IEEE-CIS Fraud Detection ADS Final Report

Amr Al Kozbari (aak661), Saahil Chamdia (sc7069)

May 2021

1 Background

The Institute of Electrical and Electronics Engineers (IEEE) partnered up with Vesta Corporation to host a competition on Kaggle to develop a Automated Decision System (ADS). The aim of this ADS is to detect fraudulent transactions using customer transaction data.

The main stakeholders that are funding this competition are IEEE and Vesta Corporation. The Institute of Electrical and Electronics Engineers is an institute with aims in advancing nature-inspired computational paradigms in science and engineering. Vesta Corporation is a company that applies ADS systems to identify emerging threats and maximize merchant revenue and security.

Furthermore, The Kaggle competition is set up in order to create a collaborative attempt for an ADS to be created, without the input of the hosts, who already have a fraud detection ADS in place, with hopes of an ADS that is more accurate and having better efficacy rates than previously.

1.1 Competition Motives

There are several benefits to the implementation of this ADS across the different stakeholders. Vesta Corporation benefits from the ADS created in this competition as it can potentially provide more accurate results to its merchants, allowing for an increase in business confidence which would result in revenue increase. The merchants using the results of the ADS benefit as it can further prevent potential losses as a result of fraud. Lastly, the consumers benefit, as though being flagged for fraudulent charges is inconvenient, a more accurate score from the ADS decreases the potential times that a consumer would be flagged, therefore improving customer satisfaction. The consumer satisfaction perpetuates back to the merchants and Vesta corporation, therefore allowing all stakeholders to benefit. IEEE is bipartisan in this analysis as their mission statement is to solve the problems that is currently occurring by advancing the accuracy of the ADS in comparison to what was previously used. Therefore, IEEE does not benefit from the implementation of the ADS, however, their reputation as researchers would increase due to their joint funding in this competition.

2 Data Exploration

The data provided in this competition comes from Vesta's real-world e-commerce transactions and was independently collected by them. They provide 2 datasets: Identity, and Transaction. Identity pertains to user information such as device type, OS, email domain etc. while Transaction pertains to individual transaction details such as card type, transaction amount, account type etc. Both these datasets contain the same primary key called TransactionID, meaning that we can link individuals in the identity dataset to specific transactions in the Transaction table. After exploring both datasets, it appears that the Transaction table is more complete with significantly fewer null values.

In each data set that is provided, there are different input variables. In the transaction files, there are 6 types of categorical features: ["ProductCD", "card1"-"card6", "add1", "add2", "P_emaildomain", "R_emaildomain", "M1" - "M9"]. In the identity features, there are 3 types of

categorical features: ["DeviceType", "DeviceInfo", "id_12" - "id_38"]. Overall, there are 799 float columns, 38 string columns, 24 boolean columns and 10 object columns.

Due to the vast breadth of the number of input features, only a few input features were explored. While Vesta provides the real classification of transactions for the training set, they do not provide the real classification for the test set. This is understandable as they would not want participants in this competition to tune their models in accordance to their test data. Hence, to assess the accuracy of his model, the author splits the training dataset into a training set and a validation set.

Overall, the vastness of this dataset makes it seem sufficient to assess if a transaction is fraudulent or not.

2.1 Feature Engineering

Due to the vast breadth of the number of input features, only a few input features were explored. For example, input feature "id_03" contains almost 98% values that are NULL or 0, while "id_07" contains 76% NULL values, which is less than "id_03". Furthermore, even categorical data such as "DeviceType" contained a count of 449,730 NULL values, while the other values desktop and mobile contained only 85,165 and 55,645 counts respectively.

The implementer dropped all columns that contained a NULL to object ratio greater than 90% in that column. Furthermore, the implementer dropped columns that has a frequency of unique values in that column greater than 90%. Lastly, the implementer removed columns in which only contained one unique value.

The implementer also created new columns in the train and test sets. The new columns contained the calculation of a feature in ["TransactionAmt", "id_02", "D15"] to a feature in ["card1", "card4", "addr1", "addr2"]. All the features were present in the test and training set, and the values in "TransactionAmt" and "id_02" column was divided by the mean and standard deviation of "card1" and "card4" in each respective dataset. Moreover, the values in the "D15" column was divided by the standard deviation and mean of all ["card1", "card4", "addr1", "addr2"] in each respective dataset as well. When running the ADS with this information, it can be seen from the feature importance graph that the calculated values are highly important in the prediction of the fraudulent score.

