FRAUD DETECTION

USING LOGISTIC REGRESSION

**TABLE OF CONTENTS**

[**ABOUT LOGISTIC REGRESSION**](#_nofq3q26n0uc) **3**

[**ASSUMPTIONS OF BINARY LOGISTIC REGRESSION**](#_3mi5qovumhyv) **4**

[**GENERAL EXPLORATORY DATA ANALYSIS**](#_8s56z5v2r1n9) **5**

[Inspect the Data Structure](#_i5m78ksfs3kd) 5

[Analyse Data Distribution and Skewness](#_n8sfza5un719) 6

[Detect Outliers](#_dnqjutt0u7q9) 7

[**EXPLORATORY DATA ANALYSIS FOR REGRESSION**](#_y30p8hvdzo4) **7**

[Assess Target Variable Class Imbalance](#_k1g2sse7n4qi) 7

[Correlation Matrix](#_k9m8ir6nkhbn) 8

[Detect Multicollinearity](#_o09ln0q9xnfc) 8

[Analyze Independent Relationships](#_cdspdns95wui) 9

[Assess the Independence of Observations](#_ishtnayf5p12) 9

[**DATA TRANSFORMATION**](#_3ejfkfcubxk8) **10**

[Creating Dummy Variables](#_2ktfqgjos2ty) 10

[Normalizing the Data](#_kgf12omri5g) 10

[Random Over-Sampling using SMOTE](#_hltj7wwsta2x) 11

[**MODEL IMPLEMENTATION**](#_2wvkxe7u67j5) **11**

[Logistic Regression](#_9xu5vbhmpbt4) 11

[AIC Score](#_qfsd5o9ex145) 11

[Evaluating Classification Performance](#_q6qc5mkqc62j) 12

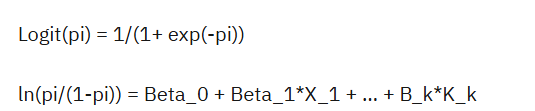
[**CONCLUSION**](#_f36jfbprxvs9) **13**

[**REFERENCES**](#_fozdrviop768) **14**

# 

# ABOUT LOGISTIC REGRESSION

“Logistic regression is a supervised machine learning algorithm used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not” (Geeks for Geeks,2024). The model is also known as the logit model where the logit transformation is applied to the log odds - the probability of success divided by the probability of failure (IBM, n.a). The following formula represents the logistic regression function:



In this equation, logit(pi) is the target variable and x is the independent variable. The beta parameter (coefficient) is estimated using the maximum likelihood estimation. This method tests different values of beta through multiple iterations to find the optimal coefficient for the best fit of log odds. Given the optimal coefficient/coefficients, the conditional probabilities are calculated, logged, and summed together to get a predicted probability. In binary classification, if the probability is less than 0.5 the prediction will be 0 and a probability greater than 0 will yield a prediction of 1 (IBM,n.a).

There are three different types of logistic regression models which are defined based on their target variables. We have binary, multinomial, and ordinal logistic regression. In this project, i conducted a binary logistic regression on our dataset given that our target variable is dichotomous in nature.

# ASSUMPTIONS OF BINARY LOGISTIC REGRESSION

To implement the logistic regression model, it is critical to ensure that certain assumptions about the data are met to get the model’s optimal performance. These are:

1. *The response variable is binary*

The model assumes that the target variable is dichotomous in nature. For example: ‘yes or no’, ‘0 or 1’ or ‘pass or fail’. Therefore, the model can only take two values.

1. *Independence of observations*

The model assumes that the predictor variables are independent, which means there should be no correlation between these variables. Therefore, the value of one observation should not give any information about the value of another.

1. *No multicollinearity among independent variables*

The model assumes that there is no severe multicollinearity among the independent variables. Multicollinearity means when two variables are highly correlated, typically with a correlation coefficient greater than 0.8, that they cannot provide independent information in the regression model.

1. *No extreme outliers*

The presence of outliers or influential observations will affect the model’s performance.

