UNIVERSITY OF CALGARY

BTMA 531 Final Project: Analyzing Credit Card Fraud

Brandon Chiu, Strahinja Mirkovic, Nolan Ruzicki, and Doug Strueby

Haskayne School of Business, University of Calgary

BTMA 531: Data Analytics Tools for Business

WINTER 2024 L01

Professor Hooman Hidaji

April 9th, 2024

Problem Formulation

The business problem we have chosen to address concerns credit card fraud, focusing specifically on identifying potential signs of fraud and predicting its occurrence before it happens. This issue is particularly relevant from a business perspective because, in Canada, the Fair Credit Billing Act (FCBA) is designed to protect credit card users from unfair billing practices. Essentially, this means that credit card issuers bear the financial burden of reimbursing cardholders for fraudulent charges and are responsible for investigating and resolving cases of fraud. In 2020, there was a reported loss of $31.67 billion (Figure 1). Consequently, credit card companies are highly interested in this problem, as addressing it could save them billions.

Data

We selected a dataset from Kaggle (Kaggle, 2024) that consists of transactions from the year 2020. This dataset comprises 555,719 complete entries with 22 attributes that cover various aspects of credit card transactions. The attributes include a mix of categorical, numerical, and date/time values, as listed below:

Transaction Information:

- Trans_date_trans_time: Timestamp of the transaction
- Trans_num: Transaction identifier
- Unix_time: Transaction timestamp (Unix format)
- Customer Information:
- CC_num: Customer identification number
- First: Cardholder's first name
- Last: Cardholder's last name
- Gender: Gender of the cardholder
- Dob: Cardholder's date of birth
- Job: Cardholder's job title
- Street: Cardholder's street address
- City: Cardholder's city
- State: Cardholder's state
- Zip: Cardholder's ZIP code
- Lat: Latitude of cardholder's location
- Long: Longitude of cardholder's location
- City_pop: Population of the cardholder's city

Merchant Information:

- Merchant: Merchant involved in the transaction
- Category: Transaction type
- Amt: Transaction amount
- Merch_lat: Merchant's latitude
- Merch_long: Merchant's longitude

Target variable:

- Is_fraud: Fraudulent transaction indicator (1 = fraud, 0 = legitimate)

Exploratory Analysis and Data processing

Initially, we examined the distribution of true versus false fraud cases and discovered that over 99.6% (553,574) of our entries were non-fraudulent, while the remaining 0.4% (2,145) entries represented fraudulent transactions (Figure 2). Additionally, we sought to analyze the geographical distribution of

fraudulent transactions to determine if location was a significant factor in credit fraud. By utilizing the longitude and latitude of the cardholders' locations along with the count of credit fraud cases (Figure 3), we found that the eastern regions had a higher density of fraudulent transactions compared to the western regions, where credit fraud seemed to be less frequent.

We tackled the issue of data skewness through under-sampling, a process where we extracted all true credit fraud cases and paired them with an equal number of randomly selected non-fraudulent transactions, thus creating a balanced dataset. To counteract the bias that this approach might introduce, due to the significant number of entries removed, we executed multiple runs for each model, each time with a newly randomized dataset. Our final results were derived from the average of the outcomes across these various iterations. This method ensured more consistent results and minimized the variability between runs.

Decision Trees

The first model we decided to implement was a decision tree, primarily because of its ease of understanding. This simplicity facilitates explaining the model's predictions to stakeholders, which is crucial given the significance of the problem at hand. Being able to elucidate the process to less technically inclined individuals is a substantial advantage. After running the model, our objective was to ensure that it correctly predicted the type of error we aimed to minimize, focusing on reducing type 2 errors. To achieve a higher sensitivity, we adjusted the model accordingly. Refer to figure 4 for the resulting tree structure.

The outcomes achieved by this model included an accuracy of 92%, a sensitivity of 93%, and a specificity of 92%. This model demonstrated good accuracy with minimal overfitting. As mentioned, the model was adjusted to prioritize sensitivity, hence predicting more type 1 errors than type 2, aligning with our goal to effectively identify fraudulent transactions.

