TELECOM CHURN PREDICTION

Waqas Khan

AGENDA

01 **INTRODUCTION**

Background and Problem Statement

02 **EDA**

Exploratory Data Analysis

03 **MODEL TRAINING**

Training multiple ML models

04 **MODEL EXPLAINING**

Feature Importance and Shap Analysis

01

INTRODUCTION

Background and Problem Statement

CHURN RATES IN TELECOM

KEY PERFORMANCE METRIC

Measured as the percentage of customers who discontinue their services

VOLUNTARY / INVOLUNTARY

Focus is on voluntary churn, which occurs due to company-customer relationships

CHURN PREDICTION

Use of Predictive analytics to predict churn and create focused customer retention programs

PROBLEM STATEMENT

Objective: Help Orange Telecom develop focused retention programs by creating churn prediction models

Model Metric

Precision

- Of all the users that the algorithm predicts will churn, how many of them do actually churn?

Recall

- What percentage of users that end up churning does the algorithm successfully find?

F1-Score

- In this problem, both precision and recall are important. So we will use F1-Score for model evaluation

02

EDA

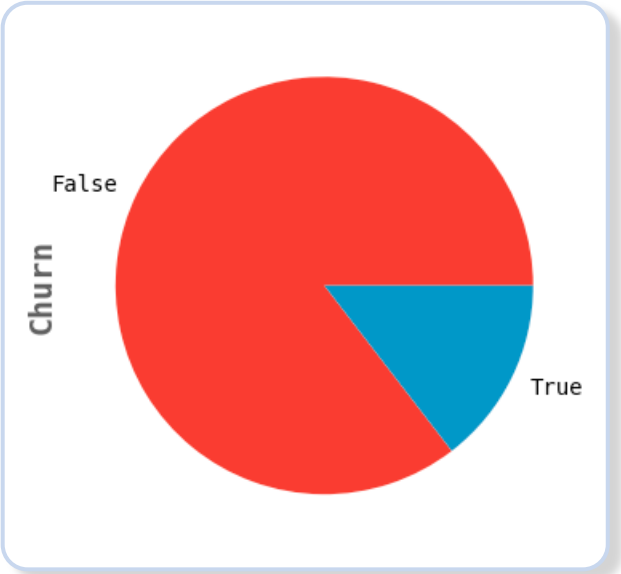
Exploratory Data Analysis

DATA SUMMARY

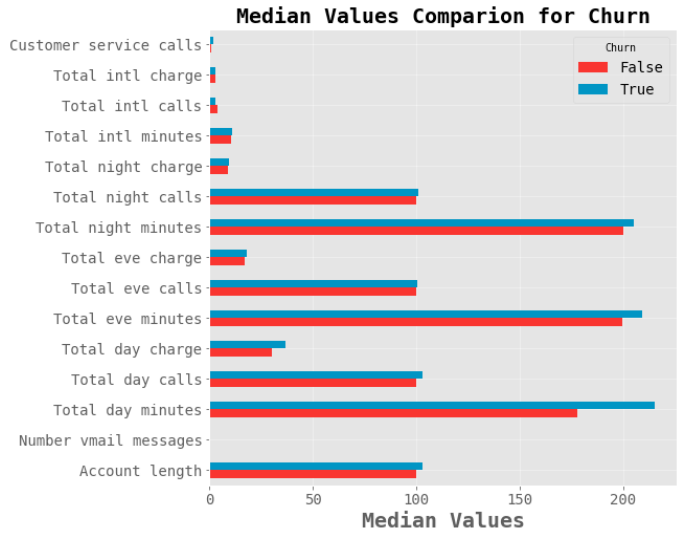
DATA SET

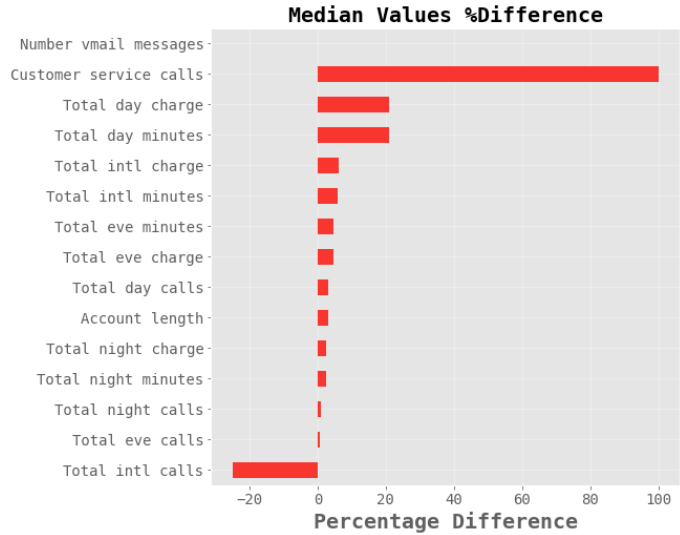
Feature	Description	Range/Unique Values Stat	Data Type
State	State Code	Unique values 51	Categorical
Account length	Length of Account	Range 1-243	Integer
Area code	Area Code	Unique values 3	Categorical
International plan	Yes/No Flag	Unique values 2	Categorical
Voice mail plan	Yes/No Flag	Unique values 2	Categorical
Number vmail messages	Number of Voice mails	Range 0-50	Integer
Total day minutes	Total Day call minutes	Range 0-351	Float
Total day calls	Total Day number of calls	Range 0-160	Integer
Total day charge	Total Day charge amount	Range 0-60	Float
Total eve minutes	Total Evening call minutes	Range 0-364	Float
Total eve calls	Total Evening number of calls	Range 0-170	Integer
Total eve charge	Total Evening charge amount	Range 0-31	Float
Total night minutes	Total Night call minutes	Range 44-395	Float
Total night calls	Total Night number of calls	Range 33-166	Integer
Total night charge	Total Night charge amount	Range 0-18	Float
Total intl minutes	Total International call minutes	Range 0-20	Float
Total intl calls	Total International number of calls	Range 0-20	Integer
Total intl charge	Total International charge amount	Range 0-5	Float
Customer service calls	Total Number of Customer Service Calls	Range 0-9	Integer
Churn	Yes/No Flag of Churn	Unique values 2	Categorical

IMBALANCED CLASSES



DISTRIBUTION COMPARISON

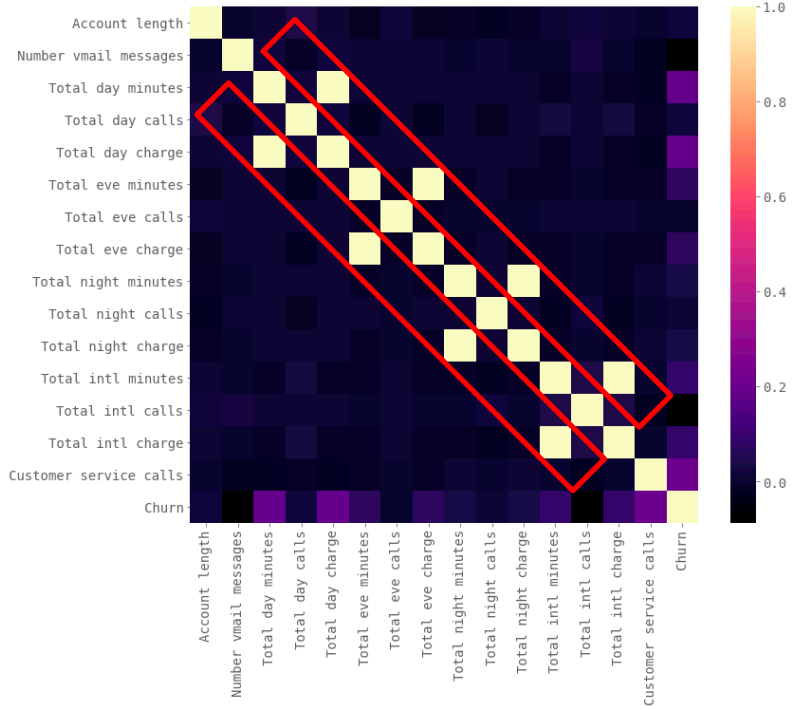




Customer who Churned:

- Made **2x** more service calls
- Had **20%** more Bill
- **25%** Less international calls

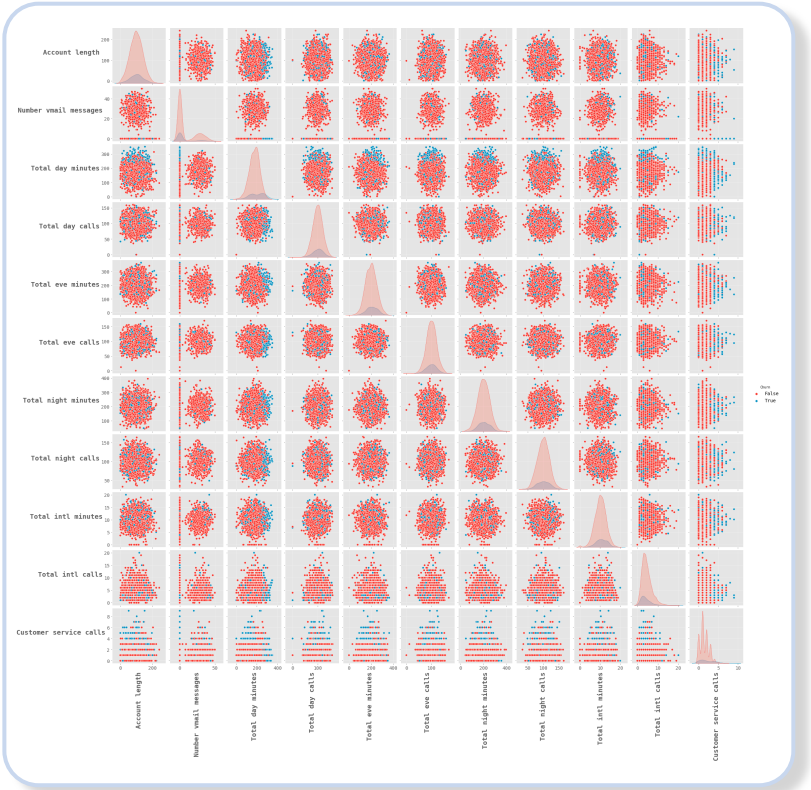
CORRELATION MATRIX



Redundant Feature

- **Call minutes** and **Call Charge** are strongly correlated with a correlation factor of 1

