Credit Card Fraud Detection

Machine Learning Project- Phase 2

Malak Gaballa
900201683
Data Science
The American University in Cairo
malakhgaballa@aucegypt.edu

Masa Tantawy
900201312
Data Science
The American University in Cairo
masatantawy@aucegypt.edu

Introduction

In this phase, we aim to describe the dataset as well as the data wrangling and feature engineering process. As mentioned in the previous phase, the aim of this project is to develop a credit card fraud detection system using the best-performing machine learning algorithms. This system can be applied at financial institutions that issue cards, such as banks, to monitor credit card activity and hinder fraudulent transactions.

The dataset used in this project consists of simulated credit card transactions over the span of 2 years, from the 1st of January 2019 to the 31st of December 2020 [1]. As explained in the previous phase, it was found on Kaggle and was generated using a simulation tool in python, created by Brandon Harris, where transactions for 1000 customers with 800 merchants were created based on factual information. This dataset was considered the optimal choice among the other available datasets because of its appropriate size (instances and features) as well as containing no anonymized columns. Furthermore, it was also found to contain no missing values.

The original dataset is divided into a training set of 1,296,675 instances and a testing set of 555,719 instances. This dataset contained an index column and 22 features, one of which is a binary label that classifies each transaction as fraudulent (1) or legitimate (0). During this phase of data, these 2 sets were combined together to be preprocessed , such as handling missing data and feature engineering. The merged dataset consists of 1,852,394 instances. After preprocessing, it contains 15 features.

Feature Description

As mentioned above, the original dataset consists of 22 distinct features each of which will be described in this section to illustrate their importance. There are initially a total of 9 numeric columns, 11 categorical columns including the label, and 2 datetime columns. There are neither missing values nor duplicated instances in the dataset. The statistical distribution of these numeric features before preprocessing are shown in Table 1 and the number of unique values of the categorical features are shown in Table 2.

The 1st feature is *trans_date_trans_time*. It is a datetime feature that states the exact day, month, year, hour and minute at which the transaction occurred. This is relevant as it describes exactly when a transaction has occurred. During preprocessing, this feature will be split into several features.

The 2nd feature is *cc_num* which is the credit card number of the customer. It is a numeric column that acts as a unique identifier to the customer as no 2 customers have the same card number.

The 3rd feature is *merchant, which* is a categorical feature of the merchant's name. This feature seems not very important as there is a set of specific merchants from which this data was simulated, yet the decision of removing it can only be taken after preprocessing and data analysis.

The 4th feature is a categorical variable named *category* that represents the classification of the credit card transaction, such as personal care, travel, or health and fitness. The 5th feature *amt* is a numeric column that shows the amount of the transaction. This is extremely relevant as it is important to find out

All the tables and figures mentioned can be found at the appendix at the end of this document.

whether fraudulent transactions are more likely to occur within a specific range of amounts, such as very large transaction, or not.

Features 6 to 17 contain the information of the cardholder. These include: *first* and *last* which are both categorical variables for the cardholder's name, *gender* which classifies the cardholder as F or M for female and male respectively, *street*, *city, state* and *zip* are all categorical variables for the cardholder's address details, *lat* and *long* are numeric features that describe the latitude and longitude of the location of the cardholder, and lastly *city_pop* is also a numerical feature that contains the population of the cardholder's city at the time of the transaction. The cardholder's job and date of birth are provided as 2 separate features, a categorical *job* and a datetime *dob.* This group of features are very relevant as they dictate a large amount of the transaction's details since most of the information about a transaction is related to the customer who did it.

As for the rest of details of the transaction, a categorical feature *trans_num* contains the transaction number as a combination of letters and numbers. As shown in Table 2, each transaction has a unique transaction number that is specific to only it. This means that there are 1852394 unique values in this feature making it extremely sparse thus it needs to be removed as it is not of any importance.

and *unix_time* contains the transaction's unix time. This is the total number of seconds that have elapsed since January 1st 1970 at UTC until the transaction. The latitude and longitude of the merchant's location are stored in 2 numeric features: merch_lat and *merch_long.* The importance of these features will be clarified during preprocessing and data analysis.

Finally, the label: *is_fraud* is a binary feature where 0 refers to a legitimate transaction and 1 fraudulent. The statistical distribution of the label is shown in Table 1.

