Credit Card Fraud Detection using Big Data

Ronald Garcia
Rayna Ilieva
Urooj Khaleeli
Kristy Kwong
Dimple Sachdev
Mohammad Tahsin

University of Waterloo Data Science
Group 2
December 14, 2020

Credit Card Fraud Detection using Big Data

A credit card is a thin rectangular piece of plastic or metal issued by a financial institution (FIs) or financial service company that allows cardholders to pay for goods and services. It relies on the cardholder's promise to pay for these goods and services at a later period. If you have purchased items in a shopping mall or paid for coffee at a local shop, then it is likely you have used a credit card. By the 20th century, the demand and convenience of credit cards have increased dramatically and became the most popular methods of payments worldwide, particularly in "North America due to the existence of a widespread point of sale (POS) network" (Paypers, 2019). The modern design and technology of credit cards created many problems such as, getting into credit card debt, missing payments, spending over credit limit, experiencing fraud and many more.

Fraud is "when someone steals your credit card, credit card information or personal identification number (PIN) and uses it without your permission to: (1) make a purchase at a place of business; (2) make a purchase or transaction online; (3) make a purchase or transaction by telephone; (4) withdraw money from an automated teller machine (ATM)" (Government of Canada). Every year, FIs bear millions of dollars of losses due to fraudulent activities, especially VISA and MasterCard suffer over $1 billion (Paypers, 2019). They continue to find new ways to prevent fraud such as declining magnetic stripe transactions, using computer chip and PIN to make transactions more secure (Government of Canada), three-dimensional holograms, and CVC (card validation codes). A potential way is to replace credit cards with Smart Cards, but the replacement cost is very expensive due to the widespread POS network in North America and the huge number of credit cards in circulation worldwide (Paypers, 2019).

FIs constantly monitor customers' behavior in order to estimate, detect, flag, and/or stop unwelcome purchasing patterns using data. There are many fraud detection techniques being used and implemented through data such as, artificial intelligence (AI), machine learning, and many more. In order to detect credit card fraud, Naive Bayes (Gaussian), Generalized Linear Model and Logistic Regression are the models used for identification in this report. This will create a robust algorithm through analyzing cardholders' transactions data, in hopes to reduce and detect fraudulent activities for both FIs, merchants and cardholders.

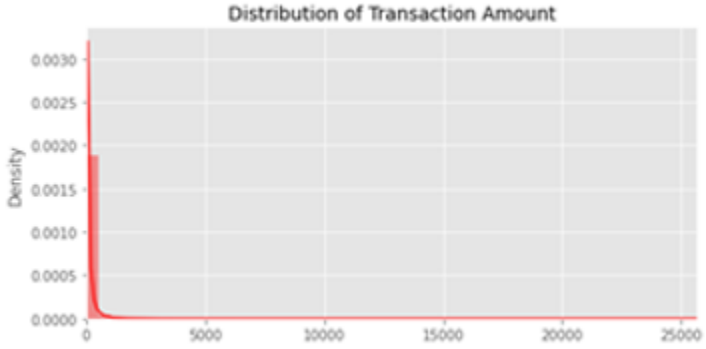
Materials

Data Preparation

The dataset contains transactions made by European cardholders in September 2013. A total of 284,807 transactions documented, and 492 frauds were flagged for transactions that occurred within two days. Below, are the descriptions of the dataset.

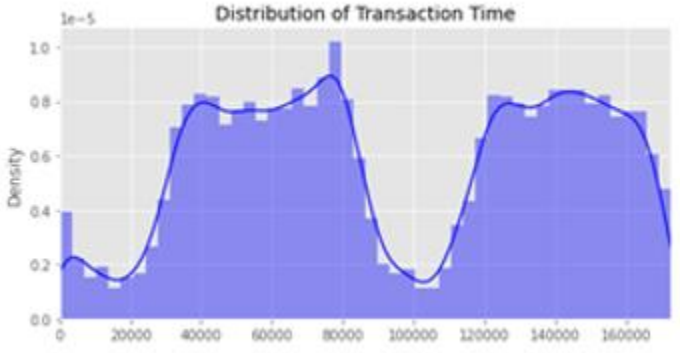
- No "Null" values
- **Variables**: The data (i.e., V1, V2, V3, […], V28) has gone through PCA transformation (Dimensionality Reduction Technique), except for 'Time' and 'Amount'. The purpose was to maintain anonymity of the cardholders' information.
- **Amount**: Most transactions amount is relatively small. The Mean of all amounts made is approximately $88 USD. Figure 1 shows a distribution of transaction amount from data.

