Group Project-Credit card fraud

GROUP 21
Ding Jiayi
Huang Binqian
Huang Chengdong
Wu Tianchi
Zhang Miao

1. Background and objectives

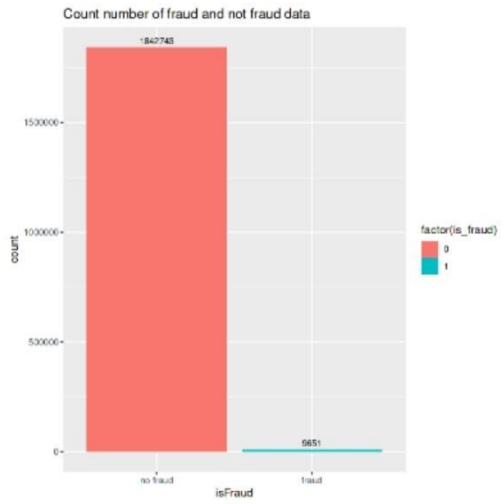
- **Background of the selected case:**
  This is a simulated credit card transaction dataset containing legitimate. This simulation was run for fraud transactions from the duration of 1st Jan 2019 - 31st Dec 2020. It covers the credit cards of 1000 customers doing transactions with a pool of 800 merchants.
- **Define scope of fraud data analytics:**
  Create algorithms and predictive models for credit card fraud detection based on historical transaction data from customer and merchant pools, provide cost-benefit analysis models to stakeholders, and provide them with appropriate recommendations to mitigate fraud risks
- **Identify fraud scenarios:**
  a.Fraudsters create fraudulent merchant accounts after stealing an e-commerce business's identification.The newly created "business" then posts charges to customers' credit cards, collects the money and closes the accounts.
  b.Fraudsters successfully impersonates the legitimate owner of a credit card by providing enough accurate personal data about a cardholder to convince merchants and payment processors the cardholder placed the order.
  c.Fraudsters uses the stolen personal information of account holder to fraudulently gain access to the account. They then use the account to make purchases the actual account holder did not authorize.

2.Fraud data analytics method

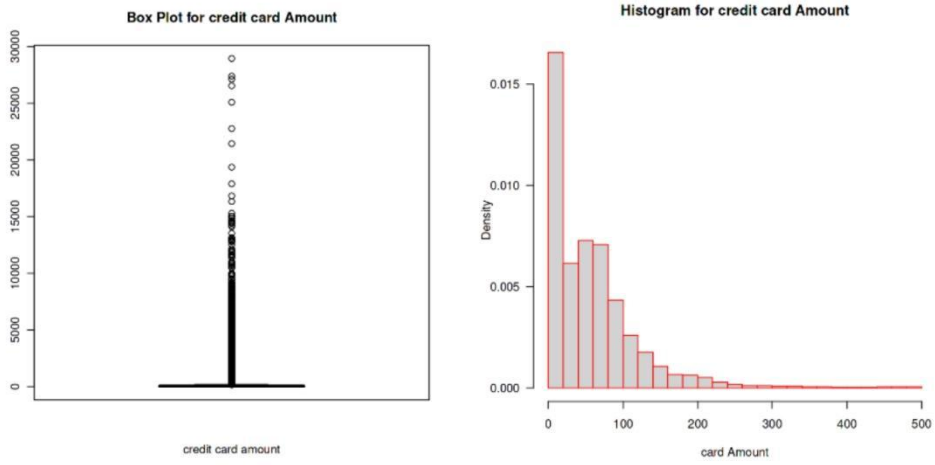
2.1 EDA(more details in appendix)

First, create new features from date column for better analysis. Combine train and test into a new dataset – creditcard

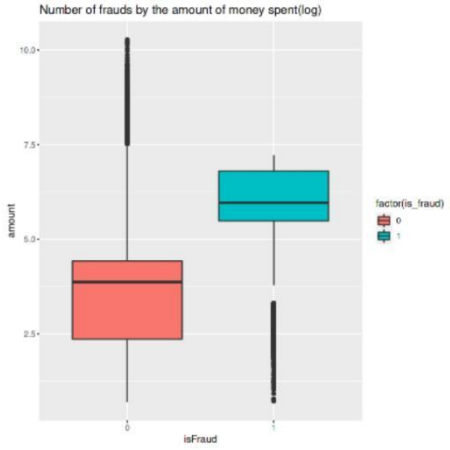
- **isFraud**
  Count number of fraud and not fraud data. Total number of data point is 1852394 in which 9651 is fraud and 1842743 is not fraud data point. The data only contains a relatively small number of frauds. From a logical perspective, it makes sense and is also advantageous for society that there aren't many frauds committed.



