Term Project 3 - Milestone 3:

White Paper

Fraud Based Detection and Prediction

DSC 680

Prof. Amirfarrokh Iranitalab

Carlos Cano

Business Problem:

In our digital age electronic transactions can reach into the billions daily, this in part lays opportunities to would be criminal activity through the use of fraudulent transactions that can cost merchants, and global conglomerates billions of dollars in loss annually.

Background/History:

The history of fraud is one that may very well be as old as recorded time, as we evolved from a barter to a fiat based transactional system. Fraud has too evolved into a high paced sophisticated system in which identities, debit card and credit cards are a staple in the criminal underworld. These criminals would be victims' damages reverberates throughout various degrees of losses in accounting systems.

In hopes of circumventing and combatting instances of fraud, we must continue to evolve and garner increasing insight through data science and assisting in its application and further model refinement as a mean to aid in its detection and possibly prediction.

Data Explanation (Data Prep/Data Dictionary/Etc.):

Data processing for this EDA consisted of importation, exploration, modification, and implementation. It was evident through this dataset research that there were many variables located within that coded for place holder column headers. As expressed in Figure 1. there exist direct correlations among the variable classes.

Methods:

This EDA will focus on Random Forest Classifier for its use of supervised machine learning through the use of a targeted variable to identify and predict instances of fraud within the model. A linear relationship may be explored within the model as a means of standardizing the efficacy of the model at hand.

Analysis:

The first analysis of this EDA would be explored with an overall expression of instances of fraud found. As seen in Figure 2. there exists an unproportionate number of cases of non-fraud in relations fraudulent transactions. While this may prove insignificant in the overall percentage, the argument here is that the magnitude of each transaction is not captured.

The second graphical analysis showcases the concentration of cases of fraud as it relates to time in seconds from one instance of fraudulent transaction to the next, in this instance we can explore the bi-modal concentration which showcases the varying time increases as we explore cases of fraud.

Conclusion:

This research into instances and the predictability of fraud is one of constant evaluation and monitoring. While being given a snapshot may suffice, being able to create and model in which new information is the end goal for some corporations when it comes to predicting fraud before it happens in hopes of staving off loss. As explored through supervised machine learning and the application of a random forest classifier to analyze the dataset that was processed it was evident that this model proved extremely accurate with a 99.95% return. Although this is extremely well laid out, this technique is applicable out in other areas of use for its shear efficacy in working with variable classification in understanding deeper relationships outside of linear and regressive models.

Assumptions:

The leading assumption found here is that of the percentage relationship of fraud vs non-fraud transactions. From a graphical analysis, it demonstrated that the given dataset had minimal instances of fraudulent transaction in comparison and that it would be an impressive task to establish a working model given the disparity. That said it was assumed through varying models that there would indeed be a positive correlation.

Limitations:

The data presented consisted of various unknown and possibly proprietary based which were re-coded by the initial publisher of the dataset. This single change leaves us with a limited understanding of what each variable is, and the relationship established therein. If not for these recoded variables a deeper understanding could be achieved.

Challenges:

The initial dataset inquiry was the first challenge undertaken, of which locating a dataset that would not focus heavily on regressive modeling and would further need to be refined with continued processed technique would need to be undertaken. As such the subsequent challenge undertaken was that of a more advance application of machine learning techniques.

Future Users/Additional Applications:

Future applications within this space will continue to exist as there is a demand to avoid losses by merchants and corporations the world over. As we continue to decrease our reliance on cash based transactions and move into a newer and more refined space of electronic transactions, the understand and continued development of fraud detection is paramount.

Recommendations:

Although the variables presented within this EDA were recoded, it would be interested to further develop a working model in which the variable connection would be presented to garner a better understanding of the relationship within. This alongside geolocation to detect increased fraud by locations would further expand the application of the current model.

Implementation Plan:

A role out of this model in its current state would serve within a basic function, that said it could further be expanded on through multi-steps. Of which, data capture and processing in real time to detect on going transaction to detect trends and patterns as they arise. The previously mentioned capture of geolocation data to coincide with data processing would be some but not all the needs to implement this EDA into a working model.

Ethical Assessment:

The main ethics dilemma of this EDA would be the datasets original variable collection. This process clearly was used to protect the identity and transactional information of the credit card users. Being able to blind study the clientele allows the protection of said clientele, it is in that protection that we have an ethical obligation to focus on a science-based approach. Data removal was very limited in its nature and afforded a more robust application. This approach provides for an ethical implementation and outlook when analyzing personal information.

Graphical Representations

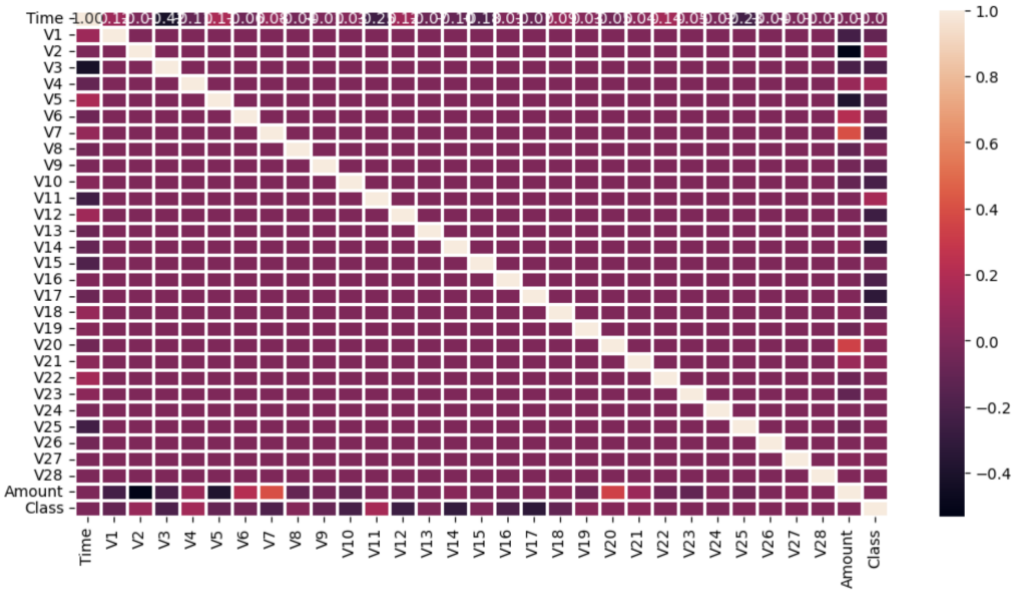

Figure 1.