As visible from Table 1, the numeric features do not follow the same distributions. They have different means and standard deviations. This means that the data requires scaling. It was observed from the box plots of the numeric features that there are outliers in some of them; the percentages of outliers in each feature was computed. Since we will normalize the features to make them all follow the standard normal distribution, mathematicians commonly state the normal amount of outliers 3 standard deviations, which is 0.03% in the case of standard normal distributions. Hence, using this threshold, none of the features contained an amount of outliers less than 0.03%, so they were treated as normal instances and thus not removed. Normalization was done on these numeric columns and Table 3 shows the statistical distributions of the numeric features after normalization.

Feature Relationships

Next, we examine the relationships between the features. The correlation of the numeric features with the label (binary *is_fraud*) are shown in Table 4. It is vital to identify if the dataset contains collinear features as their results in noise and affects the performance of the model.

For numeric features, the Pearson correlation coefficient was used to identify collinearity between the features. The threshold used was absolute 0.8; this means that if the absolute value of correlation coefficient between 2 features was greater than or equal to 0.8, they are considered collinear. The pairs plots were also examined. From the heatmap in Figure 1, it was found out that there are three sets of collinear features which are: *zip, long,* and *merch_long; lat* and *merch_lat*; *long* and *merch_long*. Hence, to solve this problem, the features *merch_lat*, *merch_long,* and zip will be dropped. This is also recommended as the box plots of *merch_lat* and *merch_long* demonstrated a large number of outliers.

For categorical features, the chi square test of independence was used. All the categorical features were dependent on the label using an alpha of 0.05 to the p-value test. However, this value of alpha also concluded that all the categorical features are dependent on each other. One explanation for this is

All the tables and figures stated in the sections can be found at the appendix at the end of this document.

that the size of the dataset and the chi-square statistic is very large which yields a p-value that always returns to zero.

Feature Engineering

After inspecting the features and analyzing their statistical distributions, the data was prepared for usage in future phases. First, the features *zip, merch_lat,* and *merch_long* were dropped from the dataset to solve the collinearity issue explained in the previous section. Additionally, *cc_num* and *unix_time* were dropped from the dataset due to their very large column variance as shown in Table 1.The feature *trans_num* was also removed due it it being very sparse as previously stated.

After inspecting the features and analyzing their statistical distributions, the data was prepared for usage in future phases. First, the features *zip, merch_lat,* and *merch_long* were dropped from the dataset to solve the collinearity issue explained in the previous section. Additionally, *cc_num* and *unix_time* were dropped from the dataset due to their very large column variance. The feature *trans_num* was also removed due it it being very sparse as previously stated.

Next, datetime features, which are *trans_date_trans_time* and *dob,* had to be transformed in order to facilitate the machine learning process. Because they were classified as string variables, they were converted into datetime. For the feature *trans_date_trans_time*, five new features were created by extracting the day, month, year, hour, and minute at which the transaction took place. As for the variable *dob,* it was transformed by creating another feature called *age* that would replace it; this feature was computed by subtracting the credit card holder's date of birth (*dob*) from the transaction date (*trans_date_trans_time*) and dividing the difference by time delta of one year which is a function in python that is used for dat mainpulations. Since the yielded values had decimal points, they were rounded down in order to accurately display the credit card holder's age. All the features extracted from *trans_date_trans_time* contain no outliers, but *age* has 0.0235% outliers which are visible in Figure 2, so according to the aforementioned threshold 0.03%, these instances were removed from the dataset. Figures 2 and 3 display the feature age before and after removing outliers respectively.With these new features, there appears to be no collinear sets in the data which made it appropriate to remove the features *trans_date_trans_time* and *dob* now that they are considered redundant. To ensure that all the numeric features in the dataset follow the same distribution, these new numeric features were similarly normalized.

For the remaining eight categorical variables in the dataset, label encoding was used to transform *gender* from a categorical to a binary variable; 'F' was replaced by 1 and 'M' was replaced by 0. Given that the other categorical features are not ordinal, other encoding methods had to be used. For instance, frequency encoding was implemented on the features *category* and *state* because no 2 classes shared the same frequency, which eliminated the issue of collisions. As for the variables *merchant, first, last, street, city* and *job,* the use of frequency encoding leads to collisions therefore cannot be used. It is also important to note that one-hot encoding, binary encoding, and hashing are also not suitable for these features as each has over 350 classes, so it will cause a problem of curse of dimensionality. Hence, other encoding methods were resorted to such as target encoding and leave-one-out encoding; yet, the problem of collisions still arose. Thus, after trying all feature encoding techniques, these features were dropped due to their sparsity.