Figure 1



- **Time**: The 'Time' data is recorded in seconds since the first transaction is captured, meaning the data was recorded over two days (172,800 seconds). Figure 2 shows there was a dip in the evening time, which indicates a change from Day 1 to Day 2.

Figure 2



- **Class**: ~99.83% of the transactions were identified to be non-fraudulent; 0.17% of the remaining transactions were identified as fraudulent activities (see Figure 3). To analyze this further with 'Class' as a dependent variable, the data requires cleaning and its described in the next section.

Figure 3

Data Cleaning

As seen in Figure 3, there is an imbalance data – a classification problem where 'Class' is not represented equally. This small dataset of fraudulent activities flagged creates a problem for the classification algorithm, which cannot learn from. In order to collect the best value from this dataset and have a model to classify as accurately as possible, we created a new set of data from this current set for extract to further investigate.

Balancing of Dataset

To remove the biasness towards most of the 'Class' (non-fraudulent activities) from this current dataset, we need to create a balanced dataset from it (Frei, 2019; George, n.d.). This is an approach called, Sampling. In this analysis, Undersampling technique is used to adjust and the distribution of data.

Undersampling will help to create a more balanced dataset from an imbalanced dataset (George, n.d.) by resampling most of the 'Class' to make it all equal to the minority 'Class' points. Then we created a sub-sample of the data where it will have an equal amount of fraudulent activities versus non-fraudulent activities. This technique will assist our models to better understand different transactions patterns and determine whether a transaction is indeed a fraudulent activity or not.

A random selection of 492 cases from the current dataset are taken in order to build the new sub-dataset. Therefore, totaling to 984 cases (adding the original 492 cases to it). The disadvantage of using Undersampling as a technique is the loss of relevant information. We reviewed the data after removal of the non-fraudulent transactions, there are no major pattern changes as seen in Figure 4.

Figure 4

Data Exploration

Correlation matrices are used to better understand the make-up of our dataset. We want to understand if there are variables that caused a transaction to be fraudulent or not, so it is important to compare to the original dataset. This will help us determine which variable had a high positive or negative correlation (i.e., Class = 1) (Frei, 2019, George, n.d.). The original unbalanced dataset (as shown in Figure 5a) does not have much correlation relationship to fraudulent activity, which falls within our expectation as stated previously. However, in Figure 5b, there is more variability.

