Customer Churn Prediction
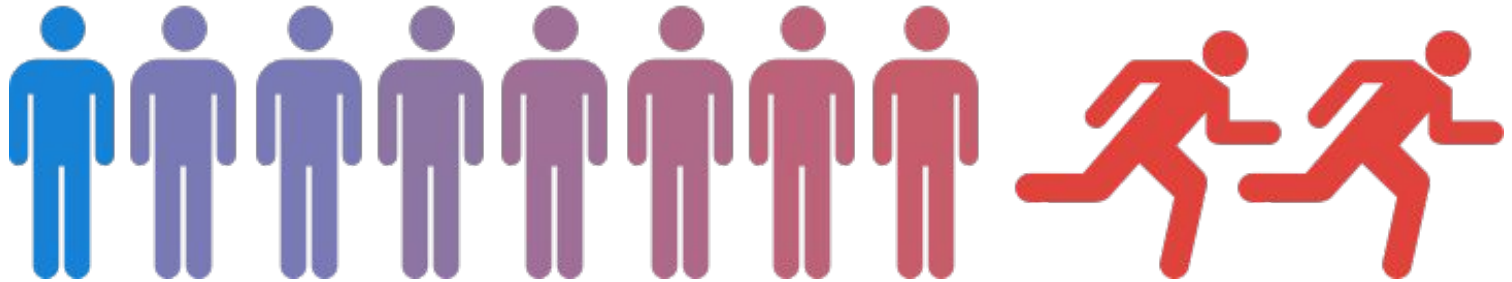
Waruni N. Wijayasinghe

Data Science Career Track
Springboard

March 2022

Business Problem

CUSTOMER CHURN

Data

Telco Customer Churn Data from Kaggle.

	0	1	2	3	4
customerID	7590-VHVEG	5575-GNVDE	3668-QPYBK	7795-CFOCW	9237-HQITU
gender	Female	Male	Male	Male	Female
SeniorCitizen	0	0	0	0	0
Partner	Yes	No	No	No	No
Dependents	No	No	No	No	No
tenure	1	34	2	45	2
PhoneService	No	Yes	Yes	No	Yes
MultipleLines	No phone service	No	No	No phone service	No
InternetService	DSL	DSL	DSL	DSL	Fiber optic
OnlineSecurity	No	Yes	Yes	Yes	No
OnlineBackup	Yes	No	Yes	No	No
DeviceProtection	No	Yes	No	Yes	No
TechSupport	No	No	No	Yes	No
StreamingTV	No	No	No	No	No
StreamingMovies	No	No	No	No	No
Contract	Month-to-month	One year	Month-to-month	One year	Month-to-month
PaperlessBilling	Yes	No	Yes	No	Yes
PaymentMethod	Electronic check	Mailed check	Mailed check	Bank transfer (automatic)	Electronic check
MonthlyCharges	29.85	56.95	53.85	42.3	70.7
TotalCharges	29.85	1889.5	108.15	1840.75	151.65
Churn	No	No	Yes	No	Yes