1. *There exists a linear relationship between independent variables and the logit of the target variable*

The model assumes that there is a linear relationship between the independent variables and the log odds of the target variable.

1. *The dataset is sufficiently large*

The dataset should be large enough to make valid conclusions (Bobbitt, 2020).

# GENERAL EXPLORATORY DATA ANALYSIS

## Inspect the Data Structure

To get a better understanding of the data and its features i inspected the structure of the data using the *.info()* features to check data types, column names, and target null values. I noted that the dataset did not consist of any null values and had four (4) categorical columns and three (3) numerical columns. Given this insight, there is no need for mechanisms to handle missing values. The column names were changed, using the *.str.replace()*, function to include an underscore ‘\_’ to replace each space for ease of callability. Additionally, a check was done for duplicated rows which involved using the *.duplicated()* function and it was noted that there were no duplicated rows in the dataset. Also, using the *.unique()* function on the target variable *‘Fraudulent’* it was identified that it only consists of binary values. This satisfies our assumption that our target variable only has binary values.

## Analyse Data Distribution and Skewness

The numerical columns *‘Transaction\_Amount’*, *‘Fraudulent’*,and *‘Risk\_Score’* were isolated using a for loop, and the skewness was calculated for each. It was noted that *‘Transaction\_Amount’* had a skewness of 4.38; *‘Fraudulent’* had a skewness of 0.87; ‘Risk\_Score’ had a skewness of 0.45. Therefore, it can be deduced that columns *‘Transaction\_Amount’* and *‘Fraudulent’* were positively skewed and hence do not have a normal distribution as their skewness value is greater than 0.5. This can also be visualized in the graphs shown below (Figure 1.1 and Figure 1.2)

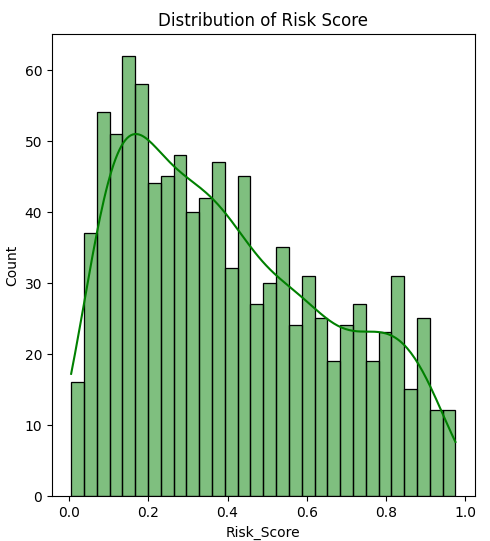
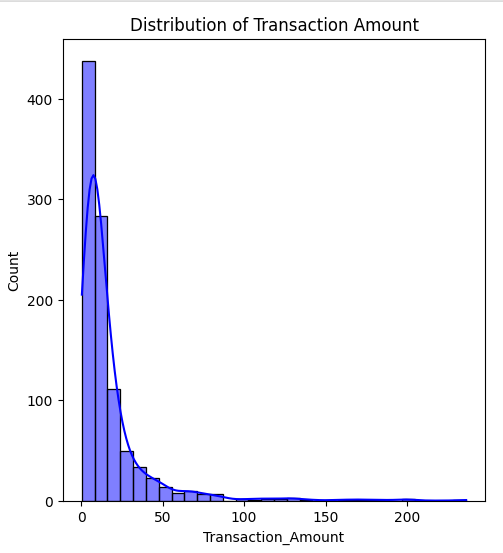


Figure 1.1 Figure1.2

Figure 1.1 and Figure 1.2 show that *‘Transaction\_Amounts’* and *‘Risk\_Score’* are clustered at lower amounts, with higher transaction amounts and risk scores being less frequent which pulls the distribution to the right causing it to be positively skewed. The solid blue and green line represents the density curve of the data distribution. This skewness can cause biased predictions in the dataset where it either underfits the minority class or overfits the majority class.