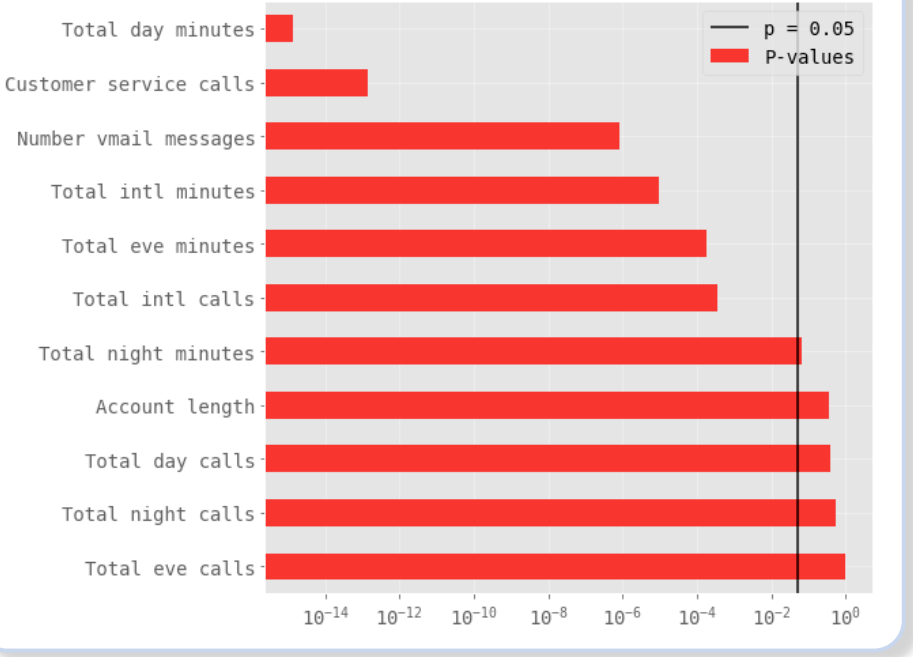
PAIR PLOTS



- Visual bivariate analysis suggests that **Customer service calls** and **Total day minutes** split the data much better as compared to other features
- This confirms what we saw before when we did median value analysis between two groups

STATISTICAL ANALYSIS – T-TEST

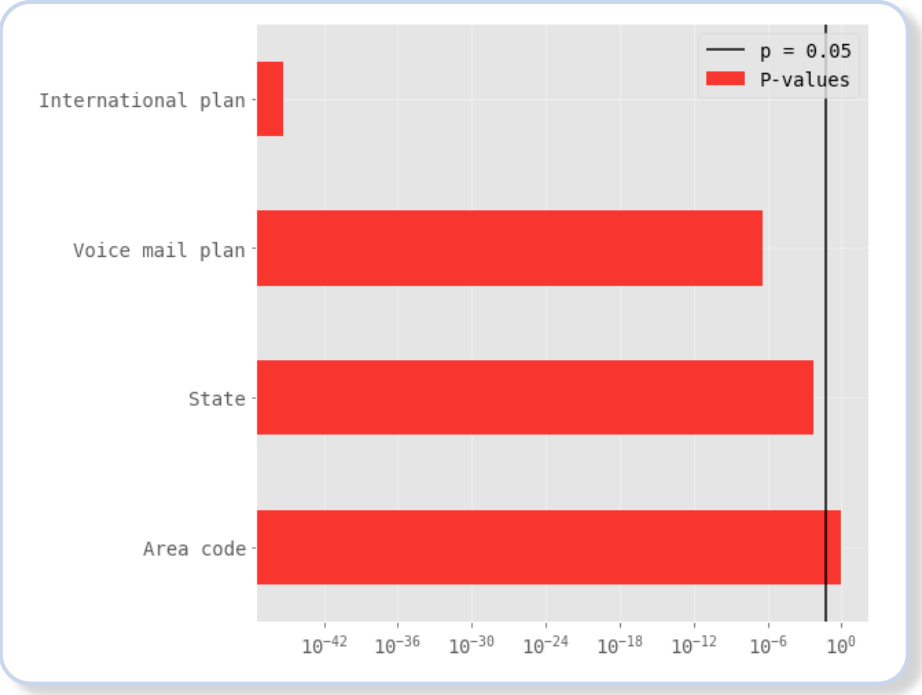
Used to test the difference between two groups on some continuous variable.