3 ADS Exploration

As accuracy of the model on the test dataset cannot be determined (due to the absence of the real classifications of the test set), the author uses the accuracy of the validation set to assess the model's overall accuracy. This accuracy is measured as the area under the curve (AUC) for a Receiver operating characteristic curve (ROC Curve).

3.1 ADS Model

The ADS allows users to choose the type of model to implement when classifying transactions as fraudulent or not as long as the model is part of the 'sklearn' package. However, the author has set the default to be a Light Gradient Boosting Machine as that provides the best accuracy for his validation set and is not as resource intensive as the other options. The author of the code has also tuned the model to optimize its classifications. These parameters include features like the maximum number of children for a tree, the maximum depth, etc. While some of these hyperparameter tuning are performed to optimize accuracy, the author states that some are also incorporated to decrease runtime and reduce the resources utilized by the model during runtime.

The ADS runs the model over 5 folds, allowing for out of fold predictions to be made as well to determine the accuracy of the ADS with data not used in the training set. The results that are procured from this model is a float ranging between 0 and 1 that classifies if a transaction is fraudulent or not. The value can be interpreted such that the closer the prediction is to 1, the greater the probability that the transaction is fraudulent. For this report, we chose the cut-off to

be 0.8 under the assumption that for the sake of good user experience, card companies would not want to flag transactions they are not very sure are fraudulent as fraudulent.

3.2 Model Output

The original code by the author calls onto a method called 'train_model_classification' which initializes a sklearn model, runs it over the folds, and then returns a dictionary with relevant information on the classification of the best model. However, for the sake of analysis, this code was slightly modified to obtain the model and the variables created by the function that were then used for a deeper analysis of this ADS. This code also develops a feature importance graph which is displayed below:

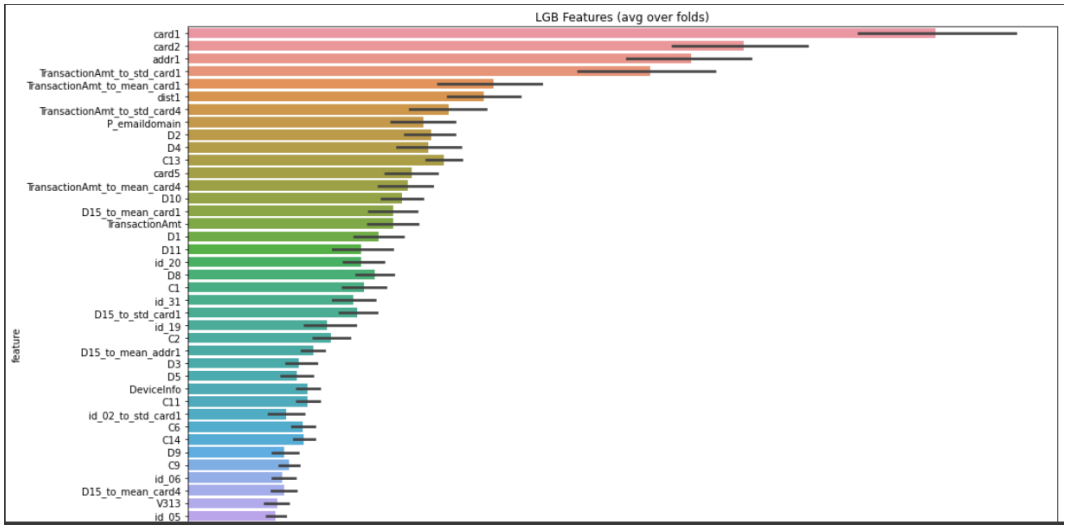

Figure 1: Ranking of The Important Output Features

As seen in Figure 1, the the card type, address, and transaction amounts play a big role in the classification process of the model. However, due to the lack of variable descriptions and the broad range of card1 and card2 (0-17090), it is difficult to decipher the meaning behind the those features. The role transaction amounts and their relations to the mean transaction amount in classification is understandable as transaction amounts are critical to the comprehension of a transaction and its nature.

