INTELLIGENT SYSTEMS
ECE 579
WINTER 2020

Project Title:
Fraud detection on credit card transactions using transactional and user data

Group members:
Sanjyot Thete (48844546)
Vishnu Priya Velamuri (77120506)
Vinayaka Manjunatha Malya (02338153)

Introduction:
What is the project trying to accomplish, give a brief overview of what the main problem is and why it is useful to investigate this problem.

The project is trying to predict whether the given transactions are fraudulent or not, given the data related to the transaction and the user. This approach is a preventive measure to avoid fraudulent transactions from going through, which is very essential to the working of modern markets.

Objective:
What, specifically, do you propose to do?

Given the dataset, which has each transaction as a separate row and transaction details in the columns, we propose to build a predictive model that will predict the probability of a given transaction being fraudulent given the features we have in the columns of the dataset. We also want to investigate which factors/features are prominent in concluding a transaction as fraudulent or not.

How is success defined?

Our baseline success would be to come up with a model that performs better than chance prediction. The actual success of the project would be determined by the maximum F1 score we can achieve with any of our models.

Background summary:

What has already been done in this area?

The dataset is a part of, now closed, Kaggle competition where people have attempted to make their own models.

Are you extending existing work or applying standard techniques to a data set? Either approach is valid, but need to be clear on what will be done.

We are going to apply standard techniques to a new dataset. We will try to improve on how we use the existing techniques with innovative preprocessing and data analysis, feature engineering ideas, and original prediction pipelines.

Initial approach:

What techniques will you begin with, how will you determine if your approach is working well?

We will start preprocessing and getting the data into a format where we can iterate through some machine learning algorithms. We will try with the simplest models and then shift to more involved ones noting the performance of each. We will monitor the training error and validation set error for gauging how well our approach is doing. We will also use dimensionality reduction methods to select the features and try to determine the features that attain the optimum f1 scores.

Are there existing software libraries you plan to use or do you need to design and code an algorithm?

Yes, we plan to use existing software libraries like sci-kit-learn, pandas, seaborn, matplotlib and other libraries, if the project demands it.

Are there any problems or limitations you are aware of now? How would you work around any expected problems?

We are aware of the class imbalance present in the target labels, as often is the case in fraud detection. To work around this problem, we would primarily work with model performance metrics like precision, recall and, F1-score that take into account class imbalance when determining the performance of a model.

There is also a problem with missing values in many features. We will try to figure out an appropriate way to deal with this problem, like imputing the missing values with the median of feature or removing the missing records altogether.

If there are multiple members in this project, briefly describe what each team member will work on

Since all the parts of a machine learning project heavily interact with others, each of us will contribute to the design of each of the following 3 parts of the project. Still, we would like to assign one person as the final owner of their respective part. For now, here is the list of the parts of the project and their owners:

1. Exploratory data analysis - Vishnu
2. Preprocessing and feature engineering - Vinayaka
3. Model development and calculating performance metrics - Sanjyot

Literature review:

Provide at least four (4) references. At least two (2) of the references must be academic papers or technical reports. Two may be taken from Internet sites or more general information papers or articles or tutorials.

1. https://www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-07-015.pdf
2. https://link.springer.com/chapter/10.1007/978-3-642-04003-0_10
3. https://ieeexplore.ieee.org/abstract/document/6784638
4. https://ieeexplore.ieee.org/abstract/document/618940

Data sources:

What data will you use?
We will use a credit card transaction dataset that was hosted on a Kaggle competition.

Is there a sufficient amount of data?
Yes, the training dataset consists of around 600,000 records with a total of around 450 features. Testing data has a similar size.

Will you need to pre-process or transform your data in any way?
Yes, preprocessing of categorical labels and dealing with missing values will be needed to be done, in addition to other preprocessing which will be needed as the project develops.