Figure 5a

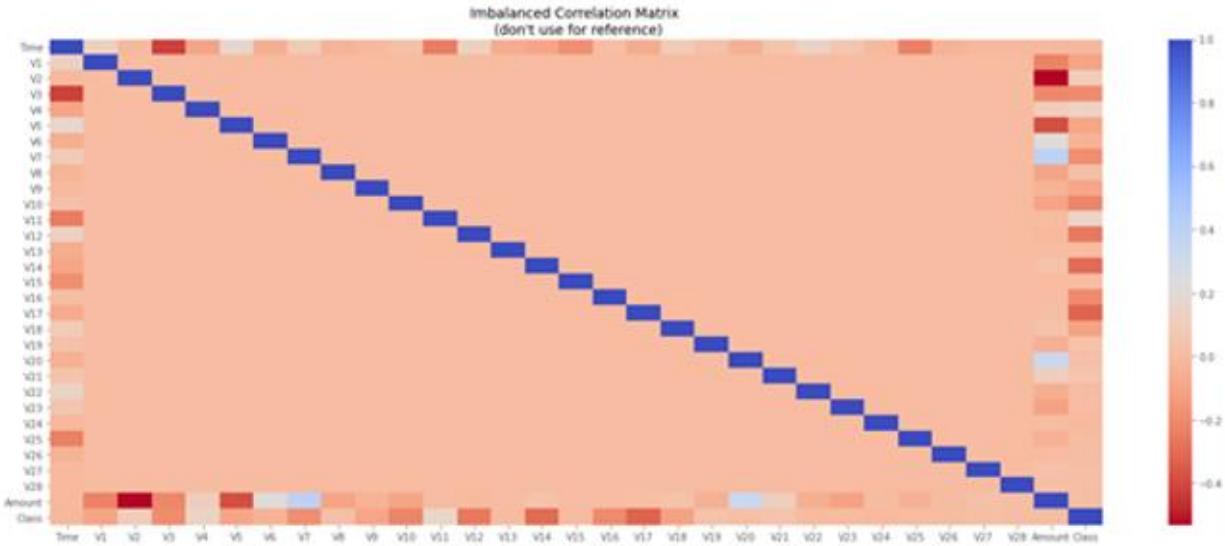
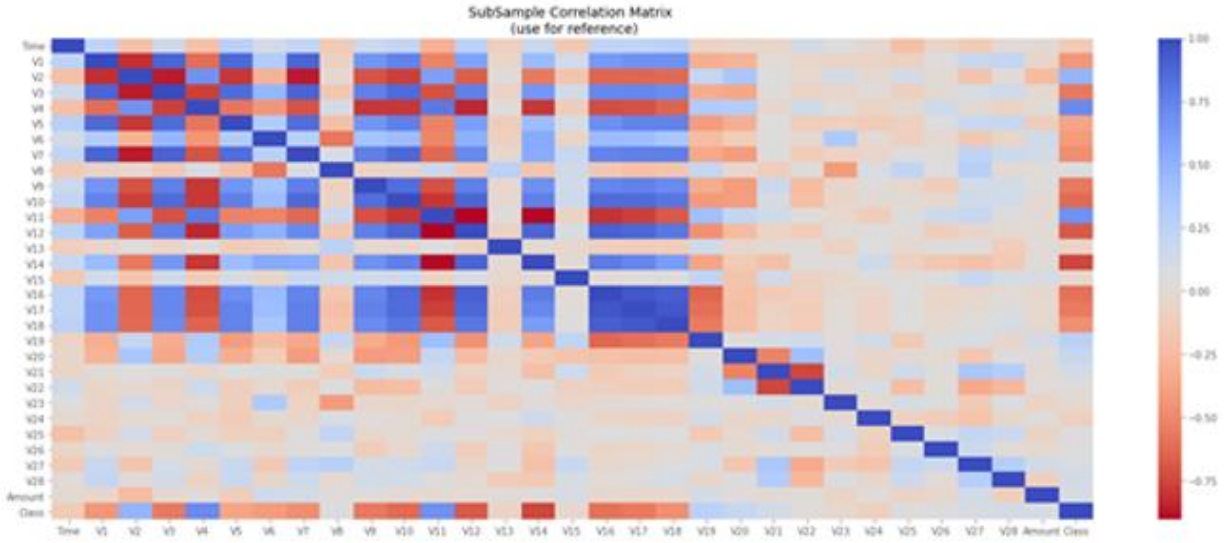


Figure 5b



Note. V10, V12, V14 and V17 are highly negatively correlated. It means, the lower these values are, the more likely the result would be coded as a fraud transaction. In comparison, V2, V4, V11 and V19 are highly positively correlated. It means, the higher these values are, the more likely the result would be coded as a fraud transaction too.

Challenges

The key challenge in using this dataset was the variables. It is identified to been anonymized, so it may indirectly impact the result. Also, the variables: 'Amount' and 'Time' are used in review for analyzation.

Models

This section proposes to answer whether the cardholders' transactions are considered as a fraudulent or non-fraudulent activity. Different approaches and models have been considered prior to finalizing with using the following models on the dataset – Naïve Bayes (Gausian) and Generalized Linear Model (GLM). After that, the Logistic Regression Analysis can help us identify and evaluate the dataset more thoroughly.

Naïve Bayes (Gausian)

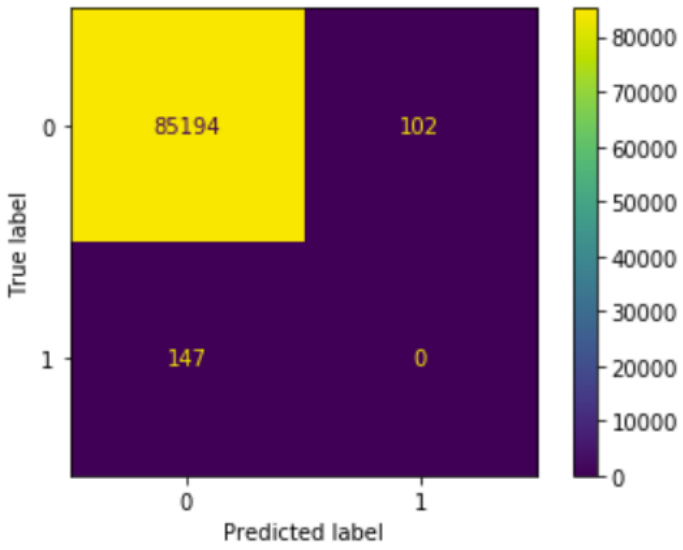
The Naive Bayesian classifier operates under the two assumptions: (1) generated through a Gaussian algorithm and (2) it operates under that the distribution is normal. This method works best when dealing with continuous variables which is applicable to this dataset. Like the standard Bayesian algorithms, this dataset will be split into training set and testing set (70% to 30% respectively). To test the accuracy of this model, the training set will be used first then follow by

the testing set. The Bayesian algorithm resulted in a 99% accuracy in prediction based on the two sets. Despite the high accuracy, this model is not our final evaluation due to many reasons.

1. Misrepresentation – The 99% is an inaccurate representation of the effectiveness of the classifier. The dataset has a total of 284,315 cardholders' transactions, and only 492 are flagged as fraudulent activities which is 0.173%, making most of the Gaussian classifier was training from was indeed honest transactions. It raises a question, "The model can classify honest transactions well, but can this be trusted when testing for fraudulent transactions?"
2. Unknown – The unknown data was taken out from this dataset (i.e., V1-V28). These are sensitive data points that are crucial information in contributing to the classification decision. The reason the data has been kept veiled for the security of the cardholders. As the exact contents were not known they had been dropped. However, there are still two other variables available to use – 'Amount' and 'Time' – to provide a more diluted model.
3. Confusion Matrix – By analyzing the Confusion Matrix, Figure 6, there are 147 transactions that are classified as not fraudulent. Many non-fraudulent transactions are predicted correctly. This strengthens our previous assumption that, this model creates an overfitting effective and it does not make it a desirable model to continue for data analysis.

Figure 6

Confusion Matrix - Bayes Gaussian



Generalized Linear Model (GLM)

GLM is a flexible linear regression that allows for variables to accommodate for error distribution other than a normal distribution. This is a viable model that can help to understand the relationship between different variables. To correctly predict the 'Class', we worked under the assumption that the probability of it being a fraudulent transaction is a function of 'Time', 'Amount' and V1-V28 variables. After using Undersampling technique, as stated above, this can compensate for the larger number of genuine transactions. The summary() function, Figure 7, also shows the change effect each variable has on the probability of 'Class'. For example, a = 1 increase in V28, meaning that the odds of fraud will increase by 9.4%. Conversely, Unit = 1 increase in V26 will have the odds of a fraudulent transaction go down by 6.9%.

Figure 7

Model Summary

	odds_ratio	percentage_effect
Intercept	1.246077	24.607673
Time	1.000000	0.000000
Amount	1.000000	0.000000
V1	0.980199	-1.980133
V2	1.001001	0.100050
V3	0.987084	-1.291586
V4	1.053376	5.337574
V5	1.019182	1.918165
V6	0.982161	-1.783897
V7	1.024290	2.429032
V8	0.987084	-1.291586
V9	1.010050	1.005017
V10	0.981179	-1.882064
V11	1.003005	0.300450
V12	1.015113	1.511306
V13	0.980199	-1.980133
V14	0.948380	-5.161999
V15	0.998002	-0.199800
V16	0.996008	-0.399201
V18	1.013085	1.308487
V19	1.010050	1.005017
V20	1.006018	0.601804
V21	1.016129	1.612869
V22	1.047074	4.707441
V23	0.980199	-1.980133
V24	1.012072	1.207229
V25	1.001001	0.100050
V26	0.931462	-6.853811
V27	0.989060	-1.093972
V28	1.094174	9.417428

Also, a posterior predictive model is used to showcase the change in 'Amount' or 'Time'. This affects how the probability of a transaction being flagged as fraudulent. Both Figure 8 and 9 demonstrates the relationship between 'Amount', 'Time' and 'Class' visually. The accuracy of this model can be shown in the confusion matrix below.