- **P-value** is the probability that the results from sample data occurred by chance
- Low p-values are good, they indicate the data did not occur by chance. In most cases p-value of 0.05 is accepted to mean data is valid
- **Total day minutes**, **Customer service calls**, **Number vmail messages**, **Total intl minutes**, **Total eve minutes**, **Total intl calls** are all statistically significant with p-values < 0.05

STATISTICAL ANALYSIS – CHI2-TEST

Determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying

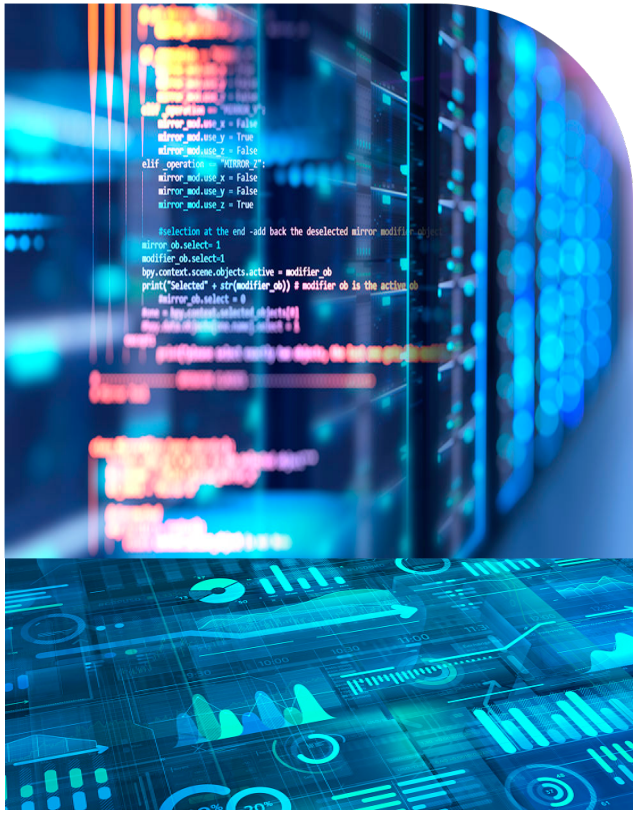


- **International plan**, **Voice mail plan** and **State** are all statistically significant with p-values < 0.05

