Telco Customer Churn

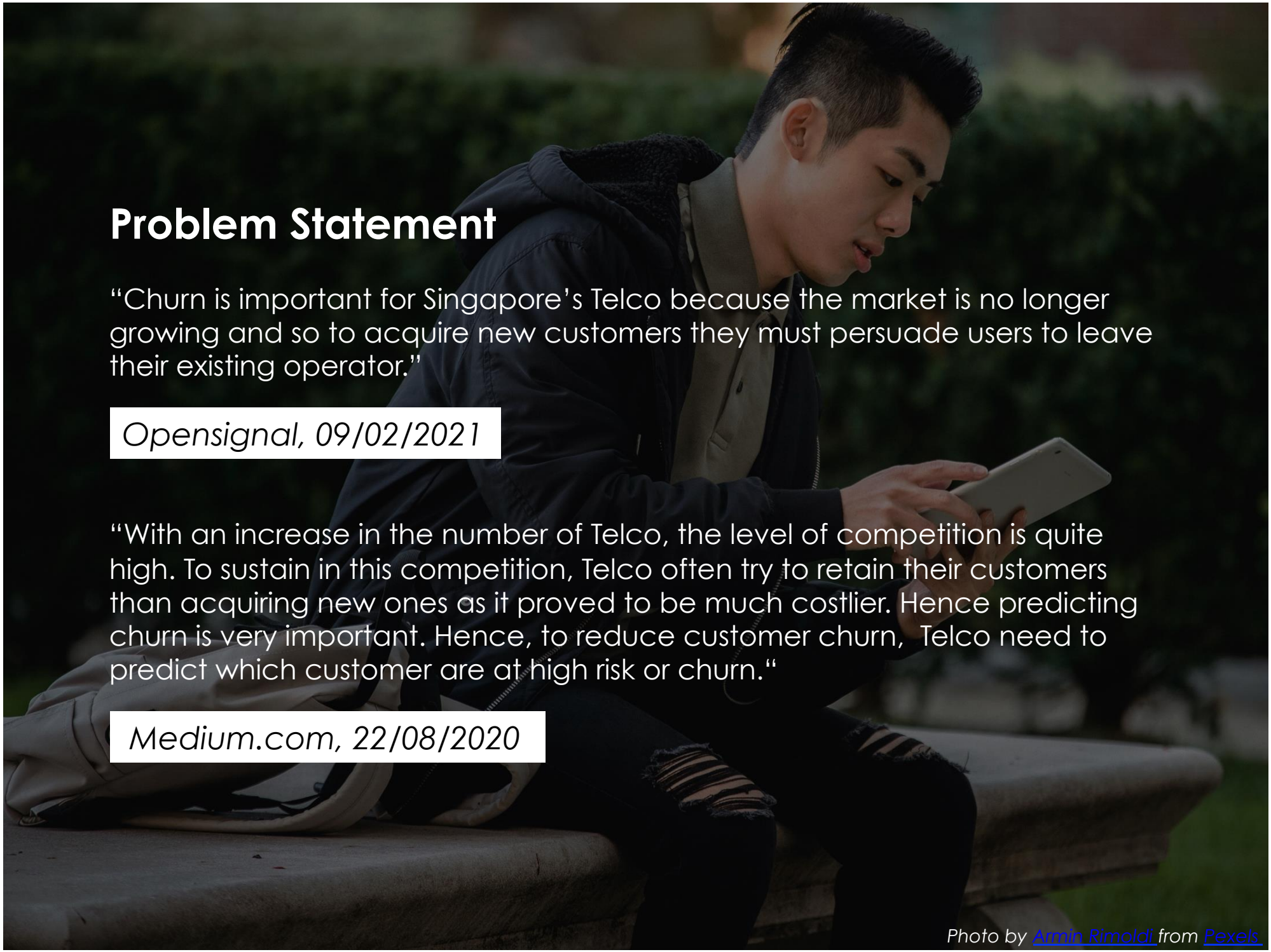
DS106 – Capstone Project Presentation

By Jia Yang

Photo by Cristian Dina from Pexels

Contents

Photo by Ketut Subiyanto from Pexels



Problem Statement

"Churn is important for Singapore's Telco because the market is no longer growing and so to acquire new customers they must persuade users to leave their existing operator."

Opensignal, 09/02/2021

"With an increase in the number of Telco, the level of competition is quite high. To sustain in this competition, Telco often try to retain their customers than acquiring new ones as it proved to be much costlier. Hence predicting churn is very important. Hence, to reduce customer churn, Telco need to predict which customer are at high risk or churn."

Medium.com, 22/08/2020

Goal

To find out the most striking behaviour of customers through EDA and train the most predictive machine learning (ML) model to determine the customers who are most likely to churn.

Questions

- What's the % of churn and customers that still in active services?
- Can we see different patterns in churn customers based on the type of service provided?
- Any difference pattern of churn between demographics?
- Any difference between customers that pay monthly and by year?
- Any other questions will be raise through EDA…
- What‘s the best classification ML model for prediction?
- What’s the best hyper-parameters for selected ML model?

Photo by Michael Herren from Pexels

Dataset

File :
Dataset_(Jia Yang).csv

From :
www.kaggle.com

Description :
Each row represents a customer, each column contains customer's attributes such as Churn/No-Churn, Signed-up Services, Account Info and Demographic Info

21 Columns & 7043 Rows

Dataset

Total: 21 Columns & 7043 Rows

Churn/No-Churn

Target
D.Type: **Character**
Var.Type**: Categorical**
Column: **1**

Churn

Services

Predictor
D.Type: **Character**
Var.Type**: Categorical**
Column: **9**

Phone Service
Multiple Lines
Internet Service
Online Security
Online Backup
Device Protection
Tech Support
Streaming TV
Streaming Movies

Account Info

Predictor
D.Type: **Num./Char.**
Var.Type**: Cont./Cat.**
Column: **7**

Customer ID
Tenure
Monthly Charges
Total Charges
Contract
Paperless
Payment Method