	Not-Fraudulent(predicted)	Fraudulent(predicted)
Not-Fraudulent(Real)	113945	99
Fraudulent(Real)	3024	1040

False Positive Rate: 0.0008680860018940058
False Negative Rate: 0.2559055118110236
Accuracy: 0.9735580993666814

Figure 2: Overall Model Performance

Figure 2 depicts the overall performance of the model using a confusion matrix, followed by the false positive rate, false negative rate, and accuracy of the model. Overall, the model boasts a very low false positive rate of 0.0009 as only 99 transactions out of the immense validation set were falsely flagged as fraudulent. However, the false negative rate was sub-par at 0.2560 as 3024 out out the total of 4064 transactions flagged as fraudulent were falsely flagged.

The accuracy of the model was a high 0.9736 but this does not mean that the model performs well. We say this because the accuracy is reinforced by a high rate of true negatives, but this can be attributed to the fact that very few transactions in the training set are fraudulent. When looking at just the number of fraudulent transactions caught by the model, only 26% of all fraudulent transactions in the validation dataset are caught. Hence, despite the high accuracy of the model, the model does not do a very good just capturing fraudulent transactions.

Aside from assessing the model holistically, we chose to explore certain categorical features that could be the basis for bais. Within the features listed in Figure 1, 'card4' and 'card6' caught our attention as 'card6' is a categorical feature that depicts the Company that provides the card and 'card4' depicts the binary classification of the transaction as a credit transaction or a debit transaction. Hence, we further explored the implication of those features and how the model used them to influence it's' outcome.

3.3 Exploration of feature: 'card4'

'card4' refers to the providers of the card that was used to make a transaction. There are 5 categories within this feature and they are "American Express", "Discover", "MasterCard, "Visa", and "null". We thought that having null fields could hinder our exploration but the count of null values are fairly low and hence, we believe it has a negligible impact on the outcome. First, we checked to see if there was any disparity in the rate of fraud for the different cards in the validation set (we use validation as the organisers did not provide the real classification of transactions in the test set).

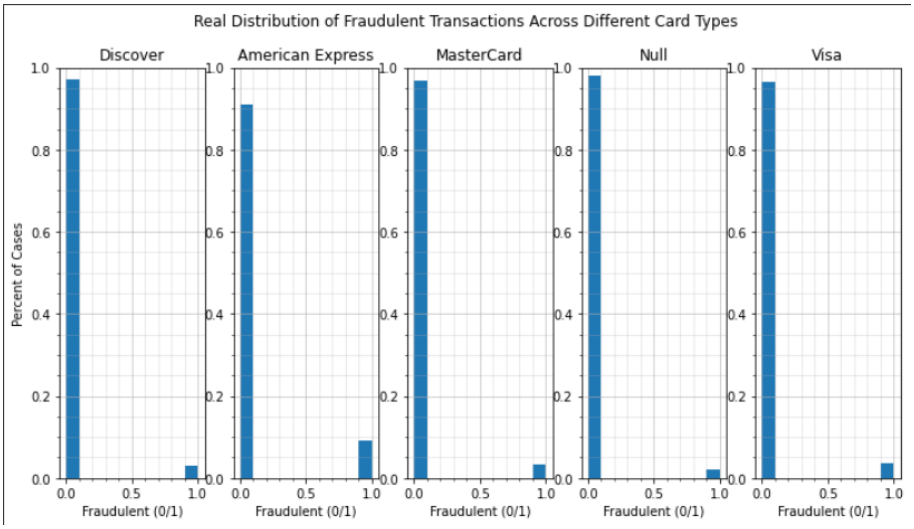

Figure 3: Real Distribution of Fraudulent Transactions Across Different Cards

As seen in figure 3, American Express seems to have a higher rate of fraudulent transactions when compared to the other cards in this dataset. Almost 0.1% of all transactions made with an American Express were in reality fraudulent while the rate for other cards are consistently close to 0.03%. This seemed like a big disparity and hence, we wanted to check if the model would use this as a basis to establish a bias against transactions on American Express cards. Despite American Express having a noticeably higher rate in fraudulent transactions, the fraction of fraudulent transactions on either card is very low. Hence, we believed that this disparity was not a valid reason to discriminate on the service provider. Therefore, we obtained the indices for all transactions for each card and plotted the percent of fraudulent transactions predicted by the model for each to see if it used American Express as an indicator of fraudulent transactions.

