Machine Learning Based
Telco Customer CHURN Prediction

Capstone Project Report

Name : Chathura Peiris

Course : Machine Learning Foundations

Date : 21st Nov 2021

Table of Contents

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. Due to direct impact on market share & revenue, telcos strive very hard to sustain in this competition. Primarily to sustain this competition they often 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.

During my capstone project I made an attempt to build a Machine Learning based Telco CHURN Prediction model to handle above problem.

2. DATA

2.1 Summary

Data set for this project is taken from Kaggle data repository.

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

Raw data set comprises of 7043 Rows & 21 Columns. Each Row represent a customer and each column represent feature attributes for respective customer.

The data set includes information about:

- 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

Following table describe all features along with respective data types.

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

Table 1 : Feature Analysis

Original data set comprises of features falling under both categorical and numerical data types.

2.2 Data Pre-Processing

During data pre-processing pipeline following actions were taken.

- 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’

2.3 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**	

Table 2 : Churn vs Non Chun Analysis

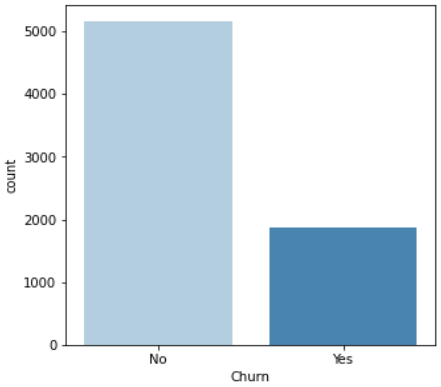


Figure 1 : Churn Analysis

To address class imbalance following approaches would be recommended including : Gather new data set with more data points, Resampling, Data Augmentation. However, to keep this case simple due to limited time, class imbalance was kept forward.

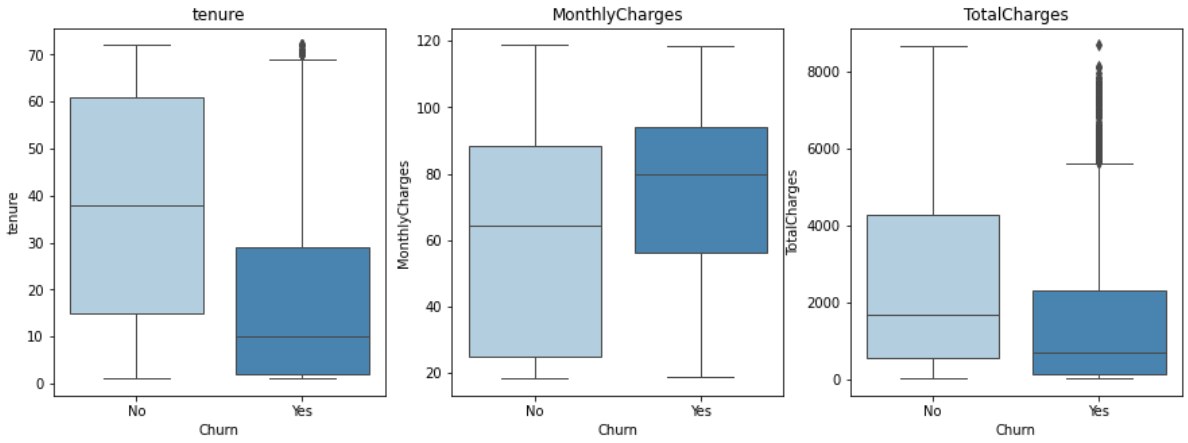


Figure 2 : Box Plot Analysis

Box -Plot analysis indicate following:

- 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.