Demographics

Predictor
D.Type: **Character**
Var.Type**: Categorical**
Column: **4**

Gender
Partners
Dependents
Senior Citizen

All columns are important as they are used as predictor and target

Dataset

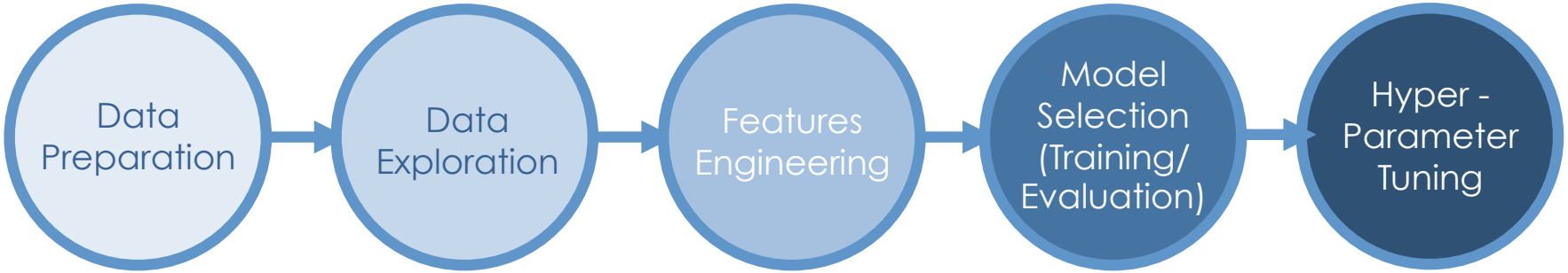
Snapshot

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No

5 rows × 21 columns

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

Approach & Challenges



- Null values treatment
- Duplicate records
- Joining of datasets

- Relationship between variables
- Answer most of the identified questions

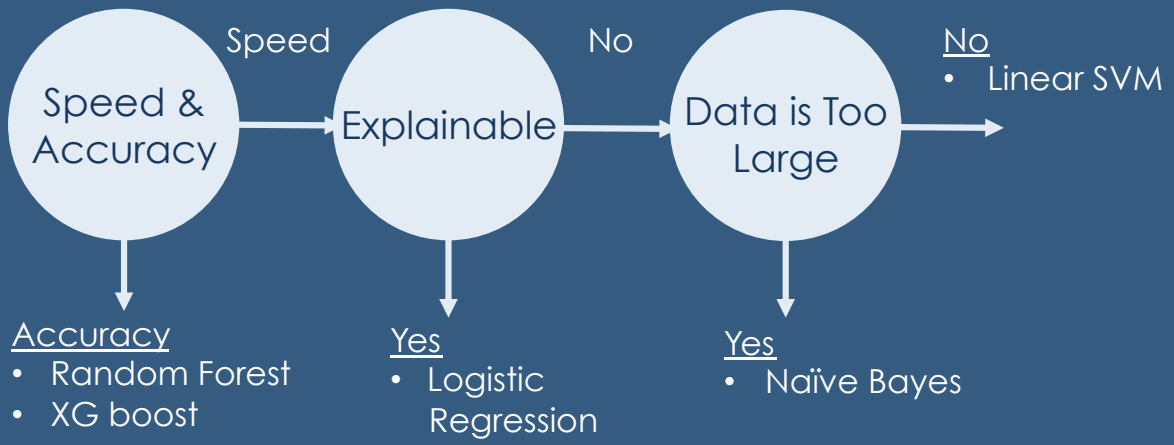
- Create new meaningful features
- Reduce unused features
- Data unbalance.
- Encode categorical features
- Scale overall dataset

- Choose the best Classification ML model based on different evaluation method

- Choose the best hyper parameter

Approach & Challenges

Classification Models Selection



Evaluation Methods

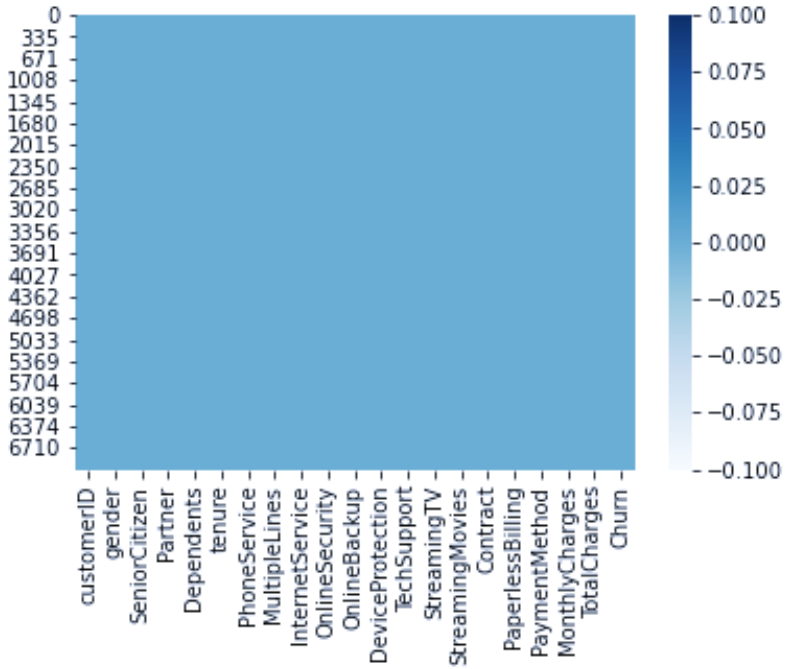
- Confusion Matrix
  - Accuracy
  - F1-Score
  - ROC / AUC

Machine Learning Process

1. Preparation | 2. Exploration | 3. Feature Eng. | 4. Select Model | 5. Tuning

1.1) Converting data type for 'TotalCharges' column **from object to float** for continuous features.

1.2) **No Null Values** in all columns



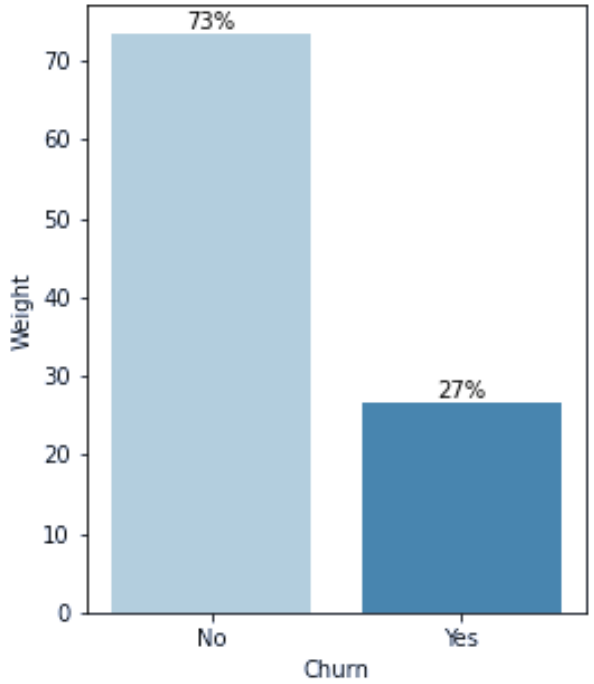
1.3) **Unique Customer IDs** are in 'cusomterID' column

Machine Learning Process

1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.1 Target Variable

- We are trying to predict if the customer churn. This will be a binary classification problem.
- Slightly unbalanced of the target churn/No Churn

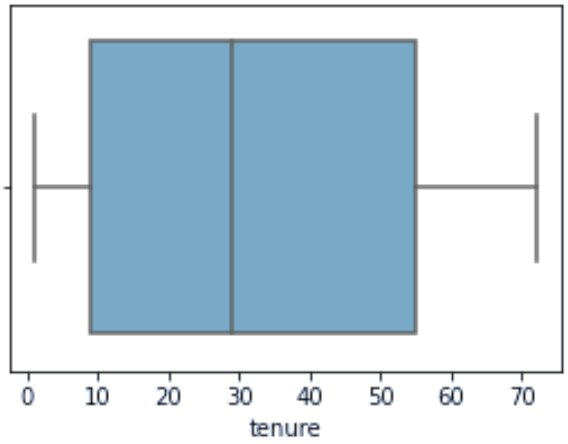


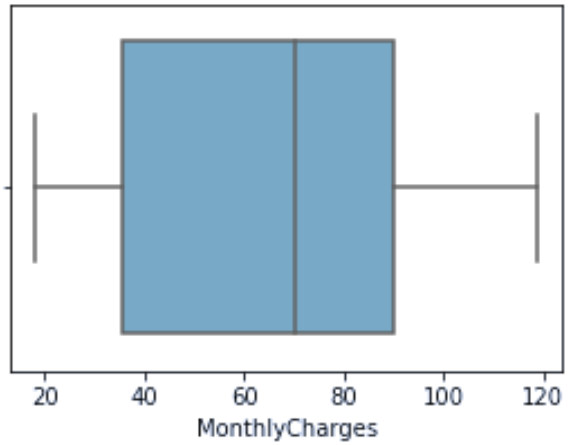
Machine Learning Process

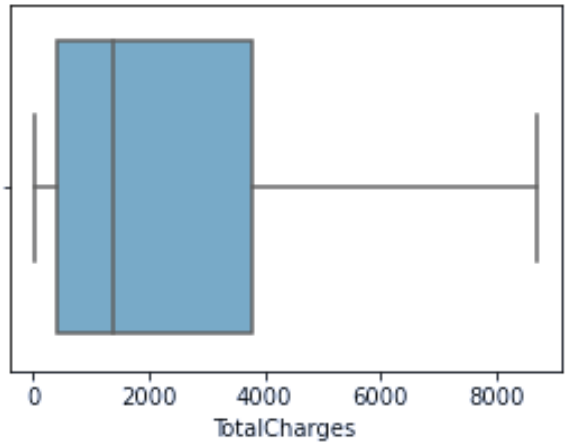
1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.2 Continuous Features