Logistic Regression

Next, we opted to apply logistic regression modeling to our data. We chose logistic regression because it is an effective model for predicting a dependent variable with two possible outcomes. In our context, this involves determining the presence or absence of credit card transaction fraud. Logistic regression is known for its simplicity, efficiency, and ease of computation. Following some preliminary exploratory analysis, we developed a logistic regression model that compared the "is_fraud" variable with all other relevant variables to identify which ones were statistically significant.

Statistically significant variables identified were:

- Category
- State
- City Population
- Amount
- Merchant Latitude
- Merchant Longitude
- Latitude

We proceeded to refine our logistic regression model, incorporating only the variables identified previously to determine their true statistical significance. Ultimately, the variables of category, city population, and transaction amount were found to be statistically significant in determining the presence or absence of credit card transaction fraud.

Given the nature of credit card transaction fraud, we recognized that type 2 errors, failing to detect a fraudulent transaction, would be more costly. Therefore, in our logistic regression modeling efforts, we aimed to minimize the occurrence of false negatives, striving for the highest possible sensitivity. To achieve this, we adjusted our Bayes boundary from 0.5 to 0.15.

With the Bayes boundary set at 0.15, our model yielded an accuracy of approximately 0.72. It demonstrated a specificity of 0.44, indicating an increase in type 1 errors compared to the model with a Bayes boundary of 0.5, but it also achieved a sensitivity of 0.99, showing a decrease in type 2 errors relative to the model with a Bayes boundary of 0.5.

It is important to note that, given our analysis is based on a sample from a larger dataset, slight variations in numbers might occur during logistic regression calculations. However, the identification of significant variables and the overall calculations are expected to average out as anticipated.

Cluster Analyses

The cluster analysis presented employs the K-means algorithm with three clusters to categorize transactional data based on three attributes: the transaction amount 'amt', the transaction timestamp 'unix_time', and the transaction category 'category'. Each data point corresponds to a transaction defined by these attributes, colored based on the cluster assignment (Figure 6).

In the 'amt' versus 'category' plot, cluster 1 (red) aggregates data points around lower transaction amounts across various categories, suggesting these may represent routine, low value transactions. Cluster 2 (black) is distributed more evenly across the amount range but is particularly dense in specific categories, indicating transactions that vary in amount but are potentially linked to certain types of purchases or services. Cluster 3 (blue), characterized by fewer data points, spreads out over higher transaction values and spans multiple categories, indicating occasional but high value transactions.

Observing the 'unix_time' attribute, it's noticeable that the time of the transactions doesn't provide a clear distinction among the clusters, suggesting that the transaction timing might not be a primary factor in differentiating between the types of transactions captured by the clusters.

The categorization by 'category' shows some differentiation, especially between clusters 1 and 2, although cluster 3 does not present a substantial variance from the others in this attribute. This lack of distinct separation in the 'category' dimension for cluster 3 might imply that high value transactions are not constrained to specific categories.

There are limitations to this cluster analysis, such as the overlap of clusters, particularly between clusters 1 and 2 in the 'amt' versus 'unix_time' graph. This could suggest that these features alone are not sufficiently discriminative for the clustering or that there's a subset of transactions that exhibit characteristics of both clusters. Additionally, the presence of outliers, especially noticeable in clusters 2 and 3, underscores the need for further preprocessing to refine the clusters for clearer distinctions and interpretations.

The K-means clustering represented in Figure 7 categorizes merchants by examining patterns in their location and transaction categories. The algorithm considers four features: a numeric representation of the merchant's identity ‘merchant_code’, the merchant's latitude ‘merch_lat’, the merchant's longitude ‘merch_long’, and a numerical code for the transaction category ‘category_code’. Each cluster, differentiated by color, encapsulates a group of merchants who exhibit similarities across these features.

Cluster 1 (red) denotes merchants concentrated in a specific geographical area, as deduced from their tight clustering in the latitude and longitude dimensions. Their diverse distribution across the category_code axis suggests that these merchants conduct a wide range of transaction types. Cluster 2 (green) suggests a different set of merchants who, while also spread across a particular geographical area, seem to specialize in fewer transaction categories compared to the first cluster. Cluster 3 (blue), distinctly separate in both the geographical plots and the category_code plot, represents a group of merchants that are not only region specific, but also cater to a unique segment of transaction categories distinct from the other two clusters.