- **Amount**

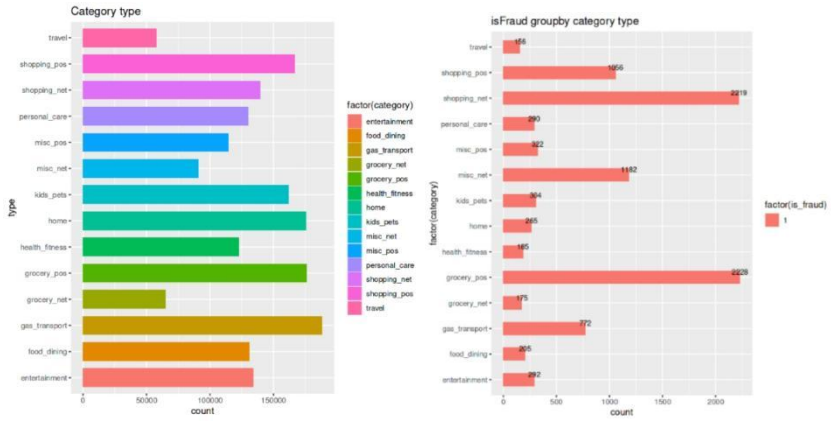


Box Plot for credit card Amount and Histogram for card Amount. The transaction amount of the credit card is about 0-30000, most of which are concentrated in 0-15000. It can be seen that most of the credit card transactions are small transactions.



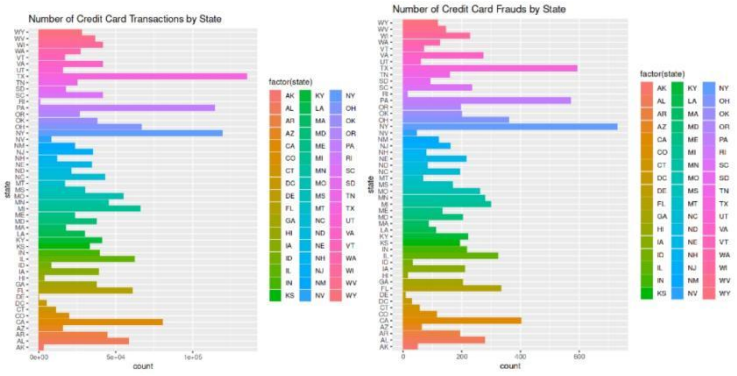
Number of frauds by the amount of money spent(log). Most not fraud are small transactions, while fraud is mostly large transactions.

- **Category**



Category type statistics. isFraud groupby category type. Most frauds occurred in categories of shopping_net and grocery_pos.

- **State**



Number of Credit Card Transactions by State. Number of Credit Card Frauds by State. States OH, TX and LA have the highest number of transactions and report the most number of credit card frauds.

2.2 Clean,Transform and Sampling the data

- **Clean the data**

  First, we need to confirm whether there are empty values or missing values in the dataset. Fortunately, there are no missing values and empty values in the dataset.

- **Transform the data**

  According to our EDA above, we found that the transaction amount, the age of the credit card holder, the consumption category, and the transaction time and location are all related to credit card fraud to varying degrees. This helps us choose which features to include in the data model. Since the data model requires numeric input, we need to convert some categorical observations and observations of non-numeric type to numeric ones. For age of the credit card holder, the original data is the date of birth (yyyy-mm-dd), we covert the dob to credit card holder's age. Age is calculated based on the accurate time of transaction in 2020 (trans_date_trans_time). For transaction date and time, we convert them to month, hour and week, For transaction location and merchant location, we already have vertical and horizontal data. For shopping categories, we convert string to numeric variables. Finally, for is_fraud, we convert the original labels Y and N to 1 and 0.

dob	age
1988-03-09	31
1978-06-21	41
1962-01-19	57

Change date of birth(dob) to age.

trans_date_trans_time	hour	week	month
2019-01-01 00:00:18	0	3	1
2019-01-01 00:00:44	0	3	1
2019-01-01 00:00:51	0	3	1

Change transaction date and time to hour week and month.

```
$ amt       : num  4.97 107.23 220.11 45 41.96 ...
$ category  : num  9 5 1 3 10 3 4 3 10 5 ...
$ state     : num  28 48 14 27 46 39 17 46 39 43 ...
$ lat       : num  36.1 48.9 42.2 46.2 38.4 ...
$ long      : num  -81.2 -118.2 -112.3 -112.1 -79.5 ...
$ merch_lat : num  36 49.2 43.2 47 38.7 ...
$ merch_long: num  -82 -118.2 -112.2 -112.6 -78.6 ...
$ age       : num  31 41 57 52 33 58 25 71 78 45 ...
$ hour      : int  0 0 0 0 0 0 0 0 0 0 ...
$ week      : int  3 3 3 3 3 3 3 3 3 3 ...
$ month     : int  1 1 1 1 1 1 1 1 1 1 ...
```

Change category and sate to numeric variable.

```
$ amt       : num  0.0284 0.6125 1.2572 0.257 0.2397 ...
$ category  : num  1.095 0.608 0.122 0.365 1.217 ...
$ state     : num  0.898 1.54 0.449 0.866 1.476 ...
$ lat       : num  0.928 1.258 1.085 1.189 0.988 ...
$ long      : num  -0.889 -1.295 -1.23 -1.228 -0.871 ...
$ merch_lat : num  0.926 1.265 1.11 1.21 0.995 ...
$ merch_long: num  -0.899 -1.295 -1.229 -1.233 -0.862 ...
$ age       : num  0.63 0.834 1.159 1.057 0.671 ...
$ hour      : num  0 0 0 0 0 0 0 0 0 0 ...
$ week      : num  0.7 0.7 0.7 0.7 0.7 ...
$ month     : num  0.142 0.142 0.142 0.142 0.142 ...
$ is_fraud  : num  0 0 0 0 0 0 0 0 0 0 ...
```

Nrmalization of the numeric data.

- **Sampling the data**

  ROSE, SMOTE and ADASYN were used for sampling and processing respectively.

  ROSE : Random Over Sampling Example

  SMOTE: Synthetic Minority Over-Sampling Technique

  ADASYN: Adaptive Synthetic sampling

TrainSet_Adasyn	2578229 obs. of 12 variables
TrainSet_Rose	1296675 obs. of 12 variables
TrainSet_Smote	2572695 obs. of 12 variables

2.3 Classification models

We have selected five classification models, namely

- Logistic Regression Model
- Neural Network Model
- KNN Model
- Random Forest Model
- XGBoost Model

We use grid search to adjust the parameters, using 70% of the Train_Dataset as Training Set,30% of the Train_Dataset as Validation Set.

2.4 Evaluation

We use six targets for Performance Evaluation, namely

- Confusion Matrix
- Recall (The most important performance measure for fraud detection models)
- Accuracy
- Precision
- F1 score
- AUC (important)

Then, we will select the best model, mostly based on the performance of Recall and AUC.

Evaluation: ROSE dataset

Model	Recall(most important)	Accuracy	Precision	F1 score	AUC(important)
Logistic Regression	0.74	0.96	0.07	0.13	0.85
Neural network	0.74	0.97	0.08	0.15	0.93
KNN	0.60	0.98	0.13	0.21	0.87
Random forest	0.74	0.98	0.14	0.23	0.93
XGBoost	0.75	0.97	0.09	0.16	0.92

Evaluation: SMOTE dataset

Model	Recall(most important)	Accuracy	Precision	F1 score	AUC(important)
Logistic Regression	0.75	0.95	0.06	0.11	0.85
Neural network	0.80	0.96	0.08	0.15	0.95
KNN	0.49	0.99	0.24	0.32	0.79
Random forest	0.93	0.89	0.03	0.06	0.97
XGBoost	0.96	0.82	0.02	0.04	0.91

Evaluation: ADASYN dataset

Model	Recall(most important)	Accuracy	Precision	F1 score	AUC(important)
Logistic Regression	0.77	0.87	0.02	0.04	0.85
Neural network	0.91	0.85	0.02	0.04	0.95
KNN	0.49	0.99	0.23	0.32	0.80
Random Forest	0.93	0.87	0.03	0.05	0.96
XGBoost	0.98	0.77	0.02	0.03	0.96