03

MODEL TRAINING

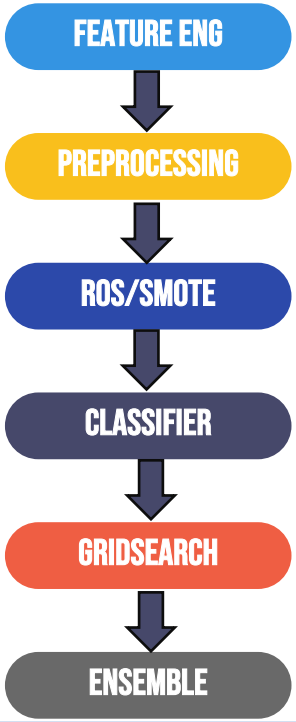
Training multiple ML models



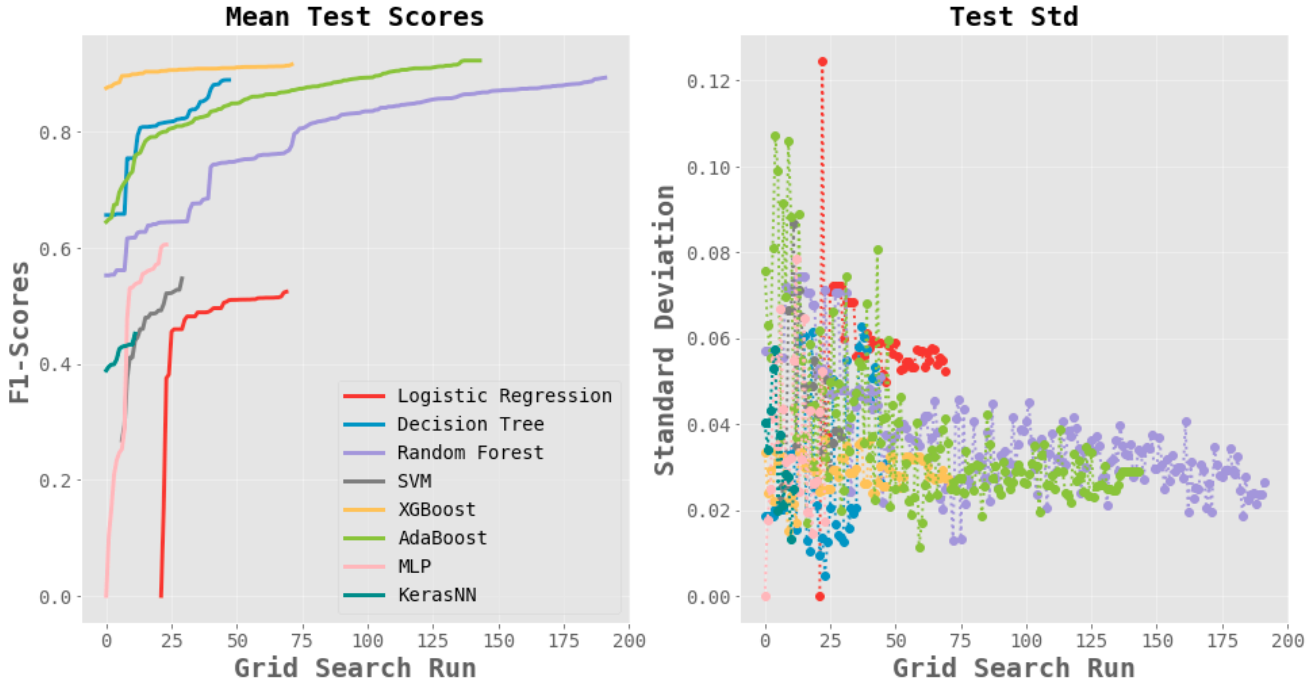
MODELS TRAINED

	FEATURE ENGINEERING	F1-SCORE
BASELINE	✗	0.12
LOGISTICS REGRESSION	✓	0.54
DECISION TREE	✓	0.92
RANDOM FOREST	✓	0.93
SVM	✓	0.59
XGBOOST	✓	0.92
ADABOOST	✓	0.93
MLP CLASSIFIER	✓	0.62
KERAS CLASSIFIER	✓	0.46

MODELING STEPS



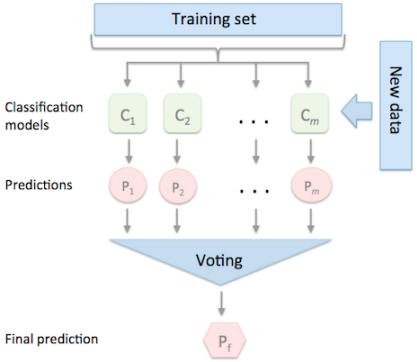
MODEL COMPARISON



- **Tree based** models outperform all other models

MODEL ENSEMBLES

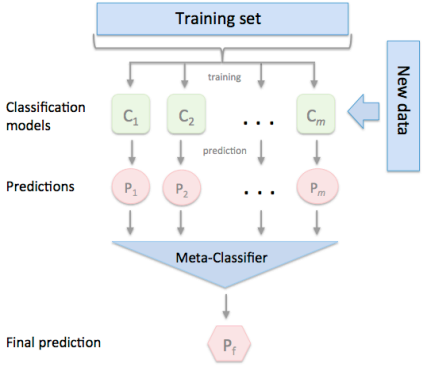
VOTING CLASSIFIER



Voting Ensemble (**F1-score** 0.93)

Decision Tree | Random Forest | XGBoost | AdaBoost

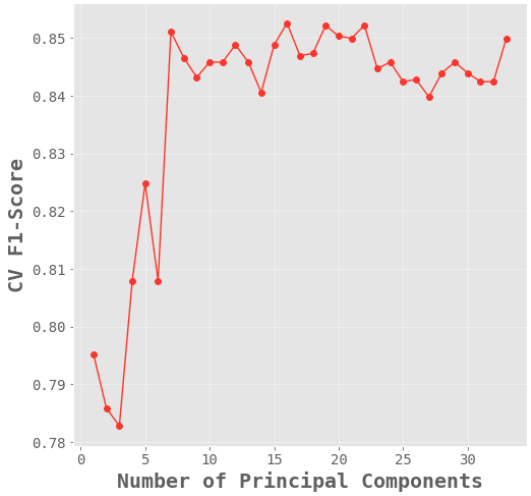
STACKING CLASSIFIER

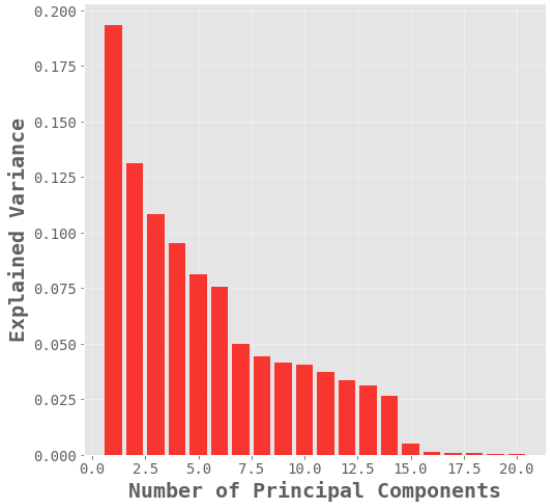


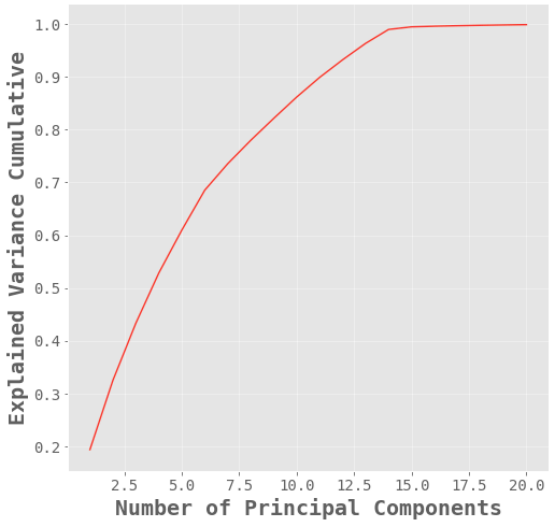
Stacking Ensemble (**F1-score** 0.93)

