TELECOM CUSTOMER CHURN PREDICTION

GROUP 2

SUBMITTED BY:

PRABAKARAN V
VIGNESH J
ULLAS D P
GUNASEELAN B

About Telecom Industry

OBJECTIVES & BACKGROUND

- One of the fastest growing industry providing infrastructure for data transmission and communication.
- It includes wireless service, cellphone service, Internet service and fiber optic networks among other things.

BUSINESS CONTEXT

- Worldwide, the telecom industry is facing increasing competitive pressure and new disruptive services offered by OTT players.
- Wireless penetration is reaching a saturation point in major markets and it is getting difficult to provide differentiating products/services.
- Telecom companies are forced to respond through more competitive offers, bundle offerings and price cuts.

PROBLEM STATEMENT:

Customer churn or customer attrition is one of the key business metrics for telecom industry. The cost of retaining an existing customer is far lesser than acquiring a new one Customer churn can be due to one of the following factors:

- Price
- Service quality
- Lack of customer service
- Billing disputes
- New competitors etc.

GOALS:

- ➢ To predict the customer churn by assessing their propensity of risk to churn.
- ➢ To focus on the customer base who are more vulnerable to churn

ABOUT DATASET:

- The dataset has the past behavior of more than 7000 customers with 21 features.
- Dataset is taken from Kaggle(https://www.kaggle.com/blastchar/telco-customer-churn)
- The dataset has information about customers who left within the last month, customer account info, demographic info and services that each customer signed up for.

MODEL SELECTION AND EVALUATION METRICS:

We built various algorithm models to predict the customer churn and evaluated the best model by comparing key metrics like accuracy score, confusion matrix, precision, recall and F1 score.

ALGORITHMS USED TO PREDICT CHURN:

- Logistic Regression
- Decision Tree Classifier
- Bagging Classifier
- Random Forest Classifier
- Gradient Boosting
- Ada Boost

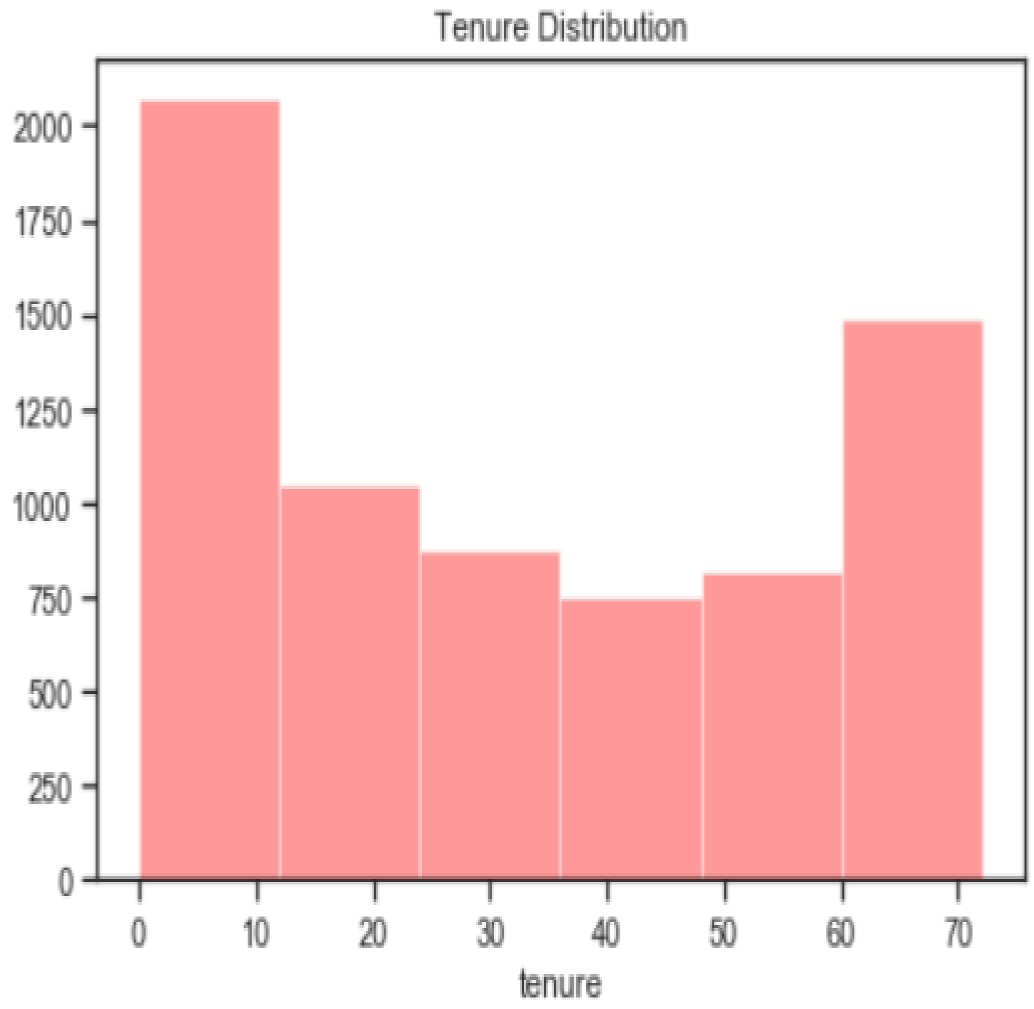
PACKAGES USED:

