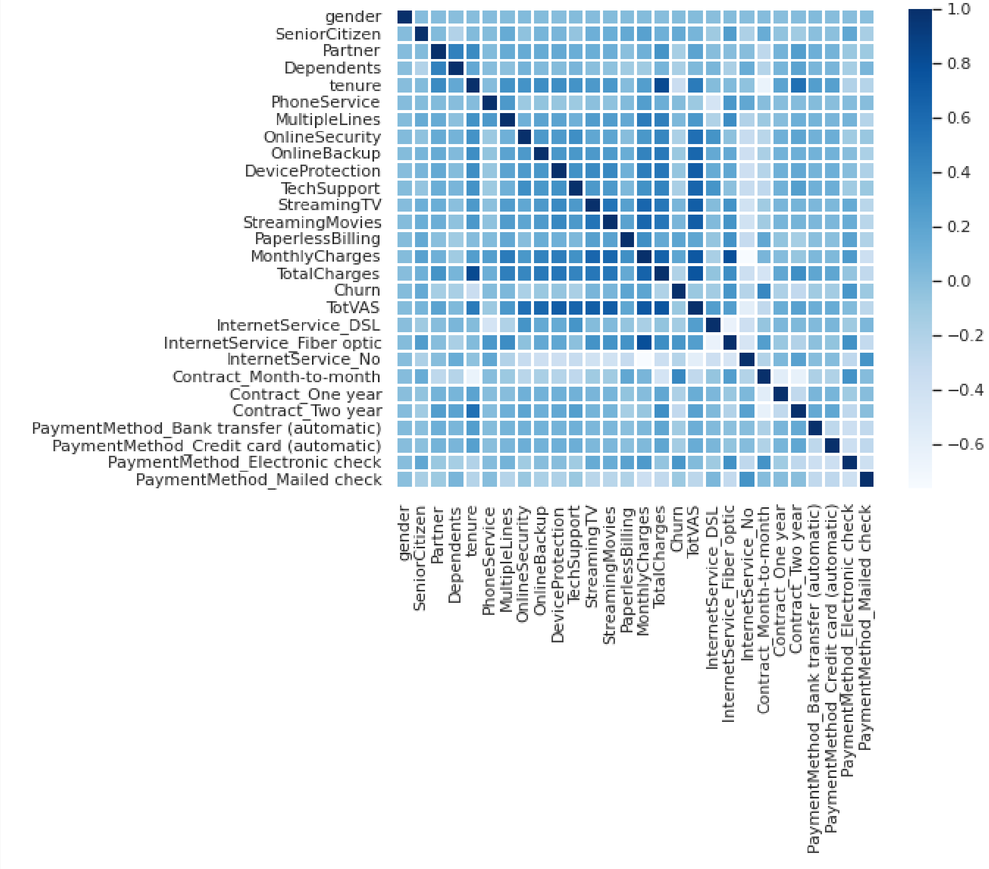
- Senior citizens churn rate is much higher than non senior churn rate.
- Churn rate for month to month contracts much higher that for other contract durations.
- Moderately higher churn rate for customers without partners.
- Much higher churn rate for customers without children.
- Payment method electronic check shows much higher churn rate than other payment methods.
- Customers with InternetService, fiber optic as part of their contract have much higher churn rate.

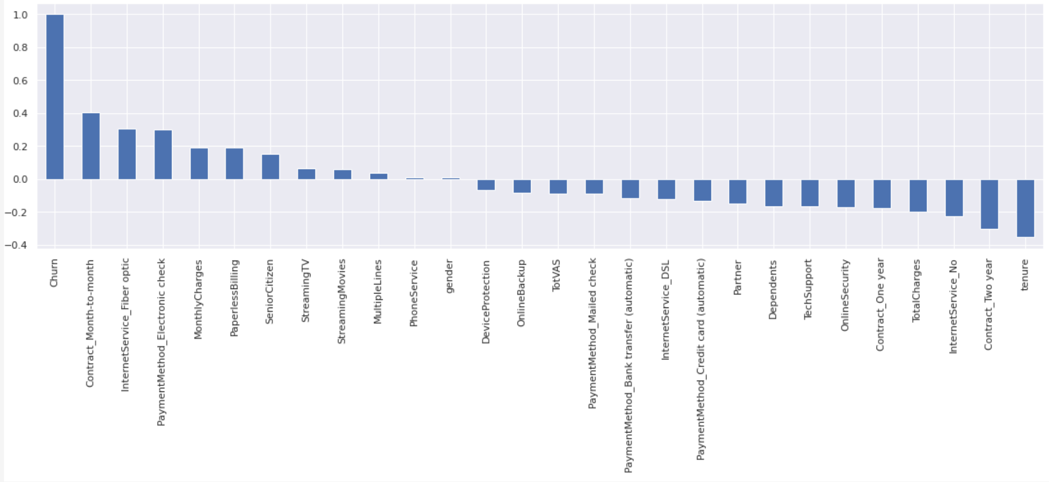
2. DATA : Exploratory Data Analysis (EDA) *Contd..*



- Additional VAS count plot shows a very high churn rate for customers that have 1 additional service.
- Customers with a very high number of additional services do have a low churn rate.

