Ryan Long
DSC 680-T301
7.2 Project 2: Whitepaper and Q&A

Business Problem
The objective of this project is to use an XGBoost model to predict fraudulent credit card transactions to prevent losses to financial institutions.

Background/History
Fraudulent transactions and costs associated with those transactions continue to rise. It is claimed it costs an organization approximately $3 for every $1 of a fraudulent transaction (1). Additionally, fraud costs the global economy upwards of 5 trillion annually (2). Given the prevalence and scale, organizations should focus on preventative measures. While not every instance can be prevented due sheer volume and ever-changing techniques of fraudsters, even a marginal investment would have a short payback period if fraudulent transactions could be halted prior to clearing.

Data Explanation (Data Prep/Data Dictionary/etc.)
A dataset of credit card transactions was obtained from Kaggle. The selected dataset contains 1,000,000 observations and is relatively pre-processed with 8 variables, 7 of which are independent and 1 containing the target or dependent variable of fraud. The below table provides an overview of the dataset features. See Appendix 1 for additional information on the dataset.

Feature	Feature Description
Distance from home	Distance from home where the transaction happened.
Distance from last transaction	Distance from last transaction happened.
Ratio to median purchase price	Ratio of purchased price transaction to median purchase price.
Repeat retailer	Whether the transaction happened from same retailer
Used chip	Whether the transaction through chip (credit card)
Used pin number	Whether the transaction happened by using PIN number
Online order	Whether the transaction an online order.
Fraud	Whether the transaction fraudulent

For the fraud target value, the following illustration depicts the distribution within the dataset:

Value	Count	Frequency (%)
0.0	912597	91.3%
1.0	87403	8.7%

Methods

For language and IDE, Python in Jupyter Notebook was used. Libraries leveraged were Pandas, NumPy, Seaborn, XGboost, sklearn, and pandas profiling.

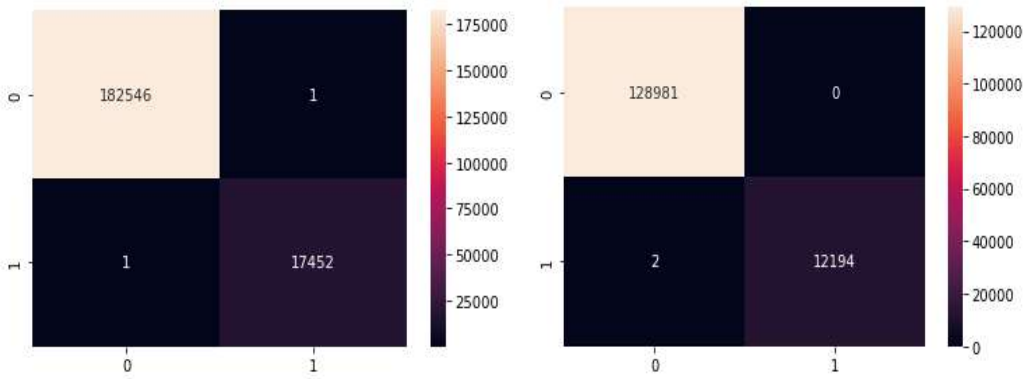
An XGBoost algorithm, a supervised learning decision tree model, was utilized for this project due the high execution speed and general model performance. The data was split into training (65%), test (20%), and validation (15%) sets. All seven of the above independent variables were used to predict the target fraud value. To evaluate the model, various metrics and tools were used, such as accuracy, confusion matrix, and classification reports for both the test and validation sets.

Analysis

Accuracy of the model using both the training and validation sets were very favorable as can be seen below, though this shouldn't be considered only metric to evaluate the model.

```
Accuracy on training set: 1.000
Accuracy on validation set: 1.000
```

The below confusion matrices illustrate the results from both the test set (left) and the validation set (right) of the model. Very few false positives and false negatives predicted by the model in both the testing and validation sets. True label is depicted on the y-axis, with predicted value on the x-axis.



The classification report for the validation set further suggests the model's strong performance for all metrics of precision, recall, and f1-score.

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	128981
1.0	1.00	1.00	1.00	12196
accuracy			1.00	141177
macro avg	1.00	1.00	1.00	141177
weighted avg	1.00	1.00	1.00	141177

Conclusion

Overall, the model shows strong performance and results in both the testing and validation sample sets. Both sets had relatively immaterial false positives and false negatives as displayed by the confusion matrices. Additionally, accuracy, precision, recall, and f1-score were all strong. With the strong performance of the model in its current state, it is ready for deployment with the assumption of all features being readily available in production.

Assumptions / Limitations / Challenges

The model results are almost too good to be true, which warrants further exploration of the dataset and confirmation of the modeling approach prior to deployment. Financial institutions may not initially obtain the features contained within this dataset. Distance from home, distance from last transaction, and ratio to median purchase price are all computed features. While all great indications of potential fraudulent transactions, they would need to be established and calculated prior to deployment of this model by a financial institution.

Future Uses/Additional Applications

As the model performs relatively well with the given dataset, future uses would include a deployment and implementation in a real world setting to predict and prevent fraud. This dataset is focused on fraudulent transactions; however, other fraud schemes exist which financial institutions would want to prevent. A similar approach could be taking with credit card and loan application processes or check kiting scheme as well. Outcomes would be heavily influenced on the independent variables and features which could be captured.

Recommendations

Additional consideration should be made to fine tune the model to determine if there can be additional features or a reduction in features without significant detriment to the current performance of the model. This could enhance applicability and reduce the independent variables needed to be gathered in the field for prediction.