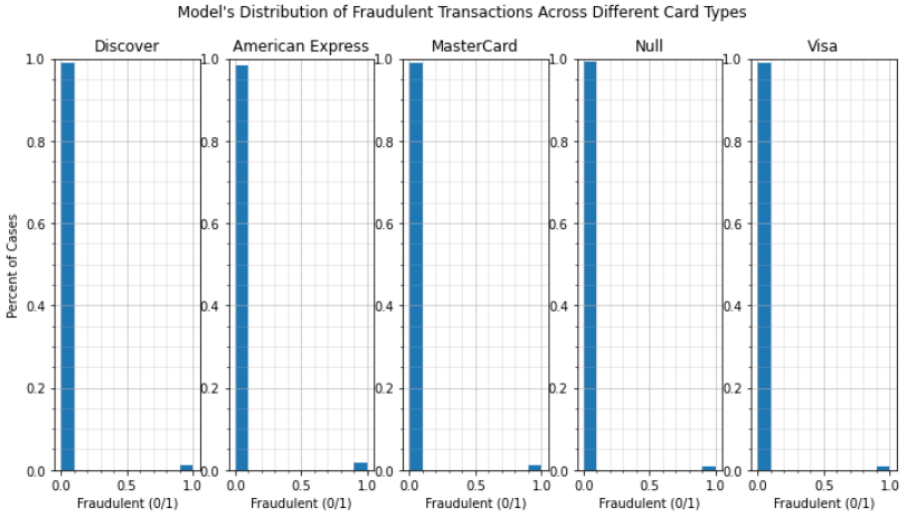

Figure 4: Model's Distribution of Fraudulent Transactions Across Different Cards

As seen in figure 4, the model did not capture the trend that was noticed in the distribution of fraud across different service providers. To verify this fairness, we developed confusion matrices for all card services and calculated the False Positive Rate, False Negative Rate, and Accuracy. The following table reflects the aforementioned metrics:

	FPR	**FNR**	**Accuracy**
American Express	0.0009	0.1858	0.92601
Discover	0.0009	0.3636	0.9805
MasterCard	0.0009	0.3024	0.9768
Visa	0.0008	0.2361	0.9725
Null	0.0027	0.2667	0.9825

Figure 5: Fairness Metrics Across Different Service Providers

As seen in figure 5 (neglecting 'null' as a very few percentage of transactions had that label), the False Positive Rate was consistent throughout the different card providers. However, transactions made using an American Express card saw a lower false negative rate and a lower accuracy than the rest. The lower false negative rate suggests that the model is not capturing fraudulent transactions on American Express Cards at a higher rate than the other cards. This is expected because- like we saw in figure 3- the model does not reproduce the same distribution of fraudulent transactions across cards as the real distribution. Nevertheless, we believe that this disparity is acceptable as the volume of fraud transactions on the card are not sufficient to justify discriminating on the card type. As the the model already does not rank 'card4' very high in its feature importance, we believe no intervention is required to combat a potential bias.

3.4 Exploration of feature: 'card6'

'card6' is a binary classification of transactions as credit and debit. There are instances where a transaction is labelled as 'null' but those are minimal. For the sake of this analysis, the null transactions were treated as debit transactions.