1. Pandas
2. Numpy
3. Matplot library
4. Seaborn
5. Scikit-learn

DATA PREPARATION

- Converted datatypes of Variables (predictors)
- Identified and Treated Missing Values
- Replaced irrelevant observation in variables
- Converted Binary categorical variables
- One-hot encoding of the categorical variables
- Dropped the irrelevant variables to build the model
- Did Label Encoding for ordinal variable

EXPLORATORY DATA ANALYSIS

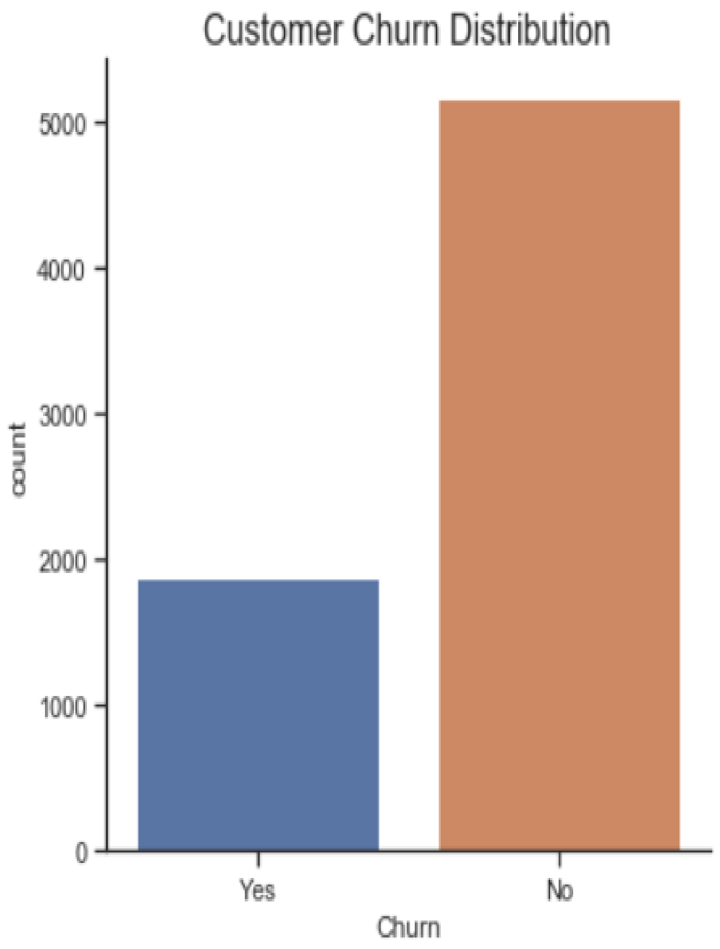


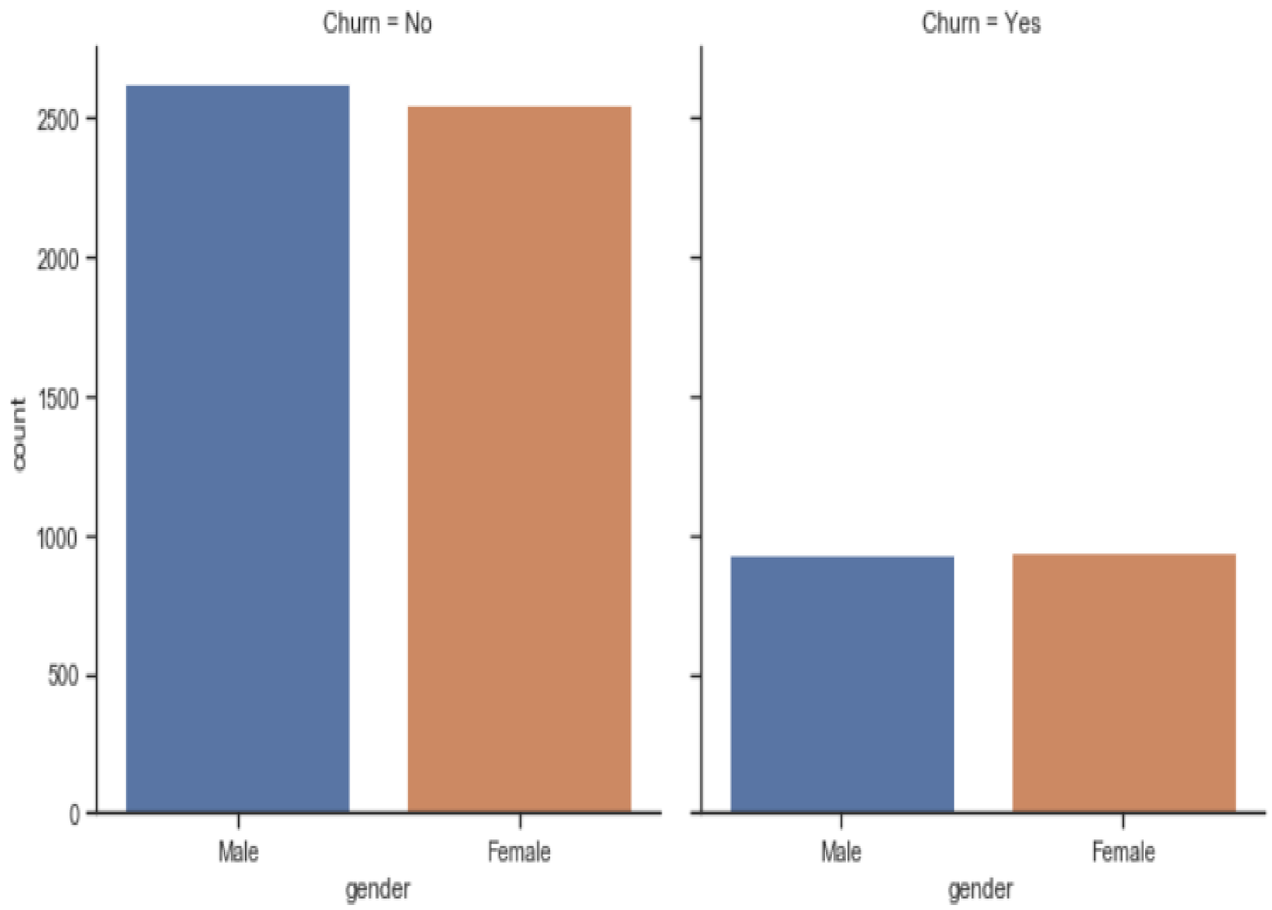
CONVERTED 'TENURE' VARIABLE INTO A CATEGORICAL COLUMN

Tenure values can be divided into 6 categories of 12 months duration in each.

No of customers are more in 2 categories: less than 12 months and greater than 60 months