## Detect Outliers

Boxplots were used to identify and visualize outliers within the independent variables *‘Transaction\_Amount’* and *‘Risk\_Score’* using the *sns.boxplot()* function. It was noted that the column *‘Transaction\_Amount’* had outliers which would be winsorized, using *thestats.mstats.winsorize()* function, to reduce its effect on the model’s performance (Figure 1.3 and Figure 1.4).

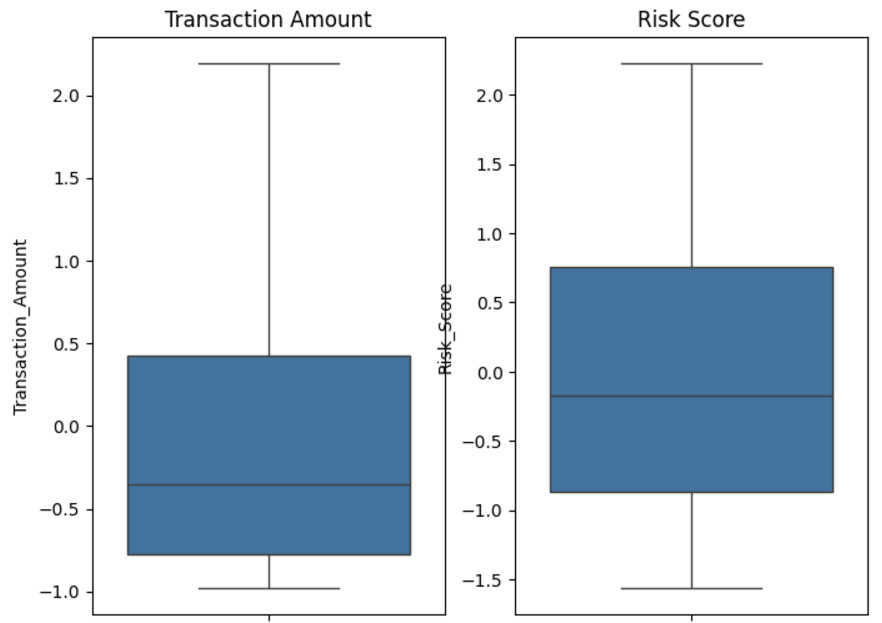
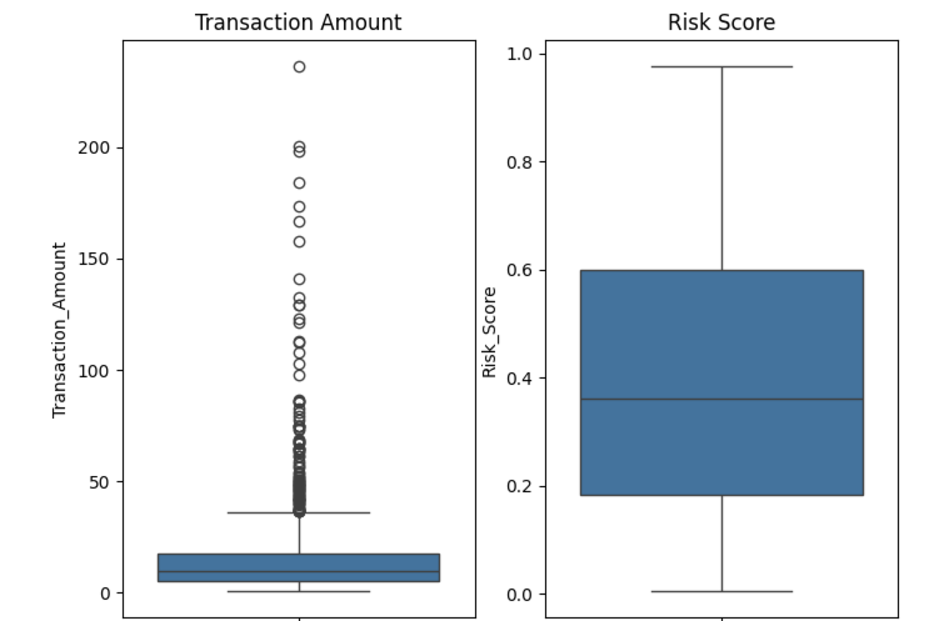


Figure 1.3 Before Winsorization Figure 1.4 After Winsorization

Winsorization reduces the effect of outliers by capping extreme values. A threshold of 0.1 was used so that only the most extreme outliers were affected and to retain the original variability of the dataset. After using this technique, the assumption of not having extreme outliers is now satisfied.

# EXPLORATORY DATA ANALYSIS FOR REGRESSION

## Assess Target Variable Class Imbalance

In cases such as fraud detection, it is imperative to ensure that the target variable *‘Fraudulent’* has balanced data among fraudulent and non-fraudulent cases (1s and 0s). This was done by using the function *.value\_counts()* to identify the count of each binary value. This resulted in ‘0’ having a count of 700 and ‘1’ a count of 300. Therefore, the data is imbalanced.

## Correlation Matrix

The heatmap visually represents the correlation coefficients between the independent variables and the target variable (*‘Fraudulent’*) which was created using the *.corr() function*. It was noted that the independent numerical variables *‘Transaction\_Amount’* and *‘Risk\_Score’* have a moderate positive correlation of 0.37 with each other which is not ideal for the logistic regression model as this violates the independence of observations assumption (Figure 1.5).

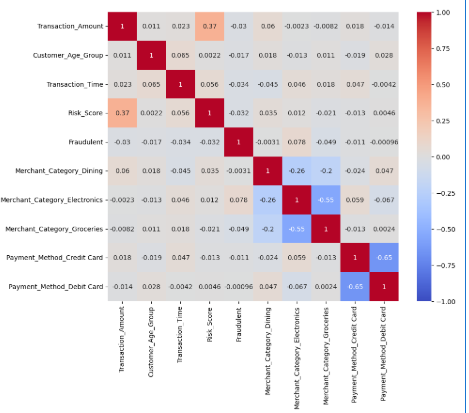
**

Figure 1.5

## Detect Multicollinearity

After creating the correlation matrix, there were no columns with a correlation greater than 0.8 which would suggest high multicollinearity. Also, a second test was done by calculating the variance inflation factor (VIF), using the *variance\_inflation\_factor()* function, for each independent variable. A VIF value greater than 5 or 10 would suggest a significant level of multicollinearity, suggesting that the variable is highly correlated with one or more variables in the model (Geeks for Geeks, 2024). I noted that the VIF values for all our predictor variables were less than 5 (Figure 1.6). This satisfies our assumption that multicollinearity does not exist among the independent variables.

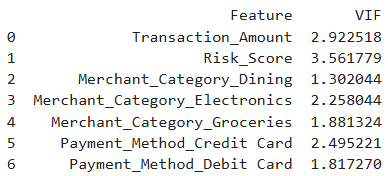


Figure 1.6

## Analyze Independent Relationships

To assess the linearity of the independent variables and the logit of the target variable *‘Fraudulent’* a regression plot was created using the *.regplot()* function (Figure 1.7 and Figure 1.8). The logistic curve struggles to fit the data points well as it deviates from the expected S shape, which suggests nonlinearity and other complexities in the data. Therefore, this violates the assumption of linearity of the independent variables to the logit of the target variable.

