Detecting Customer Churn

Prepared By: Erum Akmal

1. **Problem Statement for Project:**
   In this project, I have predicted the customer churn based on customer age, subscription length, monthly bill, total usage, support calls, contract type, payment method and additional services. Logarithmic and logistic regression has been used for Multivariate Dataset.

2. **Data Sources and Description:**
   For data sources I used the data set for customer churn list shared by the instructor for the assignment.

   For this project, I worked on a customer churn project with logarithmic and logistic regression for multivariate data set. My data set includes customer churn (dependent variable) and customer age, subscription length, monthly bill, total usage, support calls, contract type, payment method and additional services as independent variables.

3. **Data Mining Operations:**
   -Data wrangling:
   After loading and describing the data set, I proceeded with preprocessing of data. Handling of data was not required in this project since there was no missing data. I converted the Converting categories into numbers using Label Encoder. The correlation matrix was calculated next.

   After that data was normalized using min max scaler. In the next step, splitting of data was done. Data was partitioned into 80% training and 20% testing.

   Logistic regression was performed, confusion matrix and accuracy were calculated. In the next step, model evaluations were done; followed by the generation of graphs.

   -Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

   -Data Modeling: I have performed the analysis below in my project:

   - Preprocessing of data
   - Correlation metrics
   - Min Max Scaler
   - Splitting dataset into 80-20
   - Modelling with Logistic and logarithmic regression
   - Evaluation: MSE, MAE, r2_score
   - Scatter plot graph for both logarithmic and logistic
   - Correlation heatmap
   - Confusion matrix for both logarithmic and logistic
   - Histogram of diff or target and prediction
   - Box plot for dataset
   - ROC Curve
   - Sigmoid Curve

4. **Model Outputs:**
   Below is my evaluation of the model performance for logarithmic:

   - mean squared error: 0.1646
   - mean absolute error: 0.3276
   - r2score: -0.0069

   Below is my evaluation of the model performance for logistic:

   - mean squared error: 0.206
   - mean absolute error: 0.206
   - r2score: -0.2594

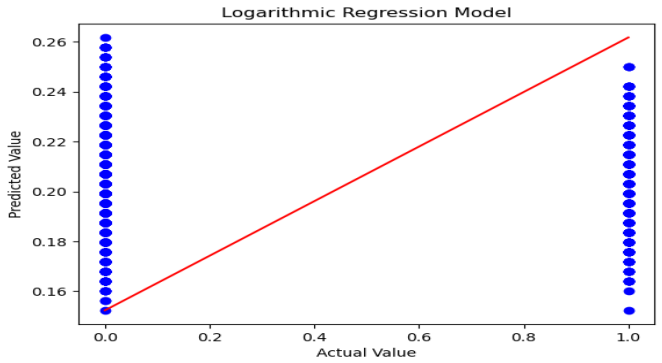
After analyzing the logarithmic model, it can be observed that the squared difference between predicted and actual values is 0.164. The average error in the prediction here is 0.327. By looking at

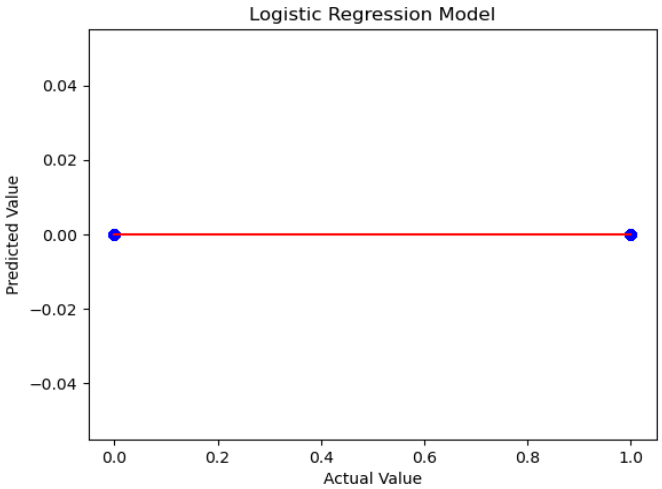
the R^2 value (-0.0069) of this model, it is observed that there is no variance in the target variable.

After analyzing the logistic regression model, it can be observed that the squared difference between predicted and actual values is 0.206. The average error in the prediction here is 0.206. By looking at the R^2 value (-0.259) of this model, it is observed that there is a very poor fit and negative R^2 implies that model performance was not good here, even though the accuracy is good in confusion matrix.