Implementation Plan

After the model is fine-tuned and finalized it should be deployed with live data to evaluate transactions before they are fulfilled. The predictive output of the model should flag transactions for review by a human analyst. The analyst can then determine whether to allow the transaction to be fulfilled or follow-up with the customer based on information provided.

A post implementation feedback loop should also be considered. The purpose would be to evaluate the live model performance and potential changes to consider if fraudulent transactions are not being flagged appropriately. Additional human review decision tree matrices could be augmented based on the model performance and changes in behaviors of fraudsters.

Ethical Considerations

Credit card information typically is protected by PCI (payment card industry) standards which provides safeguards for consumer information. This dataset does not contain PCI data, but future developments should keep this restriction in mind. Additionally, it is possible for fraudsters to evaluate the methods use here or by other financial institutions to determine more sophisticated manners of committing fraud to avoid detection. Methods to determine fraud should be safeguarded by financial institutions to prevent this from occurring.

Appendix

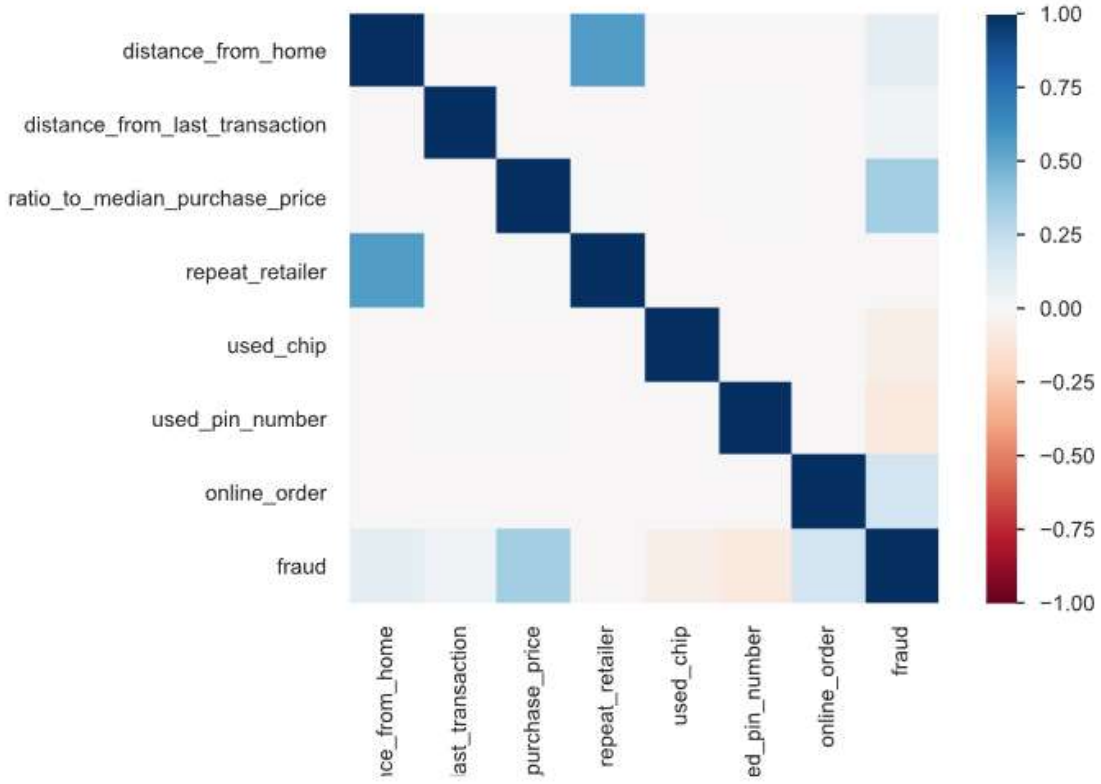
Dataset statistics

Number of variables	8
Number of observations	1000000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	61.0 MiB
Average record size in memory	64.0 B

Variable types

Numeric	3
Categorical	5

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 8 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   distance_from_home              1000000 non-null  float64
 1   distance_from_last_transaction  1000000 non-null  float64
 2   ratio_to_median_purchase_price  1000000 non-null  float64
 3   repeat_retailer                 1000000 non-null  float64
 4   used_chip                       1000000 non-null  float64
 5   used_pin_number                 1000000 non-null  float64
 6   online_order                    1000000 non-null  float64
 7   fraud                           1000000 non-null  float64
dtypes: float64(8)
memory usage: 61.0 MB
```


distance_from_home
distance_from_last_transaction
ratio_to_median_purchase_price
repeat_retailer
used_chip
used_pin_number
online_order
fraud
ıce_from_home
ast_transaction
purchase_price
repeat_retailer
used_chip
ed_pin_number
online_order
fraud
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Appendix 2 – References

(1) Lucas, P. (2021, August 24). *The latest LexisNexis report finds the cost of fraud is on the rise since the pandemic set in*. Digital Transactions. Retrieved October 1, 2021, from https://www.digitaltransactions.net/the-latest-lexisnexis-report-finds-the-cost-of-fraud-is-on-the-rise-since-the-pandemic-set-in/

(2) *Fraud costs the global economy over US$5 trillion* . Crowe Global. (n.d.). Retrieved October 1, 2021, from https://www.crowe.com/global/news/fraud-costs-the-global-economy-over-us$5-trillion

(3) DATA REFERENCE/SOURCE: https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud