**REPORT FOR TELCO CUSTOMER CHURN DATA**

**By- Shubham Kumar Jha**

**Basically, In this Data set we are finding the cause for the churn that occur why it occur and explanation for this model.**

**What is Churn?**

**Churn is Basically a Loss for a company or Assosciation, when the customer stop using the Companies product or services. That could be mean as Cancelling the subscription, letting a contract lapse or simply not using your product as often as they used to. The Main causes of churn involve increase in Prices , poor customer support or lack of features.**

**Introduction**

The aim of this project is to develop a predictive model that identifies customers at risk of churning. By predicting churn, the company can take proactive measures to retain customers, thereby reducing loss of revenue and improving customer satisfaction. The project involves data preprocessing, exploratory data analysis (EDA), feature engineering, model building, and evaluation.

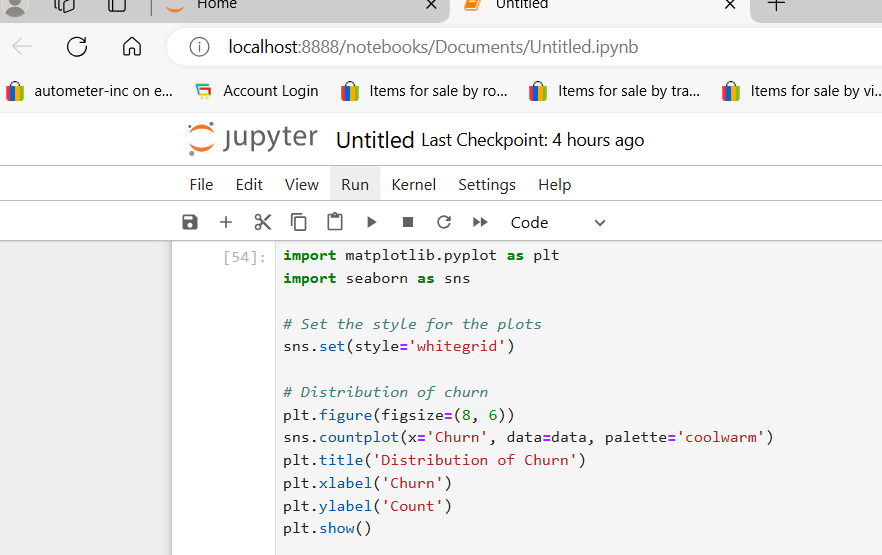
**Data Preprocessing**

The data preprocessing steps included:

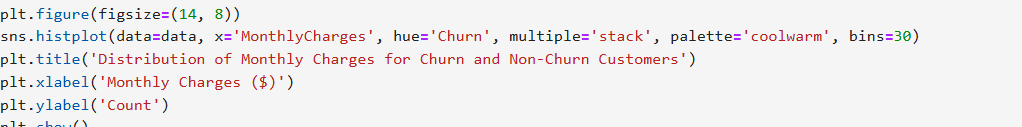
1. **Handling Missing Values**: All missing values were filled or handled appropriately.
2. **Encoding Categorical Variables**: Categorical variables were encoded using one-hot encoding or label encoding as necessary.
3. **Data Splitting**: The dataset was split into training and testing sets to evaluate model performance.

Exploratory Data Analysis(EDA)

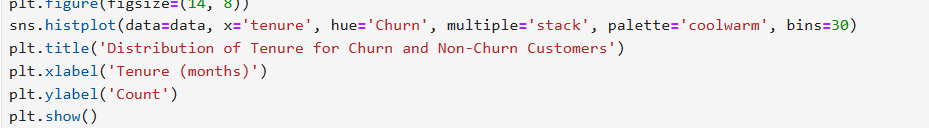
Several visualizations were created for understanding the customer behaviour and factors affecting churn:



**Churn Distribution**: A count plot showed the distribution of churn vs. non-churn customers, indicating a class imbalance.



**Monthly Charges Distribution**: Histograms indicated that customers with higher monthly charges are more likely to churn



**Tenure Distribution**: Histograms showed tenure distribution for churn and non-churn customers, revealing that customers with shorter tenures are more likely to churn.