Data Wrangling

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   customerID        7043 non-null    object
 1   gender            7043 non-null    object
 2   SeniorCitizen     7043 non-null    int64
 3   Partner           7043 non-null    object
 4   Dependents        7043 non-null    object
 5   tenure            7043 non-null    int64
 6   PhoneService      7043 non-null    object
 7   MultipleLines     7043 non-null    object
 8   InternetService   7043 non-null    object
 9   OnlineSecurity    7043 non-null    object
 10  OnlineBackup      7043 non-null    object
 11  DeviceProtection  7043 non-null    object
 12  TechSupport       7043 non-null    object
 13  StreamingTV       7043 non-null    object
 14  StreamingMovies   7043 non-null    object
 15  Contract          7043 non-null    object
 16  PaperlessBilling  7043 non-null    object
 17  PaymentMethod     7043 non-null    object
 18  MonthlyCharges    7043 non-null    float64
 19  TotalCharges      7043 non-null    object
 20  Churn             7043 non-null    object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```
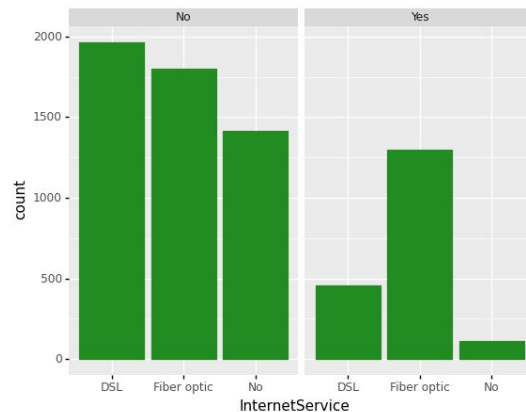
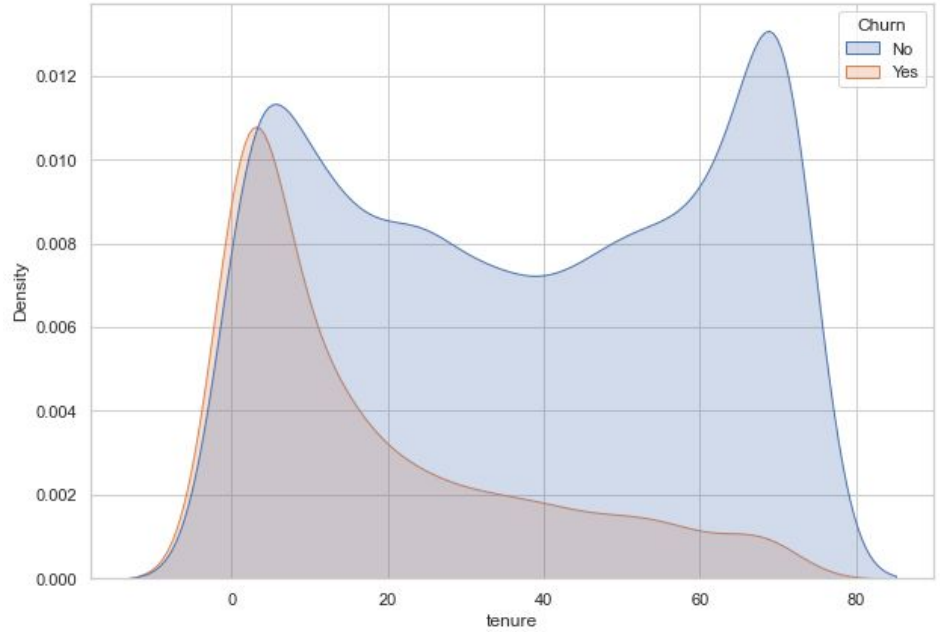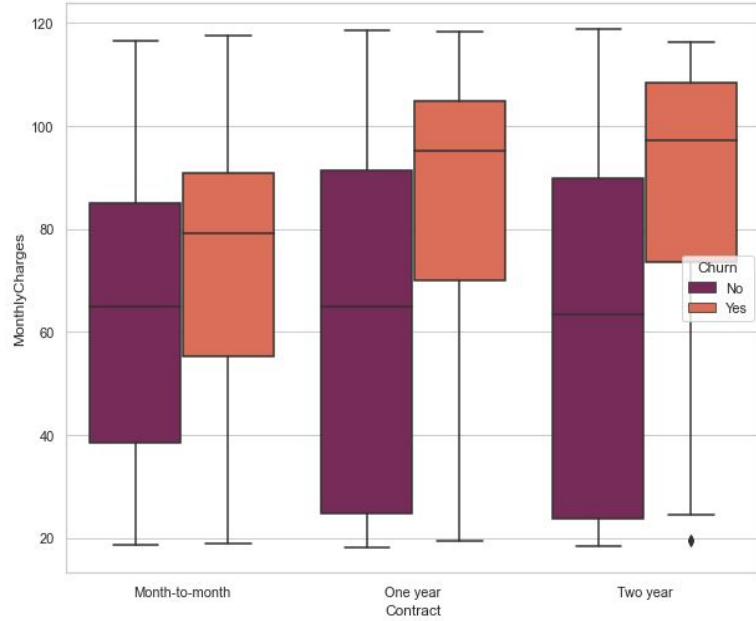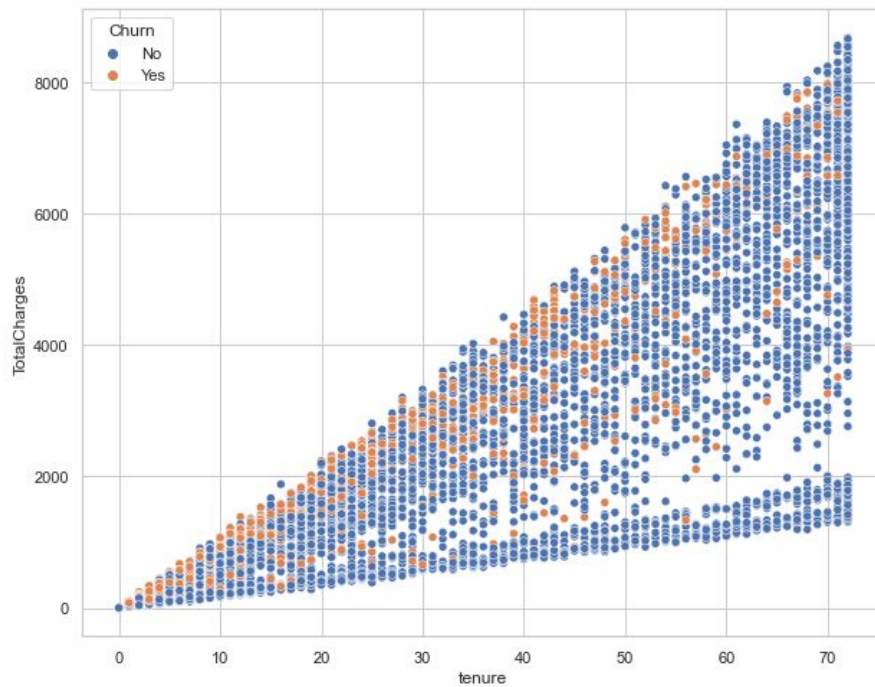# Data Exploration

