Capstone Propostal - Machine Learning Nanodegree

Rodney Sales Nogueira Jr

November 2017

1 Project Domain

I work for a company focused on software solutions to the financial business, in which I've also choose a similar domain for the current project. Our goal is to look for potential credit frauds, or find potential credit scores for the new costumers [1].

As stated: "The need for credit analysis was born in the beginning of commerce in conjunction with the borrowing and lending of money, and the purchasing authorization to pay any debt in future. However, the modern concepts and ideas of credit scoring analysis emerged about 70 years ago with Durand" [2]. Which is a good technique to evaluate potential credit for new data entries.

A credit score is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of an individual. A credit score is primarily based on a credit report information typically sourced from credit bureaus [1]

So the main idea of credit scoring models is to identify the features that influence the payment or the non-payment behavior of the customer as well as his default risk, occurring as the classification into two distinct groups characterized by the decision on the acceptance or rejection of the credit application [2].

Some of the problems are to solve a common supervised learning approach to credit scoring, as the scores can be a numerical expression in a certain range, such as a FICO Score: 300-850. Our particular goal is to comprehend how the machine learning design process works for applications in this domain. So credit scoring is a good way to start.

2 Problem Statement and Datasets

As stated by [2] sometimes the decision is binary based on an acceptance/rejection rate, meaning that this problem can be treated as a classification problem.

[1] I'm doing the study alone most of the time, this project is also interesting for my company and coworkers that's why I wrote this document as a "plural (we, our)" report.

Meaning that the datasets provided might have common features such as: age, credit status, expenses, income, assets, marital status, etc ...

From our research we found a few open ideas and datasets:

- https://github.com/gastonstat/CreditScoring
- https://www.kaggle.com/c/GiveMeSomeCredit/
- http://archive.ics.uci.edu/ml/datasets/credit+approval
- https://onlinecourses.science.psu.edu/stat857/node/215

We decided to focus on the "Give Me Some Credit" competition from Kaggle, firstly because it's also a binary decision problem. But it aims to find a possible financial distress in the next two years as quoted: "Improve on the state of the art in credit scoring by predicting the probability that somebody will experience financial distress in the next two years [3]" as it's a Kaggle competition the dataset is publicly published, and the data is properly secured. The competition goal is a little bit different from our original goal, but the study it's worthwhile.

3 Solution

The solution is to look for a good approach to the supervised learning classification problem. It will include:

- Understanding the data, pre processing it, etc ...
- Select a few candidate algorithms that might be able to address the problem.
- Find the optimal algorithm/parameters to best address the problem.

As the survey [2] states there are several classification approaches that are used to credit scoring, such as: neural networks, support vector machines, logistic regressions, decision trees, and other AI based approaches such as fuzzy logic, and genetic algorithms.

As we did in the project 2, the idea is to explore the data and the approaches given the competition and look for three good algorithms that might be good to this study. But in this case there's also the need for exploring correctly the data, looking for outliers, best features, and so on.

4 Benchmark and Metrics

The metrics in the competition is the AUC [3]. The competition is finished, and there's a lot of submissions, and discussions on the forum around the solutions, benchmarks, and the data. It will be easy to know if the models are fitting the data well. The best models for the competition are around > 0.8, so we have a good criteria for comparison of our own results.

5 Project Design

The project can be designed as a supervised learning classification problem. With these steps:

- Problem Evaluation
- Data Analysis
- Data Visualization
- Initial Algorithm Evaluation
- Data Selection and Feature Selection
- Data Pre Processesing
- Algorithm Selection
- Validation Selection (K-Fold, and other)
- Algorithm Validation and Parameter Tunning
- Study conclusions

We'll start with a more focused research on the *credit scoring* area, to look for possible ideas and solutions used to this kind of problems, and also the *Kaggle Competitions submissions* and *Discussion Thread.* Than we focus on understanding our available data, and how it behaves in comparison to the ones we might have found on our research. The next step is to choose a simpler model, like a logistic regression, or a Naive Bayes and see how it reacts to the data we have, we can also plot the results in a 2d/3d graphic too see how linear/non-linear the data is separated. After that we can look for the feature distribution, and understand how each feature might affect our previous evaluation, and how we should pre process each feature.

Having an overall idea of how the problem is, what features we are dealing with, we can look for the best *machine learning* algorithms to choose for our final solution. Then we can evaluate them, by selecting the validation method, and the parameter tunning method. After this we score the results and compare them with the actual results obtained by the *Kaggle Competition Leaderboard.*

References

[1] Wikipedia. Credit Score, *https://en.wikipedia.org/wiki/Credit$_s$core*

[2] Louzada Francisco, Ara Anderson, Fernandes B. Guilherme. Classification methods applied to credit scoring: Systematic review and overall comparison. Surveys in Operations Research and Management Science.

[3] Kaggle Competition, Give me Some Credit *https://www.kaggle.com/c/GiveMeSomeCredit*