Capstone Proposal for Machine Learning Nanodegree

By Shobhit Mishra

Definition

Project Overview

Financial companies who provide loans to their customers face risk of defaulted loans. Companies often collect a vast amount of data about their customers and the loans. Using Machine Learning techniques, we can predict the outcome of a loan. We can analyze the data to identify the relevant features. We can then use this data to train the Machine Learning model. Having a reliable and accurate model reduces the risk of loan default significantly.

In this project, I use the publically available loan data of Lending tree to predict the loan outcome. I clean up the data and do exploratory data analysis to identify the useful features. I then create a model to predict the loan outcome. I use the f1 score as a metric to evaluate the performance of the model.

Problem Statement

Although the data provided by Lending tree is rich and has multiple loan status, we can reduce the problem to a binary classification problem. The goal is to build a loan classifier which takes various features of a loan application as input and predicts if the loan will default or not. The project involves the following steps:

1. Download the Lending tree loan data
2. Get rid of empty columns and rows
3. Do exploratory data analysis to identify the trends and useful features
4. Do data cleaning and transform categorical data to numerical data.
5. Split the data in train and test set
6. Train a model on the training data set
7. Predict the outcome on test data set and measure the model’s performance.

Metrics

We can use several metrics to measure the performance of binary classification models. F1 score is a widely used metric to measure the performance of binary classifiers and that is what I am going to use in this project. F1 score considers both precision and recall of the model. According to Scikit [documentation](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score), F1 score can be defined as:

F1 = 2 \* (precision \* recall) / (precision + recall)

We’ll use the scikit’s built in method to calculate the f1 score.

Analysis

Data Exploration

The lending tree dataset is an unprocessed data set with hundreds of columns. The dataset is so big that lending tree provides a separate excel to explain what individual columns stand for. Here is the summary for the unprocessed data set:

RangeIndex: 42542 entries, 0 to 42541

Columns: 111 entries, id to total\_il\_high\_credit\_limit

dtypes: float64(86), object(25)

memory usage: 36.0+ MB

Out of these columns, 60 columns have either 1 or 0 values. These columns are of no use to us and we drop them.

The ‘loan\_status’ column stores the outcome of loan. This column has several possible values. Here are the possible loan\_status values:

array(['Fully Paid', 'Charged Off', 'Late (31-120 days)', 'Current',

'Late (16-30 days)', 'In Grace Period', 'Default', nan,

'Does not meet the credit policy. Status:Fully Paid',

'Does not meet the credit policy. Status:Charged Off'], dtype=object)

The respective counts are as follows:

Fully Paid 34108

Charged Off 5662

Does not meet the credit policy. Status:Fully Paid 1988

Does not meet the credit policy. Status:Charged Off 761

Late (31-120 days) 10

Current 3

In Grace Period 1

Late (16-30 days) 1

Default 1

We’ll consider only “Fully Paid” and “Charged Off” for our project. We’ll drop the rows with other values to simplify our analysis.

There are too many columns in the dataset and it is not possible to describe what individual columns stand for. Please take a look at the ‘LCDataDictionary.xlsx’ in the root folder if you want to understand what individual columns stand for.

Exploratory Data Analysis

First, we want to see the ratio of ‘paid loans’ and ‘defaulted loans’ to see if the data is skewed. Here is a count plot of the loan\_status.