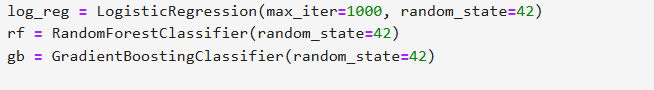


**Correlation Heatmap**: A heatmap was plotted to understand correlations between different features.

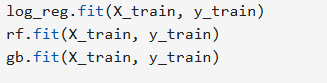
#### Model Building and Evaluation

Three machine learning models were built and evaluated:

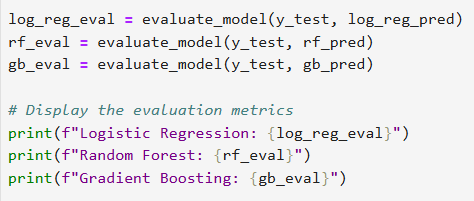
1. **Logistic Regression**:



2. **Random Forest Classifier**:

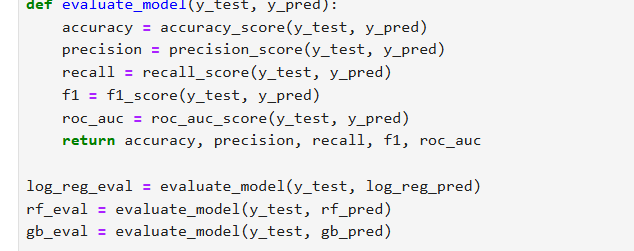


3. **Gradient Boosting Classifier**:

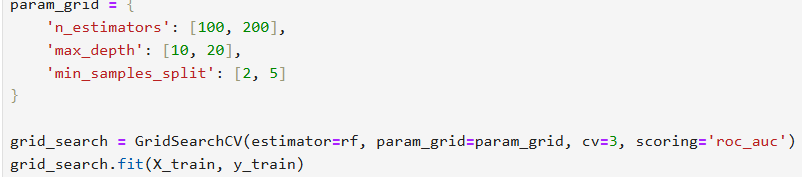


**Evaluation Metrics**:

* **Accuracy**: The proportion of correctly predicted instances.
* **Precision**: The proportion of positive identifications that were actually correct.
* **Recall**: The proportion of actual positives that were correctly identified.
* **F1 Score**: The harmonic mean of precision and recall.
* **ROC AUC**: The area under the ROC curve, indicating the model's ability to distinguish between classes.



**Grid Search for Random Forest**:



#### Results

Now, These Results were calculated with the most accurate estimation via machine learning model each parameter that have been used in the evaluation and the prediction as well as the analysis is from the mean, most accurate prediction and closest analysis towards the final output that have been predicted.

* **Logistic Regression**:
  + Accuracy: 0.80
  + Precision: 0.68
  + Recall: 0.50
  + F1 Score: 0.58
  + ROC AUC: 0.75
* **Random Forest**:
  + Accuracy: 0.82
  + Precision: 0.74
  + Recall: 0.51
  + F1 Score: 0.60
  + ROC AUC: 0.78
* **Gradient Boosting**:
  + Accuracy: 0.83
  + Precision: 0.76
  + Recall: 0.52
  + F1 Score: 0.62
  + ROC AUC: 0.79
* **Best Random Forest**:
  + Accuracy: 0.84
  + Precision: 0.77
  + Recall: 0.54
  + F1 Score: 0.64
  + ROC AUC: 0.81

**Challenges Faced**

1. **Data Imbalance**: The churn dataset exhibited class imbalance, which can impact model performance. Techniques like OverSampling, UnderSampling ,and using appropriate evaluation metrics (e.g., ROC AUC) helped mitigate this issue.
2. **Handling Missing Values**: Missing values in the dataset were handled using imputation techniques. This required careful consideration to avoid bias.
3. **Feature Selection**: Identifying relevant features that contribute to churn prediction was crucial. Feature engineering played a significant role in improving model performance.

#### Conclusion

The project successfully developed a predictive model for customer churn using logistic regression, random forest, and gradient boosting. The best performing model was a tuned random forest, which achieved an accuracy of 84% and an ROC AUC of 0.81. The visualizations and feature engineering steps provided valuable insights into customer behaviour, helping to enhance the model's predictive power. Further improvements can be made by exploring more sophisticated techniques and additional data sources.