- Correlation analysis indicate degree of respective feature correlation towards Churn vs Non-Churn decision.
- “total chargers” feature is highly corelated with monthly chargers and tenure. Hence total chargers feature was not considered a X feature variable.





- Feature weights: Indicates the top features used by the model to generate the predictions

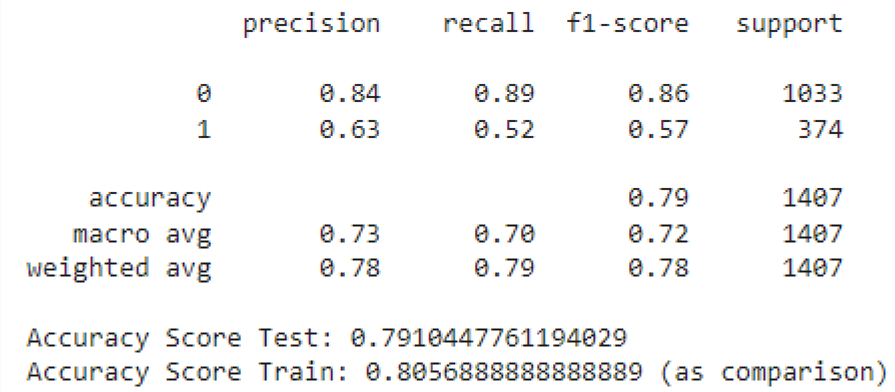
3. METHODOLOGY

- **Approach :** Based on exploratory data analysis it concluded that case should follow "Supervised Machine Learning Classification" approach.
- **Models Considered :** Following models were taken into consideration.
  - Logistic Regression
  - Random Forest
  - Support Vector Machine
  - Deep Learning – Artificial Neural Network
- **Train Test Split :** For model training and testing, the data set is divided into 80% training and 20% test data. The "Churn" column is defined as the class (the "y"), the remaining columns (except TotVAS) as the features (the "X").
- **Scoring Metrices :** Following metrics were considered for model performance assessment :
  - a. Feature weights:
  - b. Confusion matrix:
  - c. Accuracy score:
  - d. F1 Score:
  - e. ROC Curve:
  - f. AUC (for ROC):
  - g. Precision-Recall-Curve:
  - h. AUC (for PRC):
- **Hyperparameter Tuning :** Based on the analysis, respective models were further optimized by tuning its hyperparameters using Random Search method.

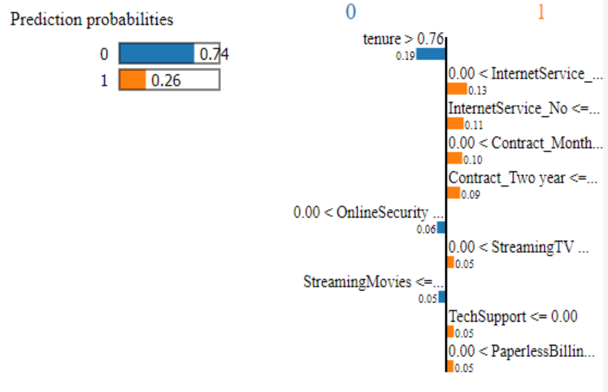
4. RESULTS – Model : Logistic Regression Classifier

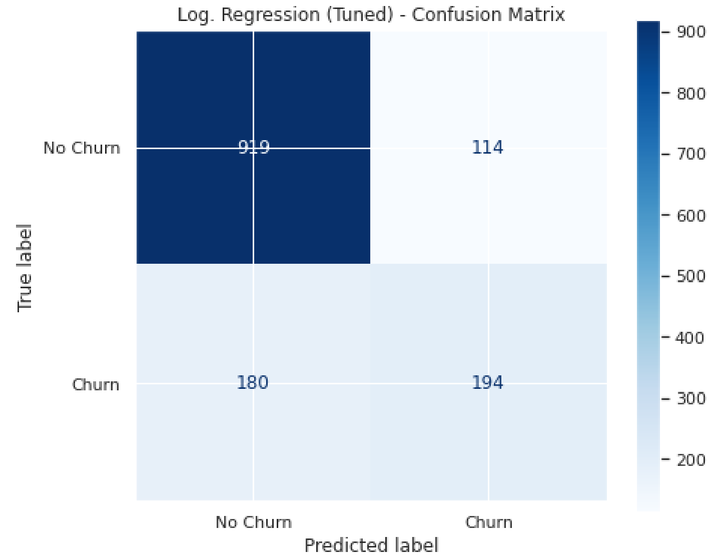
Hyperparameter Tuned

- Model : Logistic Regression
- Best Parameters : {'C': 0.30000000000000004, 'penalty': 'l2'}
- Hyperparameter Model : GridSearchCV

