**Customer Churn Prediction (Telco)**

**About the Dataset**

The Customer Churn Prediction (Telco) is taken from Kaggle (<https://www.kaggle.com/blastchar/telco-customer-churn>).

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

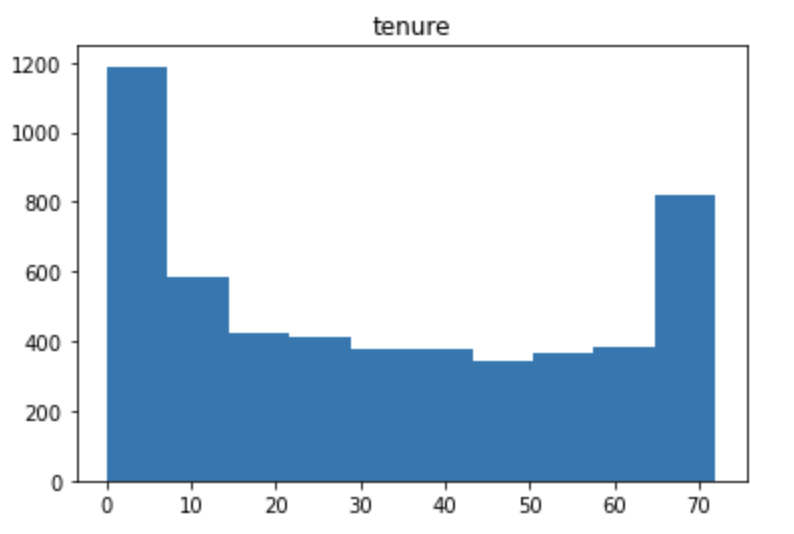
**1.Data Exploration**

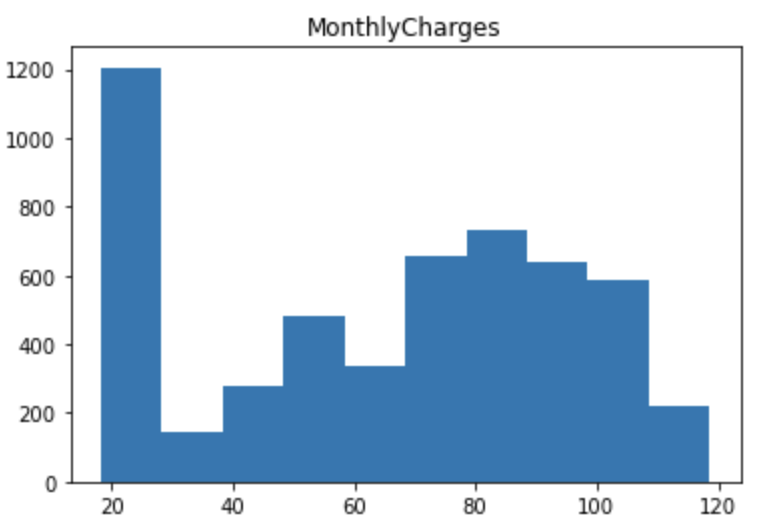
The dataset contains customer features of Telco customer like how long the customer is staying with company (tenure), is the customer is senior or not, whether the customer have streaming movie service, how billing has been made by the customer etc. and it contains column whether the customer churn or not.

**Numerical Features**

This dataset contains few continuous features like Tenure, Monthly Charges, Total Charges. The distribution is bimodal which tell us that most of the customer stayed with company for (0-10) months or around 70 months.

