Capstone Proposal

Machine Learning Engineer Nanodegree

Sherif Ramadan Mohammed

Dec-2018

Domain Background

Loans from banks are everywhere and considered a major component of banks business and their interests are a main resource of banks revenue.

As customers requested loans from banks, there is a risk that some of these loans are defaulted and their account holders are not able to pay off the loan in time.

If these defaulted loans exceed a specific threshold, banks could get into risk.

Problem Statement

Banks would like to predict the ability of a new customer to pay off the requested loan in time, so that the bank could better agree or not agree to grant that customer the requested loan.

Banks already have historical data of customers and their loan status (Paid Off/ Defaulted) and would like to make use of it to predict a new customer ability to pay off his loan in time.

Datasets and Inputs

The datasets for this project are two excel sheets (Train and Test) containing data about customer information:

- Age: Age of applicant.
- Gender: Male or Female.
- Education: Education of applicant.
- Loan Effective data: When the loan got originated and took effects.
- Loan Status: Whether a loan is paid off on in collection.
- Effective date: Since it's one-time payoff schedule, each loan has one single due date.

These datasets will be used to predict whether a new customer could pay off his requested loan or not.

The dataset was provided by IBM through a machine learning course on Coursera.

Course Url: https://www.coursera.org/learn/machine-learning-with-python

Data source link:

https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/ML0101ENv3/labs/loan_train.csv

https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/ML0101ENv3/labs/loan_test.csv

Solution Statement

As we already have labels “loan status”, this problem will be solved as a supervised learning problem. We will implement classifiers to classify the new customer expected loan status and will it be paid off or defaulted based on available old customer data.

We will implement four classifiers:

- K Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

For each classifier we will tune the hyperparameters (i.e. k value in KNN) to get the best possible accuracy.

We will then measure the accuracy for each classifier using accuracy metrics (i.e. F1-Score).

Based on accuracy metrics, we will promote which classifier(s) is recommended to use.

Benchmark Model

We will use the accuracy scores resulted from our training set as a benchmark to our classifier and we will compare this benchmark values with the scores resulted when applying the models on our test set.

Currently, scores from training set are around 0.78 %.

We expect the scores of the models when applied to test dataset to not be less than 0.74%.

Evaluation Metrics

For this problem, we will use Jaccard similarity score to compare our classifiers accuracy from both training data set and testing set.

The Jaccard similarity score function computes the average (default) or sum of Jaccard similarity coefficients, also called the Jaccard index, between pairs of label sets.

The Jaccard similarity coefficient of the i-th samples, with a ground truth label set yi and predicted label set y^i, is defined as:

$$\boxed{J(y_i, \hat{y}_i) = \frac{|y_i \cap \hat{y}_i|}{|y_i \cup \hat{y}_i|}.}$$

Project Design

For this problem, we will use the following work flow:

Data cleaning and exploration

1- Load the training dataset.
2- Doing required data conversion (i.e. convert date time values to datetime objects).
3- Explore value counts for our labels:” loan status”.
4- Visualize “loan status” per “Gender” and “Age” to better understand the data.
5- Convert categorical values to numerical values (i.e. Gender) through one hot encoding.
6- Normalizing our data.

Model Implementation

We will implement four classifying models:

- KNN.
- Decision Trees.
- SVM
- Logistic Regression

For each model, we will try to tune the parameters of the model to get the best accuracy.

Model Evolution

We will evaluate the models (using best parameters from previous step) using Jaccard similarity score.