**Machine Learning Final project**

**Marjan Rezvani, Ayub ali Sarker, Maryam Akrami**

**Spring 2020**

**Realtime Fraud Detection on Credit Card Dataset**

**Abstract:**

The financial industries and services that required financial transactions are suffering from fraud related loss. The challenge of financial industry is to development of real-time claim assessment and improve the accuracy of the fraud detection. The machine learning approach to fraud detection has got popularity in recent years and industry are moving from rule-based approach to ML-based approach.

In this project work, we will development a real time machine learning model that will identify a transaction is fraud or not and improve accuracy of fraud detection.

**Introduction:**

In this work we will preprocess the dataset. We will do features selections and features extraction by PCA. We will train machine learning models (SVM, Random Forests, K-nearest, Voting-classifier) with best parameters and will selection best model best performance. Best Model will be deployed as REST Service.

Dataset: transactions.txt

Total number of rows: 641,914

Total number of columns: 29

**Dataset preparation and preprocessing:**

Our dataset has missing values we will clean up data. We will first clean data then we will convert categorical data to numeric. We scaled data to standard level so that it can feed into model We will select importance features form the dataset using RandomForestRegressor. By using PCA or t-SNE we will visualize dataset.

**FlowChart:**

Raw data --> Clean up --> convertCategorialToNumeric --> scaled standard --> Select important features --> visualize dataset --> splitting data into train, validation and test set --> fit models --> validate models -> evaluate model on test sets --> find model with best accuracy --> deploy model as REST Service

**Dataset splitting:**

Clean dataset will be partition into tree subsets - train, test and validation sets.

**Modeling:**

Train different models and define which one provides the most accurate prediction.

we will train each model on data. we will cross validate model and then evaluate the model

performance test set. We will use voting classifier to see it they improve over individual model.

**Model deployment:**

Best model will be deployed on server and prediction services will be provided

as REST services. Any user can use rest api and can get realtime response.