2.2.1 **No outlier** found for all continuous features





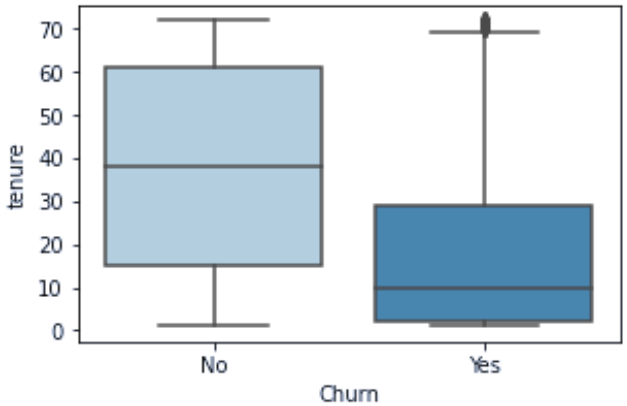


Machine Learning Process

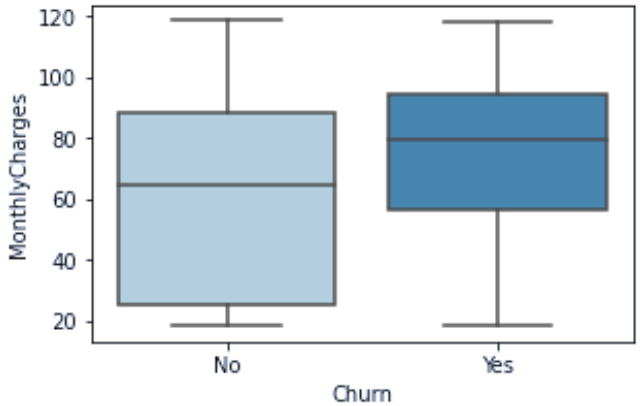
1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.2 Continuous Features

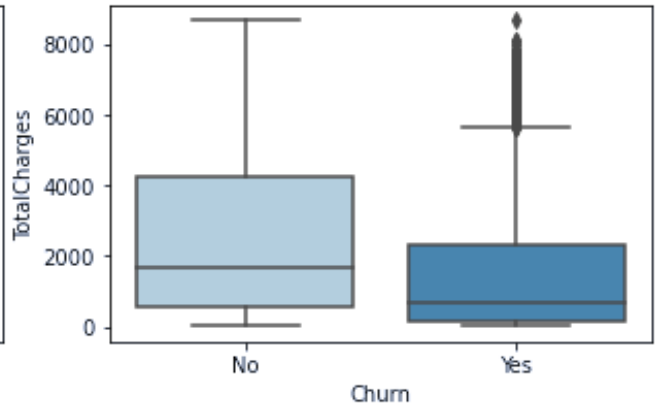
2.2.2 Continuous Variables & Churn



- Customer that do not churn tend to stay longer tenure with Telco.
- Most churn at the start of the tenure.
- Customer might not satisfy with the services, hence, drop out at the beginning.



- Higher churn with high monthly charges. This might be the contributor of the early churn



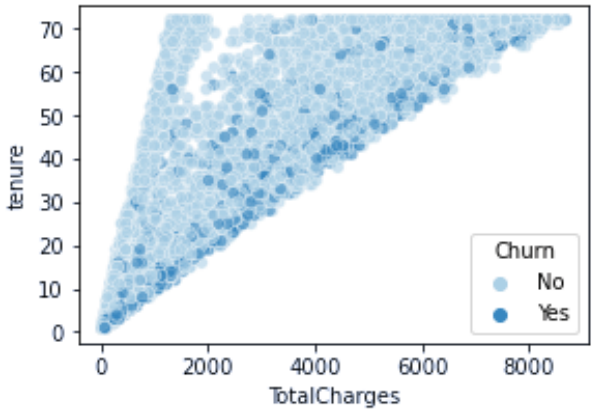
- Higher churn when total charges are lower.
- This further indicate that most churn at start of tenure as total charges increase across the tenure

Machine Learning Process

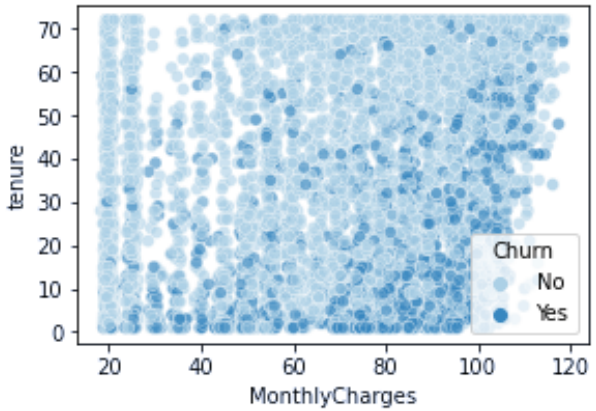
1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.2 Continuous Features

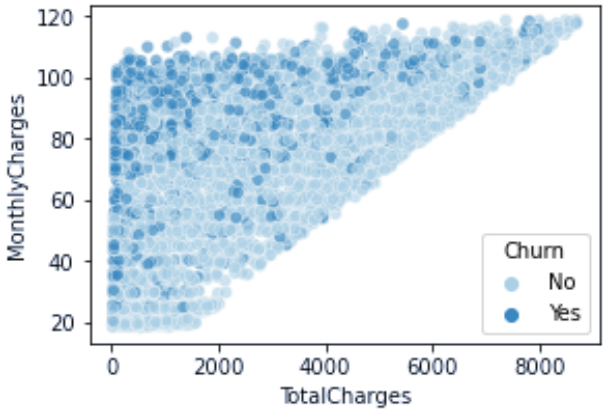
2.2.3 Continuous & Continuous Variables



- Total charges increased when tenure increase.
- It make sense as more year in contract, more charges the customer need to pay.
- Churn were quite distributed across the TotalCharges at each Tenure



- More churn at the high side of the MonthlyCharges at the beginning of the tenure.
- This answer to our initial finding that high monthly charge contribute to high churn at the start of the tenure



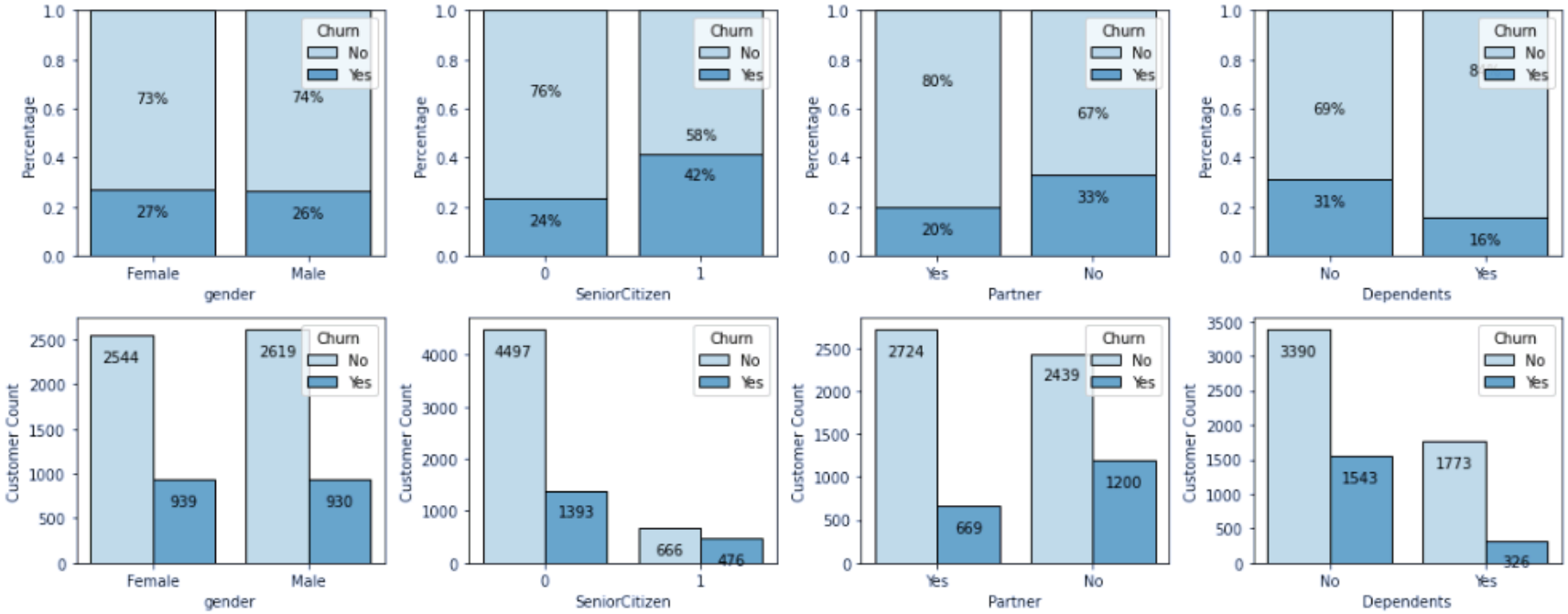
- More churn at the start of the total charges,
- This could be due to the high monthly charges at the beginning of the tenure, hence customer churn higher.

Machine Learning Process

1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.3 Categorical Features

2.3.1 Demographics



- The churn rates and customer count are almost equal in the case of Male and Females
- There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people.
- SeniorCitizen have a much higher churn rate at 42% against 23% for non-senior customers.
- Customers that doesn't have partners & dependents are more likely to churn

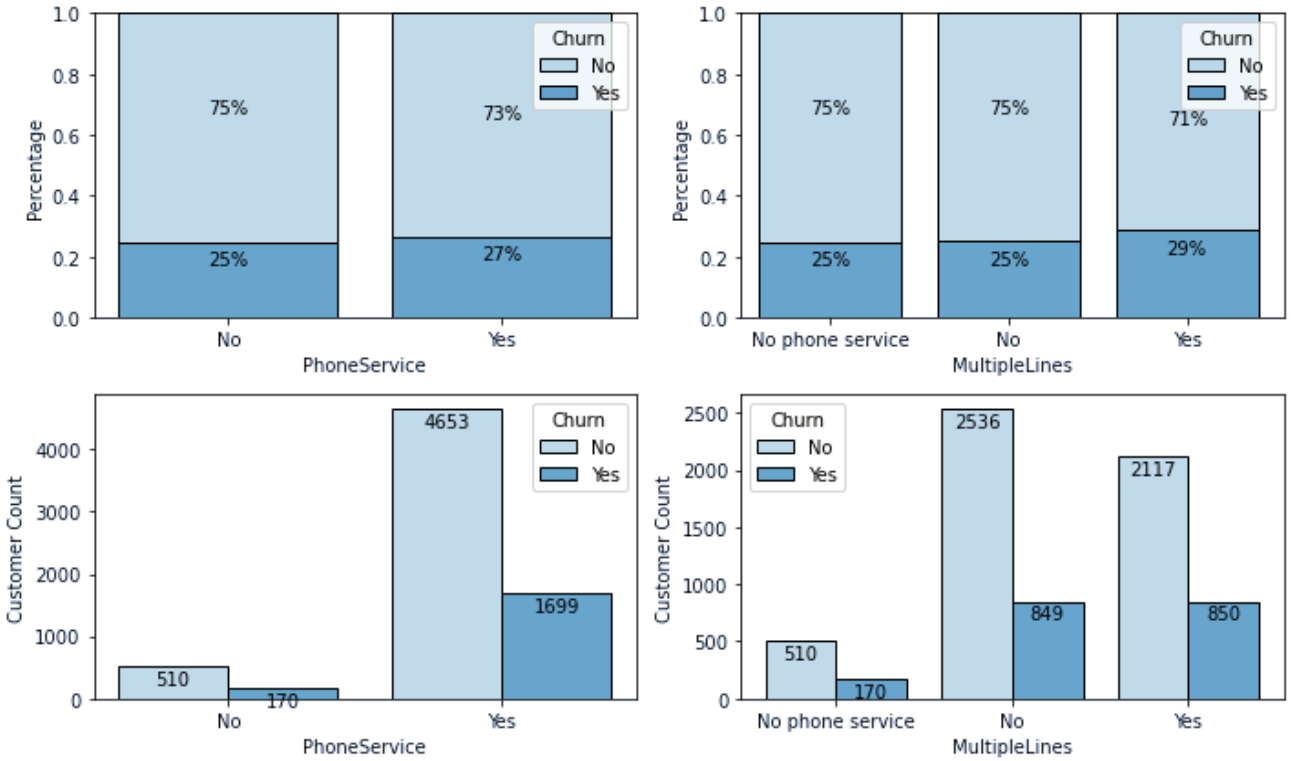
Machine Learning Process

1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.3 Categorical Features

2.3.2 Phone Services

- Churn rates are almost the same across all categories
- Almost 90% of customers holding phone service..



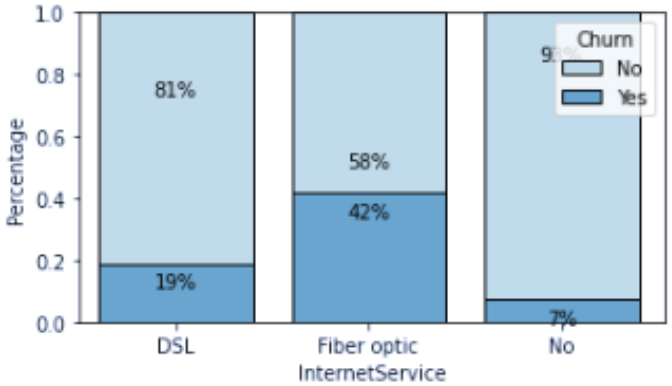
Machine Learning Process

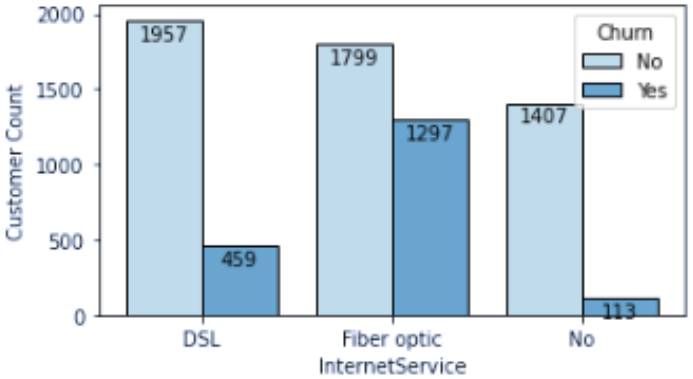
1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.3 Categorical Features

2.3.3 internet Services

- Churn rate is much higher with customer subscribed to Fiber Optic InternetServices.
- This might because customers subscribed to fiber optic are looking at internet speed, hence if it not to performance, customers will choose to churn.
- Customers without internet services will have lesser churn. As customer most likely into only mobile services.



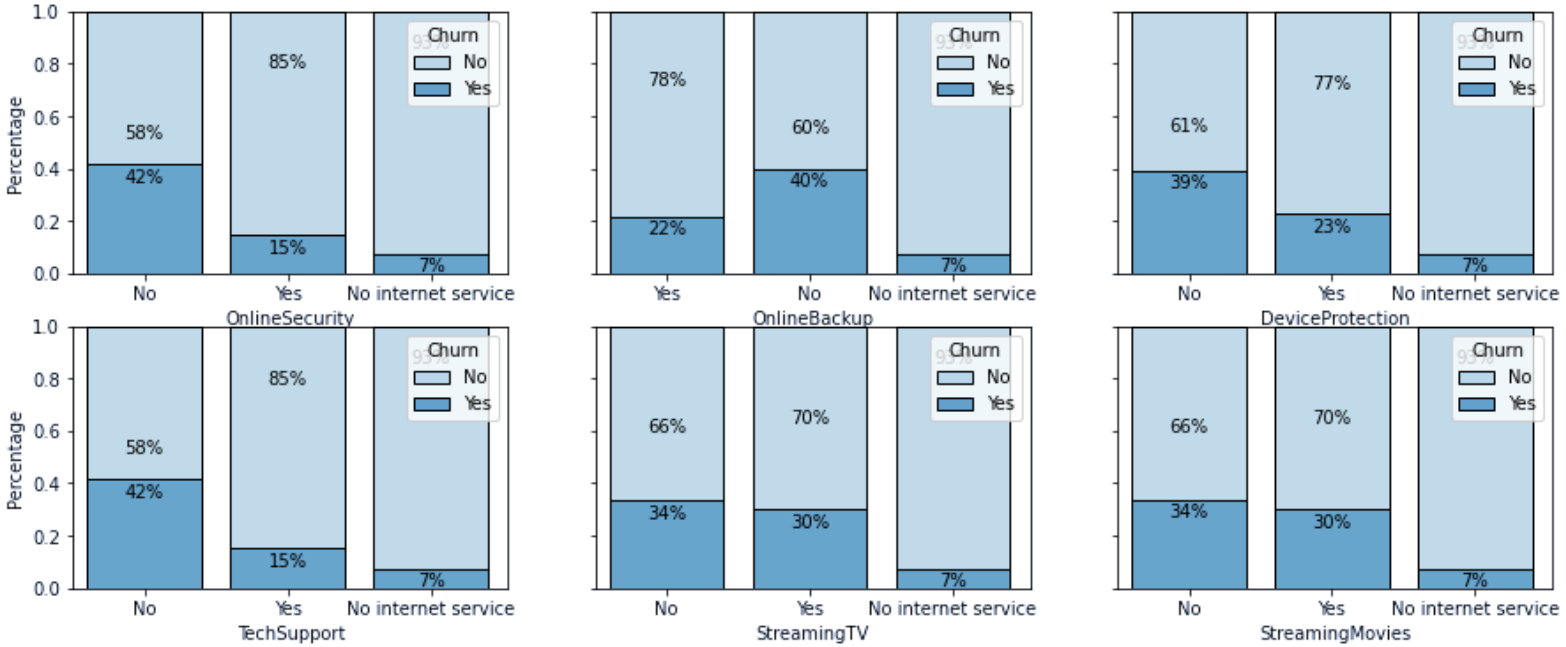


Machine Learning Process

1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.3 Categorical Features

2.3.4 Additional internet Services



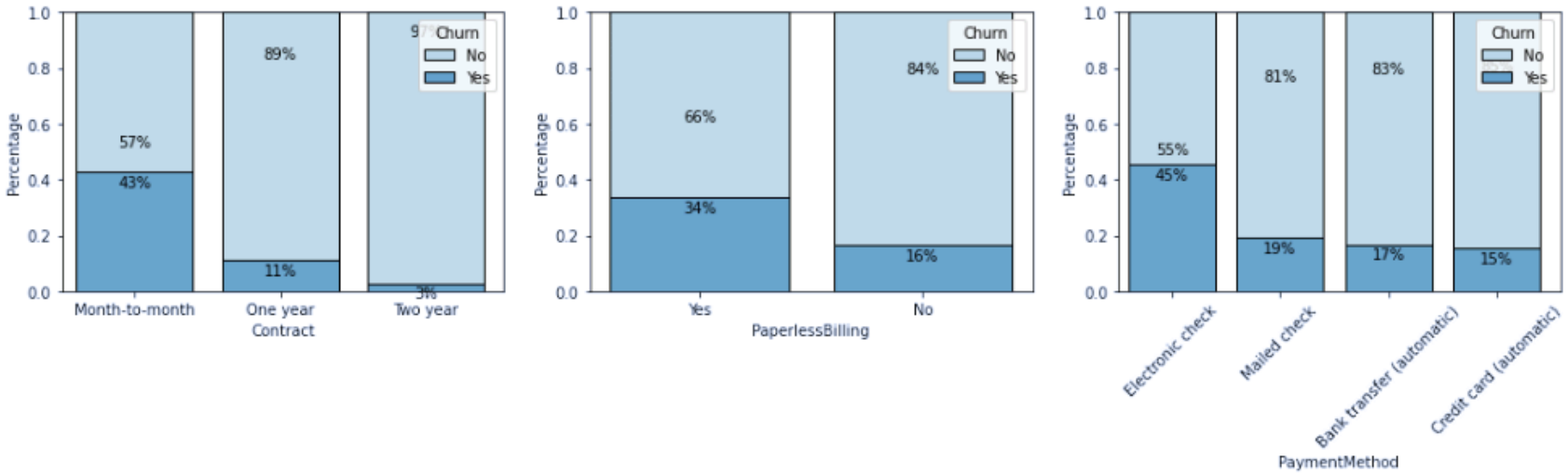
- Customers without internet services will have lesser churn. Customer may only to look into mobile services.
- Customers who do not subscribed to additional service have the highest churn.
- Customers may not feel secure & support without subscribing to the additional services, hence, churn more.

Machine Learning Process

1. Preparation | **2. Exploration** | 3. Feature Eng. | 4. Select Model | 5. Tuning

2.3 Categorical Features

2.3.5 Customer Payment Preferences



- Customers with monthly subscription have higher churn as compared to Customers with one or two year contract.
- Customer that signed monthly basis may tend to compare plan with other Telco and change to any better Telco plan.
- Churn rate is higher for customers having paperless billing option. This maybe customer not getting use of the paperless billing.
- Customers who uses Electronic Check payment churn more as compared to other payment options. This maybe customer do not feel comfortable while using check online..

Machine Learning Process

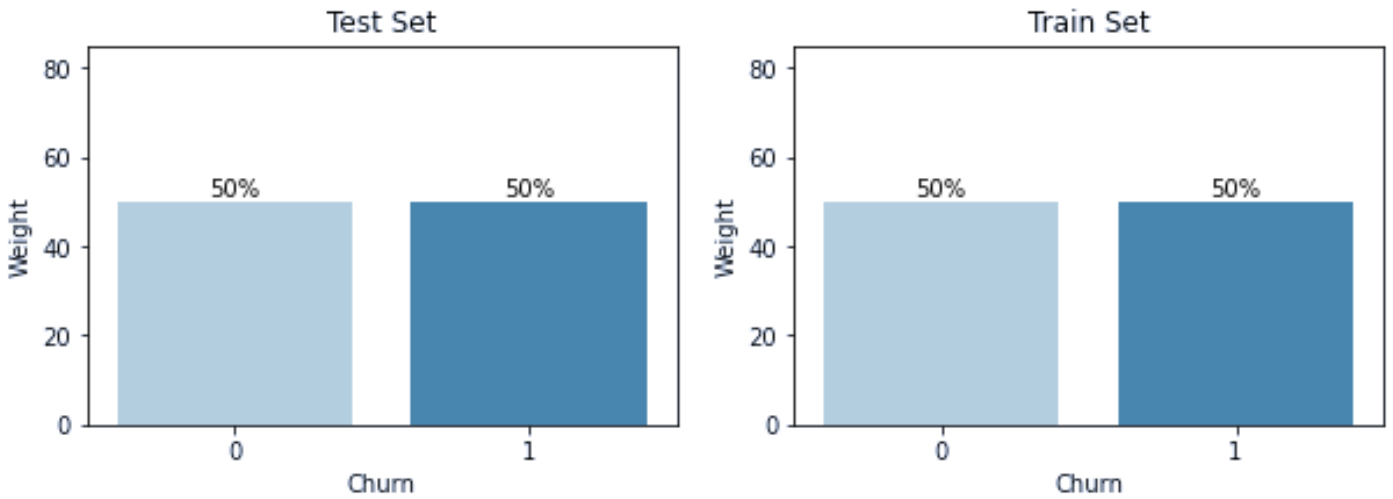
1. Preparation | 2. Exploration | **3. Feature Eng.** | 4. Select Model | 5. Tuning