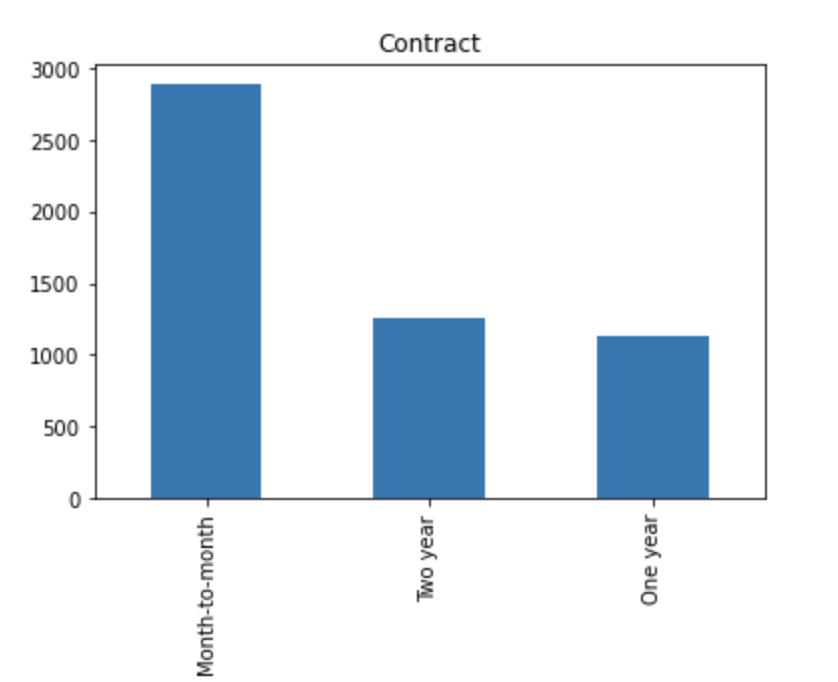
Monthly Charges distribution shows that most of customer been charged around (20 - 30) dollars and second most set of customers are charged around 90 dollars.

