Title: Fraud Detection and Clustering Analysis

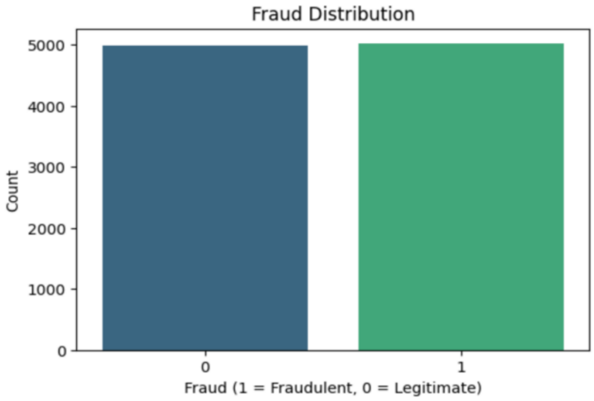
Author: Anu Bhukya
Student Number: 24004718
GitHub Repository: https://github.com/anu46464/Clustering-and-Fitting

Introduction

This report deals with the analysis of credit card fraud detection and unsupervised clustering. Characteristics included in the dataset relate to profession, income, CC Number, expiration date, and security code; a binary description for fraud was also provided in this context. This effort tries to apply clustering and regression to knowledge extraction about fraudulent activities using the given information. In this report are visualizations through K-Means clustering and Linear Regression fitted data.

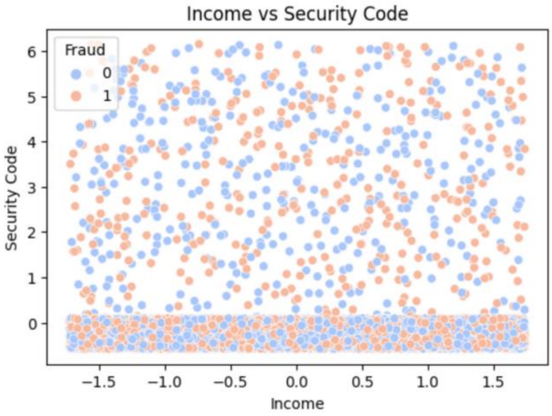
Distribution of Fraud Bar Chart

The plot shows that the data set is imbalanced and that there are more cases of legitimate transactions than fraudulent ones. Class Fraud = 0 (legitimate transaction) dominates the dataset. This sort of imbalance is quite normal in fraud detection problems and perhaps needs techniques such as over-sampling or anomaly detection to enhance model performance.



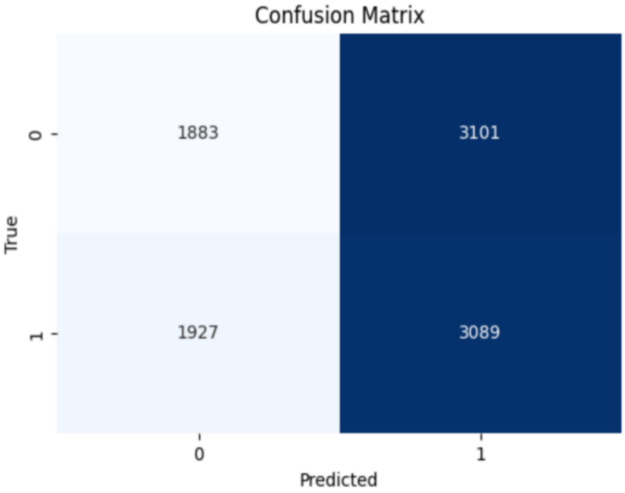
Income vs Security (Scatter Plot)

From the scatter plot, we see that there is a loose correlation between Income and Security Code, but no apparent linearity. On the other hand, fraud points are scattered all over the range of both features. It is, therefore, not clear that income and security code alone will be sufficient to distinguish fraudulent from non-fraudulent transactions. The absence of any clear clusters suggests that extra features or more advanced modeling may be necessary.



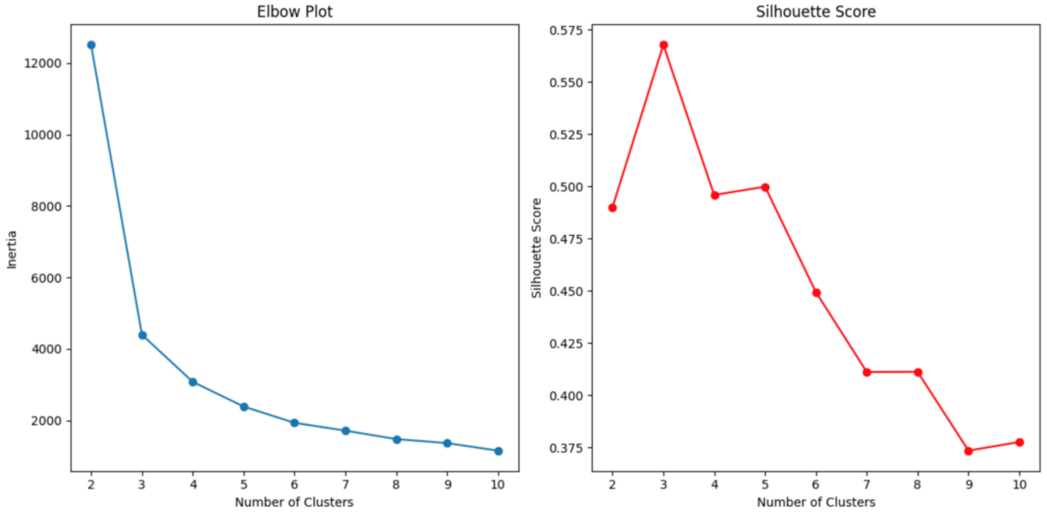
Confusion Matrix(Heatmap)

It must therefore consider the confusion matrix to ascertain the performance level of the linear regression model at providing the fraud transaction forecast. This shall be ideal if the models correctly classify the fraudulent ones as true positives and the others correctly as true negatives. The opposite may be observed in many cases, where, out of the imbalance in our dataset, the model provides more correct classifications for instances that are legitimate rather than fraudulent.



Elbow and Silhouette Plot

The Elbow Plot shows the number of clusters versus inertia, a measure of the within-cluster sum of squares. The "elbow" in the plot, or the point at which the rate of decrease in inertia begins to slow down, is typically considered to indicate the optimal number of clusters. The Silhouette Score Plot gives the measure of each point with respect to its own cluster compared to other clusters. High silhouette scores mean better-defined clusters.



Conclusion

In this analysis, K-Means clustering and linear regression have been applied to explore relationships among income, security code, and fraudulent transactions. K-Means clustering showed that three clusters give the most meaningful separation of data, though further refinements could be achieved if there were more features. This linear regression model has given basic insights into fraud prediction; however, improvements must be done on this since the dataset is highly imbalanced.

Future work could also focus on: Use more advanced fraud-detecting machine learning models, such as Random Forest or XGBoost. Trying different techniques in the process of sampling for class imbalance. Addition of more features that could better reflect the fraudsters' pattern of behavior.