Predicting and Separating Credit Card Defaulters

Aditya Thakur
11/10/2016

Domain Background:

Loaned money getting defaulted is not a new thing. Lending in excess, not having control over your expenditure and many more such factors lead to the loaned money getting defaulted. This branch can be subdivided into multiple sections; one of this is credit cards. 1 out of 20 Americans are late on the payment of their credit card dues. These things are bad for both the credit card holders as their credit score goes bad and they get followed by collection agents and for Credit card companies as the debt goes bad and they lose money. Most of the time this happens because of the unawareness from the customer's side. There have been some statistical ways to address the prediction part of defaulting. But the motive of this project is to address the prediction and classification phase using machine learning.

Problem Statement:

To avoid defaulting a loan and avoiding its consequence some precautionary measures can be taken to alert either the credit card company, or the consumer or even both, as this is bad for both the sides. There have been research in predicting the defaulter using some statistical approaches to obtain a person if is defaulter or not [http://fic.wharton.upenn.edu/fic/papers/11/11-34.pdf] and even some machine learning approaches [http://mitsloan.mit.edu/media/Lo_ConsumerCreditRiskModels.pdf] where with linear regression and some conservative assumptions were combined to form a good enough model and to estimate the reduced cost by cutting limits at the right time. In this project, a Principal Component Analysis based approach will be used to predict and classify the defaulters beforehand.

Dataset and Inputs:

The dataset was collected by Department of Information Management, Chung Hua University, Taiwan and Department of Civil Engineering, Tamkang University, Taiwan and was donated to UCI machine learning repository in year 2016. The analysis on this dataset was done using neural networks and linear regressions. The dataset consists of 23 features and is labelled as defaulter and non-defaulters. There are major features such as income, age, sex, marital status, education etc. and then there is history of few months for past payments and past bill statements. This should be considered good enough features to provide a trend of loans getting defaulted.

Solution Statement:

The dataset available has multiple features and thus using a Principal component analysis the dimensions will be first reduced to lesser number of dimensions. The principal components will be obtained to see the trends of the consumers. The data then will be separated into two, (Or possibly three) using a clustering algorithm. The data for multiple months is also available for statements and payments so there is a possibility of visualizing an increasing tendency of loss of trust. The GUI will be developed to identify the tendencies, if belonging to either of the groups.

Benchmark Model:

Earlier it was seen that the Linear Regression approach was able to forecast 85% of the defaulters. This approach, though a bit different is supposed to be giving a better or at least equal accuracy in the detection. The data available for that model is a little different than this data because of the demographics and the other reasons. Thus there is a chance of obtaining different trends in this model than the other one.

Evaluation Metrics:

The dataset labels are present already with the dataset so it is possible to verify the working of our model. The data then can be verified if categorized correctly or not. The false positives and false negatives can be taken together to obtain precision, recall and their R2 score. The same R2 score was obtained for the earlier cited model so it will be easier to compare this one with the benchmark model.

Project Design:

- Data Collection
    - Data has been collected in an excel sheet and can be retrieved into the Python workspace.
- Data Normalization
    - The data we have is having different scales so it is necessary to normalize the scale to the same level.
- Data Visualization
    - Data will be visualized into 2D and 3D spaces and with multiple plots with different X-Y scales to observe the correlation between the features and use those properties for further processing
- Principal component analysis
    - Principal component analysis will be done on the data to observe the trends in the credit card users by the use of components.
    - Dimension reduction will be done for this multi-dimensional data.
- Clustering

    - Reduced Data will then be sent through the clustering algorithm. It is not yet observed how the data will look like in the 2D space but the initial plan is to use K-means and K-Nearest Neighbor algorithm to separate the User nodes.
- Validation
    - To check if the model is functioning correctly the clusters will be subjected to the comparison with the actual labels of the users. The error will be calculated.
- GUI Implementation
    - A simple GUI will be developed where a user can subject parameters to the trained model and the model will tell if the user is under risk or not.