Data Analytics 350 Credit Card

Sally Hyde and Matthew Rinker

***Abstract*— In this report we outline our analysis of extensive credit card transaction data. Much of the analysis is centered around time and transaction amount for the transactions, as these were the only identifiable variables in our data.**

I. INTRODUCTION

Credit card fraud is one of the most important financial issues of the 21st century, and in this lab, we used exploratory data analysis, feature engineering, and unsupervised learning techniques to explore this issue. Using data of over 280,000 credit card transactions, we executed analyses such as min-max normalization, principal components analysis, and K-means clustering algorithm. Interestingly, we found that fraudulent credit card transactions have a skewed distribution, with most fraudulent transactions occurring at very low dollar amounts.

II. DATA

In this lab, we worked with credit card data from Kaggle.com. According to the documentation on Kaggle, the data comes from two days of credit card transactions in the European Union in September of 2013, and there are 492 frauds out of 284,807 transactions. For each transaction there are 30 explanatory variables. (Shown in Fig. 1)

Variable	Definition
Time	Time of the transaction
V1-V28	Numerical data w/ context obscured for privacy
Amount	The amount of the transaction

Fig. 1. The variables in the dataset

Due to the fact that we have 28 variables with all context obscured to preserve privacy, we have to do some operations on the data to figure out what variables are most important. To do this, we looked at the average differences in means for our variables. This gave the following result: we found that V4, V11, V2, V21, and V19 would be the most meaningful. The median and mean values of these variables are found in Fig. 2 and Fig. 3 respectively.

	Time	Amount	V4	V11	V2	V21	V19
Fraudulent	75,568.5	9.25	4.177147	3.586218	2.717869	0.592146	0.646807
Non-Fraudulent	84,711	22	-0.022405	-0.034923	0.064070	-0.029821	0.003117

Fig. 2. The median values for our important variables

	Time	Amount	V4	V11	V2	V21	V19
Fraudulent	80746.806911	122.211321	4.542029	3.800173	3.623778	0.713588	0.680659
Non-Fraudulent	94838.202258	88.291022	-0.007860	-0.006576	-0.006271	-0.001235	-0.001178

Fig. 3. The mean values for our important variables

III. RESULTS

To further analyze our data we used min-max normalization to ensure that our results are not thrown off by the data's magnitude. This involved taking the minimum and maximum of the data and using it to scale the columns. The formula used is as follows: $\overline{x} = \frac{(x - \min)}{(\max - \min)}$. We then took the medians and means of the data and compared the data in the same manner as before. This gave the surprising result that in our transformed data the most important columns are V7, V9, V27, V15, and V10. The median and mean values can be found in Fig. 4 and Fig. 5 respectively.

	Time	Amount	V7*	V9*	V27*	V15*	V10*
Fraudulent	75,568.5	9.25	0.820969	0.668666	0.742393	0.637235	0.699149
Non-Fraudulent	84,711	22	0.041138	0.001230	0.064070	0.048294	-0.091872

Fig. 4. The median values for our new important variables

	Time	Amount	V7*	V9*	V27*	V15*	V10*
Fraudulent	80746.806911	122.211321	0.769625	0.646486	0.720645	0.632113	0.660782
Non-Fraudulent	94838.202258	88.291022	0.009637	0.004467	-0.000295	0.000161	0.009824

Fig. 5. The mean values for our new important variables

We then used two dimensionality reduction techniques on our data: PCA and t-SNE. For the PCA, we set the number of principal components to two. A scatterplot of these results can be seen in Fig. 6. There are two clear clusters based on the principal components analysis. The second dimensionality reduction technique we used was t-Distributed Stochastic Neighbor Embedding, otherwise known as t-SNE. A scatterplot of those results can be seen in Fig. 7. This technique resulted in very distinct clusters. For this dataset, PCA is likely the more appropriate dimensionality reduction technique. t-SNE, though an incredibly valuable tool for image and speech processing, genomic data, or natural language processing, is more suitable as a data exploration and visualization technique. This is because the input features at the end are no longer recognizable, and t-SNE results are not usable for making inferences. In this way, PCA is the better method with our dataset, because not knowing what columns values represent makes data visualization unnecessary. In addition, t-SNE is much more computationally inefficient than PCA.