3.1) **Convert binary classification** features to 0 and 1

3.2) **Convert non binary classification** features with frequency encoder

3.3) Train-Test Split: **70% Train** & **30% Test**

3.4) Make dataset balance with **Oversampling method**

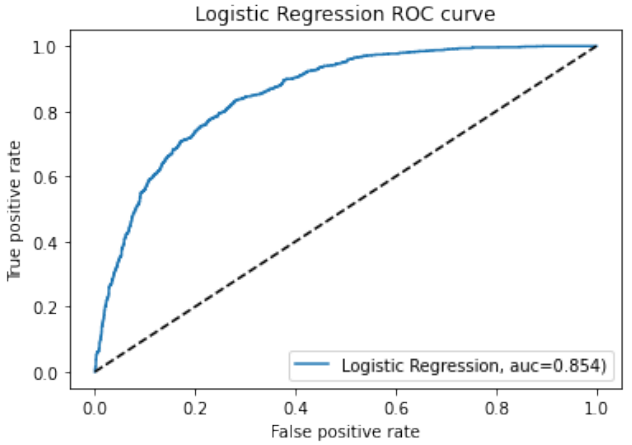


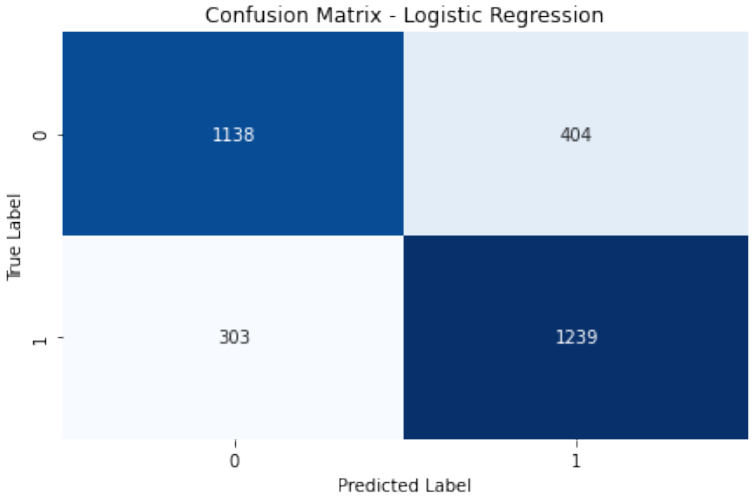
Machine Learning Process

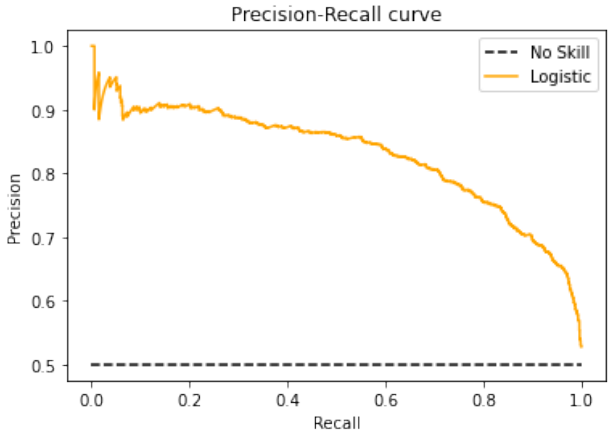
1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.1 Logistic Regression

Metrics	Result
Accuracy	0.77
Precision	0.75
Recall	0.80
F1-Score	0.78
AUC	0.85





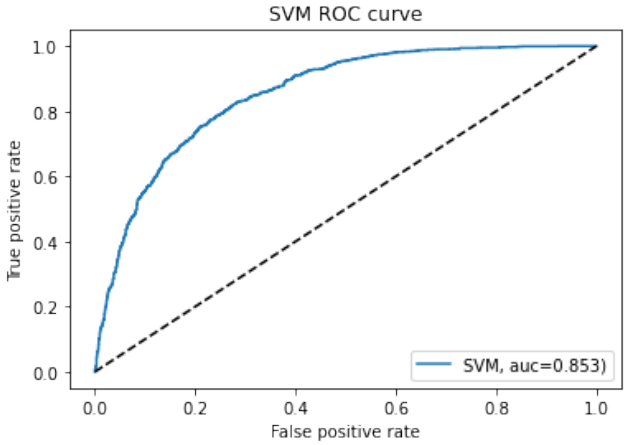


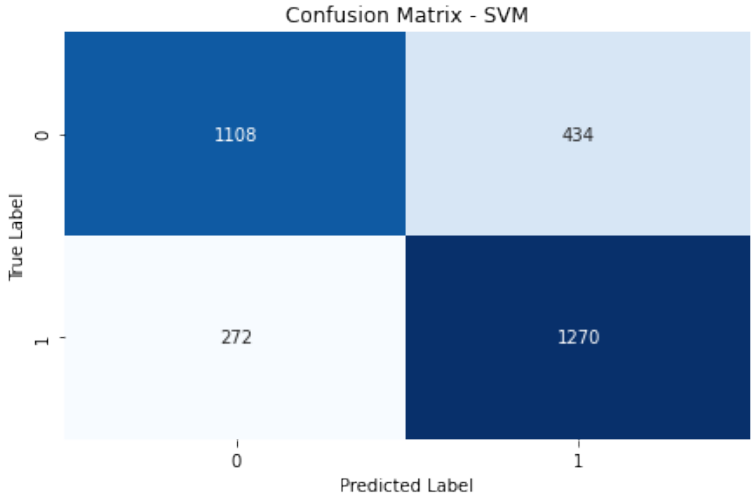
Machine Learning Process

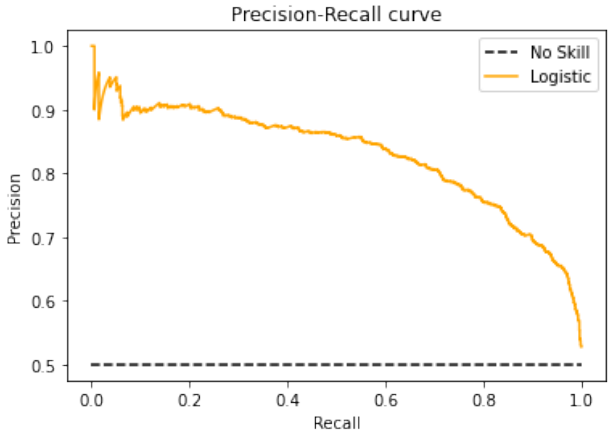
1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.2 Support Vector Machine

Metrics	Result
Accuracy	0.77
Precision	0.75
Recall	0.82
F1-Score	0.78
AUC	0.85





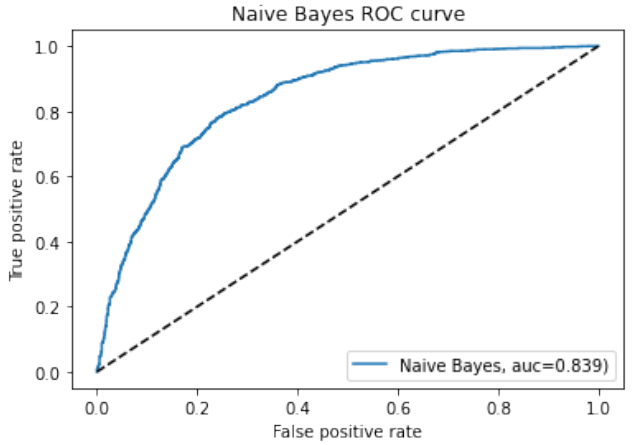


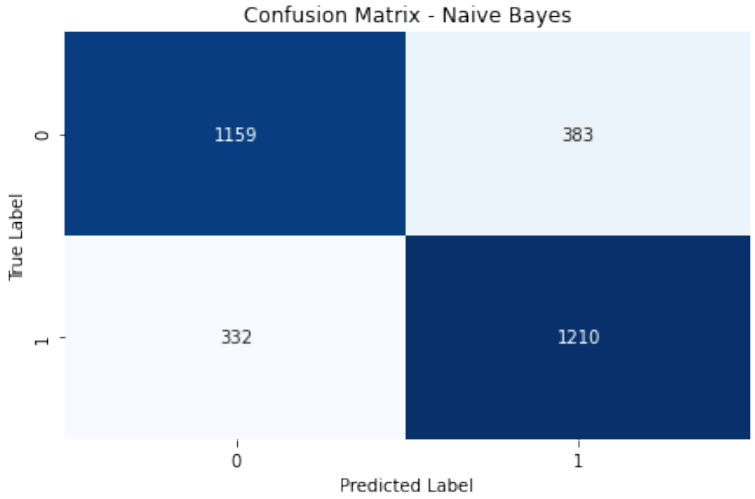
Machine Learning Process

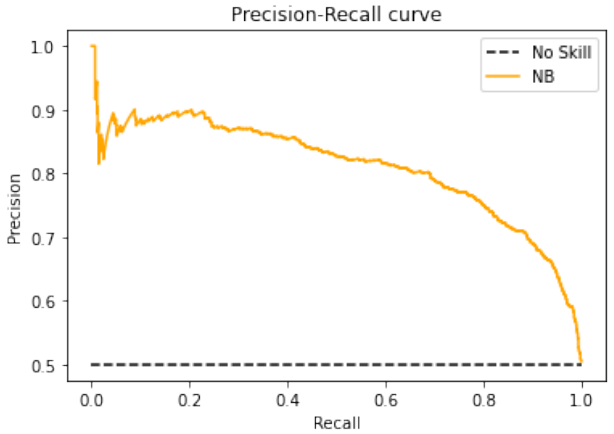
1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.3 Naive Bayes