Figure 8

GLM Confusion Matrix

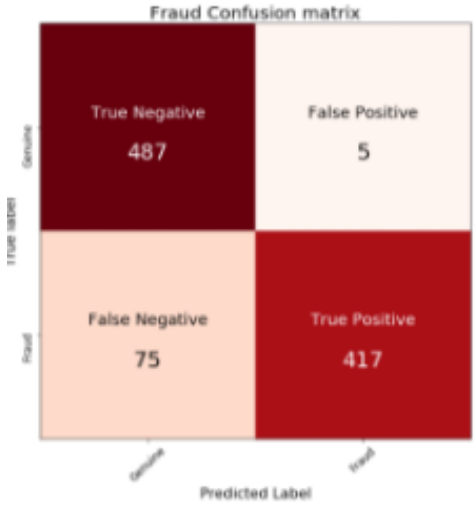
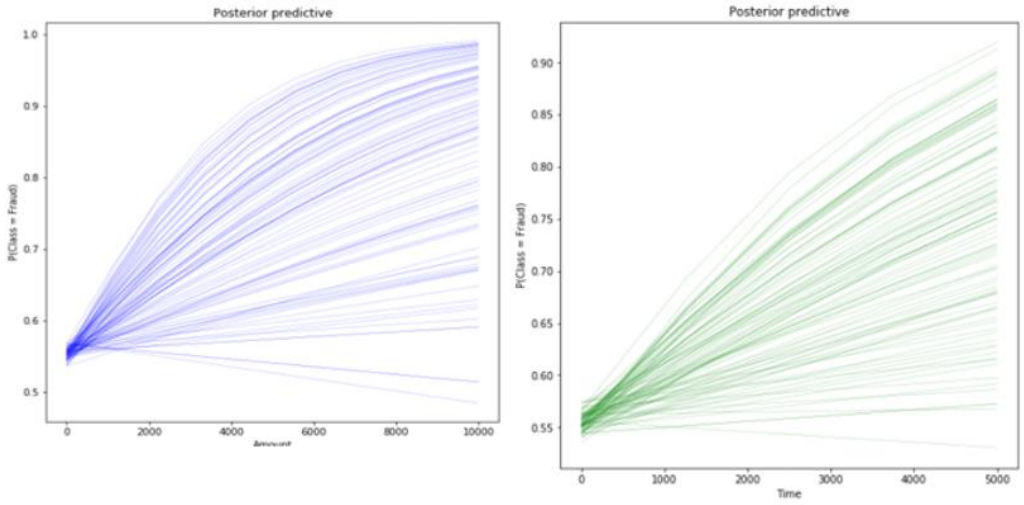


Figure 9

Posterior Predictive Model



Logistic Regression

The Logistic Regression model is used to "examine the association of (categorical or continuous) independent variable(s) with one dichotomous dependent variable," which in comparison to "linear regression analysis in which the dependent variable is a continuous variable" (Hoffman, 2019). This model uses the probability to compare certain event such as, pass or fail, win or lose, etc.,

To correctly predict the dependent variable 'Class', we used the function value with these predictor variables (see Appendix B), which is at 0.217686 and with 10 iterations. This implies a

good fit. From this, we noted the three points. Figure 11 further strengthens our understanding by showcasing the relationship between the values of logit(p) and odds (p), along with the actual values of the response ('Class'). It shows a good distribution of values along logit(p)=0. Also, when the data is grouped by 'Class', it shows a distribution of the estimated odds for both values of the actual response. Figure 12 illustrates the difference in transactions that have been identified as fraudulent activity (i.e., 'Class' = 1) and non-fraudulent activity (i.e., 'Class' = 0).

1. Converged = 1.0000. This means the regression analysis was successful.
2. Iterations = 10.000. The algorithm took 10 iterations to find the solution.
3. Pseudo R-squared: 0.686. This implies that the assessment of our model's quality was reasonably good.

