Alessia Taboga

February 2019

Udacity Machine Learning Engineer Nanodegree

CAPSTONE PROJECT PROPOSAL

Domain Background

One well known application of machine learning algorithms is predicting loan default [1, 2, 3, 4]. When borrowers apply for loans, they provide personal and financial information to the Banks or Lending companies. The Banks/Lending companies then evaluate such details in order to decide if the borrowers will be creditworthy or uncreditworthy. Although decisions can be made on the basis of statistical methods for the more obvious loan acceptance/reject cases, decisions may become difficult in less clear cases. Denying loans in all the uncertain cases would cause the loss of potential good and active customers.

Available data of past loans (data which contain the personal and financial details of the borrowers and if they were able to pay back their loans) can be used to build machine learning models capable of predicting if the borrowers will default or not, or eventually their probability of default.

Problem Statement

The aim of my project will be to build a supervised machine learning model able to predict if borrowers will be able to pay back their loans or not.

Data from past loans will be used to train several classification algorithms which will be evaluated in terms of accuracy, precision and especially recall.

Datasets and inputs

For this project, I will work with the Lending Club data, which can be downloaded from Kaggle website [5]. The Lending Club is an American peer-to-peer lending company, which connects borrowers with investors online [6].

The dataset contains personal and financial information of borrowers together with their loan status for the years 2007-2015. The loan status column can be used to create a simple binary target of good (0) or bad (1) loans. The data contain 887379 records and 74 variables and have already been analysed by several other people for example on Kaggle website to simply gather insights about the Lending Club customers or to predict good/bad loans.

An additional excel file, which provides a description of the several features, is also available for download on the same web page [5].

Solution Statement

Supervised machine learning algorithms can be used to build models for predicting good/bad loans. It is a binary classification problem and the target can easily be built from the ‘loan_status’ feature.

Logistic Regression has for example commonly been used for predicting loan default. However, also other classification algorithms can be tested.

The main challenge will be to build a good performing model, able to recall the bad loan customers, starting from data which are really imbalanced. The percentage of borrowers with bad loans is very small compared to the ones with good loans.

Benchmark Model

The more basic benchmark model will be the naïve predictor (prediction obtained just assuming that all the borrowers will behave as the majority, i.e. all good loans).

Than a good benchmark model will be a Logistic Regression run with basic/default parameters. As previously stated, Logistic Regression is commonly used for loan default prediction and is known to provide a good reference result.

As in Kaggle there are several kernels on the Lending Club data, I will also look at some of the models built there.

Evaluation Metrics

Performance comparisons between benchmarks and my models (I will run several tests) will be based on accuracy, precision and recall.

As the target is really imbalanced (the number of borrowers with bad loans is very small), accuracy on its own will not be a good metric. The model could be missing most of the bad loans and still have high accuracy. I will have to pay more attenuation to precision and most importantly to recall.

Recall will tell me what proportion of borrowers that were not actually able to pay back (bad loans) were classified by the model has being not able to pay back.

$$recall = TruePositves/(TruePositives + FalseNegatives)$$

Precision will tell me what proportion of the borrowers classified as not able to pay back (bad loans) were actually not able to pay back.

$$precision = TruePositives/(TruePositives + FalsePositives)$$

Project Design

For this project, I would like to test apart from Logistic Regression also other three classification supervised machine learning algorithms: Gaussian Naïve Bayse, Random Forest Classifier and AdaBoost Classifier.

I anticipate that for me there will be two main challenges:

a) Analysing and dealing with so many features (74 columns) in a domain far from my expertise;
b) Finding a model which performs well (especially on recall) with imbalanced data or find a way to modify the data to be more balanced.

The workflow of my project will be as follows:

1) Download and read Lending Club data;
2) Analyse all the features (identifying also missing values, outliers, features requiring transformation, engineering or hot-encoding, features with high values of correlation);
3) Data pre-processing (dealing with all the anomalies identified in the previous step);
4) Split the data in training and testing;
5) Feature scaling;
6) Build the first models with basic/default parameters;
7) Evaluate the first models;
8) Identify ways to improve them (ex. by parameter tuning or by making the target data more balanced – maybe using SMOTE algorithm, which I should research [8, 9]).
9) Evaluate and compare all the models;
10) Chose the best model and specify reasons behind;
11) Suggest possible ideas for improvement for future developments.

References:

[1] I.H. Witten, F. Eibe, M.A. Hall, 2011. 'Data Mining: practical machine learning tools and techniques', 3rd edition.
[2] https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92
[3] https://github.com/topics/loan-default-prediction
[4] https://www.kaggle.com/wendykan/lending-club-loan-data/kernels
[5] https://www.kaggle.com/wendykan/lending-club-loan-data
[6] https://www.lendingclub.com/
[7] https://en.wikipedia.org/wiki/Lending_Club
[8]https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.over_sampling.SMOTE.html
[9] http://www.dataminingapps.com/2016/11/what-is-smote-in-an-imbalanced-class-setting-e-g-fraud-detection/