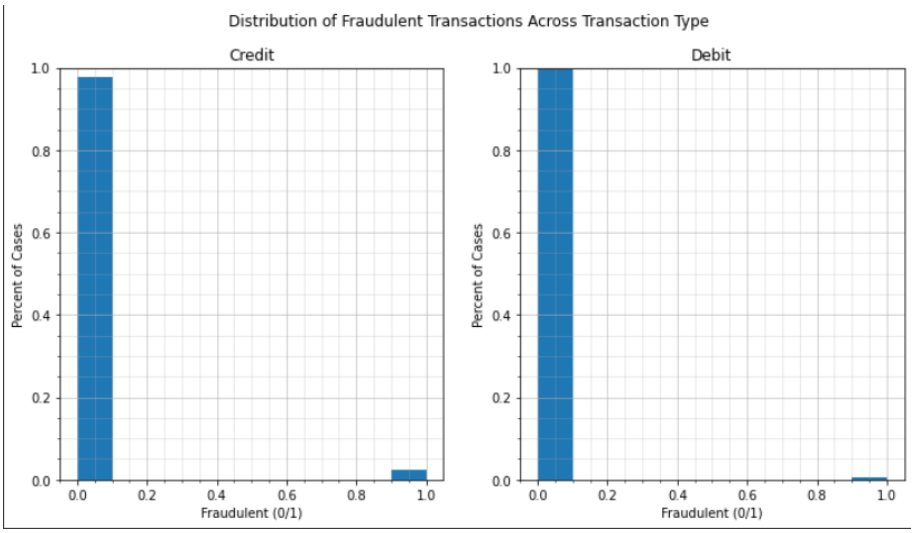

Figure 6: Distribution of Fraudulent Transactions across Transaction Types

As seen in figure 6, credit transactions saw a slightly higher percentage of fraudulent transactions in the the validation dataset. For assessing the fairness of the model in handling credit transactions versus debit transactions, we created confusion matrices for each respective type to see disparity in the accuracy, false positive rate, and false negative rate.

	FPR	**FNR**	**Accuracy**
Credit	0.0028	0.3123	0.9530
Debit	0.00030	0.2128	0.9797

Figure 7: Fairness Metrics Across Different Transaction Types

Figure 7 reveals that while the accuracy for both divisions are similar, Credit sees higher false positive rates and false negative rates. This is interesting because this insinuates that the model seems to be overall, a little worse at classifying transactions made on credit cards as opposed to debit cards.

As Credit cards are issued by financial institutions that bear the financial cost for a transaction, we reasoned that for individual consumers, it is better to fall victim to credit card fraud than debit card fraud as card issuing companies would be more vigilant and more persistent for credit fraud cases. Hence, when deciding to calculate disparate impact, we decided that our privileged class would be credit transactions and our underprivileged class would be debit transactions.

To perform in depth fairness metrics, we initially tried to make use of ai360 as they have a plethora of methods to aid with such an assessment. However, with investigation, we realised that when converting our dataset to a StandardDataset within ai360- something that is required to make use of ai360 methods- the algorithm dropped rows with null values. As the dimensionality of our dataset is very high, and there exist numerous columns with null values, converting to a StandardDataset resulted in us losing a lot of statistical power due to reduced dataset sizes. Hence, we resorted to calculating disparate impact manually using the following relation:

Outcome	$X = 0$	$X = 1$
$C =$ NO	a	b
$C =$ YES	c	d

Tab. 1: A confusion matrix

The 80% rule can then be quantified as:

$$\frac{c/(a+c)}{d/(b+d)} \geq 0.8$$

Figure 8: Disparate Impact Formula

The disparate impact we gained from this methodology was 0.2495. This value is much lower than 1, implying that the outputs of this model are generally more favorable for the privileged class as compared to the underprivileged class. However, as the 'favorable' outcome in this situation refers to a transaction being classified as fraudulent, it is actually the underprivileged group that benefit from the classification of the current model. Hence, debit transactions are less likely to be flagged as fraudulent by our model. While we could not implement any debiasing methods to address this disparity, we believe that during the actual implementation of the ADS, Vesta has the scope to address this imbalance.

4 LIME Explainer

A preface must be made before discussing the analysis of the LIME Explainer. Our data had to have artificial injection of values, such that the means of every column were put into values that were NULL. To this effect, it was not possible to use the LIME package without such manipulation of the data as our data contained NULL values within all columns, as explained in our data exploration. Furthermore, other options were attempted such as dropping all NULL values, which resulted in the entire table being cleared. Therefore, we wanted to preface that the analysis through LIME Explainer, we acknowledge that the results are not optimal due to the data imputation. However, it had to be done in order for the analysis to occur.

The explainer instance was data from X_train_filled, our X_train with imputed data, and analyzed through classification. We then wanted the explanation of three examples. Firstly, an example with the highest prediction, secondly, an example with the lowest prediction, and lastly, an example that is predicted above 0.5, however is not fraudulent.