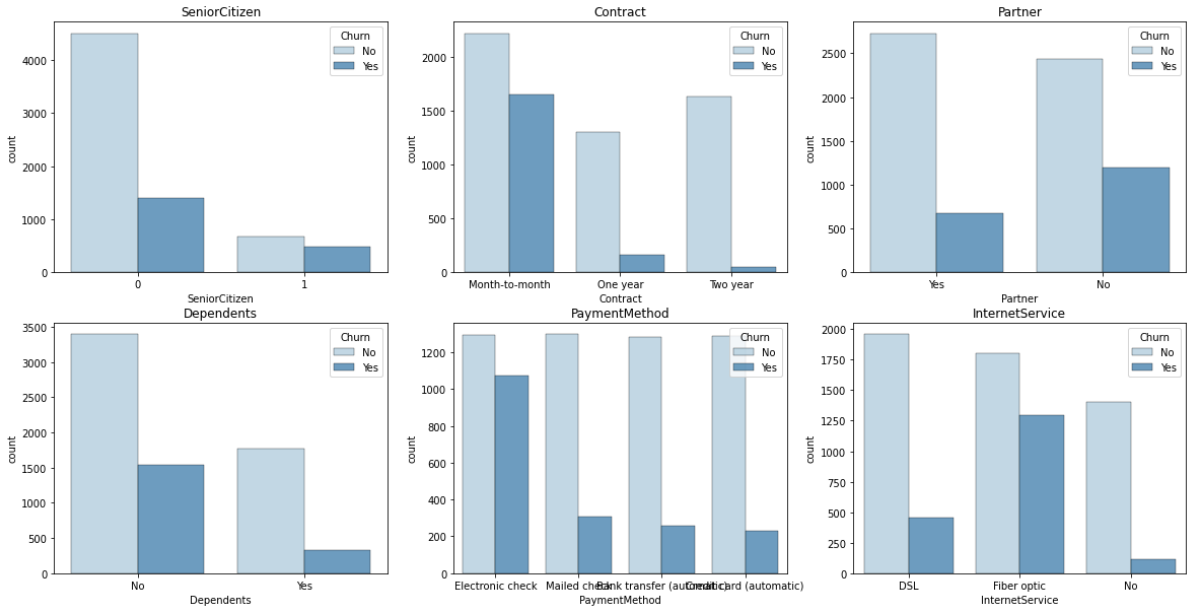
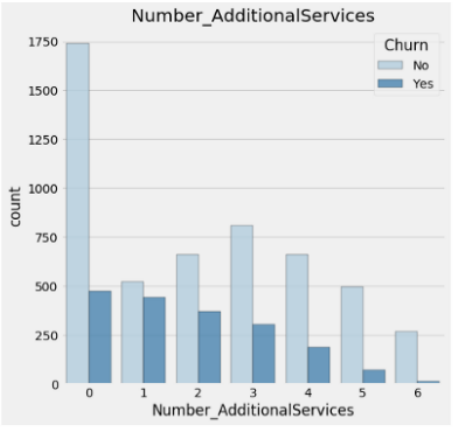


Figure 3 : Count Plot Analysis

Count-Plot analysis for selected features indicate following:

- 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.



- 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.