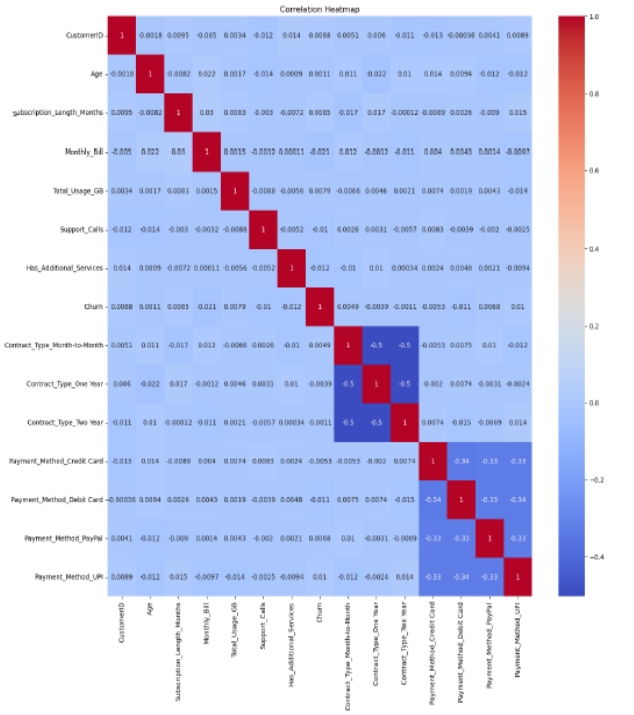
```
Confusion Matrix (logarithmic):
 [[1588    0]
 [ 412    0]]
Accuracy (logarithmic):  0.794
Confusion Matrix (logistic):
 [[1588    0]
 [ 412    0]]
Accuracy (logistic):  0.794
```

As a next step, the scatter plot was generated with predicted value on y-axis and actual value of x-axis for both logarithmic and logistic regression model.

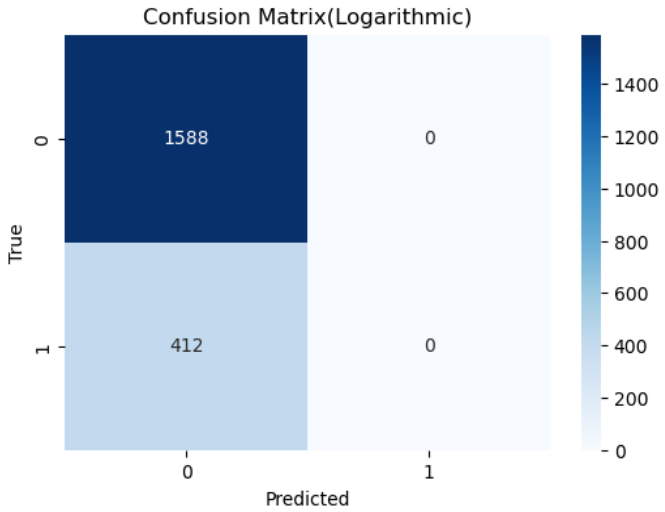




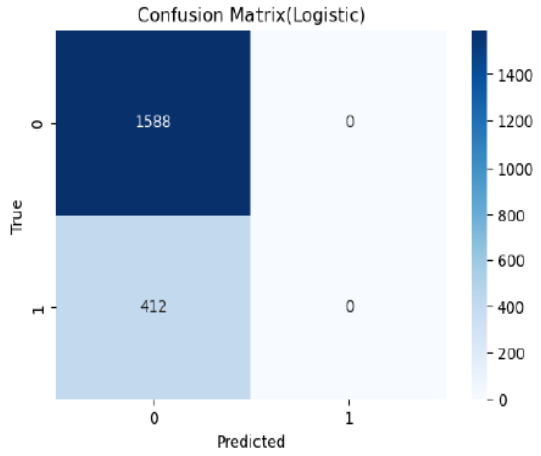
The correlation heatmap was generated next for customer churn and all other independent variables. Among all variables, the payment method UPI (0.010) is affecting the customer churn the most and monthly bill (-0.02) has the least amount of effect on customer churn.



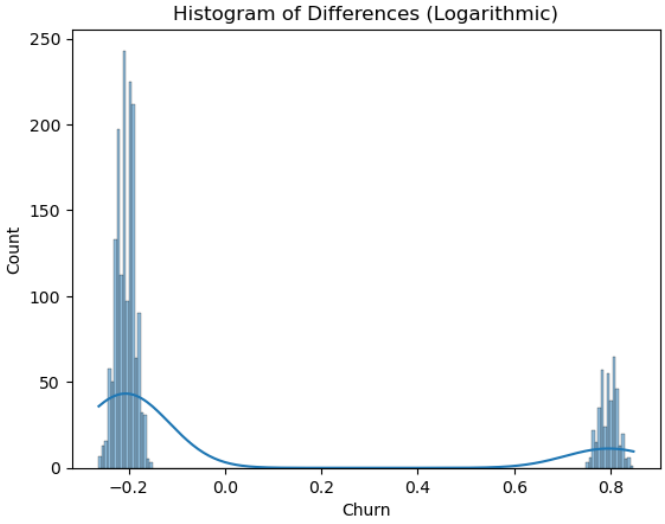
The confusion matrix for logarithmic regression was generated next.

