Whether a given credit card transaction will be fraudulent or not?

**DSE I2100 Applied Machine Learning and Data Mining Final Project**

**Marjan Rezvani  
Ayub ali Sarker  
 Maryam Akrami**

**Spring 2020**

**Abstract**

One of the most important responsibilities that a bank or financial institution has is to protect the integrity of the institution by working hard to protect the financial assets that it holds. Bank fraud can be defined as an unethical and/or criminal act by an individual or organization to illegally attempt to possess or receive money from a bank or financial institution.

Banking Fraud has been an ever-growing issue with huge consequences to banks and customers, in terms of financial losses, trust and credibility.

It is anticipated that card frauds would amount to around $30 billion worldwide by 2020. So, how banks can improve security by detecting and obstructing frauds?

Machine learning is perfect for detecting frauds! Its algorithms learn to tell fraudulent transactions from legitimate operations. Machine learning fights financial fraud by using big data – better and faster than people do.

In this project, we developed a machine learning model using card transaction data to identify fraudulent transactions (i.e. Fraud Detection)

**Introduction**

How does our model work?

In this problem (Fraud Detection), we are dealing with a classification problem. We have looked into some supervised classification algorithms such as Logistic Regression, KNN, Random Forest and VotingClassifier to solve this problem.

So our machine learning model collects information, analyzes the data gathered and extracts the required features. And simply, the more information the model gets, the better it works. Then, features that describe customers’ behavior are added, besides being fraud and not fraudulent. Some of numerical attributes are like available money, credit limit and categorical attributes like merchant name and transaction type. There are 641914 instances in the dataset. We also have few attributes which totally have missing values that we dropped these columns.

**Background**

Discuss other relevant work on solving this problem. Most of your references are

here. Cite all sources. There is no specific formatting requirement for citations

but be consistant.

**Data**

Where you go the data. Describe the variables. You can begin discussing the

data wrangling, and data cleaning. Some EDA may happen here. This includes

your data source (including URL if applicable), any articles behind the data source.

**Methods**

How did you take your data and set up the problem? Describe things like

normalization, feature selection, the models you chose. In this section, you

may have EDA and graphs showing exploration of hyper-parameters. Note:

Use graphs to illustrate interesting relationships that are important to your

final analyses. DO NOT just show a bunch of graphs because you can. You

should label and discuss every graph you include. There is no required number

to include. The graphs should help us understand your analysis process and

illuminate key features of the data.

**Evaluation**

Here you are going to show your different models performance. It is particularly

useful to show multiple metrics and things like ROC curves (for binary classifiers).

Make sure it is clear not just what the score is but for which instances in the data

one has the largest errors (in a regression), or just sample examples miss-classified.

Make an attempt to interpret the parameters of the model to understand what

was useful about the input data. Method comparison and sensitivity analyses are

absolutely CRUCIAL to good scientific work. To that end, you MUST compare

at least 2 different methods from class in answering your scientific questions. It

is important to report what you tried, but do so SUCCINCTLY.

**Conclusion**

How well did it work? Characterize how robust you think the results are (did

you have enough data?) Try for interpretation of what the model found (what

variables were useful, what was not)? Try to avoid describing what you would

do if you had more time. If you have to make a statement about “future work”

limit it to one short statement.

**Attribution**

Using the number and size of github commits by author (bar graph), and the

git hub visualizations of when the commits occured. Using these measures each

person should self-report how many code-hours of thier work are visible in the

repo with 2-3 sentences listing their contribution. Do not report any code hours

that cannot be traces to commits. If you spend hours on a 2-line change of code,

or side-reading you did, you cannot report. If you do searches or research for the

project that does not result in code, you must create notes in a markdown file

(eg. in the project wiki) and the notes should be comeserate with the amount of

work reported. Notes cannot be simply copy pasted from elsewhere (obviously).

**Bibiliiography**

References should appear at the end of the report/notebook. Again, no specific

format is required but be consistant.

**Appendex**

If there are minor results and graphs that you think should be included, put

them at the end. Do not include anything without an explanation. No random

graphs just for padding!! However, lets say you did a 50 state analysis of poverty

and demographics and your report focused on the 5 most interesting states, for

completeness you could include all in an appendex. Be sure though to provide

some (very short) discussion with each figure/code/result.