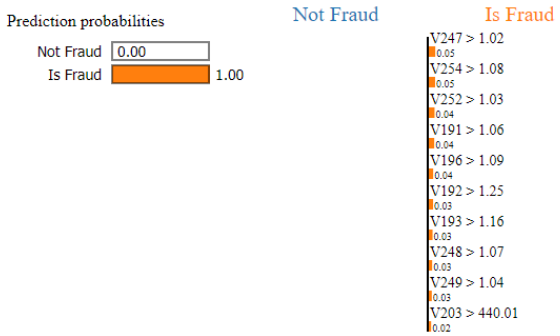

Figure 9: LIME Explainer Highest Prediction

When analyzing the transaction in which the model predicted fraudulence at 0.9993, and was

correct in its prediction when compared to the actual output. Comparing the prediction of this row with the highest feature importance list, it can be seen that features specified in this row as having high weights for fraudulence. This is because that row had values that were abnormal in the models classification of whether a transaction is fraudulent or not. Therefore, when looking at the LIME explanation in figure 9, it can be seen that the features with most importance were classified as fraudulent, which were features that had over 90% NULL values, and features that we expect the ADS to flag due to lack of training on.

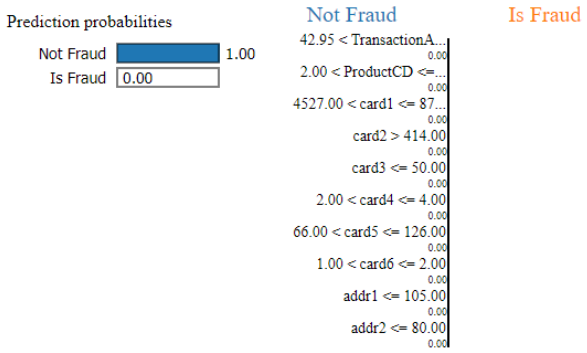

Figure 10: LIME Explainer Lowest Prediction

Furthermore, when analyzing the transaction in which the model predicted fraudulence at 0.000114, it was correct in its prediction when compared with the actual output. It can be seen from figure 10 that the features of that row had features similar to the ADS feature importance list where they were given positive weights, and given that the weights are positive and are of high importance, this transaction was classified as not fraudulent. Therefore the model was able to correctly classify transactions with features that had numerous data.

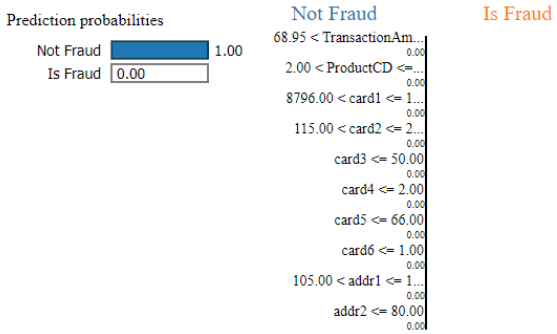

Figure 11: LIME Explainer Random Prediction

Lastly, when analyzing a random transaction from the model with a prediction of 0.676 fraudulence, it was seen that that transaction's output was not fraudulent. From figure 11, the feature importance ranking is similar to the feature importance ranking in Figure 10. Therefore, it can be interpreted that even though the transaction in Figure 11 had features that would aid in it's classifications against fraudulence, it also contained features that would enable it to be considered fraudulent, which is important for a fraud detector to be able to determine as transactions contain numerous features. Therefore, given that the model was trained to be able to classify this transaction correctly, there are certain implications to LIME and there model that are further explained in the conclusion.

5 Conclusion

Overall, the data was and was not appropriate for this ADS. We believe that the dataset was appropriate as it is reflective of the collection of data that Vesta uses when running their ADS. Such that transactions often contain unique meta data based upon the method that the transaction occurred. For example, a transaction by an individual who uses their desktop computer to purchase an item will have a different metadata than a transaction from a cafe in another city. Such information collected are the features the ADS uses to classify a transaction, and are over 90% of all the data presented in the datasets. Therefore the data forces the architect of an ADS to determine a model that uses the most important features. However, we do not believe the data is appropriate for this ADS due to the lack of an independent validation dataset. The ADS split up the training data to have a validation set and a training set. Therefore, from the data processing that was done before the ADS was deployed - calculating mean and standard deviation - the split leaks information from the categories that are being analyzed to the validation set, due to the calculations being done originally on the training set. Therefore, the accuracy that we calculated is potentially incorrect.