We also explored patterns in the deciles of the **Amount** and **Time** columns. In Fig. 8 and Fig. 9 are the counts of

fraudulent and non-fraudulent transactions for each decile of transaction amount and time. In the time table (Fig. 8), there seems to be several spikes in fraudulent transactions, with particularly high spikes in the first 10% and the 60% decile. In the transaction amount table (Fig. 9), it seems that most fraudulent transactions happen at very low dollar amounts, with spikes in the $100 - $200 dollar ranges and higher. This may be due to how fraudulent transactions are executed; scammers may test transactions at low dollar amounts to ensure success with low-stakes.

Entering the classification and prediction phase, we managed to execute a K-Means clustering algorithm using $k = 2$ clusters. The performance indices we utilized for this algorithm were the Silhouette Index and the Calinski-Harabasz Index and these results can be seen in III.

Finally, we created some visualizations of our data to explore the relationships between variables as well as some of the distributions of our variables. Figures 10 and 11 showcase histograms and distribution plots of our **Amount** and **Time** variables. From this it's clear that most of the transactions are low dollar amounts. We also created a pairplot (Fig **??**) of the five scaled variables we created, along with **Amount** and **Time**. This pairplot displays scatterplots of each variable with each other variable in a matrix format. Interestingly, V9 and V10, as well as V7 and Amount seem to be roughly positively correlated with one another.

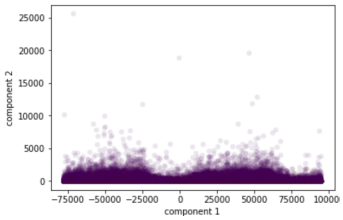

Fig. 6. Scatterplot of PCA

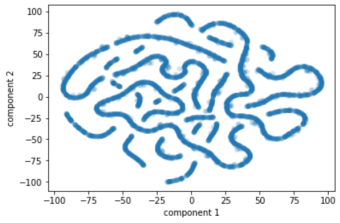

Fig. 7. Scatterplot of t-SNE

	10%	20%	30%	40%	50%	50%	60%	70%	80%	90%
Fraudulent	93	63	47	37	28	91	24	33	53	22
Non-Fraudulent	28389	28416	28436	28441	28453	28393	28456	28444	28428	28459

Fig. 8. The counts of fraudulent/non-fraudulent transactions for each time decile

	10%	20%	30%	40%	50%	50%	60%	70%	80%	90%
Fraudulent	181	27	36	13	14	17	24	50	45	85
Non-Fraudulent	30311	26446	28523	28392	28700	28538	28342	28865	28005	28373

Fig. 9. The counts of fraudulent/non-fraudulent transactions for each transaction amount decile

Index	Result
Silhouette	0.373
Calinski-Harabasz	5.317

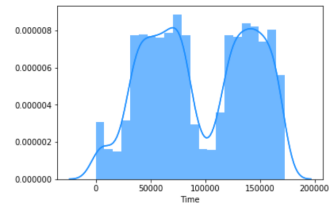

Fig. 10. Histogram and Distribution plot of Time

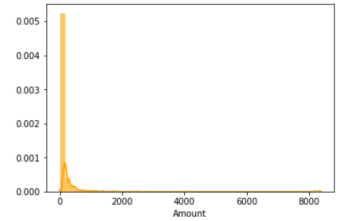

Fig. 11. Histogram and Distribution plot of Amount

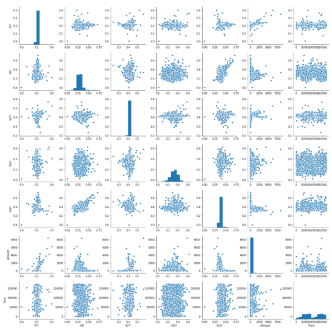

Fig. 12. Pairplot of Important Variables