Two Main Services:

-Phone

-Internet

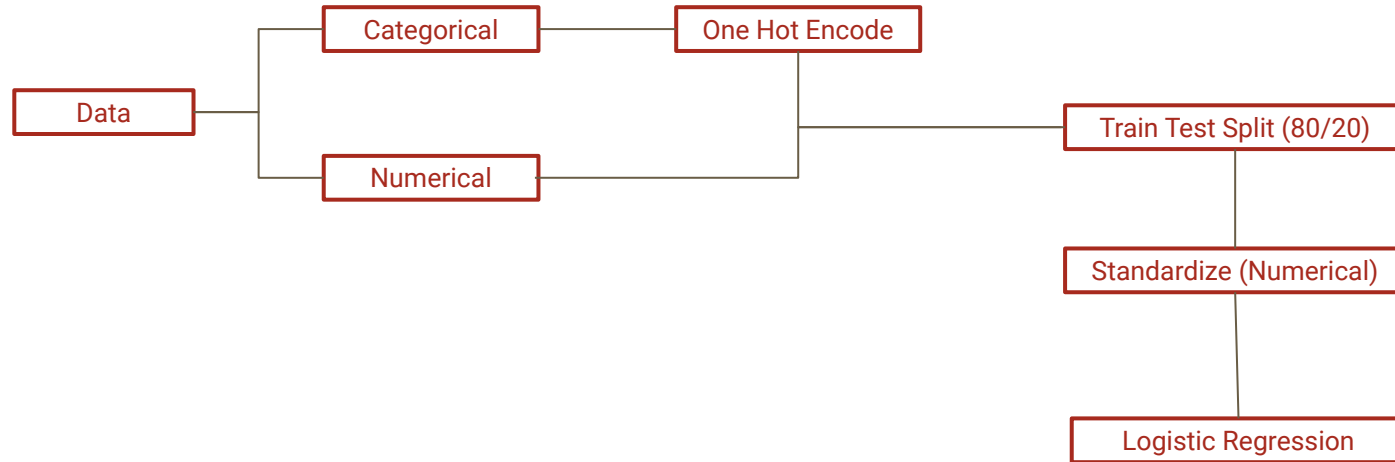| Internet | Did Not Churn | | | Churn | | |
| | Phone | | Total | Phone | | Total |
	No	Yes		No	Yes	
DSL	512	1450	**1962**	170	289	**459**
Fiber Optic	0	1799	**1799**	0	1297	**1297**
No	0	1413	**1413**	0	113	**113**
Total	**512**	**4662**	**5174**	**170**	**1699**	**1899**