Results

Through our analysis, we identified a series of variables of paramount significance in predicting fraudulent transactions. These variables, listed in descending order of significance, include: transaction amount, purchase category, city population, and state. Given that our logistic regression model achieved the highest sensitivity (~98%), effectively minimizing the number of type 2 errors, we recommend that banks adopt logistic regression models for predicting fraudulent transactions. Additionally, we advise banks to enhance monitoring of high-risk transactions, as indicated by the statistically significant variables mentioned above, by increasing communication with cardholders to confirm their awareness of the transactions.

Another recommendation for banks is to regularly review and update their fraud prevention strategies to reflect the latest trends and emerging threats. Since our model was trained exclusively on data from 2020, it is imperative for banks to continuously update their models with the latest available data. This approach will enable them to identify and respond to new patterns and potential threats in fraud, ensuring that their detection mechanisms remain robust and effective.

Discussion

In reflecting on the methodologies and outcomes of our project, it's evident that tackling credit card fraud through data analytics presents both significant challenges and opportunities. The skewed nature of our dataset, with less than 1% of transactions being fraudulent, initially hindered our ability to identify significant predictors of fraud. However, through innovative approaches such as under-sampling and logistic regression, we were able to overcome these obstacles and unearth key variables that can predict fraudulent activity.

Our results underscore the complexity of fraud detection and the necessity of adopting multifaceted strategies that encompass a variety of data points. The significance of transaction amount, purchase category, city population, and state in predicting fraud highlights the intricate patterns that fraudulent transactions follow. This complexity necessitates continuous adaptation and refinement of predictive models to keep pace with evolving fraudulent tactics.

The high sensitivity achieved by our logistic regression model (~98%) is particularly promising, demonstrating the potential of data-driven approaches to significantly reduce the incidence of undetected fraud. Yet, our research also illuminated the limitations of relying solely on historical data, emphasizing the dynamic nature of fraud and the constant evolution of fraudsters' methods.

As we propose the adoption of logistic regression models for banks, it's crucial to acknowledge that this is merely a starting point. Regularly updating models with fresh data and incorporating real-time analytics could further enhance the ability to detect and prevent fraud. Additionally, exploring alternative modeling techniques and incorporating more complex machine learning algorithms could uncover deeper insights and improve predictive accuracy.

The process of clustering transactions and merchants revealed patterns that, while insightful, also pointed to the need for more nuanced analysis. The overlap observed between clusters and the presence of outliers suggest that additional preprocessing and a more sophisticated approach to feature selection might refine these clusters, providing clearer insights into transactional behaviors and merchant categories associated with fraud.

Ultimately, our project highlights the power of data analytics in combating credit card fraud, while also pointing to the critical need for ongoing research, collaboration between financial institutions, and the development of adaptive, sophisticated analytical models. As fraudsters continue to evolve their tactics, so too must our strategies to combat them, leveraging the latest technologies and data to protect consumers and financial institutions alike.

References

Kaggle (2024). *Credit Card Fraud Prediction*.

https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction

Appendices

Figure 1: Credit card fraud worldwide

Figure 2: Breakdown of credit card fraud cases

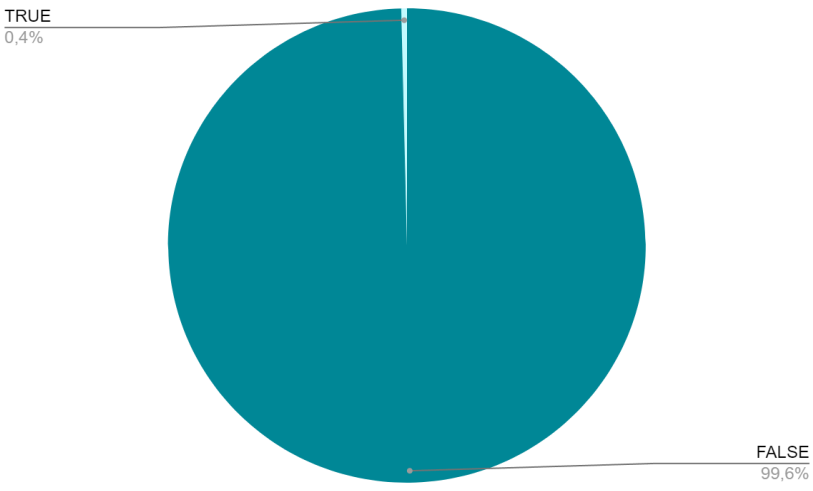

Figure 3: Count of credit fraud cases in the US

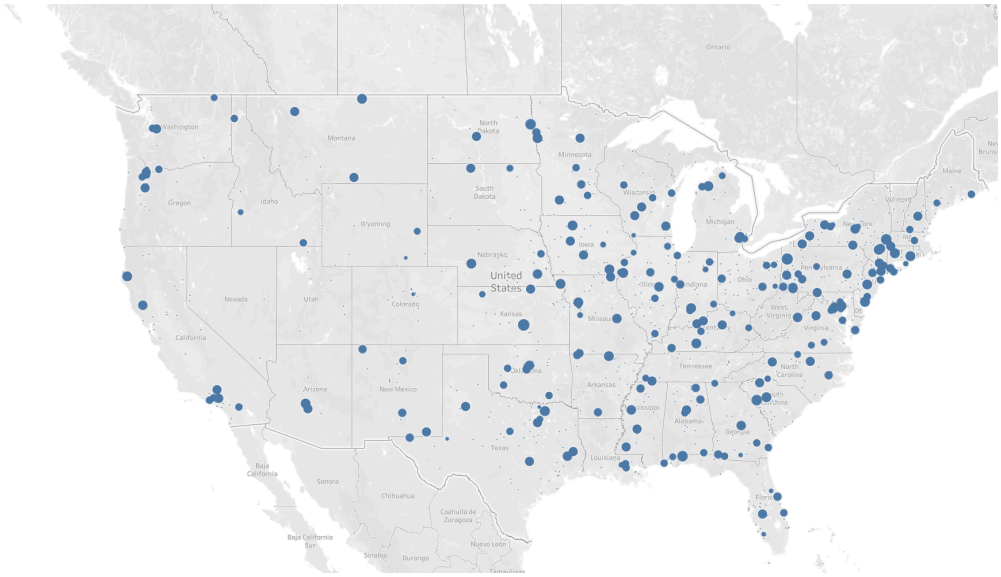

Figure 4: Decision Tree output

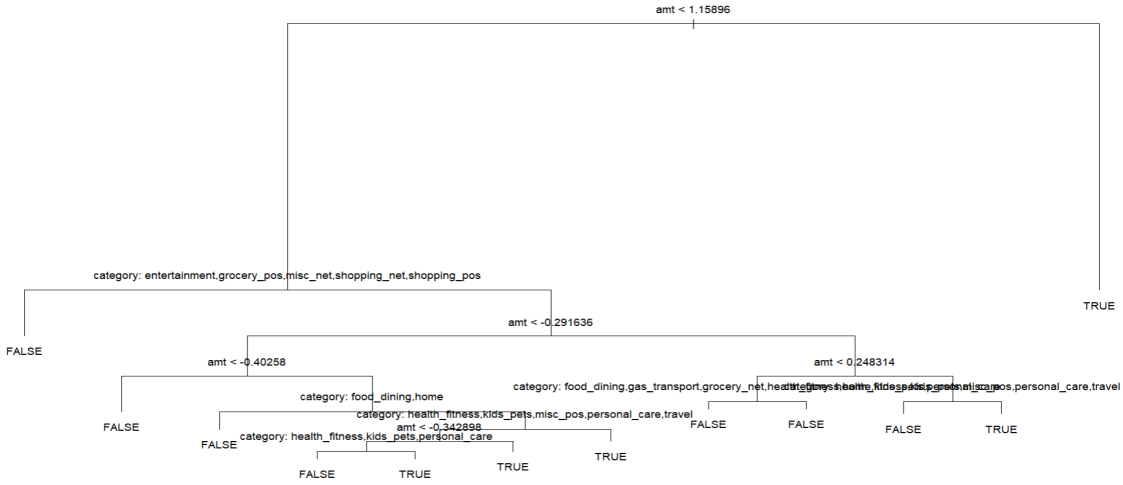

Figure 5: Logistic Regression Model Performance Metrics

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.440e+00  4.181e+00  -1.062   0.2883
categoryfood_dining    1.715e+00  3.531e-01   4.857 1.19e-06 ***
categorygas_transport  2.794e+00  3.260e-01   8.571  < 2e-16 ***
categorygrocery_net    2.616e+00  3.809e-01   6.869 6.48e-12 ***
categorygrocery_pos    1.936e+00  3.013e-01   6.427 1.30e-10 ***
categoryhealth_fitness 2.131e+00  3.564e-01   5.979 2.24e-09 ***
categoryhome           8.291e-01  3.467e-01   2.391   0.0168 *
categorykids_pets      2.001e+00  3.436e-01   5.824 5.73e-09 ***
categorymisc_net      -4.711e-01  5.048e-01  -0.933   0.3507
categorymisc_pos       2.270e+00  3.590e-01   6.321 2.59e-10 ***
categorypersonal_care  2.356e+00  3.523e-01   6.688 2.26e-11 ***
categoryshopping_net  -2.318e+00  5.315e-01  -4.360 1.30e-05 ***
categoryshopping_pos  -1.212e+00  5.112e-01  -2.371   0.0177 *
categorytravel         2.696e+00  3.932e-01   6.855 7.14e-12 ***
stateAL                1.651e-01  1.978e+00   0.083   0.9335
stateAR               -1.726e-01  1.849e+00  -0.093   0.9256
stateAZ                1.764e+00  1.663e+00   1.061   0.2889
stateCA                3.779e-01  1.427e+00   0.265   0.7911
stateCO               -1.636e+00  1.867e+00  -0.876   0.3808
stateCT                1.940e+00  2.213e+00   0.877   0.3807
stateDC               -5.190e-02  2.208e+00  -0.024   0.9812
stateFL                7.134e-01  2.129e+00   0.335   0.7376
stateGA                7.376e-01  2.052e+00   0.360   0.7192
stateHI                7.293e-01  2.329e+00   0.313   0.7542
stateIA                3.065e-01  1.743e+00   0.176   0.8605
stateID                6.127e-01  1.406e+00   0.436   0.6629
stateIL                2.544e-01  1.827e+00   0.139   0.8893
stateIN                1.337e+00  1.900e+00   0.704   0.4817
stateKS                6.425e-02  1.729e+00   0.037   0.9704
stateKY                4.407e-01  1.950e+00   0.226   0.8212
stateLA               -9.866e-03  1.954e+00  -0.005   0.9960
stateMA                5.117e-01  2.230e+00   0.229   0.8185
stateMD                4.252e-01  2.109e+00   0.202   0.8402

