Risk prediction modeling for Credit Card Defaulters

Udacity Machine Learning Engineer Nanodegree

Final Capstone Proposal

By Nishant Sharma

August 15, 2017

Domain Background:

Bank credit card default rates are (as predicted by S&P dow Jones index) at all-time high for the past 42 months, and as stated here ,delinquency rates are likely to rise throughout 2017.This possess a real challenge for banks and other lending financial institutions, as they are now more than ever are in need of a robust risk prediction model to generalize the economic behavior of their current and potential future clients. These predictive models would not only benefit the lending institutions but also the customer as it would make them more aware of their potential defaulting rate. The goal of this project is to resolve this problem by building and comparing predictive risk models using supervised learning algorithms.

Problem Statement:

The problem here is to classify which customer would default on its credit card payments, with respect to customer information already collected by the bank. This would help the bank make better credit lending decisions based on customer information they have already collected. One of the potential and most common solution to this problem is to use some type of regression technique like logistic regression, as mentioned in this publication. In this project, I will use logistic regression as benchmark algorithm and attempt to build a better classifier.

Dataset and Inputs:

Dataset here will be obtained from UCI ML repository, it was donated to them by researchers from Tamkang University in Taiwan. The dataset consists of 24 attributes and 30000 total instances. Attributes include client information like Amount of credit they borrowed, gender, education/marital status, age and past amount and bill payments. All of these attributes will act as indicators in classifying the final defaulter column in the dataset, which is either 1 or 0, signifying whether the client defaulted or not respectively.

Solution Statement:

Here I would first clean the dataset check for any missing or duplicate values. After this and some exploratory data analysis of some important features in the dataset, I will check to see if we have any correlation across different features and handle them accordingly if needed. After this I will perform a comprehensive comparison of different classification models including ensemble methods namely gradient boosting ,adaboosting, random forests, voting classifier and compare their accuracies with our base logistic regression model and if possible check to see if they perform better than a feed forward deep neural network. Finally an optimal model would be chosen which best classifies credit defaulters.

Benchmark Model:

As stated above logistic regression is generally regarded as best classification algorithm for these kind of tasks, and it has produced high accuracy rates between 70 to 75% and sometimes even as high as 90% +.So logistic regression will be good choice as base model to compare to other prediction models.

Evaluation Metrics:

Since our prediction class is imbalanced, i.e. we have more non-defaulters than defaulters, using simply accuracy for evaluation will not be enough (accuracy paradox).Thus I will consider alternate evaluation metrics namely Precision, Recall and F1 score alongside with accuracy to compare my models.

Formulas-

Precision=True Positive/(True Positives+False Positives)

Recall=True Positves/(True Positives+False Negatives)

F1 Score=2*(Precision *Recall)/(Precision+Recall)

Project Design:

- **Downloading the dataset:** First step here would be to download the dataset from UCI ML repository.

- **Exploratory data analysis:** Here we will explore each feature one at time, like graphs for distribution of different categorical features, plotting a correlation chart among features etc.

- **Data Pre-Processing & Transformation:** Since financial datasets are generally clean, there won't be much preprocessing involved, but I would check for feature scaling, missing values, duplicate values and missing values and handle it accordingly.

- **Training different predictive models:** Here I would first build a logistic regression model as base model and then work on training other more powerful ensemble methods like Adaboost,GradeintBoosted trees, Random Forests, Weighted Voting Classifiers and attempt to build a feedforward deep neural network.

- **Evaluating and selecting an optimal model:** Finally, after some parameter tuning, I will compare each model based on accuracy first and then the top models with highest accuracies will be compared with respect to precision, recall and f1 score, to ultimately give us an optimal risk prediction model.