Figure 11

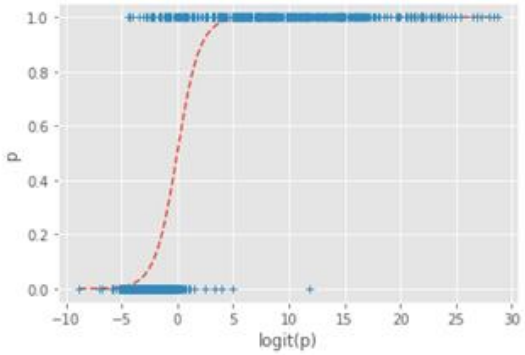
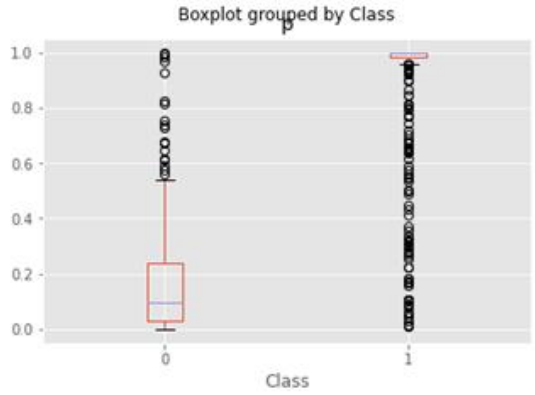


Figure 12

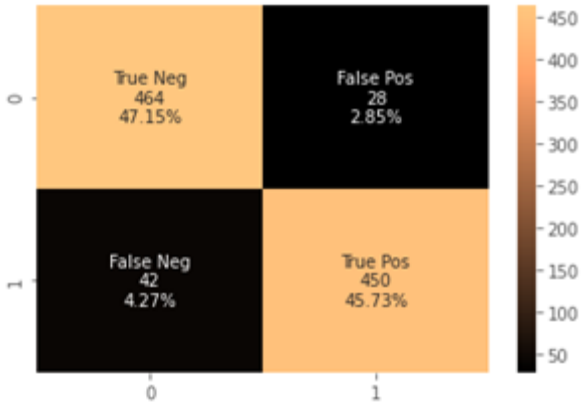


Confusion Matrix. This was also used to further analyze how well our estimates match up against the actual values of fraudulence. We summarized our findings below:

- A total of 464 transactions were estimated correctly (to be non-fraudulent); 450 transactions were also estimated correctly to be fraudulent. This represents a 92.88% accuracy.
- The remaining 28 + 42 estimates (equivalent to ~7.12%) were assigned incorrectly. The 42 transactions were falsely identified to be fraudulent and the 28 transactions were also falsely identified as non-fraudulent as shown in Figure 13.

Figure 13

Positive vs. Negative Fraudulent Transactions Assigned

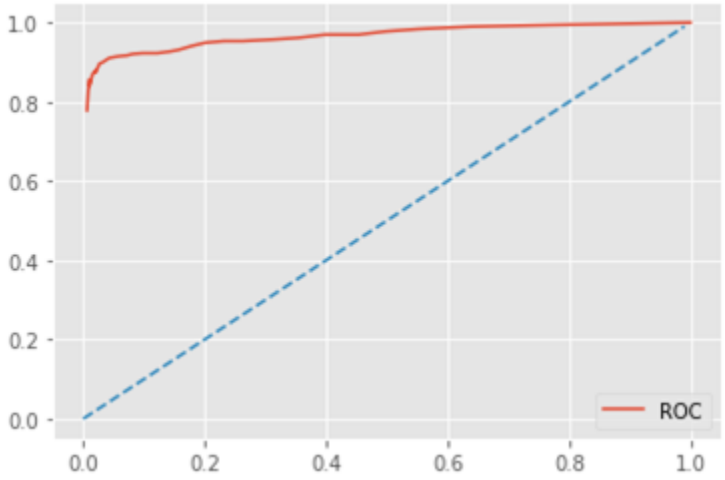


This model results in a 93% accuracy and the positive predictive value is at 95%, with a true positive rate of 91%. Of course, for the response of 'Class', 'Time', 'Amount', 'V10', 'V14', 'V12', 'V2', 'V4', 'V11', 'V19' all have an important contribution to the success of this model. Based on the datasets used, the model's prediction for fraudulence using the Matthews Correlation Coefficient (MCC) metric is also performed well – 86%. Having a Youden's J value of 0.86 indicates a very high probability of an informed decision about the classifier ('Class'). Youden's J is the likelihood of a positive identification of credit card fraud result versus non-fraudulent. It is also the probability of an informed decision, as opposed to a random guess.