stateME               -2.781e-01  2.324e+00  -0.120   0.9047
stateMI               -4.177e-01  1.916e+00  -0.218   0.8274
stateMN                6.775e-01  1.683e+00   0.403   0.6872
stateMO                3.807e-02  1.780e+00   0.021   0.9829
stateMS                7.824e-01  1.926e+00   0.406   0.6846
stateMT                8.966e-01  1.414e+00   0.634   0.5260
stateNC                3.450e-01  2.083e+00   0.166   0.8684
stateND               -1.863e-01  1.611e+00  -0.116   0.9080
stateNE                2.393e-01  1.664e+00   0.144   0.8856
stateNH                9.398e-01  2.263e+00   0.415   0.6780
stateNJ                3.425e-01  2.156e+00   0.159   0.8738
stateNM                8.450e-02  1.661e+00   0.051   0.9594
stateNV               -1.512e+01  7.217e+02  -0.021   0.9833
stateNY                3.382e-01  2.112e+00   0.160   0.8728
stateOH               -3.907e-01  1.969e+00  -0.198   0.8427
stateOK                9.488e-01  1.750e+00   0.542   0.5877
stateOR                7.020e-01  1.259e+00   0.558   0.5771
statePA                2.389e-01  2.053e+00   0.116   0.9074
stateRI               -1.476e+01  2.400e+03  -0.006   0.9951
stateSC                1.314e+00  2.061e+00   0.637   0.5239
stateSD               -8.658e-01  1.650e+00  -0.525   0.5998
stateTN                1.285e-01  1.952e+00   0.066   0.9475
stateTX                3.396e-01  1.791e+00   0.190   0.8496
stateUT               -1.590e+01  9.162e+02  -0.017   0.9862
stateVA                5.609e-01  2.083e+00   0.269   0.7877
stateVT               -1.502e+01  5.278e+02  -0.028   0.9773
stateWA                2.895e-01  1.271e+00   0.228   0.8198
stateWI                5.449e-01  1.803e+00   0.302   0.7625
stateWV               -1.519e+01  3.217e+02  -0.047   0.9623
stateWY               -8.561e-01  1.602e+00  -0.534   0.5932
city_pop              -5.257e-07  2.461e-07  -2.136   0.0327 *
amt                    1.056e-02  5.079e-04  20.789  < 2e-16 ***
merch_lat             -1.926e-03  8.711e-02  -0.022   0.9824
merch_long            -2.725e-03  2.769e-02  -0.098   0.9216
lat                    1.038e-02  9.669e-02   0.107   0.9145
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6: Transactional behaviours cluster analysis

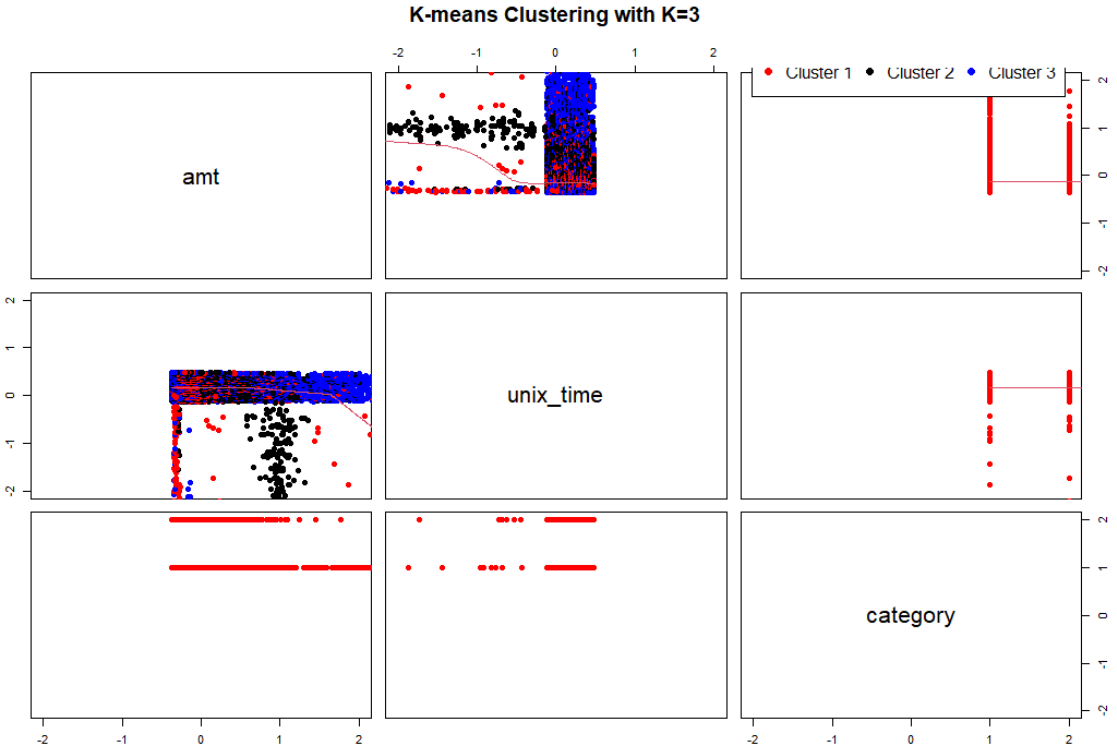

Figure 7: Merchant clusters analysis