Moreover, only 1040 transactions out of a total of 4064 fraudulent transactions are captured by the model. Hence, in essence, the model only captured 26% of all fraudulent transactions in the validation dataset. As this ADS was awarded a gold and scored a 0.9401 on the Kaggle competition, there is validity towards the accuracy of the ADS. One reason why the Vesta might still consider this a viable solution could be due to the complexity of this task. Despite a a low capture rate of fraud, Vesta attests to the quality of this model meaning that even this high dimensional dataset might prove insufficient for an accurate capture of fraudulent transactions. Hence, the high false negative rates translating into a poor capture rate of fraudulent transactions, and the existence of data leakage hinder the robustness of this model, making it less reliable.To optimise that, we believe that data collection process could be improved. Given the fact that the data contains a high amount of NULL values, a training dataset with fewer NULL values, representing a more complete dataset would allow for a better training of the model. Vesta could implement stronger requirements of data sharing when a transaction is made to ensure that more complete data is collected for each transaction. Furthermore, a validation set provided by Vesta that is independent of the training set would prevent over training, which would lead to a more robust ADS. If Vesta cannot provide a validation set, the author could perform feature engineering during the running of the model rather than at the start, as that would decrease the amount of data leakage.

The model does good in terms of fairness. A lot of the higher ranked features are not categorical and within the the categorical features we explored in our analysis, the model did not show a great degree of bias. for 'card4', which represents the service providers, we noticed that there was a disparity in the distribution of fraudulent transactions across different service providers, with American Express having the highest percentage of fraudulent transactions. As the volume of fraudulent transactions are low, we believed this was not enough justification to discriminate on the feature. After assessing the distribution of the model's prediction, we saw that the model does not discriminate on this feature, although the accuracy and false negative rate for transactions made on an American Express card were lower. We also looked at the binary feature 'card6' that captured whether a transaction was on a debit card or a credit card. We took the credit to be the privileged group because a credit transaction is incurred by a corporation and hence, they would be more likely to have robust methodologies and persistence to crack down on fraud. We calculated the disparate impact and got a value of 0.24. This signifies that credit transactions are much more likely to be flagged as fraudulent when compared to debit transactions. We did not have the industry knowledge to determine if this is discriminatory or not as it could be possible that fraud on credit cards differs from fraud on debit cards. Hence, this would be something the Vesta or the implementer of this code could consider when optimising this code prior to implementation.

We do feel comfortable deploying this ADS in the public sector. The model has a high accuracy towards non fraudulent charges, which as stated before, benefits the consumers who are interact-

ing with the vendors. In addition, the feature engineering that was done to the ADS allowed for standardization of features that were important such as 'card_1', which as explained were highly important in the determining of fraudulence. However, we would be more comfortable deploying this ADS if the false negative rate were to decrease by modifying the ADS, which could allow for a fairer risk prediction.

In conclusion, while a multitude of types of analysis was done on this ADS, the main hindrance to our analysis was the lack of complete data. Many analysis packages did not allow for the analysis of data containing NULL values.Therefore, after determining that the best solution was to impute the data, we acknowledge that the analysis would contain undetected bias. However, even with the imputation of the data, we are confident that our analysis contains valid arguments for how to fix the ADS for optimal implementation.

References

[1] Feldman, Michael, et al. "Certifying and Removing Disparate Impact." Arxiv, 16 July 2015, arxiv.org/pdf/1412.3756.pdf.

[2] Vesta Corporation.$test_identity.csv.Oregon, VestaCorporation, (Datafile), 17March2021,$ *https://www.kaggle.com/c/ieee-fraud-detection/data?select=test_identity.csv*

[3] Vesta Corporation.$train_identity.csv.Oregon, VestaCorporation, (Datafile), 17March2021,$ *https://www.kaggle.com/c/ieee-fraud-detection/data?select=train_identity.csv*

[4] Vesta Corporation.$test_transaction.csv.Oregon, VestaCorporation, (Datafile), 17March2021,$ *https://www.kaggle.com/c/ieee-fraud-detection/data?select=test_transaction.csv*

[5] Vesta Corporation.$train_transaction.csv.Oregon, VestaCorporation, (Datafile), 17March2021,$ *https://www.kaggle.com/c/ieee-fraud-detection/data?select=train_transaction.csv*