Data Exploration

Data Exploration

Data Preprocessing & Baseline Model

Other Models

1. Random Forest Classifier

2. CatBoost Algorithm

3. LightGBM Algorithm

Evaluation Metrics

1. **Accuracy**

   Logistic Regression : 0.797
   Random Forest : 0.7921
   CatBoost : 0.8034
   LightGBM : 0.8027

2. Confusion Matrix
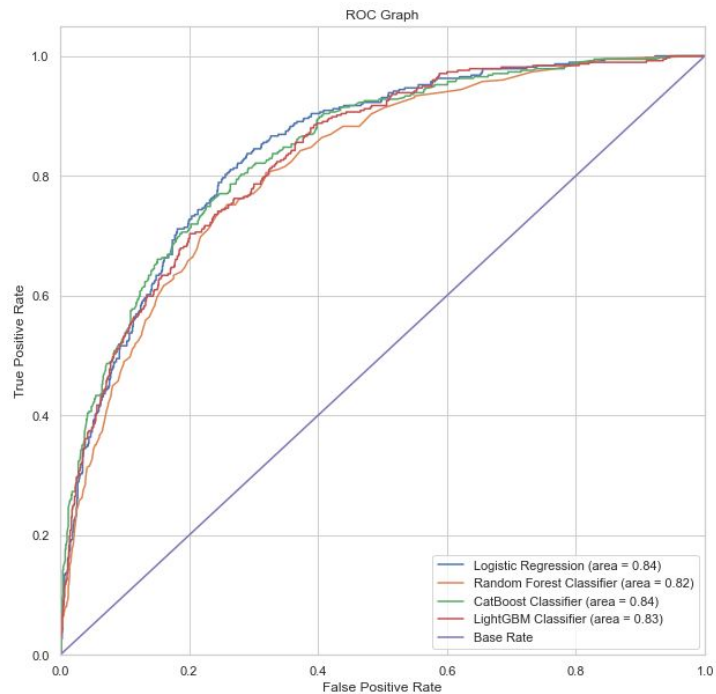
   a. Precision
   b. Recall
   c. **F1 Score**

   Logistic Regression :0.58
   Random Forest : 0.56
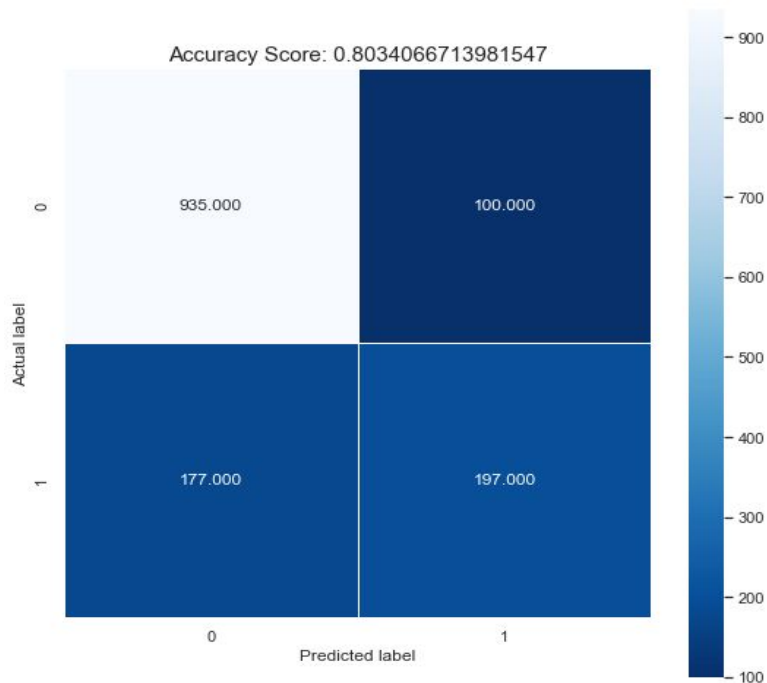   **CatBoost : 0.59**
   LightGBM : 0.59

Final Model

CATBOOST IS THE BEST!

CatBoost Model prediction

Default Threshold is 0.5!!!

Future Work

- Use resampling techniques (Under-sampling, Over-sampling, SMOTE) to minimize the effect on imbalance data issues.

- Hyper parameter tuning to increase the accuracy and efficiency of modeling.

- Strategy development on top of the model.

- Elements around deployment.

Questions?

Thank You!