Figure 4 : Count Plot Analysis - VAS

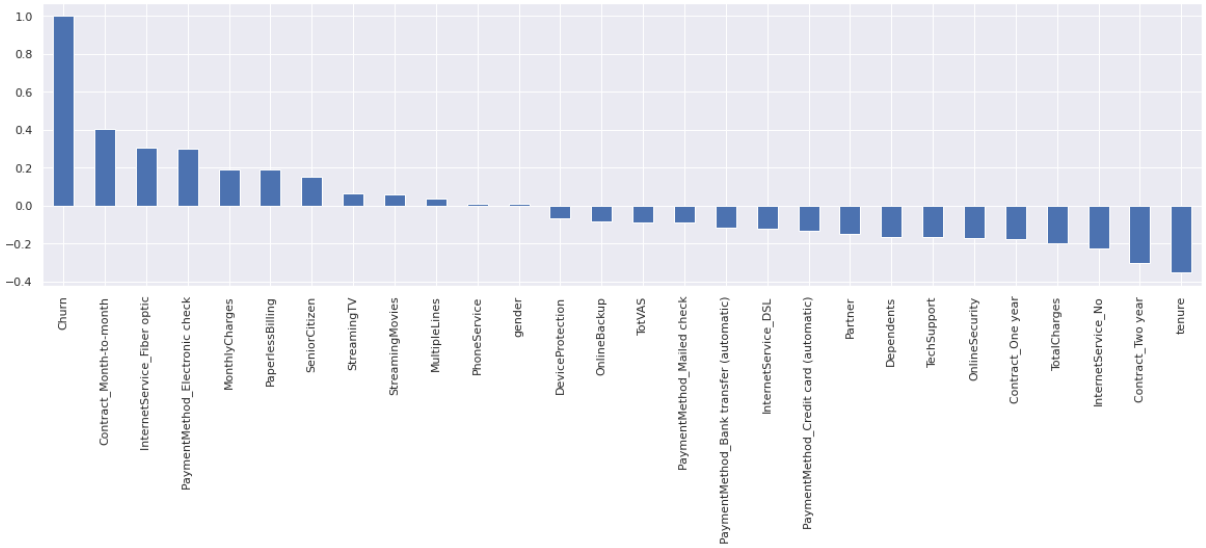


Figure 6 : Feature Weights Analysis

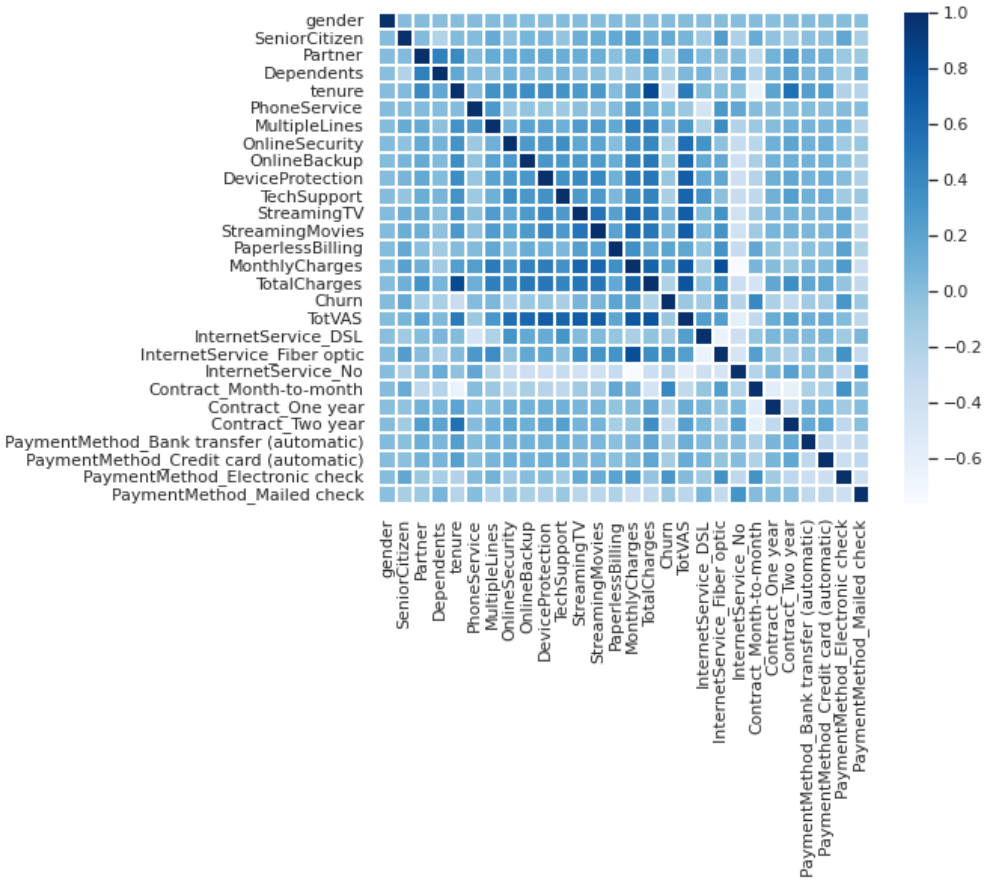


Figure 7 : Feature Correlation Analysis

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

3. Methodology

Based on exploratory data analysis it concluded that case should follow "Supervised Machine Learning Classification" approach. Following models were taken into consideration.