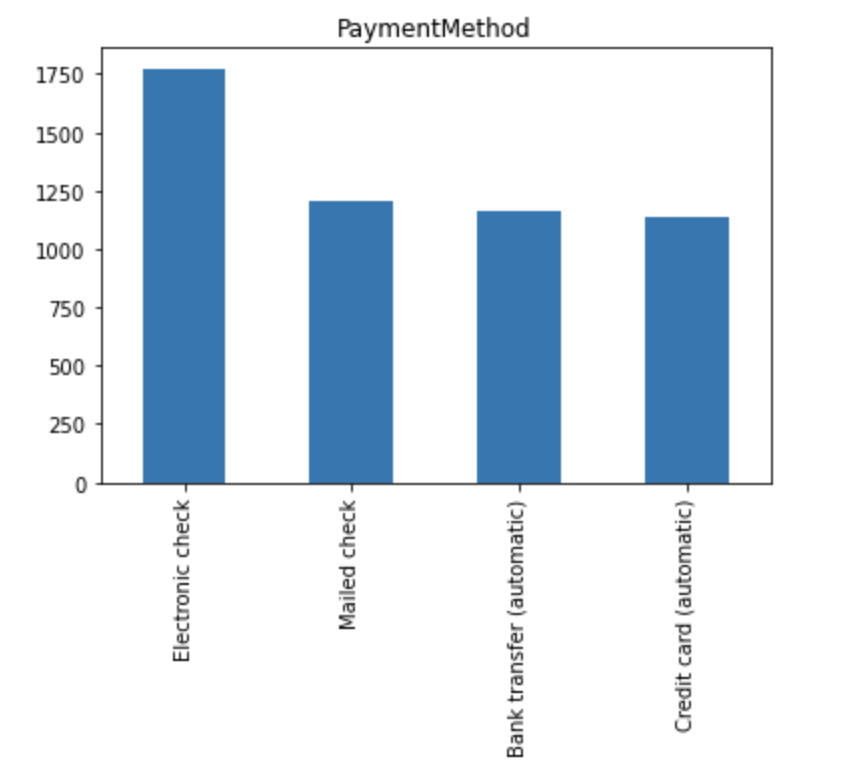




**Categorical Features**

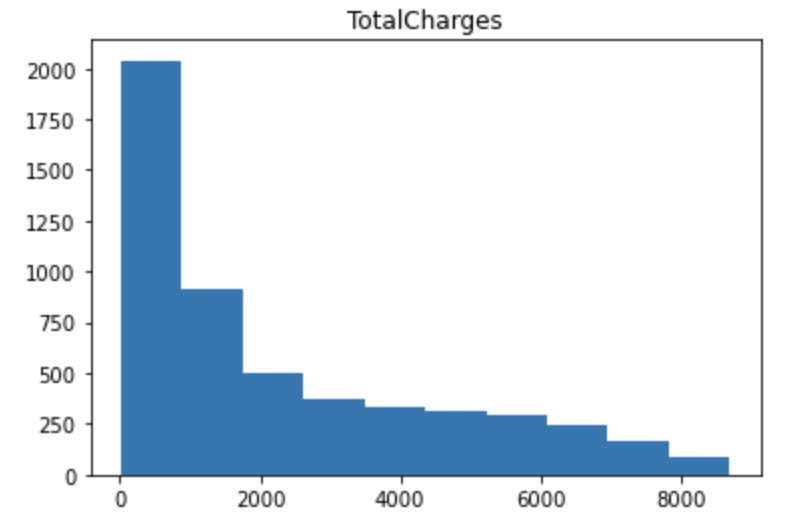
There are lot of interesting patterns to see with categorical features, here few interesting variables are Contract where most customer are with month to month contract and customer with one year or two year contract are almost same.