Final Dataset

The final preprocessed dataset consists of 1,851,959 instances and 14 features. These features are: *is_fraud, category, amt, gender, state, lat, long, city_pop, trans_day, trans_month, trans_year, trans_hour, trans_minute, age*. The label and gender are binary features whereas all the other features are numeric.

The main reasons that some features were removed are: high column variance in numeric features or

All the tables and figures stated in the sections can be found at the appendix at the end of this document.

sparsity in the case of categorical features that hinders the process of encoding, to solve the problem of collinearity, or were replaced by other more meaningful features making them redundant such as in the case of datetime features. Now, the dataset is ready for algorithm implementation. The dataset will be divided into a training and a testing set depending on the approach that will be used. Some of the possible ways are a 80%-20% training-testing split, but also bootstrapping or using K-folds are other possible ways. This will be explored in the next phase.

References

[1]
Credit Card Transactions Fraud Detection Dataset. *Kaggle*. Retrieved February 13, 2022 from https://www.kaggle.com/datasets/kartik2112/fraud-detection?resource=download

Appendix

Table 1

	mean	std	min	max
cc_num	4.17E+17	1.31E+18	60416207185	4.99E+18
amt	70.063567748	159.2539748	1	28948.9
zip	48813.25819	26881.84597	1257	99921
lat	38.539310 98	5.071470391	20.0271	66.6933
long	-90.22783229	13.74789488	-165.6723	-67.9503
city_pop	88643.67451	301487.6183	23	2906700
unix_time	1358674219	18195081.39	1325376018	1388534374
merch_lat	38.538975597	5.105603878	19.027422	67.510267
merch_long	-90.22793951	13.75969211	-166.671575	-66.950902
is_fraud	0.005210014716	0.071992175	0	1

Table 1 (continued)

	25%	50%	75%
cc_num	1800429464911 50	3.52142E+15	4.64226E+15
amt	9.64	47.45	83.1
zip	26237	48174	72042
lat	34.6689	39.3543	41.9404
long	-96.798	-87.4769	-80.158
city_pop	741	2443	20328
unix_time	1343016824	1357089331	1374581485
merch_lat	34.740122225	39.3689005	41.956263
merch_long	-96.89944	-87.4406935	-80.245108
is_fraud	0	0	0

Table 2

Column Name	Unique Values
trans_date_trans_time	1819551
cc_num	999
merchant	693
category	14
amt	60616
first	355
last	486
gender	2
street	999

All the tables and figures stated in the sections can be found at the appendix at the end of this document.

city	906
state	51
zip	985
lat	983
long	983
city_pop	891
job	497
dob	984
trans_num	1852394
unix_time	1819583
merch_lat	1754157
merch_long	1809753
is_fraud	2

Table 3

	mean	std	min	max
cc_num	2.20E-16	1	-0.318830579	3.494696365
amt	-1.98E-16	1	-0.433669348 4	181.3382459
zip	1.07E-16	1	-1.769084543	1.901199117
lat	-4.49E-16	1	-3.650265022	5.551445015
long	9.02E-17	1	-5.487710544	1.620432254
city_pop	-9.45E-18	1	-0.293944656 8	9.347171008
unix_time	-5.88E-15	1	-1.830066056	1.641111382
merch_lat	-1.80E-15	1	-3.821595728	5.674410261
merch_long	-7.77E-15	1	-5.555621075	1.6916830 2
is_fraud	0.0052100147	0.07199217	0	1

	16	5		

Table 3 (continued)

	25%	50%	75%
cc_num	-0.3186930949	-0.3161407036	-0.3152845237
amt	-0.3794163854	-0.1419968796	0.08185938557
zip	-0.8398328827	-0.02378029364	0.86410516
lat	-0.7631733364	0.1607007353	0.6706317415
long	-0.4779035456	0.2000984379	0.7324635791
city_pop	-0.2915631328	-0.2859177932	-0.2265952907
unix_time	-0.8605289941	-0.08710528964	0.8742618995
merch_lat	-0.7440557098	0.16255168 89	0.6693208316
merch_long	-0.4848582684	0.2025660157	0.7255127097
is_fraud	0	0	0

Table 4

Feature	Correlation with is_fraud
cc_num	-0.001125
amt	0.209308
zip	-0.00219
lat	0.002904
long	0.001022
city_pop	0.000325
unix_time	-0.013329
merch_lat	0.002778
merch_long	0.000999

All the tables and figures stated in the sections can be found at the appendix at the end of this document.