Decision Tree | Random Forest | XGBoost | AdaBoost

PCA



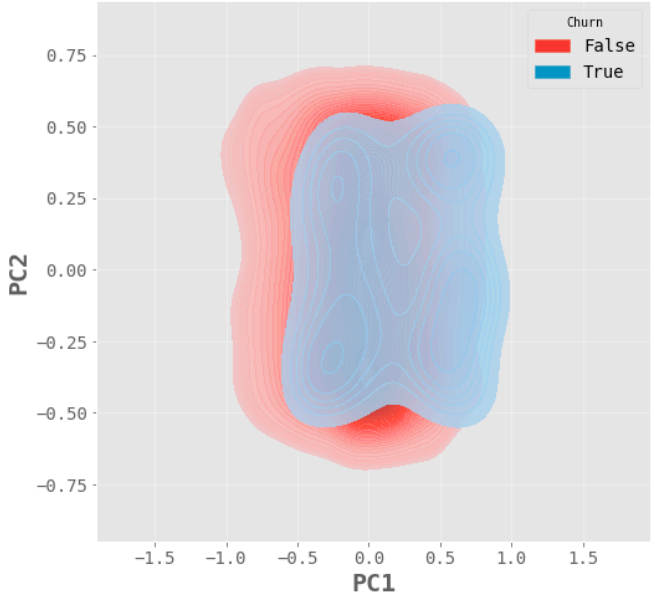


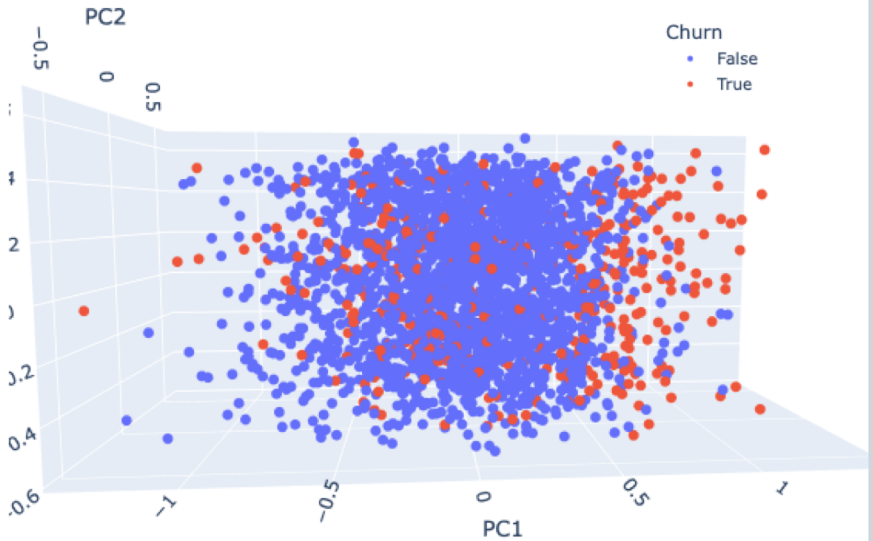


Jump in score at **8** PCAs

The explained variance plot shows that almost all the variance in the dataset(containing **33** numerical features) could be explained by **15** PCs

PCA



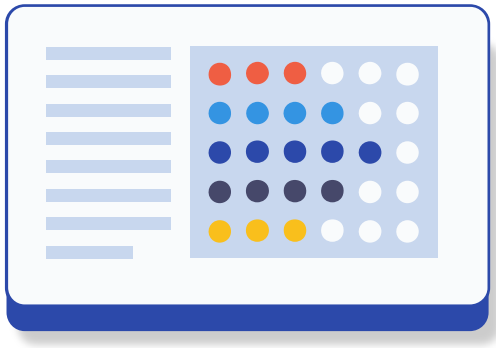


- Some level of difference between the two groups for **2** & **3** PCAs but it is not a good split
- More Principal Components needed to completely separate the two groups.

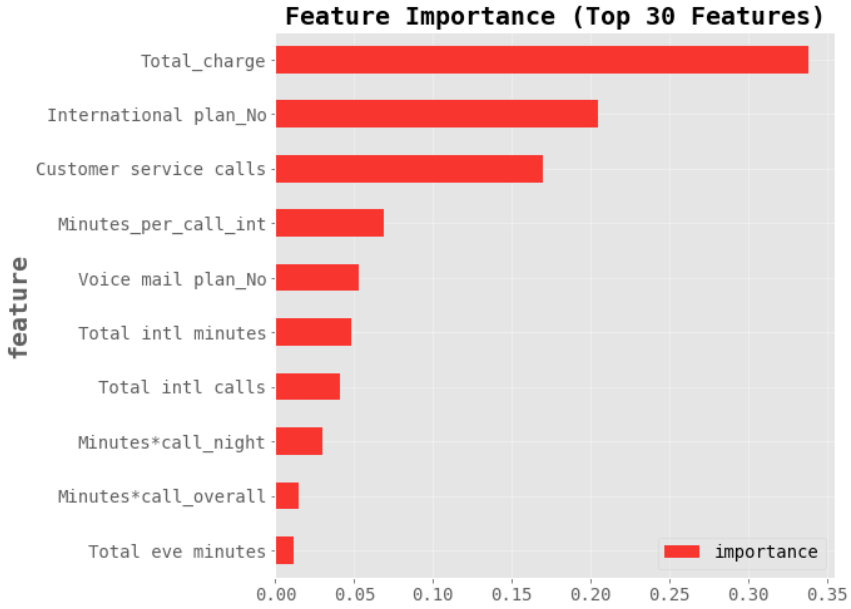
04

MODEL EXPLAINING

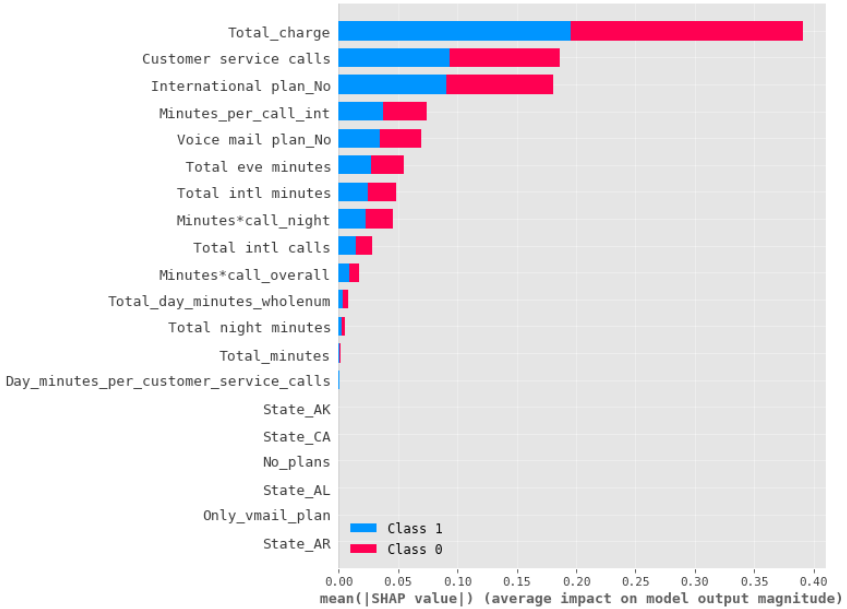
Feature Importance and
Shap Analysis



FEATURE IMPORTANCE



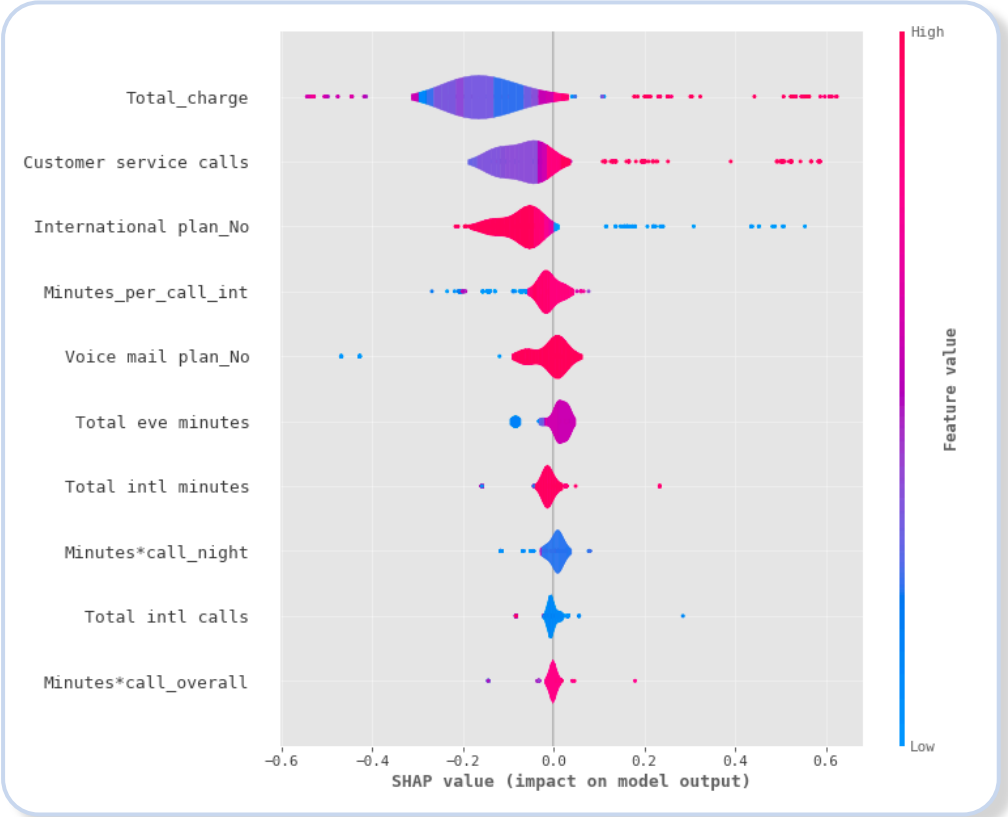
DECISION TREE FEATURE IMPORTANCE



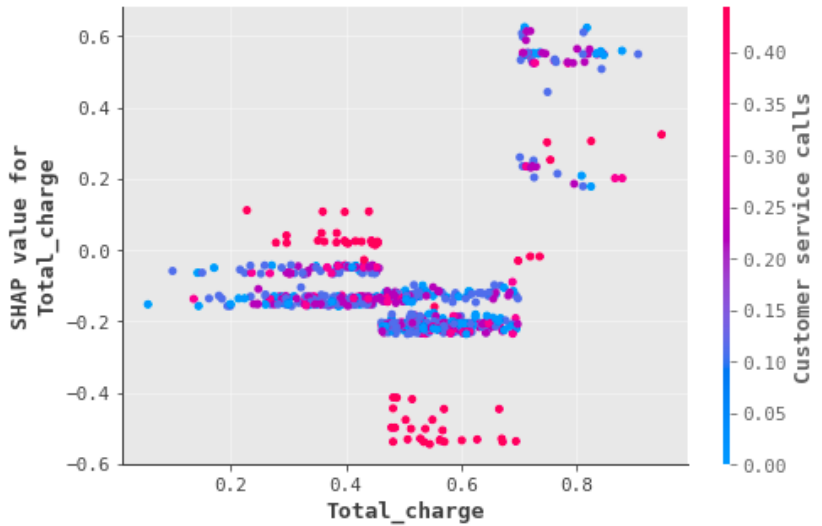
SHAP FEATURE IMPORTANCE

SHAP – SUMMARY PLOT

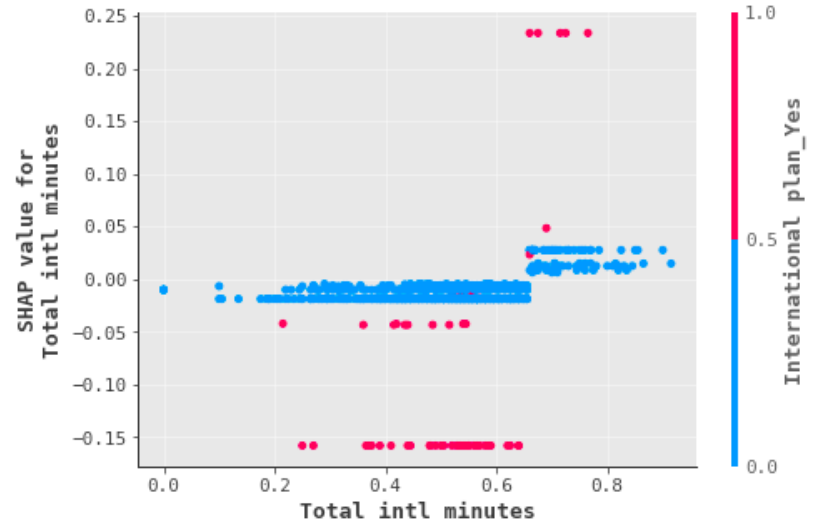
- The higher the **Total_charge** (red points) the more likely a customer will churn and vice versa.
- However, there are some customers who tend to churn less even though they have medium to high **Total_charge!** (red points towards extreme left across **Total_charge** feature). We will explore this group more in Shap dependence plot