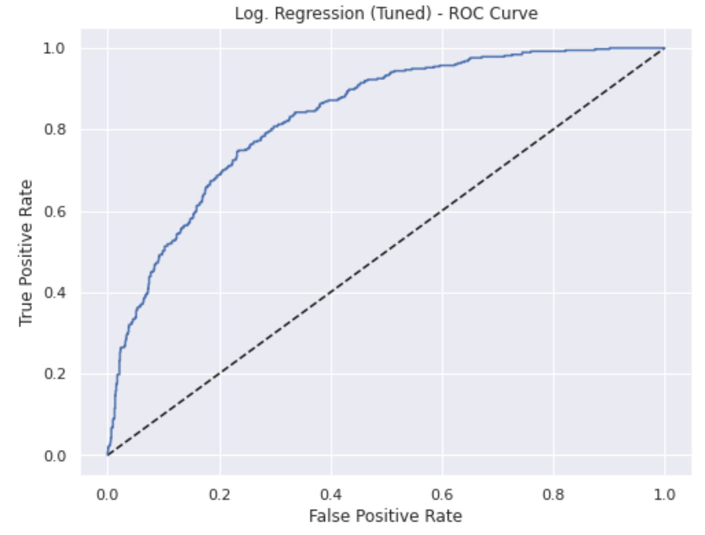


```
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1033
           1       0.63      0.52      0.57       374

    accuracy                           0.79      1407
   macro avg       0.73      0.70      0.72      1407
weighted avg       0.78      0.79      0.78      1407

Accuracy Score Test: 0.7910447761194029
Accuracy Score Train: 0.8056888888888889 (as comparison)
```



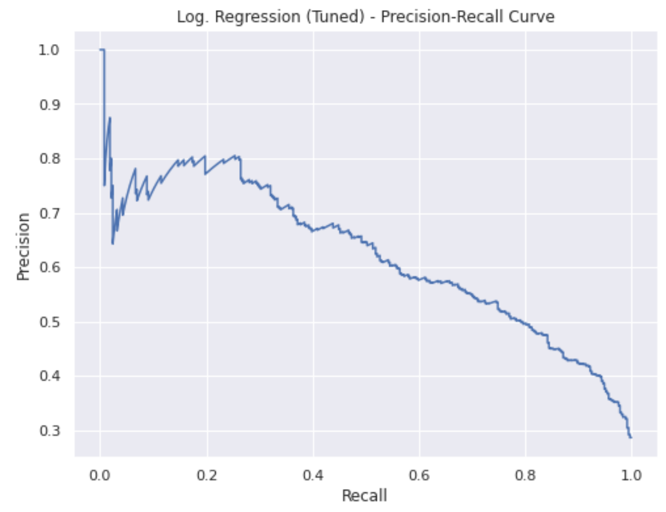


Log. Regression (Tuned) - Confusion Matrix



Log. Regression (Tuned) - ROC Curve

```
AUC Score (ROC): 0.830123310434796
```



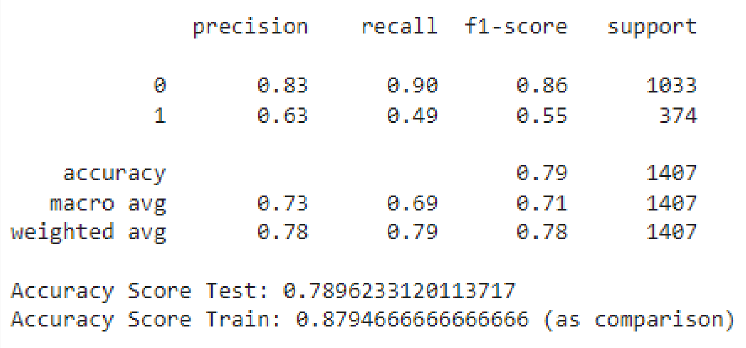
Log. Regression (Tuned) - Precision-Recall Curve

```
F1 Score: 0.5689149560117301
AUC Score (PR): 0.6228089917478592
```

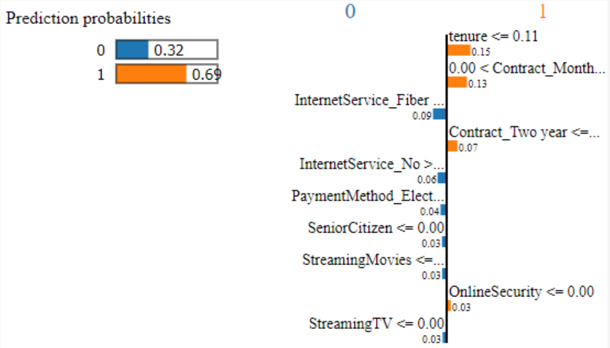
4. RESULTS – Model : Random Forest Classifier

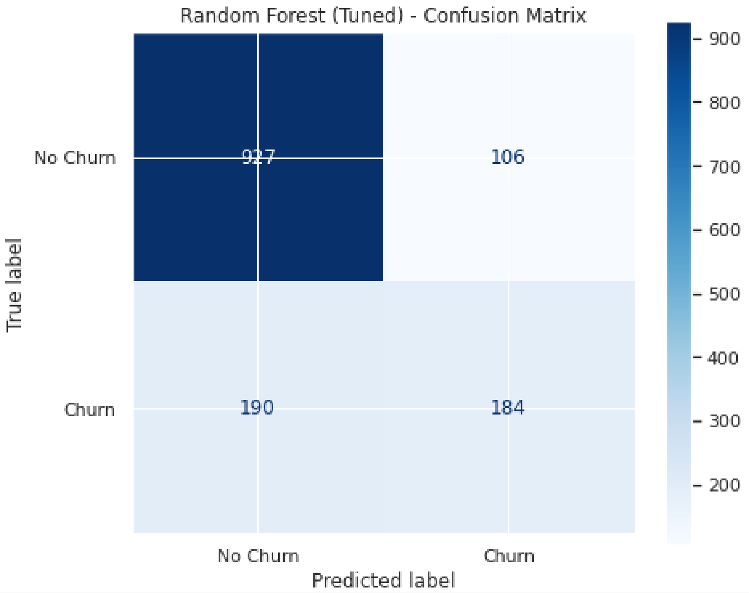
Hyperparameter Tuned

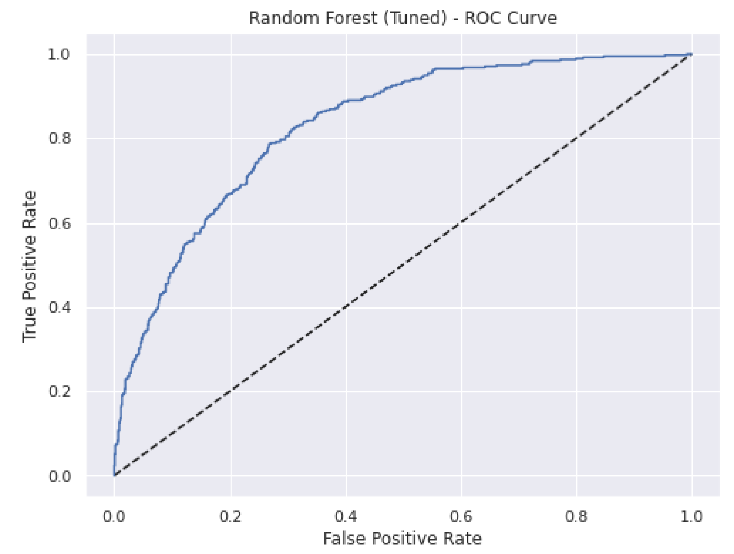
- Model : Random Forest
- Best Parameters : {'n_estimators': 1900, 'max_features': 'sqrt', 'max_depth': 10, 'criterion': 'entropy', 'bootstrap': False}
- Hyperparameter Model : RandomizedSearchCV



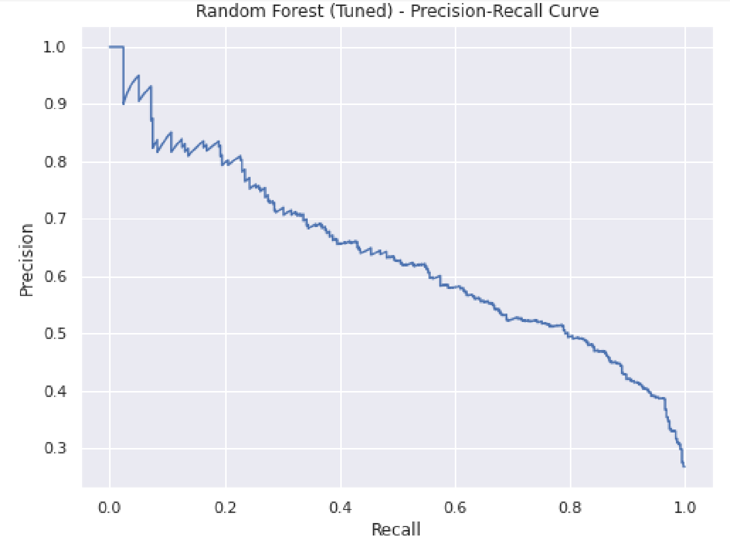
```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1033
           1       0.63      0.49      0.55       374

    accuracy                           0.79      1407
   macro avg       0.73      0.69      0.71      1407
weighted avg       0.78      0.79      0.78      1407

Accuracy Score Test: 0.7896233120113717
Accuracy Score Train: 0.8794666666666666 (as comparison)
```







```
AUC Score (ROC): 0.8289028891500276
```

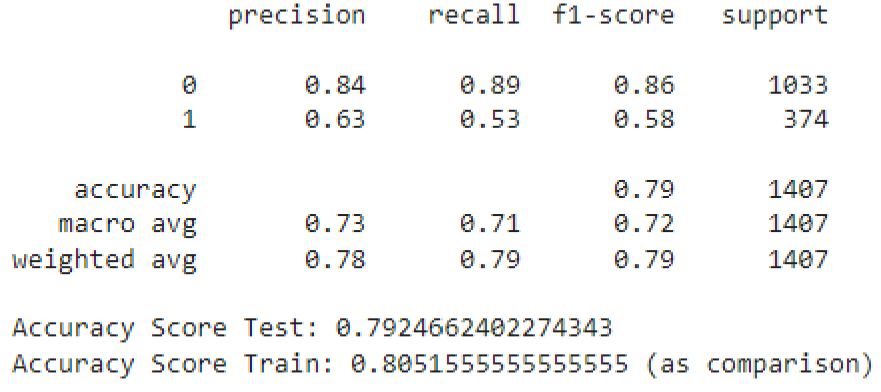


```
F1 Score: 0.5542168674698795
AUC Score (PR): 0.6361023043168073
```

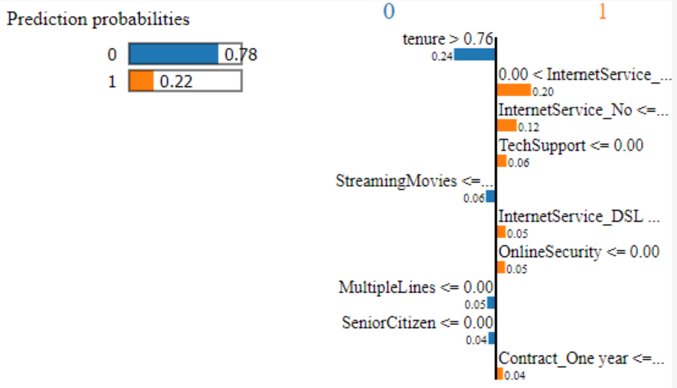
4. RESULTS – Model : Support Vector Machine

