[**DSC550-T301 Data Mining (2231-1)**](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_512555_1) **Mrs. Cindy Brinkmeyer**

**Professor Brett Werner**

Fraud detection is a great problem with a significant amount of resources dedicated to detecting fraud as quickly as possible- hopeful before too much financial damage can be done. For those of us who travel, oftentimes it can be discouraging to have a credit card blocked due to automatic fraud detection which always seems to occur at the most inopportune times. Needing to call in to the credit card when charges are declined as one is trying to check out of a hotel with a limited amount of time before getting to the airport for the next destination can be very disruptive.

My data source did not afford itself to the determination of location as a factor in a way that I could isolate it as I intended. It seems very difficult to look at this approach and determine if someone is a common traveler or even that they had already been travelling to the destination multiple times such that a detection should not be triggered.

I would need additional data and would probably need to create it at the sources somehow. Additional research into how location affects the model and how to determine cases in which one would not want fraud detection to be triggered would be a great way to influence future models. The customer frustration in having to invest additional time and effort into stopping false positives for fraud detection could be alleviated by generating a better dataset and ultimately a new model for fraud detection for segments of the user base.

There is a ton of data available for fraud detection programs in open-source data like the Kaggle sight that I found my data at. For the data science community, a multitude of approaches and strategies have been used to execute fraud detection.

**Organized and Detailed Summary of Milestones 1-3**

**EDA & Visuals**

In visualizing the data, I prepared several graphs to examine potential relationships. The first was between the amount of the purchase and potential fraud displayed below.

**Chart

Description automatically generated**

The next visualization was in categories of spending. Some areas appear to be more sensitive to fraud.

**Chart, bar chart

Description automatically generated**

Next, I visualized the gender regarding fraud detection in which the fraud detection was very similar between male and female.

**Chart, bar chart, treemap chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

And finally related to the focus area, I examined the states regarding fraudulent over non-fraudulent activity. The data would need to be collected in a different manner to get to the aspects of fraud detection regarding false positives while someone is travelling. Perhaps if the data was collected differently, one could identify instances if fraud is detected in a state that is not the card owner’s residence, it would trigger a detection but be cross-referenced over if there were transactions in that non-resident location before to not alert the card holder if the location is a repeated location signifying common travel location.

**Data Preparation**

First, we import the needed libraries to begin data preparation several iterations may be needed to contain all required libraries. The next thing to do is check for missing NA values, which were not in our dataset.

**A picture containing table

Description automatically generated**

Then I looked at the dataset in the data frame to see what columns and rows there were.

**Graphical user interface, text, application

Description automatically generated**

Then reviewed the datatypes to see what potential manipulations are required to prepare the data for modelling. Data types can be visualized below.

**Table

Description automatically generated**

Below is the code and representation of the description summary of the data including the count, mean, standard deviation, minimum and maximum and quartiles. It appears that there are no missing values.

**Graphical user interface, text, application, email

Description automatically generated**

On my first attempt I was individually removing columns that didn’t seem to be required and step by step reviewing the data frame. Unfortunately I didn’t complete the data work and still had several columns that weren’t prepared to be imported into a model. I did some additional research I found that creating the new dataframe and then inserting dummy variables was a much more efficient manner to prepare the data frame for modelling.

**Graphical user interface, application

Description automatically generated**

As you can see below, the columns are listed, and then reviewed to determine inclusion into the data frame for modelling, then the categorical columns are prepared using dummy variables. The data set is split between training and test.

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Model Building and Evaluation**

Two models (a logistic regression model and a random forest model) were created and their summaries are below including a classification and confusion matrix. Both perform similarly and are consistent with the models chosen in various research and available via the web. The random forest model performs a little better than the logistic regression model illustrated by the precision recall and f1-score.

**Table

Description automatically generated**

**Table

Description automatically generated with medium confidence**

**Conclusion**

The analysis model building results are consistent with other research available and shows very good results for the fraud detection problem set. Unfortunately, it didn’t lend itself to solve the initial problem which was false positives during travel to non-resident states repeatedly. This would not be the model that I would want to deploy for that problem set. This model is good for general fraud detection and has a high accuracy for that detection, but it doesn’t address the false positive example. I recommend a different data collection with specific focus on the geographical problem set and capturing those situations to be triggered and eliminate false positives for that scenario.

To summarize, you should submit the following two items for your term project final submissions:

1. All code, content, and analysis from Milestones 1-3 along with any updates
2. Project writeup described above