Metrics	Result
Accuracy	0.77
Precision	0.76
Recall	0.78
F1-Score	0.77
AUC	0.84





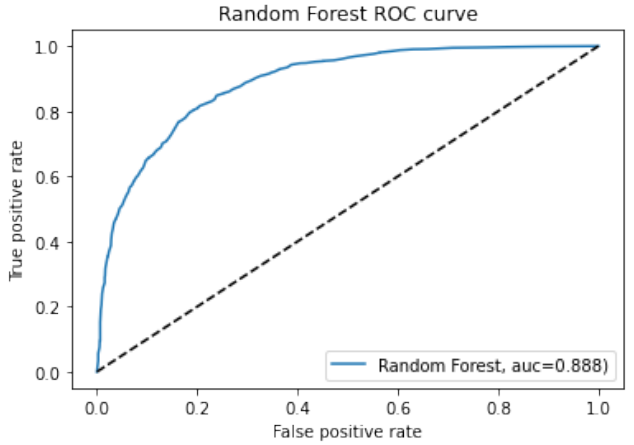


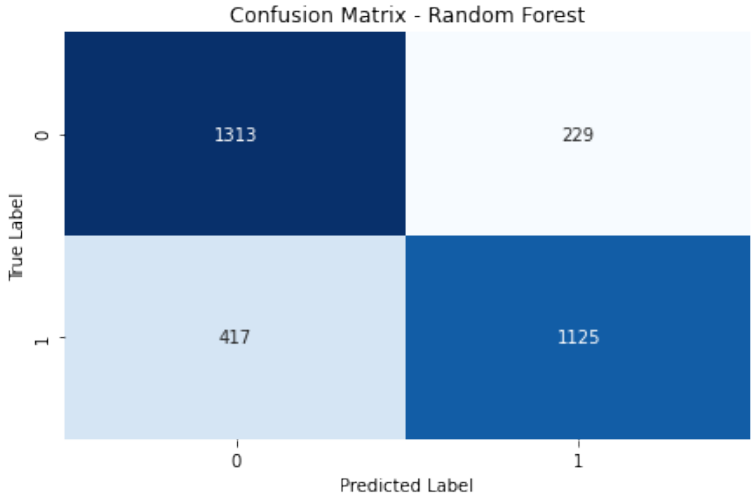
Machine Learning Process

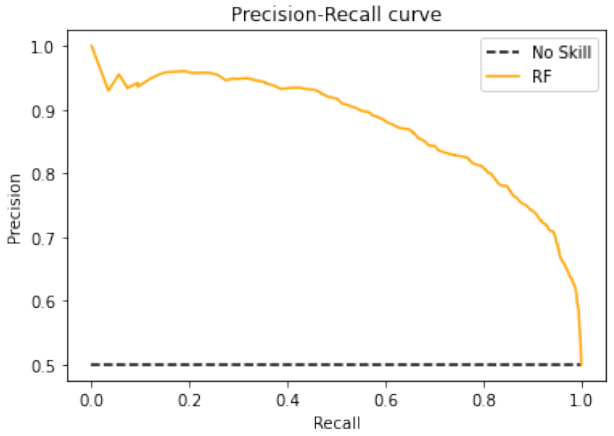
1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.4 Random Forest

Metrics	Result
Accuracy	0.79
Precision	0.83
Recall	0.73
F1-Score	0.78
AUC	0.89





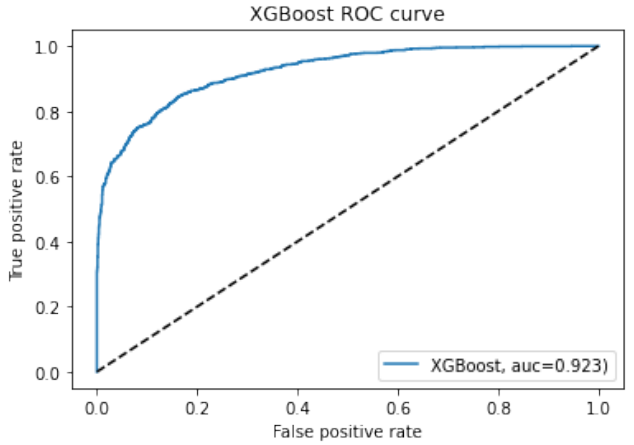


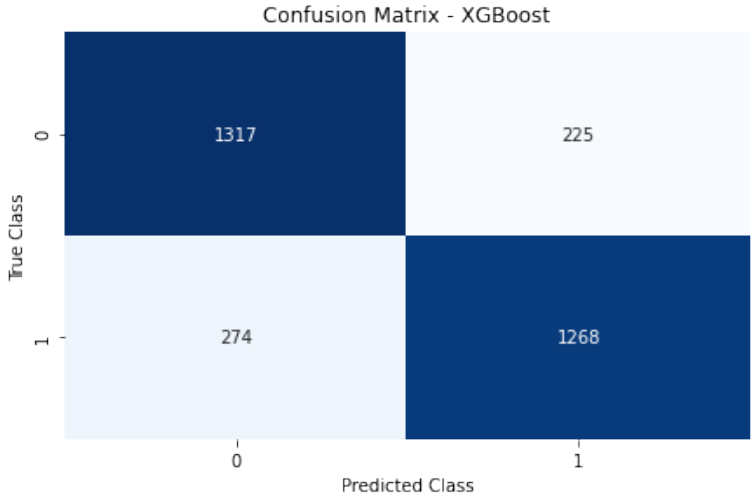
Machine Learning Process

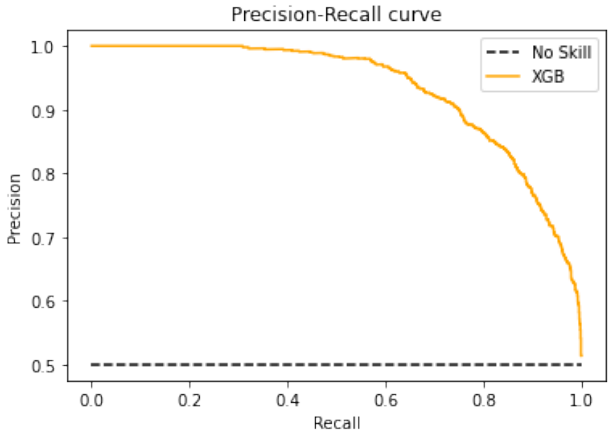
1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.5 XG Boost

Metrics	Result
Accuracy	0.84
Precision	0.85
Recall	0.82
F1-Score	0.84
AUC	0.92







Machine Learning Process

1. Preparation | 2. Exploration | 3. Feature Eng. | **4. Select Model** | 5. Tuning

4.6 Result Comparison

Random Forest and XG Boost has the best performance according to our three classification metrics (Accuracy, F1-score and AUC). Next, we can further improved the both models by tuning hyper-parameters.

RESULT COMPARISON TABLE

Test Method:	Log	SVM	NB	RF	XGB
Accuracy	0.771	0.771	0.768	0.791	0.838
F1 Score	0.778	0.783	0.772	0.777	0.836
AUC	0.854	0.853	0.839	0.888	0.923

Machine Learning Process

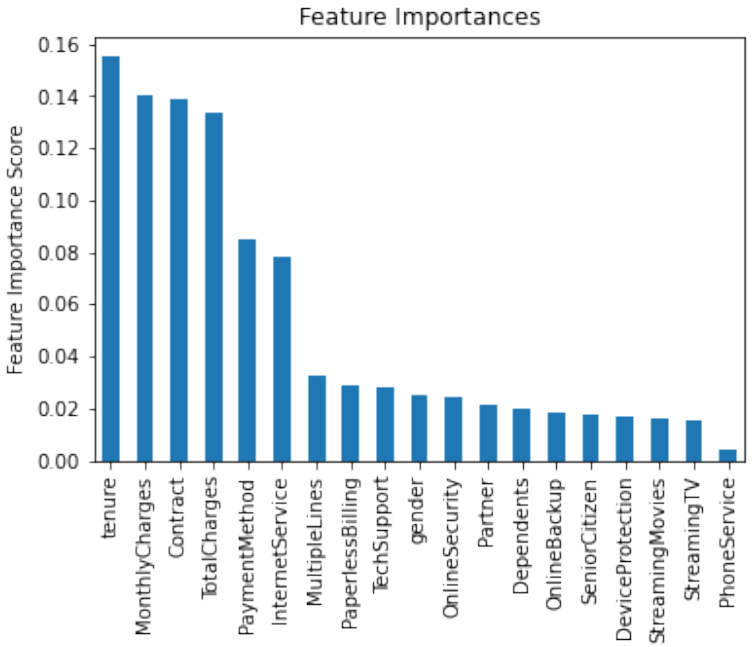
1. Preparation | 2. Exploration | 3. Feature Eng. | 4. Select Mod | **5. Tuning**

5.1 Random Forest Tuning

For tuning hyper-parameter for Random Forest, we will use and compare RandomSearch.cv and GridSearch.cv to get the best parameter.