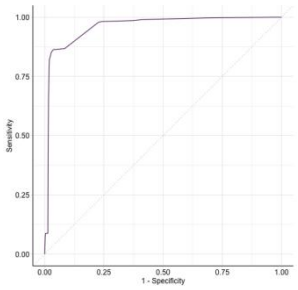
3. New fraud detection model

3.1 Final Model: XGBoost - ADASYN

Model	Recall(most important)	Accuracy	Precision	F1 score	AUC(important)
XGBoost	0.98	0.77	0.02	0.03	0.96

3.2 Confusion Matrix and ROC-AUC

		Actual	Actual
		Fruad	No Fruad
Precision	Fruad	2099	125545
Precision	No Fraud	46	428029



3.3 Advantages & Disadvantages

a.Advantages:

- **Highest Recall (98%)**

  When people use credit card, they want the deal done quickly. So for most credit card companies, they may not have enough time or ability to manually checking each deal, so they need to use a model to automatically check if the deal is a fraud one. Our final model has the highest recall. This will help the credit card companies to detect most of the fraud cases and keep the users away from losing property. In fact, recall is the most important measure for fraud detection model, our final model performs well.

- **Highest AUC (96%)**

  Second, our final model has high AUC. This means that most fraud and no fraud cases are correctly predicted, which means that our model is effective.

- **Train Fastest**

  The credit card companies have many deals every day, this means that the dataset will collect many new cases. To imporve the performance of current model, we need to use the new dataset and train the model again. It only cost about 30 seconds to train XGBoost for 1 epoch using TrainSet of over two million observations while training Random forest using the same dataset for 1 epoch cost about 1 hour. This means tha tthe company can more frequently retrain their fraud dection model, and the model can use more data to train. This can imporve the fraud detction model, which is important to the credit card company.

b.Disadvantages:

- **Low Precision**

  Our precision is quite low, about 2%. This may affect the credict card users. Many deals will be detected as fraud cases and rejected, which may affect the no fraud deals.

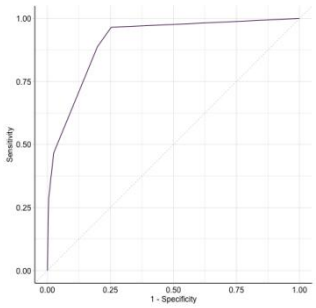
- **Why Precision is very low?**

  The most important reason is that our TestSet is not balance. There are only 2145 Fraud cases and 553574 No Fraud cases. We try to use Adasyn to balance the TestSet, and test our final model on the TestSet_Adasyn.

TestSet	Recall(most important)	Accuracy	Precision	F1 score	AUC(important)
Original TestSet	0.98	0.77	0.02	0.03	0.96
Adasyn TestSet	0.97	0.86	0.79	0.87	0.91

Confusion Matrix and ROC-AUC (Adasyn TestSet)

		Actual	Actual
		Fruad	No Fruad
Prec ision	Fruad	534359	140159
Prec ision	No Fraud	19202	413415



In fact, our final model is train on the balanced trainset. In the original trainset, the fraud and no fraud cases are unbalanced, we use adasyn to make the trainset balanced. Our final model is affected by the balanced trainset and will regard many no fraud cases as fraud cases in the unbalanced testset. When we balance the testset, the accuracy, precision, f1 score imporve significantly. Especially for the precision, it raise up to 0.79. So when we use balanced testset, the precision will be high.

3.4 Future Improvements

- Transform category and state into integers, (Try dummy variables before, but make the dimensions too high, which lead to worse recall and AUC), we can try to embed the category in future.
- Not keeping some variable to be integers while sampling, (State, Age, Hour, Week, Month ...), we can try to limit some variable to be integers while sampling in future.

4.Summary and recommendation

4.1 Risks and Red Flags

- Larger than expected orders. It occurs that there are orders that are much greater in monetary amount than the typical transaction. A thief with a stolen card number is aware that they only have a short window of opportunity before the crime is reported and the card is disabled. As a result, people frequently try to acquire as much as they can all at once.
- Multiple cards used for one order. We observe that a customer wishes to use multiple different credit cards to pay for a sizable item. The order may be being used by a fraudster to determine which card numbers would be accepted, enabling them to make more transactions with the legitimate cards.

- Different but similar card number. When analyzing fraud cases, there are some cases owing different but similar card. A list of credit card accounts may have been scraped by a criminal, who then swiftly placed orders on each card. This may be an indication of "card testing" or it might just be a careless, quick-fire fraud attempt.