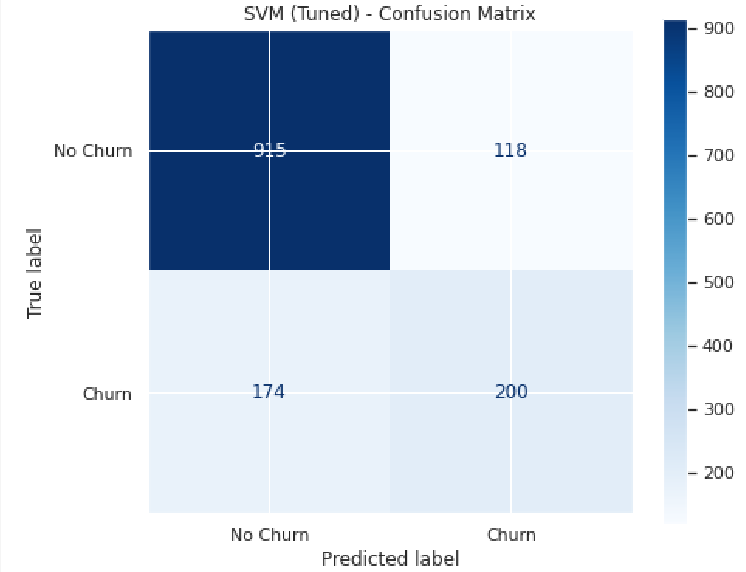
Hyperparameter Tuned

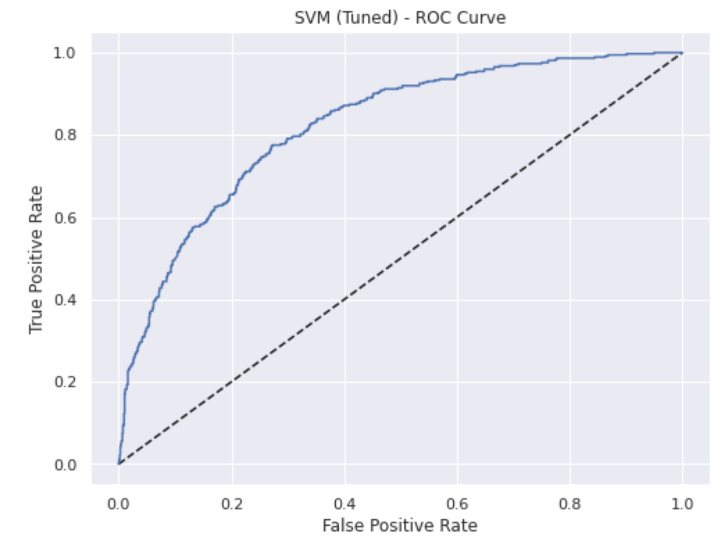
- Model : Support Vector Machine
- Best Parameters : {'C': 0.2}
- Hyperparameter Model : GridSearchCV



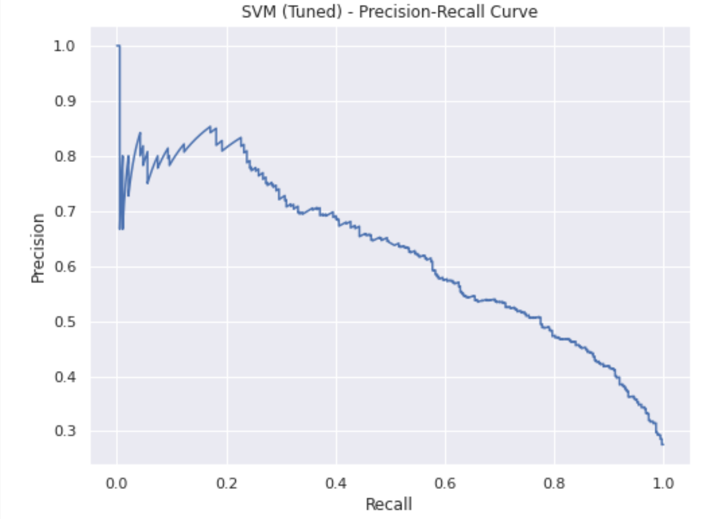
```
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1033
           1       0.63      0.53      0.58       374

    accuracy                           0.79      1407
   macro avg       0.73      0.71      0.72      1407
weighted avg       0.78      0.79      0.79      1407

Accuracy Score Test: 0.7924662402274343
Accuracy Score Train: 0.8051555555555555 (as comparison)
```







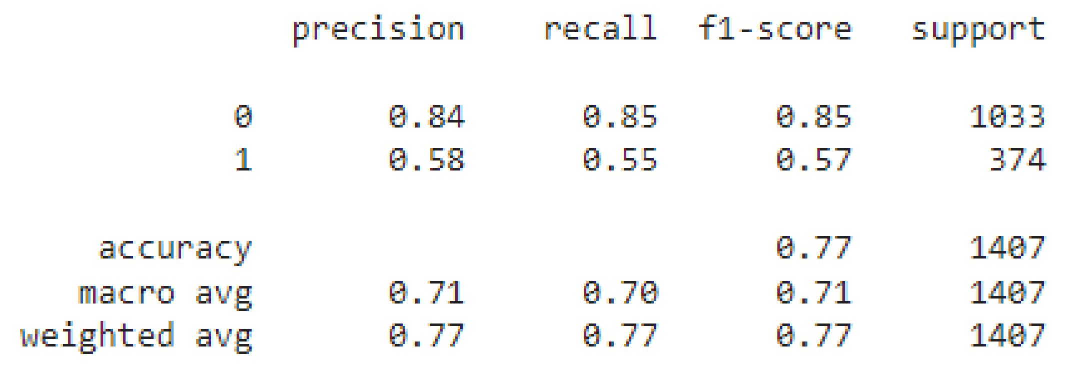
```
AUC Score (ROC): 0.8217058978832226
```



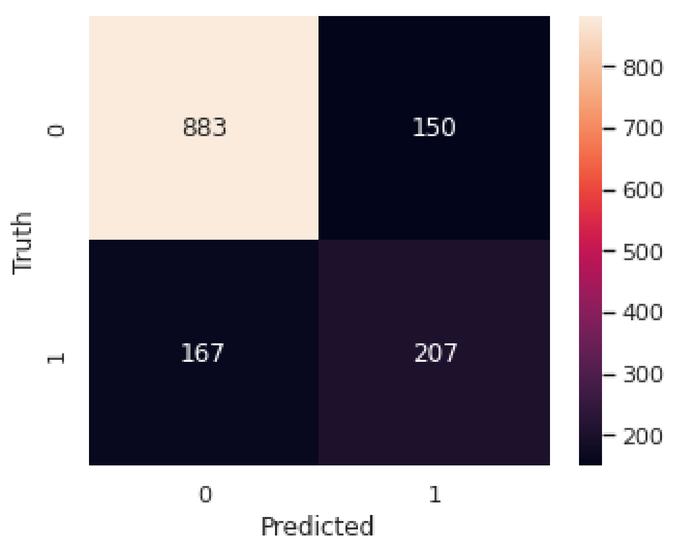
```
F1 Score: 0.5780346820809249
AUC Score (PR): 0.6247237791053385
```