- Logistic Regression — fast and linear model
- Random Forest — slower but accurate ensemble model based on decision trees
- Deep Learning – Artificial Neural Network

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").

For performance assessment of the selected models, below metrics are used:

a. Feature weights:
b. Confusion matrix:
c. Accuracy score:
d. ROC Curve:
e. AUC (for ROC):
f. Precision-Recall-Curve:
g. F1 Score:
h. AUC (for PRC):

Based on the analysis, respective models were further optimized by tuning its hyperparameters.

4. Results

Based on the modelling exercise following results achieved for respective models.

4.1 Logistic Regression

Stage	State	Accuracy	Precision	Recall	F1-Score	ROC - AUC	PRC-AUC
Pre HPT	Non-Churn	0.79	0.84	0.89	0.86	0.83	0.62
	Churn		0.62	0.52	0.57		
Post HPT	Non-Churn	0.79	0.84	0.89	0.86	0.83	0.62
	Churn		0.63	0.52	0.57		

4.2 Random Forest

Stage	State	Accuracy	Precision	Recall	F1-Score	ROC - AUC	PRC-AUC
Pre HPT	Non-Churn	0.79	0.83	0.89	0.86	0.81	0.62
	Churn		0.62	0.49	0.55		
Post HPT	Non-Churn	0.79	0.83	0.90	0.86	0.83	0.64
	Churn		0.63	0.49	0.55		

4.3 Support Vector Machine

Stage	State	Accuracy	Precision	Recall	F1-Score	ROC - AUC	PRC-AUC
Pre HPT	Non-Churn	0.79	0.84	0.89	0.85	0.82	0.62
	Churn		0.63	0.53	0.58		
Post HPT	Non-Churn	0.79	0.84	0.89	0.86	0.82	0.62
	Churn		0.63	0.53	0.58		

4.4 ANN

Stage	State	Accuracy	Precision	Recall	F1-Score	ROC - AUC	PRC-AUC
Initial	Non-Churn	0.77	0.84	0.85	0.85	0.81	0.58
	Churn		0.58	0.55	0.57		

Table 3 : Model Results

*HPT – Hyperparameter Tuning

5. Conclusion

Considering model outcomes following observations were made. Except for ANN all other three classifiers had accuracy score 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

Model has taken into consideration of following features both favourably & adversely in making final classifier outcome to predict Churners & Non-Churners.

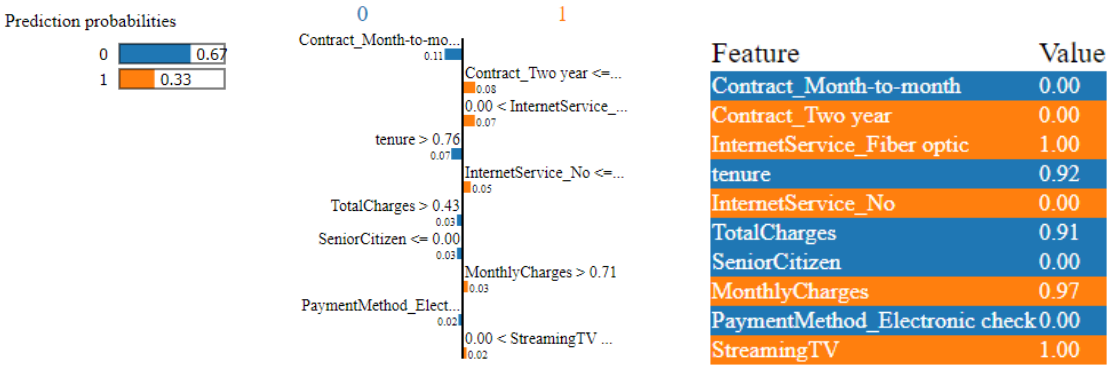


Figure 8 : Model Interpretation

Following improvements were suggested to further enhance model performance.

Telcos has access to much larger data repositories including; existing Customer Relationship Management (CRM) systems, CDR/Billing Records, KYC data repositories & other network anchors to detect Value Added Services (VAS) subscribed. Hence it is 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.

Also it is 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.