Figure 1 - Collinearity Heatmap

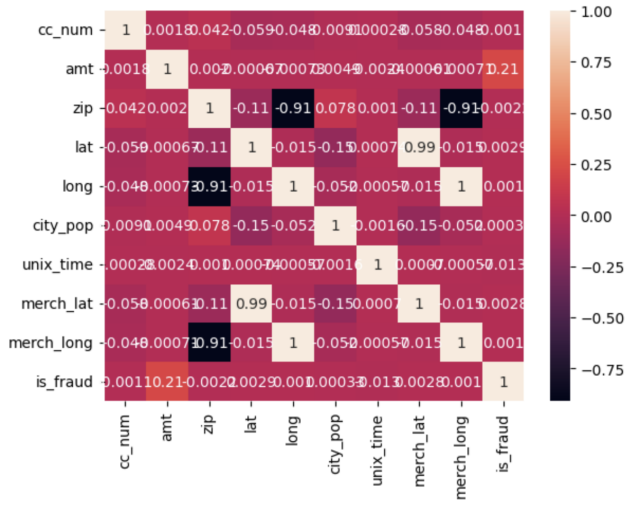


Figure 2 - Age Box Plot with outliers

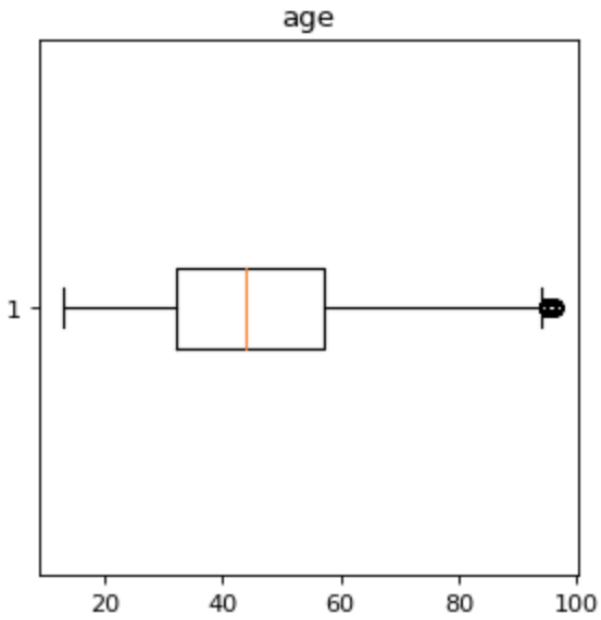
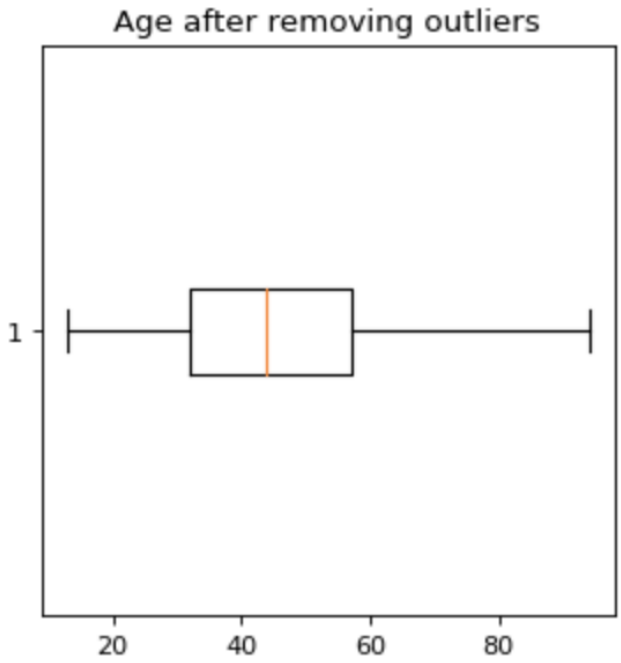


Figure 3- Age Box Plot after removing outliers



All the tables and figures stated in the sections can be found at the appendix at the end of this document.