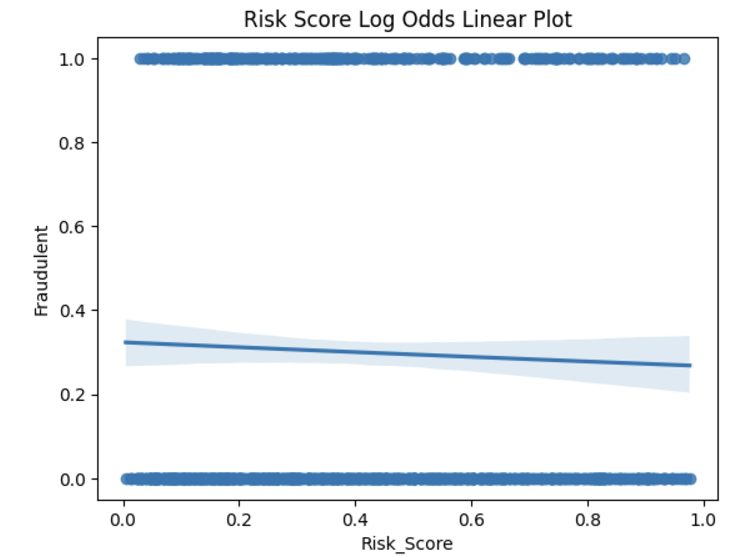
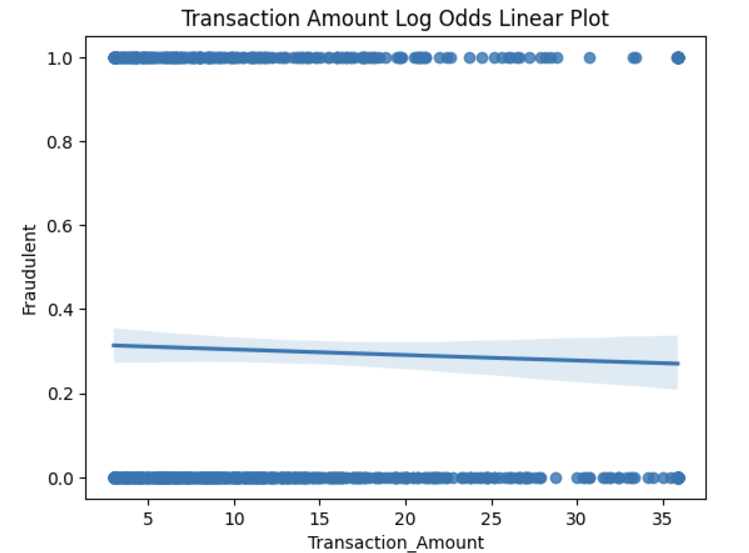


