manipal **PRO**learn

CREDIT CARD FRAUD DETECTION

By Jothi Prakash

Guide: Ramakrishna Ganesan

Context

▶ Problem motivation

▶ Project focus

▶ Data description

▶ Data limitations

▶ Project approach/scope

Problem motivation

▶ The health of the credit card industry is best measured not by the number of people with cards, but rather the number who pay their bills. Bad payment habits begin by nicking you with more fees and lower credit scores, and, in advanced cases, can lead to the loss of a vehicle or home, garnishment, and bankruptcy.

▶ Credit card defaults, after a lengthy decline, are starting to tick up slightly.

▶ Delinquencies in bank cards rose in the third quarter of 2016, but remain near historical lows, according to the American Bankers Association's Consumer Credit Delinquency Bulletin.

Project motivation

▶ Bank card delinquencies increased 26 basis points to 2.74 percent of all accounts in the third quarter, but remain well below their 15-year average of 3.68 percent. The ABA report defines a delinquency as a late payment that is 30 days or more overdue.4

▶ The trend looks similar when examining accounts that have been overdue for three months. TransUnion's Industry Insights Report found that the credit card delinquency rate reached 1.79 percent in Q4 2016, an increase of 12.6 percent from 1.59 percent in Q4 2015. The credit card delinquency rate remains more than a full point below its peak in Q4 2009 (2.97 percent).

Project focus

▸ All the banks are having loses due to illegal or defaulted transactions that happen using their network and they are trying to predict/classify a transaction to be worthy or defaulting. There are heavy losses due to illegal transactions and some of them are mentioned below.

▸ 1. Estimates created by the Attorney-General's Department show that identity crime costs Australia upwards of $1.6 billion each year, with the majority of about $900 million being lost by individuals through credit card fraud, identity theft, and scams.

▸ 2. In 2015, the Minister for Justice and Minister Assisting the Prime Minister for Counter-Terrorism, Michael Keenan, released the report Identity Crime and Misuse in Australia 2013-14.

Project focus

▸ 3. This report estimated that the total direct and indirect cost of identity crime was closer to $2 billion, which includes the direct and indirect losses experienced by government agencies and individuals, and the cost of identity crimes recorded by police.

▸ To classify/predict the defaultable transaction is the target objective.

Data description

▶ The dataset has been collected and analyzed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

▶ It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, they cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Data limitations

▸ This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Project approach/scope

▸ The scope of the project is somewhat restricted by the missing sensitive data. The PCA data might be useful for prediction but the high dimensional credit history data is really necessary in order to produce quality visualizations which are missing in the selected dataset.

▸ The customer name, a date with an exact time of transfer, location, device used for money transfer are some of the crucial information necessary for clustering/segmenting the data and the number of defaults is very low and the unbalanced data might have some impact in the prediction and insights extracted.

Project approach/scope

▶ To balance out the data and make the predictions and visuals in a justified manner t-SNE is used. To extract maximum insight variable importance is carried out in random forest algorithm. The Gini index will give us the idea of which dimensional data has the highest impact on the defaults.

▶ More than three algorithms have been carried out initially on the dataset with default parameters and random forest is coming up with the optimal output.

THANK YOU

Manipal ProLearn
#7, Service Road, Pragathi Nagar, Electronic City,
Bengaluru 560100
contact@manipalprolearn.com | manipalprolearn.com