Hyper-parameters to be tune:

- **n_estimators -** Number of trees in random forest
- **max_features** - Number of features to consider at every split
- **max_depth -** Maximum number of levels in tree
- **min_samples_split** - Minimum number of samples required to split a node
- **min_samples_leaf -** Minimum number of samples required at each leaf node



RF: BASE & GRIDSEARCH TABLE

Test Method:	Base	Grid
Accuracy	0.791	0.792
F1 Score	0.777	0.777
AUC	0.888	0.882

Machine Learning Process

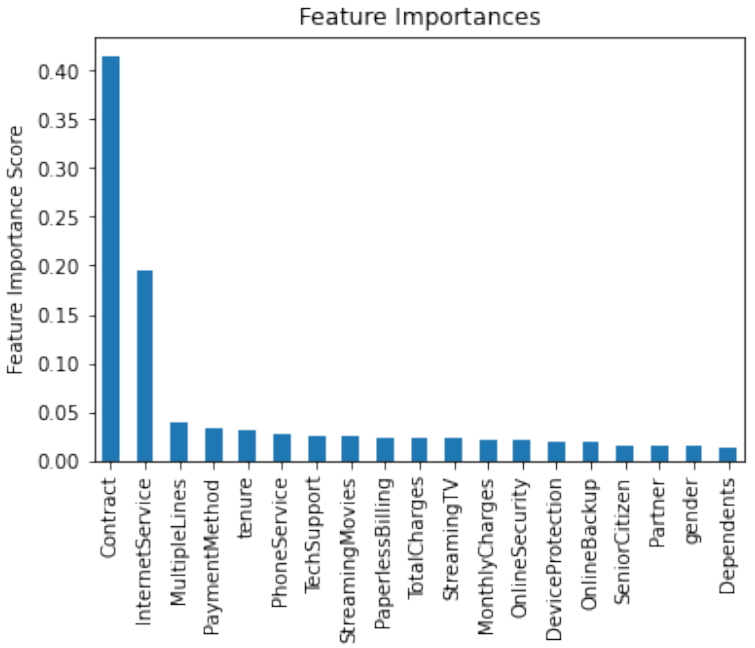
1. Preparation | 2. Exploration | 3. Feature Eng. | 4. Select Mod | **5. Tuning**

5.2 XG Boost Tuning

For tuning hyper-parameter for XG Boost, we will use and compare RandomSearch.cv and GridSearch.cv to get the best parameter.

Hyper-parameters to be tune:

- **max_depth -** Max depth of a tree
- **min_child_weight –** Min sum of weights of all observations required in a child
- **subsample -** Fraction of observations to be randomly samples for each tree
- **colsample_bytree** - Fraction of columns to be randomly samples for each tree.
- **eta** - Parameter controls the learning rate. Makes the model more robust by shrinking the weights on each step



XGB: BASE & RANDOMSEARCH TABLE

Test Method:	Base	Random
Accuracy	0.838	0.809
F1 Score	0.836	0.801
AUC	0.923	0.906

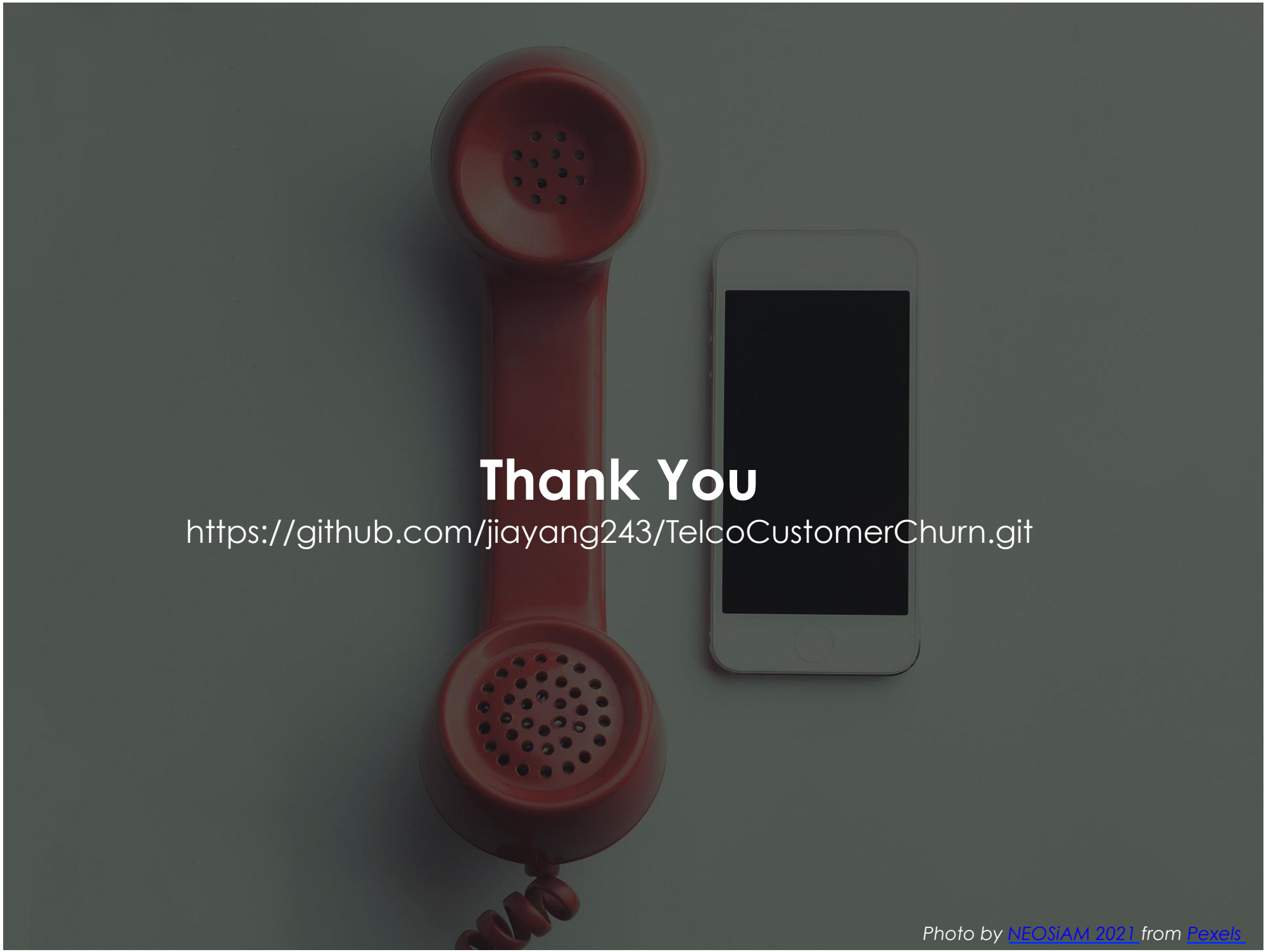
Conclusion

The best model to identify Telco Customer Churn will be XG Boost model with test of ~84% accuracy.

We will be using XG Boost to predict our values.

For future work, we can look into:

- Other value range for parameter tuning
- Look into using ensemble of different model to get a more accurate prediction
- Look into the feature importance for the top features
- Do Cost Benefit Analysis to further enhance the practical use of the model


Thank You
https://github.com/jiayang243/TelcoCustomerChurn.git
Photo by NEOSiAM 2021 from Pexels