Figure 1.7 Figure 1.8

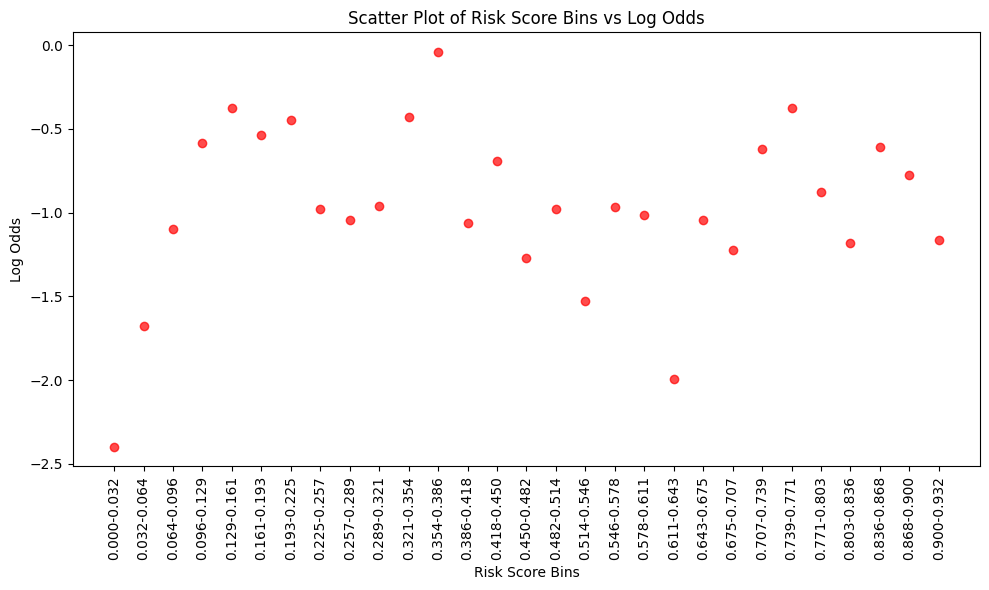
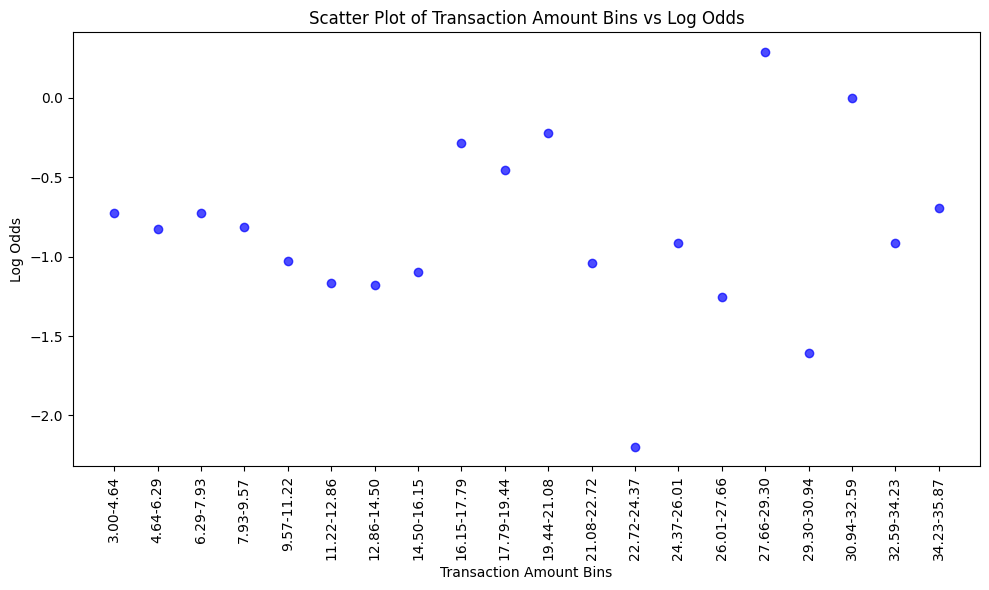
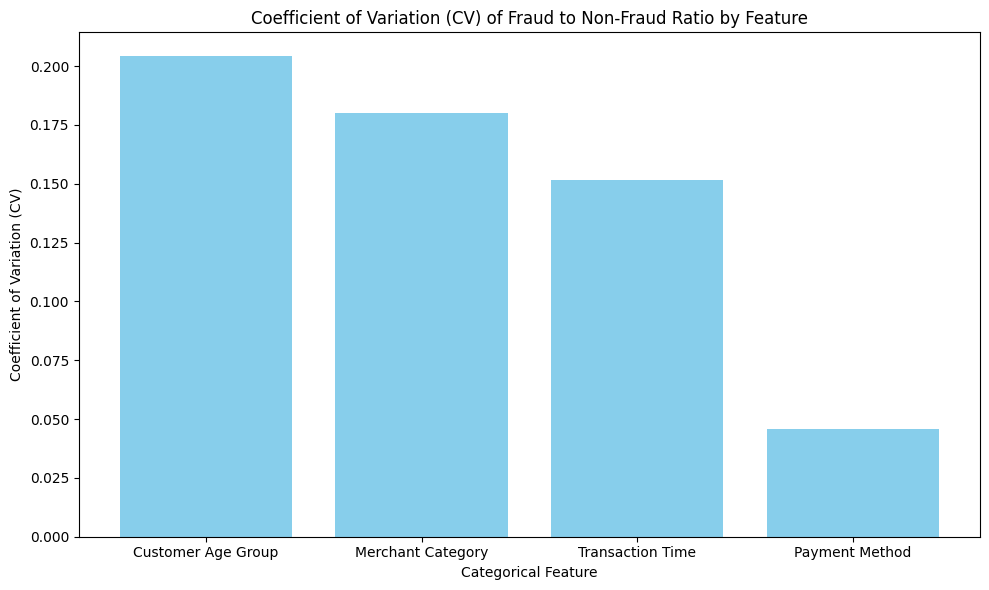


Figure 1.9 Figure 1.10

Figure 1.11

Moreover, the lack of overall correlation between numerical features and the log odds of the target is clearer (Figure 1.9 and Figure 1.10) when the plots scatter randomly.