- More **Customer service calls** (red) generally an indication of customer churning
- Customer with an **No International plan** (red) tend to churn less



SHAP – DEPENDENCE PLOT



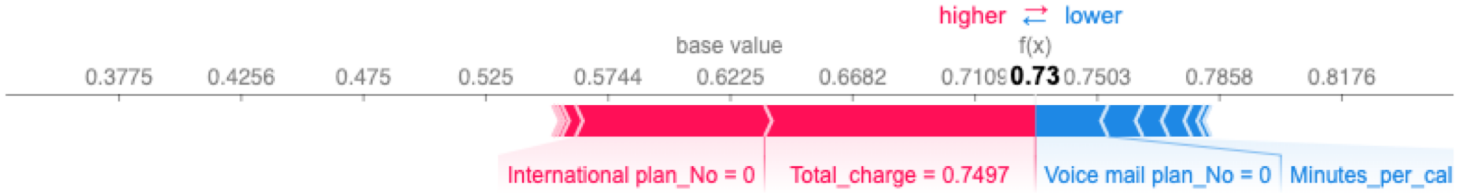
Customer service calls has strongest interaction with **Total_charge**



International plan_Yes has strongest interaction with **Total intl minutes**

SHAP – INDIVIDUAL PREDICTIONS

Predicted churn who actually **DIDN'T** churn



Model predicted customer to churn because he/she had high bill and an international plan

Predicted **NOT** churn who actually **DID** churn

Model predicted customer will not churn because he/she had low bill and low customer service calls

QUESTIONS?