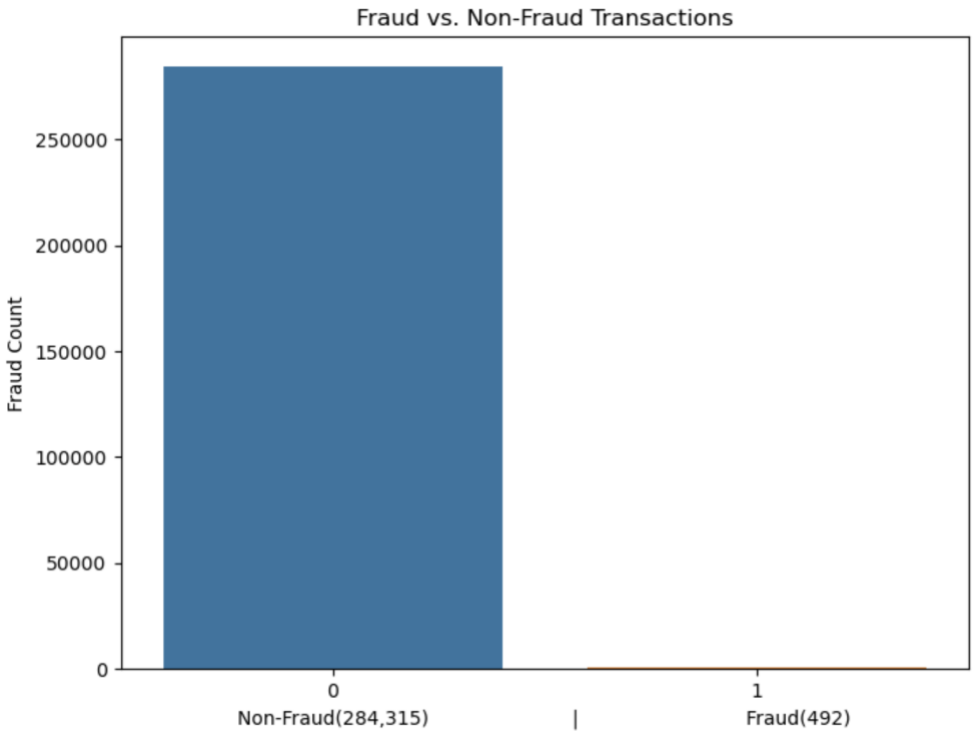

Figure 2.

Graphical Representations Continued

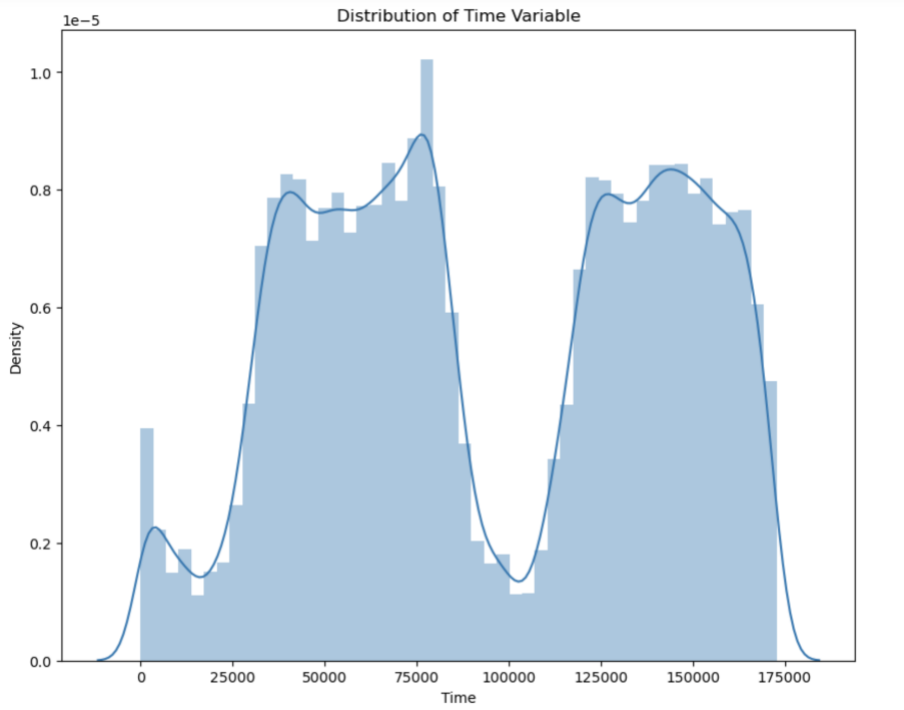

Figure 3.

References

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2017, June 22). *Credit Card Fraud Data - dataset by RAGHU543*. data.world. https://data.world/raghu543/credit-card-fraud-data

Transaction fraud: How to detect it and tips for prevention. Fingerprint. (2024, February 18). https://fingerprint.com/blog/transaction-fraud-detection/