Categorical features (Figure 1.11) with higher variability in fraud-to-non-fraud ratios, like *‘Customer\_Age\_Group’* (CV = 0.2041) and *‘Merchant\_Category’* (CV = 0.1800), are crucial for predicting fraud due to their strong differentiation potential and interaction effects. In contrast, features with lower variability, such as *‘Payment\_Method’* (CV = 0.0459), have limited standalone impact but may complement other predictors.

## Assess the Independence of Observations

To assess the independence of observations among categorical variables *(Customer\_Age\_Group, Transaction\_Time, Payment\_Method, Merchant\_Category)* a contingency table was created and a chi-square test was performed using the functions *.crosstab()* and *chi2\_contingency()* respectively. A chi-square test of independence is a statistical method used to determine whether two or more categorical variables are independent of each other. A p-value of 0.287 was obtained which is greater than 0.05, therefore, I fail to reject the null hypothesis that there is not a statistically significant relationship between the variables. Therefore, the variables are independent which satisfies the independence of observations assumption of the logistic regression model.

# DATA TRANSFORMATION

## Creating Dummy Variables

To facilitate the use of the categorical variables *(Customer\_Age\_Group, Transaction\_Time, Payment\_Method, Merchant\_Category)* dummy variables were created using the *.get\_dummies()* and *label\_encoder* function. Both label encoding and one hot encoding were used to capture the different levels in the dataset. Label encoding was used to capture the levels used in columns *‘Customer\_Age\_Group’* and *‘Transaction\_Time’* while one hot encoding was used on the columns ‘*Payment\_Method’* and *‘Merchant\_Category’* as the data contained in these columns were random and did not consist of levels.

## Normalizing the Data

The dataset consisted of numerical values that were not in scale in order to correct this the *StandardScaler()* was used to obtain values with a mean of 0 and a standard deviation of 1. The logistic regression model does not require the data to be normally distributed, however, if the data is highly skewed (columns *‘Transaction\_Amount’* and *‘Fraudulent’*) as i identified earlier this can reduce the model’s predictive performance by affecting the stability of its coefficients.

## Random Over-Sampling using SMOTE

Earlier i identified that the ‘Fraudulent’ column was imbalanced with ‘0’ having a count of 700 and ‘1’ having a count of 300. To combat this imbalance an oversampling technique called SMOTE (Synthentic Minority Over-Sampling Technique) was used. SMOTE is used to address class imbalance in classification problems. This technique was chosen because it generates synthetic samples rather than duplicating existing ones which can lead to overfitting. However, this technique should only be used on the training set in the logistic regression model otherwise SMOTE could lead to data leakage. SMOTE, implemented using the *SMOTE()* function, improves test set performance by oversampling the minority class in the training set, allowing the model to learn balanced decision boundaries. This reduces bias toward the majority class and enhances generalization to unseen data. Therefore, this resulted in the technique generating 266 synthetic samples to make the minority training class size match with the majority training class.

# MODEL IMPLEMENTATION

## Logistic Regression

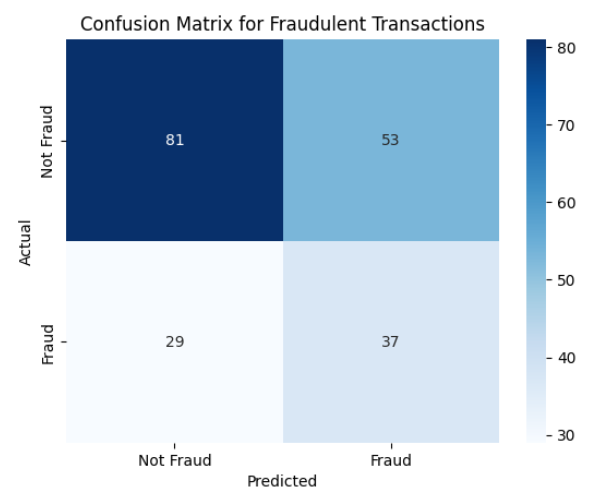
To implement the model, the dataset was first split into testing and training sets using the *train\_test\_split()* function with a test size of 20% of the original dataset and a *random\_state* of 5. The test size was selected to be 20 % of the original dataset as the original dataset only consisted of 1000 records and therefore needed sufficient data to train the model before deployment. Also, by process of elimination when the test size was >20% or <20%, the accuracy and precision of the model decreased.