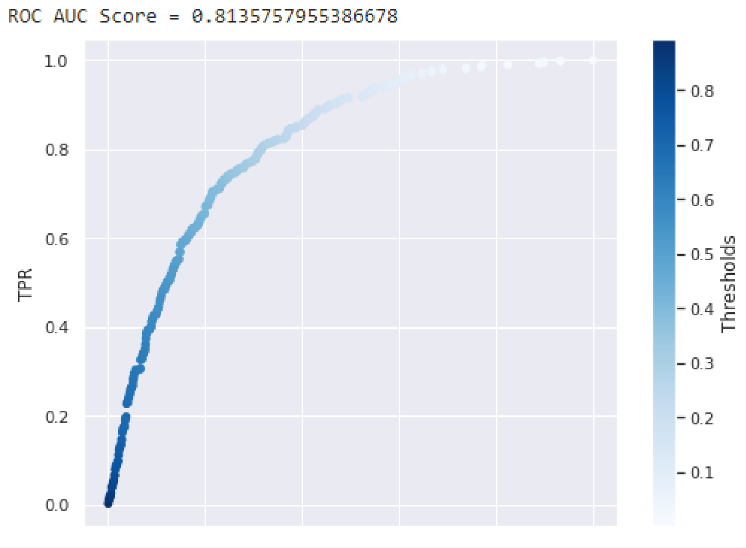
4. RESULTS – Model : Deep Learning - ANN

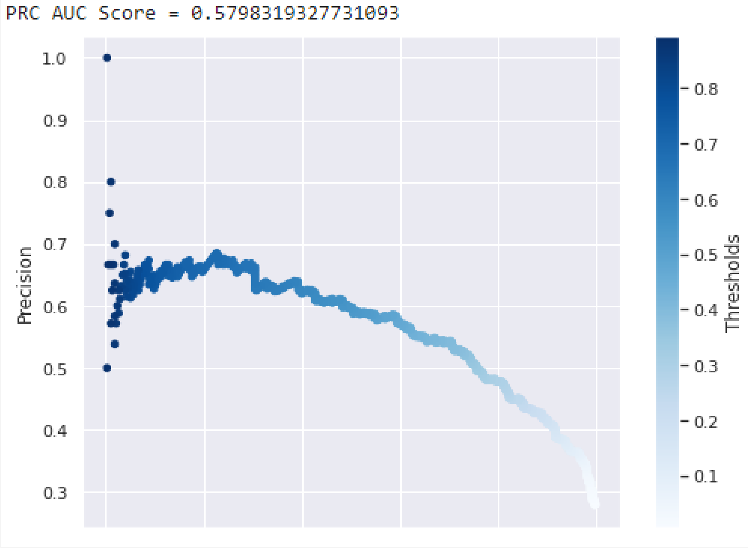
- Model : Artificial Neural Network
- No of Hidden Layers : 1
- Total Params : 39,681
- Parameters : {epochs=10}



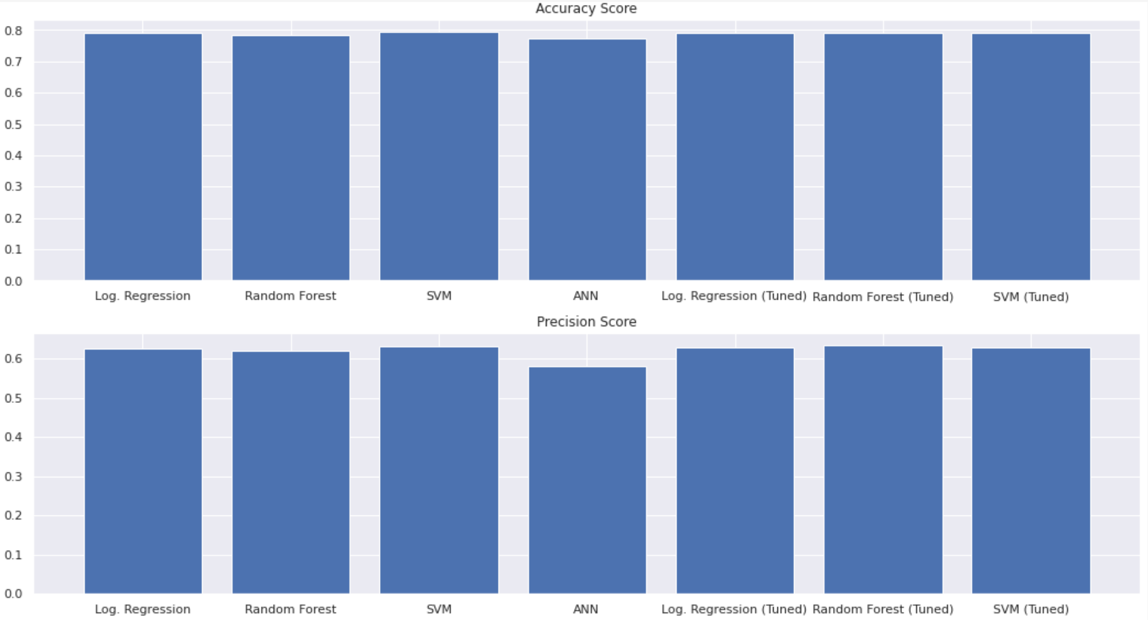
```
              precision    recall  f1-score   support

           0       0.84      0.85      0.85      1033
           1       0.58      0.55      0.57       374

    accuracy                           0.77      1407
   macro avg       0.71      0.70      0.71      1407
weighted avg       0.77      0.77      0.77      1407
```