Youden's J statistic, 'Informedness', represents the vertical distance of the ROC curve from the no discrimination line. It can take values between 0 and 1. Thus, having a value of 0.86 indicates the probability of an informed decision about the classifier ('Class'). A final validation check is where the ROC curve was plotted. Note that, the further away the ROC curve (see Figure 14) is from the diagonal (i.e., the closest it is to the top left corner), the better the classifier is (Shmueli, 2019). In this case, the classifier is the 'Class' dependent variable, which indicates fraudulent transactions vs non-fraudulent transactions. Also, further the area under the curve (AUC), it denotes the overall performance of the classifier (Shmueli, 2019). Based on all this, the analysis is that the AUC = 0.96 resulted in an excellent performance of the classifier.

Figure 14

The Probability of an Informed Decision



Conclusion

The focus of this research has been credit card fraudulent versus non-fraudulent transactions based on real cardholders' data. Although several methodologies were identified and described, the consensus is that Logistic Regression Model is and would dominate in identifying suspicious transaction activities in financial and other credit card issuing companies. Findings from our study indicate that using this model can help outline specific transactions over time based on cardholders' purchasing pattern whether it's in e-commerce or in-store.

Judging from the above, the overall security measures in fighting credit card frauds would continue to attract attention of all cardholders, FIs, merchants and others.

Implications for Future Research

After the research and data analysis we found that there is much work done on credit card fraudulent transactions. We think that this research has enough idea for future researches to develop upon, including how to identify fraudulent transactions versus non-fraudulent transactions from a FIs and/or merchants' standpoint. And further research may be one can compare other countries' cardholders' transaction data to merchants, or even FIs' in-house data for a more real-time comparsion.

References

Frei, L. (2019). *Detecting Credit Card Fraud Using Machine Learning*. Towards Data Science.

George, J. A. (n.d.). *Credit Card Fraud Detection: A Case Study for Handling Class Imbalance*. Medium.

Government of Canada. (2020). *Credit Card Fraud*. https://www.canada.ca/en/financial-consumer-agency/services/credit-fraud.html

Hoffman, J. I. E. (2019). *Logistic Regression Analysis.* Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition). https://www.sciencedirect.com/topics/medicine-and-dentistry/logistic-regression-analysis

Mishra, R. (2019). *Introduction to ROC Curve*. Data Science Blog.

Shmueli, B. (2019). *Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of*. Towards Data Science.

The Paypers. (2019). Payment Methods Report 2019. https://www.europeanpaymentscouncil.eu/sites/default/files/inline-files/Payment%20Methods%20Report%202019%20-%20Innovations%20in%20the%20Way%20We%20Pay.pdf

Appendix A

```
with pm.Model() as logistic_model:
    pm.glm.GLM.from_formula('Class ~ Time + Amount + V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V
    trace = pm.sample(950, tune = 950, init = 'adapt_diag')

az.plot_trace(trace);
```

Appendix B

Results: Logit

Model:	Logit	Pseudo R-squared:	0.686
Dependent Variable:	Class	AIC:	446.4056
Date:	2020-12-11 09:35	BIC:	490.4303
No. Observations:	984	Log-Likelihood:	-214.20
Df Model:	8	LL-Null:	-682.06
Df Residuals:	975	LLR p-value:	1.1183e-196
Converged:	1.0000	Scale:	1.0000
No. Iterations:	10.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Amount	0.0001	0.0007	0.0864	0.9311	-0.0014	0.0015
Time	-0.0000	0.0000	-11.8950	0.0000	-0.0000	-0.0000
V10	-0.1737	0.1309	-1.3268	0.1846	-0.4303	0.0829
V14	-0.8421	0.1171	-7.1943	0.0000	-1.0716	-0.6127
V12	-0.4075	0.1114	-3.6584	0.0003	-0.6258	-0.1892
V2	-0.1407	0.1168	-1.2052	0.2281	-0.3696	0.0881
V4	0.5666	0.0858	6.6062	0.0000	0.3985	0.7347
V11	-0.0793	0.1172	-0.6767	0.4986	-0.3091	0.1504
V19	0.0242	0.1439	0.1679	0.8667	-0.2579	0.3062