## AIC Score

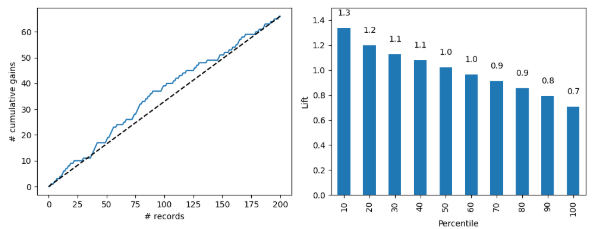
The AIC (Akaike Information Criterion) is a metric used to compare the fit of different regression models in this case the logistic regression model (Bobbitt, 2021). Given that the AIC metric is used to compare the performance of different models there is no general value to imply performance. The lower the AIC value the better the fit of the model. For our data, i determined the AIC by using the *AIC\_score()* function, which outputs an AIC of 411.26. This is considered a high AIC figure for this model which implies that the model does not fit the dataset well. Also, given this high AIC, I can predict that our precision and accuracy values will be low.

## Evaluating Classification Performance

To evaluate the model’s performance the *accuracy\_score()* and *precision\_score()* were used to assess the model’s overall accuracy and precision. This was also visualized using the confusion matrix by implementing the *confusion\_matrix()* function. The precision metric was selected to evaluate the proportion of true positive predictions (correctly identified fraud cases) out of all instances predicted as fraud. This minimizes the chances of flagging legitimate transactions as fraud. From a business perspective, minimizing these false positives would reduce or prevent arduous investigations, blocked transactions, and dissatisfied customers. The accuracy metric was used to evaluate the proportion of correct predictions out of all predictions on the test set which was 59%. Moreover, this model’s overall precision on the test set was 41%. It should also be noted that without the SMOTE technique, the precision was 0% due to the target variable being imbalanced. This can also be visualized using the confusion matrix, which shows 81 true positives and 37 true negatives (Figure 1.12).

Figure 1.12

Additionally, using the gains and decile lift chart i can also deduce that the model performed poorly which shows the model performance line being extremely close to the baseline in the gains chart and the decile lift chart lacking steepness (Figure 1.13).

**Figure 1.13

# 

# CONCLUSION

To conclude, the following assumptions of the model were not met which attributed to the model’s poor performance:

1. Independence of Observations
2. There exists a linear relationship between independent variables and the logit of the target variable

As identified earlier, using the correlation matrix there was a moderate positive correlation of 0.37 between the numerical independent variables which suggests that those observations are not independent, unlike the categorical independent variables which were proven to be independent using the chi-square test. Moreover, the logistic curve deviated from the expected S shape which violates the linear relationship between independent variables and the logit of the target variable.

# REFERENCES

GeeksforGeeks. (2024, June 20). *Logistic regression in machine learning*. GeeksforGeeks. <https://www.geeksforgeeks.org/understanding-logistic-regression/>

IBM. (n.d.). *What is logistic regression?* IBM. Retrieved December 9, 2024, from<https://www.ibm.com/topics/logistic-regression>

Bobbitt, Z. (2020, October 13). *The 6 Assumptions of Logistic Regression (With Examples)*. Statology. <https://www.statology.org/assumptions-of-logistic-regression/>

Bobbitt, Z. (2021, May 20). What is Considered a Good AIC Value? Statology. https://www.statology.org/what-is-a-good-aic-value/