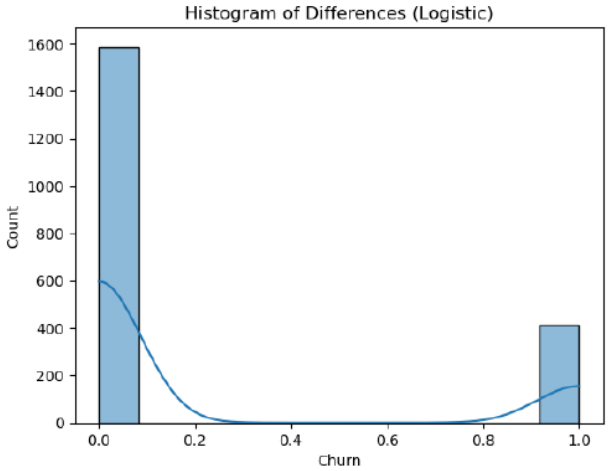


After that, the Confusion matrix for logistic regression was plotted.

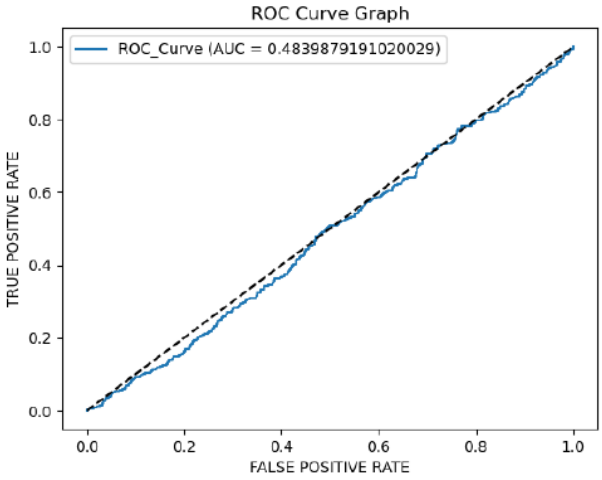


The histogram of differences was generated for both logarithmic and logistic regression.

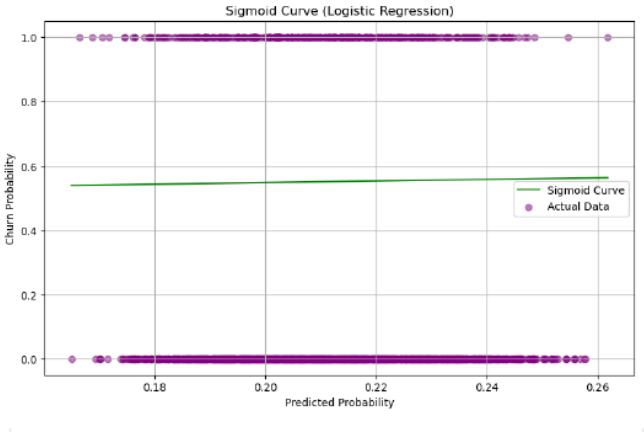




ROC Curve for logistic regression was generated next.



At the end, the sigmoid curve for logistic regression was generated.



5. Limitations and Challenges:

Some of the limitations of this study are that I have only used 8 variables in this study for predicting the customer churn. There could be multiple other factors that can potentially affect customer churn. This model worked well but there might be multiple other reasons that can affect customer churn which were not covered in the data set.

Both the models used in this project have not sufficiently predicted the customer churn with accuracy. More advanced machine learning models like Random Forest could have been used to increase the predictive power of the model.

6. Conclusion:

I was able to solve the problem partially since accuracy was good but R^2 score suggested poor performance. By adding more data and variables in this study, a better model could have been predicted for customer churn. By adding a greater number of features and rows, we can test multiple factors that can play a role in improving customer churn.

In this project, 2 predictive models have been used to detect customer churn. logarithmic regression model was slightly better than logistic regression model.

Based on the analysis above, we can conclude that logarithmic regression model indicated a weak predictive capability.

The logistical regression model was also not effective here since R^2 score is -0.2594, suggesting poor performance.