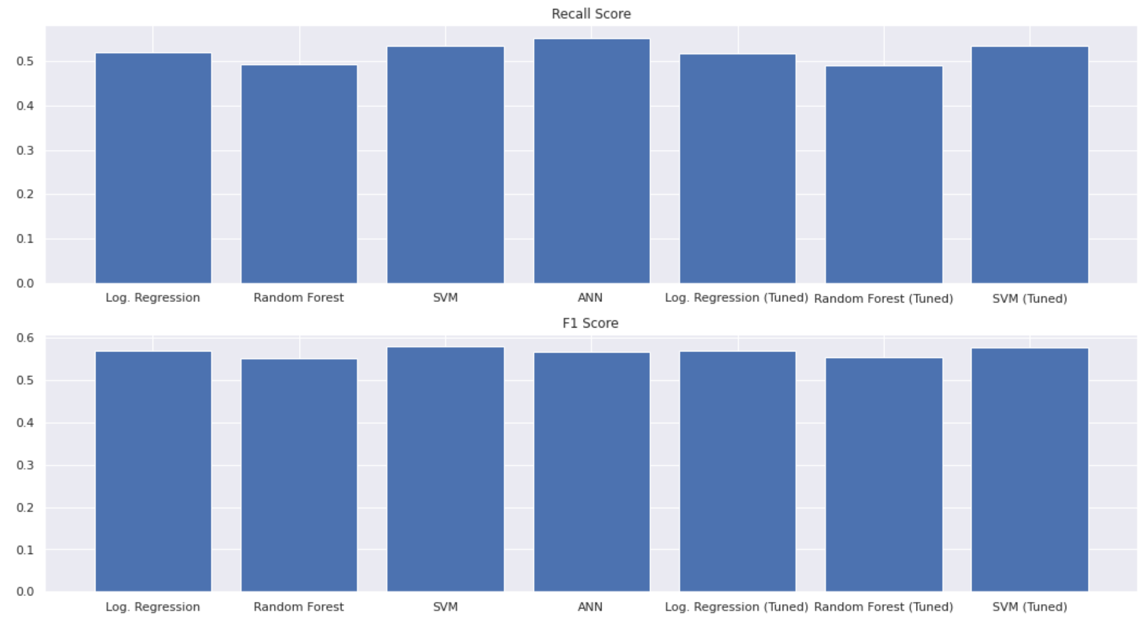






5. CONCLUSION





- Except for ANN all other three classifiers had model accuracy of 0.79
- Since data set had high class imbalance towards Non-Churners, in addition to model accuracy respective F1 Scores were evaluated along with other parameters including Precision, Recall, RoC & PRC. Considering all factors Support Vector Machine is selected as the best performing model with F1 score of 0.58 after hyperparameter tuning & predictions were made accordingly.

6. DISCUSSION

Following improvements were suggested to further enhance model performance.

- Telcos has access to much larger data repositories including;
  - Customer Relationship Management (CRM) systems, CDR/Billing Records, KYC data repositories & other network anchors to detect Value Added Services (VAS) subscribed.
- Recommended to gather a larger data set by augmenting data from multiple data repositories.
  - With larger data set it will further generalize class imbalance and improve model accuracy.
  - Further neural networks models will out perform other classifiers with larger data sets & could be properly trained to detect more complex patterns in data to perform higher accuracies.
- Recommended to further evaluate feature weights & feature correlation to eliminate non value adding features to improve model prediction efficiency during future iterations.

A high accuracy is needed to be able to identify promising customer cases where churn can be avoided, as eventually the customer returns protected need to outweigh the costs of related retention campaigns.

THANK YOU