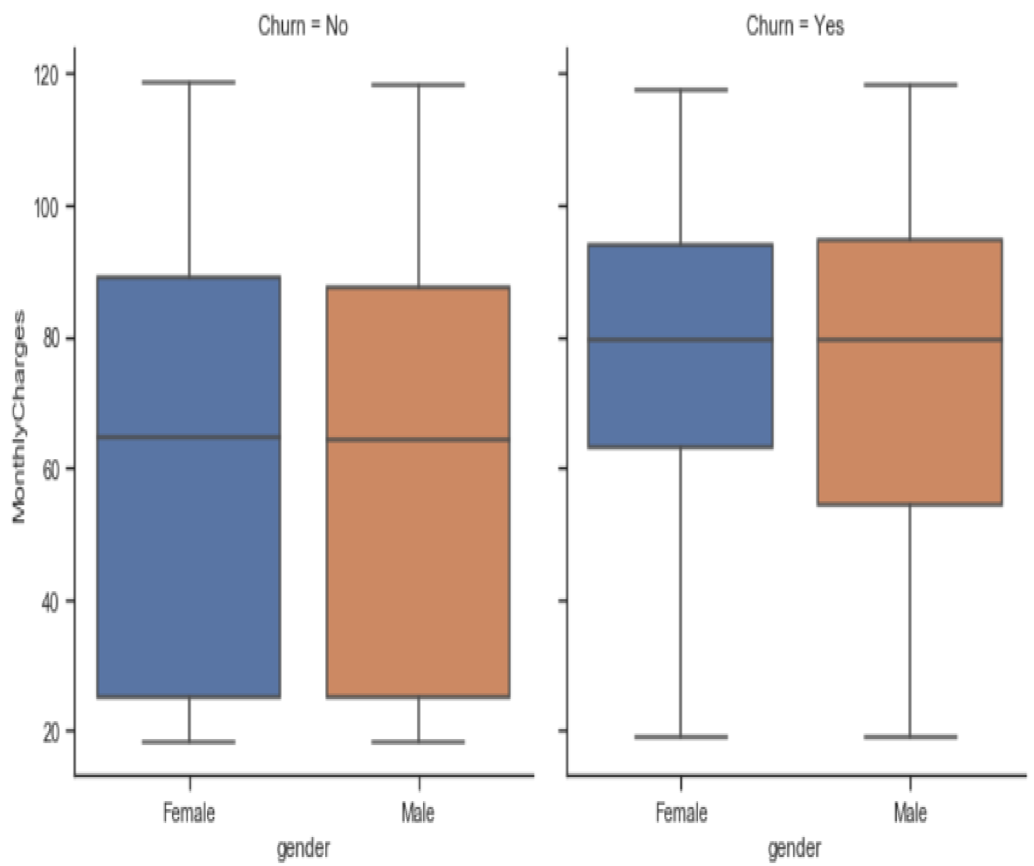
DATA VISUALIZATION

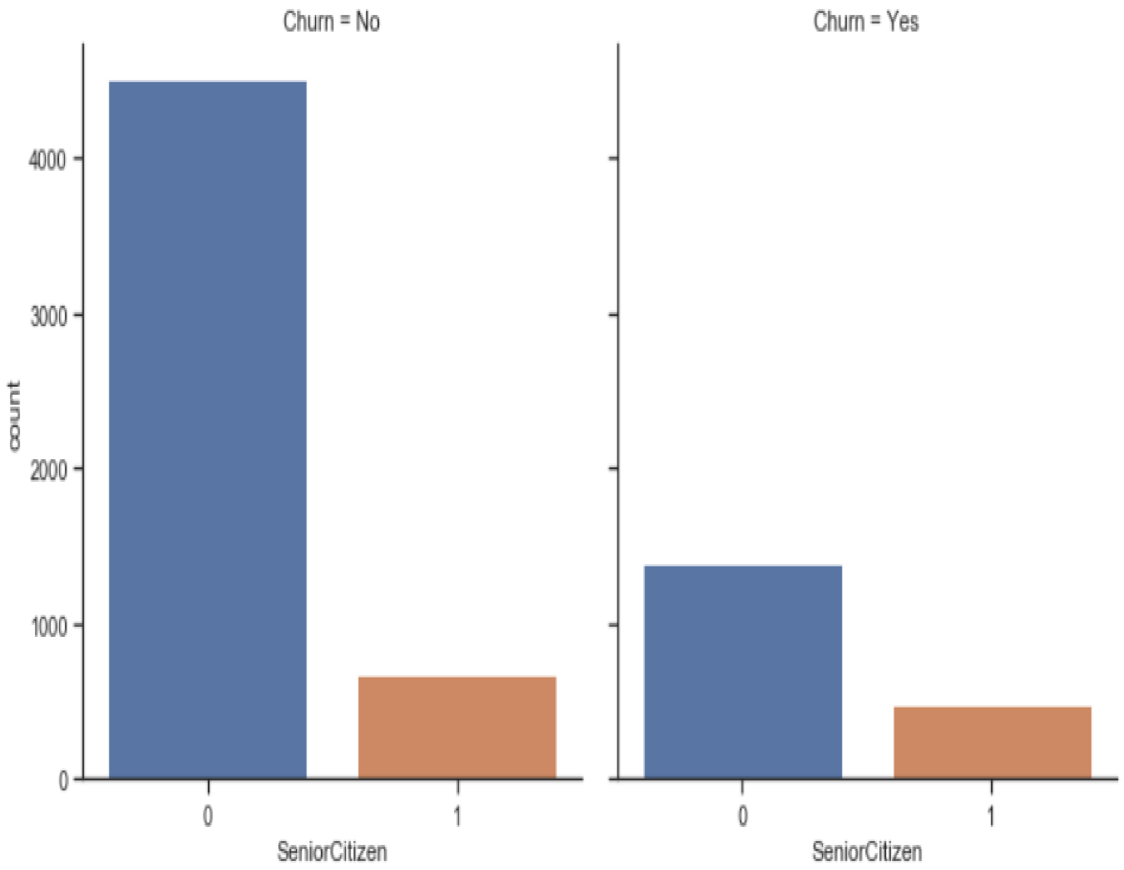




Customer churn has happened irrespective of gender.

DATA VISUALIZATION





Customers with a likely average monthly charge of 80 are likely to churn

Not much churn has happened within the Senior Citizen category.

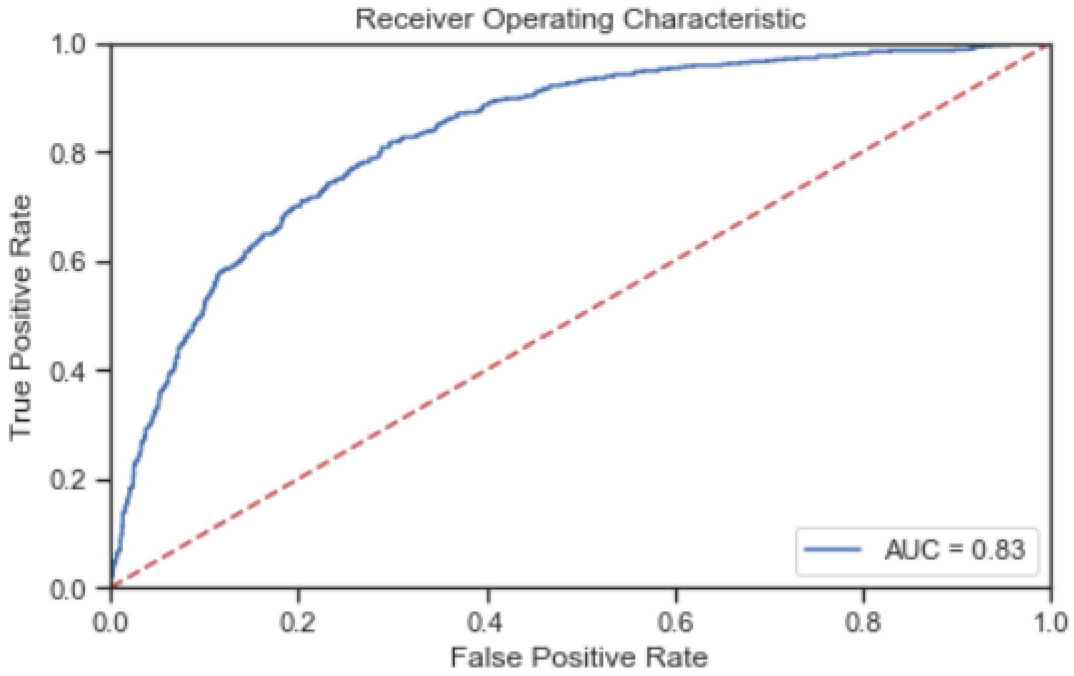
COMPARING MODEL METRICS

	Model	Accuracy_score	Precision	Recall	F1_score	Area_under_curve
0	Decision Tree using GridSearch	0.772835	0.649457	0.405085	0.498956	0.660192
1	Bagging Classifier	0.789872	0.701657	0.430508	0.533613	0.679798
2	Random Forest	0.787979	0.671498	0.471186	0.553785	0.690945
3	Gradient Boost Ensemble	0.775201	0.698962	0.342373	0.459613	0.642624
4	Ada Boost Ensemble	0.778514	0.639269	0.474576	0.544747	0.685417
5	Logistic Regression	0.737813	0.519650	0.806779	0.632138	0.758938

Logistics Regression performs well with high Recall score of 0.80 and F1 score of 0.63 compared to the other models. The AUC score is also the highest under Logistic Regression

CONFUSION MATRIX & ROC CURVE FOR LOGISTIC REGRESSION

	Predicted Not Churn	Predicted Churn
Actual Not Churn	1083	440
Actual Churn	114	476



THANK YOU