4.2 Other non-data elements

- Credit history. Credit history was not analyzed in data analysis. Through observing each transaction the applicant dealt with, we could find some of the applicants had a higher tendency to delay the transaction, who would have a higher possibility to commit fraud.
- Detail information of the applicant. Applicants' personal information was collected when applying credit cards like education, family status and employment. Relatively excellent personal conditions could lead to fewer fraud cases.
- Policy of applying credit card. In the policy of applying for bank cards, there may be a small number of loopholes. Although unlikely, it will still be found by someone. resulting in fraudulent behavior.

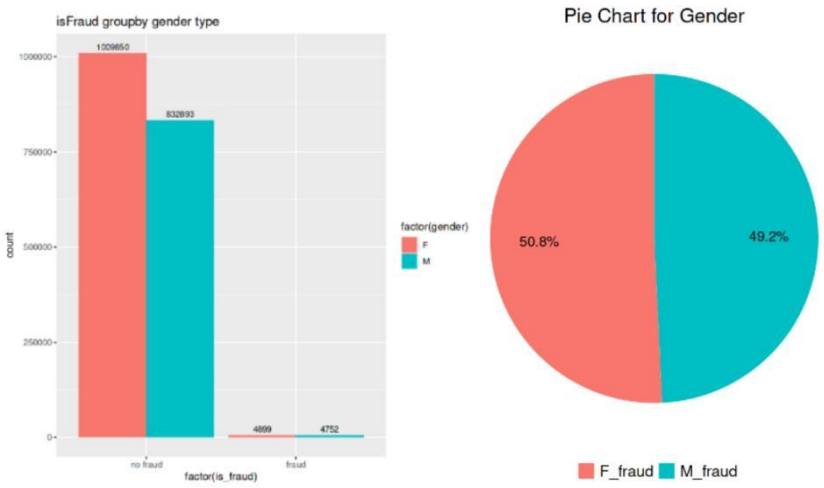
4.3 Summary

In this report, we chose credit card fraud detection. There were 3 sample functions used in training model. They were respectively ROSE, SMOTE and ADASYN. 5 models were applied to determine which one performed best. The models included Logistic Regression, Neural Network, KNN, Random Forest and XGBoost.

To draw a conclusion, XGBoost has highest recall, which is the most important standard. We recommend that XGBoost is more suitable to banks and companies for them do predict fraud. They worry about missing fraud case, his will cost them a lot, so they need to carefully check if the claims are fraud. XGBoost has higher recall, this can help them better detect the fraud cases.

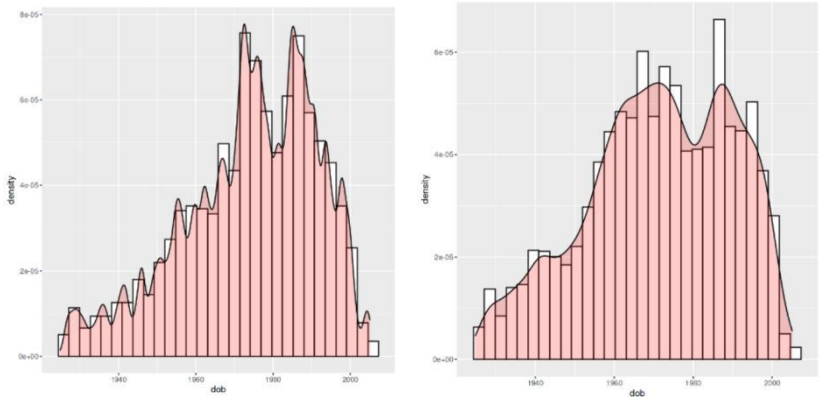
Aappendix

- **Gender**



isFraud groupby gender type. Pie Chart for Gender. Although there are more transaction incidents involving female customers, the number of fraud incidents for male and female customers is almost the same.

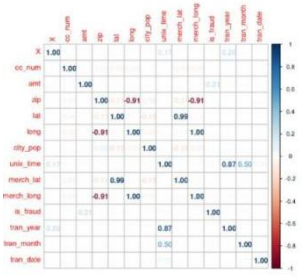
- **Date of Birth**



Number of Credit Card Transaction by Date of Birth(left). Most of the trading crowd was born between 1970 and 1990.

Number of Credit Card Frauds by Date of Birth(right). Traders born in 1970-1990 were flagged for more fraudulent trades, which may also be due to the higher volume of trades made by these age groups.

- **correlation**



The graph shows the correlation between the variables.