**2. Handling NULL/Missing Values**

This dataset contains some missing values in Total Charges column, but the distribution is skewed, so instead of choosing mean or median value I choose to fill those values with value 20.



**3. Dealing with Multicollinearity**

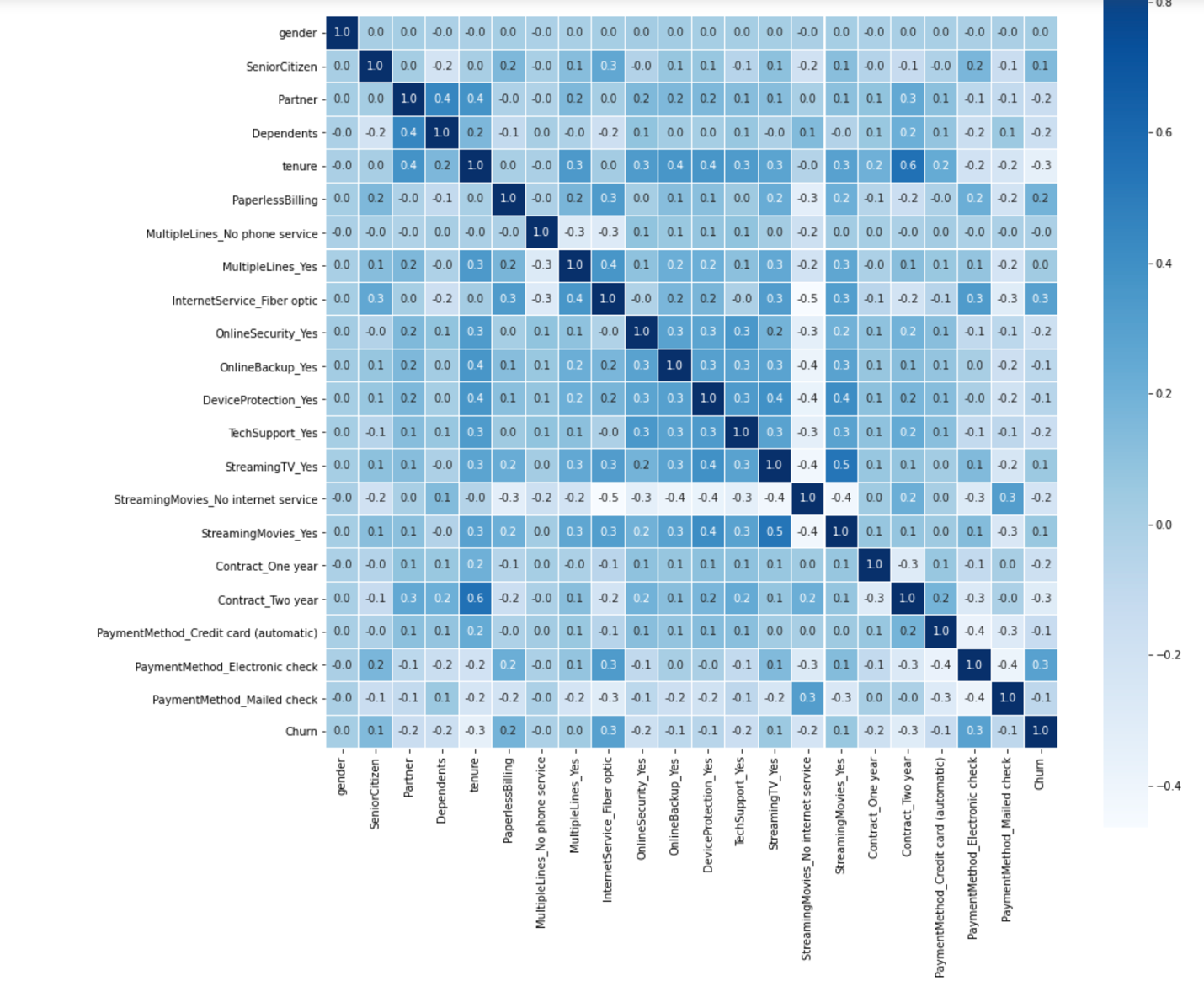
Below are the columns that are dropped because it’s VIF score is greater than 10, to avoid multicollinearity problem in the dataset.

Dropping Columns with VIF value greater than 10.

* InternetService\_No
* OnlineSecurity\_No internet service
* OnlineBackup\_No internet service
* DeviceProtection\_No internet service
* TechSupport\_No internet service
* StreamingTV\_No internet service
* PhoneService
* MonthlyCharges
* TotalCharges

**Correlation Heat Map:**

The correlation heat map shows there is no highly correlated variables except relation between tenure and contract two years.

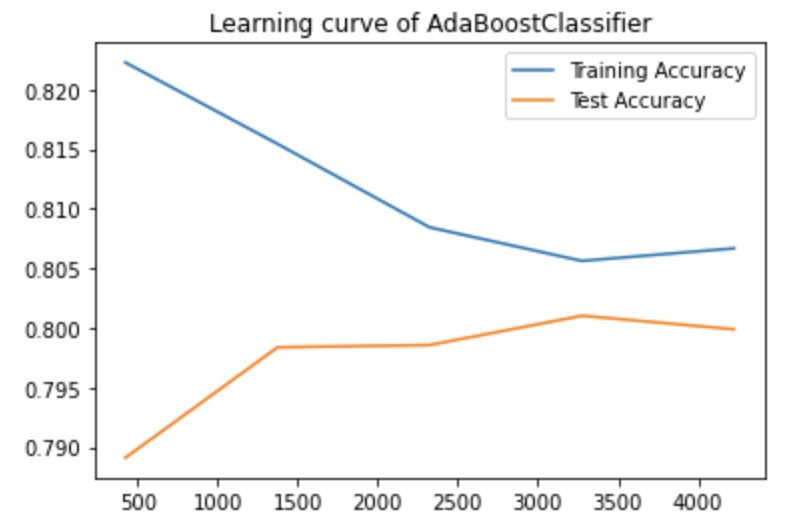


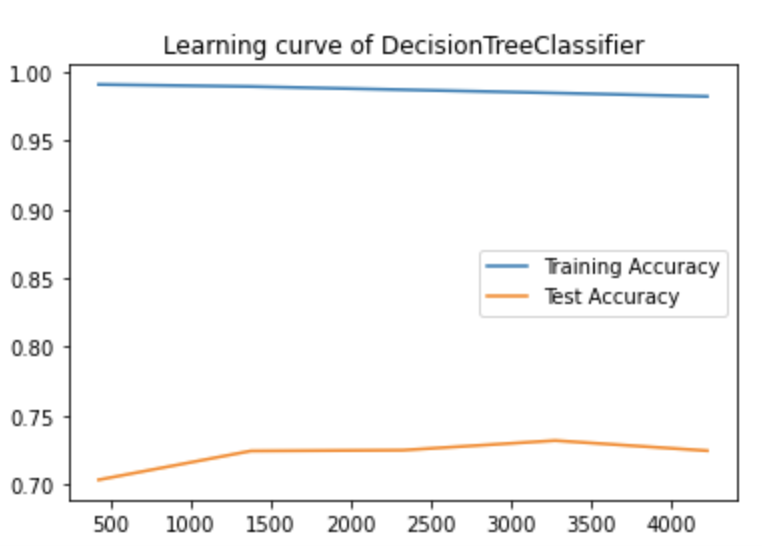
**4.Model Fitting**

Below are the models that is tried to fit the dataset

* Logistic Regression
* Decision Tree Classifier
* AdaBoostingClassifier
* Gradient Boosting Classifier
* Random Forest Classifier

After fitting models below are the learning curve for few models

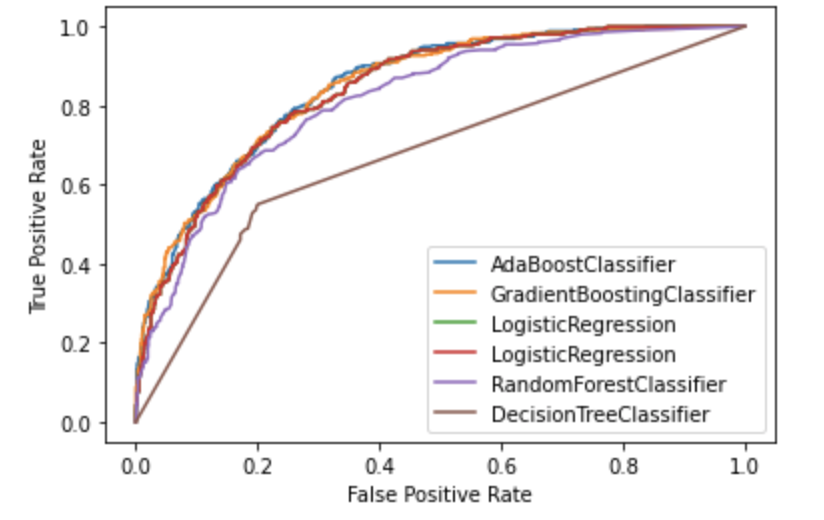




It is very clear from the graph that AdaBoostClassifier model done a good job compare to Decision tree classifier, where AdaBoostClassifier test accuracy increases as data increases but Decision Tree Classifier over fit the dataset.

**ROC Curve:**

Below are the ROC AUC score where AdaBoostClassifier did better job compare to other models.

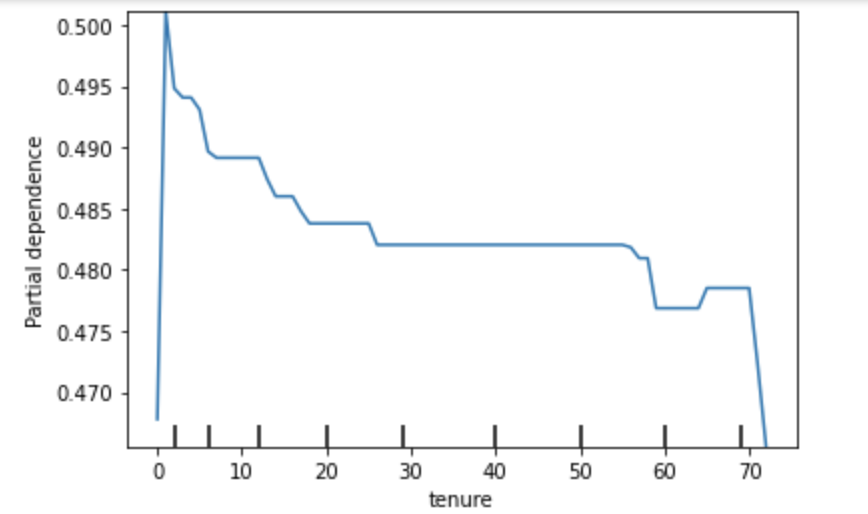


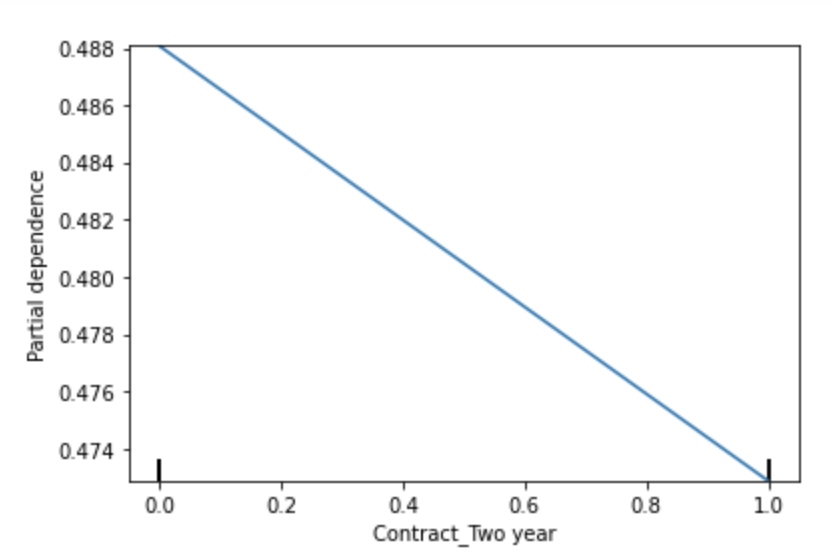
* AdaBoostClassifier - 0.844914
* GradientBoostingClassifier - 0.844507
* LogisticRegression - 0.838888
* RandomForestClassifier - 0.810109
* DecisionTreeClassifier - 0.660533

Below are few house features that choose by most of ML models.

* Tenure
* MultipleLines\_Yes (pos)
* StreamingMovies\_Yes (pos)
* Contract\_One year (neg)
* OnlineSecurity\_Yes (neg)

**Model Interpretation:**





Customer tenure seemed to impact churn rate quite a bit -- this was potentially the most

important feature. The longer a customer has been with us, the less likely they are to churn.

Additionally, these customer attributes were correlated with churn:

* Having multiple lines
* Paying with an electronic check
* Having no phone service
* Having streaming TV
* Having paperless billing
* Having fiber optic internet service

And, these customer attributes were negatively correlated with churn (in other words, correlated with customer retention)

* Having online security
* Having no internet service (streaming movies)
* Having a one-year or two-year contract

**Evaluation of Model on Test data:**

* Train Data Set 🡪 0.846463
* Test Data Set 🡪 0.861353

We evaluated our model on unseen test data and got a ROC AUC score of 0.86.

This leads us to believe that our model will generalize well to future data.

**Conclusions**

To summarize, this project shows the strength of linear model like Logistic Regression, AdaBoostClassifier model. It also shows how one-hot encoding, ordinal encoding, dropping high VIF variables, and parameter tuning improved our models. Although different model select differ feature for its predictions, if we check all model certain variables were chosen by most of the model.

Comparing training and testing ROC-AUC score, the final model did good job in Test data set then it did on Train data set which shows that final model did better learning in Train data set.