Machine Learning for Loan prediction and analysis for publication in *Journal of Physics: Conference Series*

Lic. Arnoldo Oliva

E-mail: hugo.olivac@uanl.edu.mx

Abstract. Banking institutions depend greatly on the lending activity, as this allows them to charge the customer an additional rate of the money they lend. At the same time, this fact represents a risk of loss as the client is not guaranteed that he will pay all the charged quantity. Even so it would be more difficult for a company to not pay a loan, this is also the case for those entities. Efforts have been made in this industry to develop Machine Learning models that grant more certainty of evaluating a more accurate possibility of a certain customer of not paying the charged loan, thus minimizing the risk for the banking entity of losing. In this paper ML models are deployed to help to the acheivement of this goal, resulting in some interesting findings.

1. Introduction

This paper is an attempt to discover important factors regarding loan default and also attempts to develop a model that predicts with high accuracy the probability of falling in default category. To achieve this, a dataframe was worked on, originally from the SBA (USA Small Business Administration Office).

The goal of the SBA is to encourage small businesses to access to the credit market, and one tool it has to comply with this task is a SBA loan guarantee programs. Another effort to comply with this is the tracking of the loans and their outcomes, either fully paid or default, resulting on the SBA Administration dataset. This is important to track, not only for banking institutions, because as economic theory suggest, small businesess are a primary source of job creation, thence, *"fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment".* Li et al. [2018].

2. Literature Review

This section discusses in brief about some of the work that has already been done on creating ML and DL models using various algorithms to improve the loan prediction process and help the banking authorities and financial firms select an eligible candidate with very low credit risk.

According to the paper *"Should This Loan be Approved or Denied?": A Large Dataset with Class Assignment Guidelines* from Min Li, Amy Mickel and Stanley Taylor (2018), the following procedure should be taken into account when predicting probabilities of a loan falling in default:

A logistic regression should be considered to determine that probability as it provides easy understanding outcomes, it gives great insights, and also can work with binary variables, opposite

case of the linear regression model.

The authors applied interacting variables to capture special phenomenons, and also guided by economic theory most of all to decide which variables to use.

An example of this is that they built a Recession dummy variable which indicates whether or not the loan falls in the economic crisis period of time (2008-2010); a *term* variable which means that if the month term of the loan is greater than 240 months, then the loan is covered by real state, among other examples.

The authors limited their work to only one state and only one NAICS category alleging that, if a person wanted to work with multiply categories, estimations errors could be possible.

When evaluating the results, a mean absolute error of 32 percent was obtained according to the results of the confussion matrix.

In the present paper, as a goal to isolate economic crisis variables s, to comply with the assumption of normal economic cycle (growth) and also excluding state factor (the data contains records for all US states).

3. Methodology

3.1. Data processing

Data was treated in a way it can be compatible with ML models, and also involves filtering of non relevant components. Some examples of that were that for this case study, as stated before, the dates were selected to avoid the economic crisis period from 2007 to 2011. The loans that were selected complied with approvement and disbursement dates between the year 2000 and 2006. Also, geographic variables were dropped following the assumption of most basic economic variables, *zip*, *city*, *states*. Geographic effect was captured by a dummy variable *same state*, which indicates if the bank is on the same state as the business, and *urban rural* which indicates if a business is in a rural or a urban environment.

Timestamps of approvement and disbursement dates were obtained to check for the time difference between those two variables. Data categorization, data cleaning and data replacing were made to adjust the data. As the final step, the continuous variables were scaled and the categorical variables were converted to dummies.

3.2. Unsupervised learning

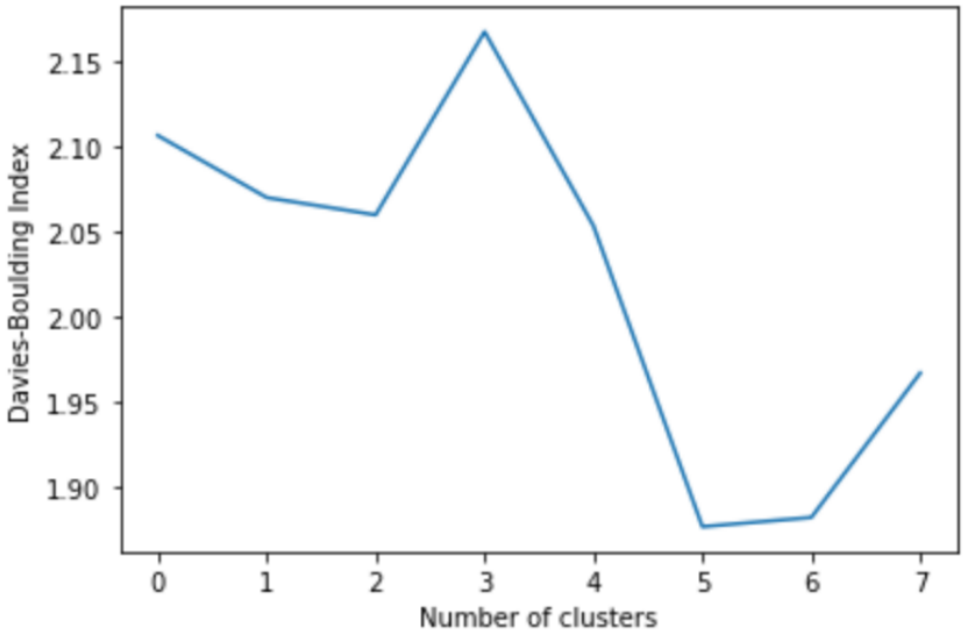
A first iteration of unsupervised learning with the objective of identifying possible patterns of data that may affect the outcome of the loan outcome was made. Before talking about this, its worth noting that some filter methods were tried to select the most important variables that would be part of the unsupervised learning clustering technique (in this case, was Kmeans algorithm).

A variance threshold model was ran in order to obtain this objective. Mainly all of its inputs were continuous variables.

A Lasso regression was made as this regression allows to work with categorical and continuous data at the same time (in this iteration the dummy variables for the most important categories of the NAICS were took as input). Both methods resulted in very different outcomes due to their differences in the methodology, but also got some similar results: variables like ChgOffPrinGr, Term, timediff, SBAAppv were marked in both techniques as the most important ones; but DisbursementGross, GrAppv, RevLineCr,SCIAN category (without being dummy) and NoEmp were either in one model or another, but not in both.

With continuous data being scaled, and categorical data (mainly most important categories of NAICS) being treated as dummies, an iteration of the KMeans algorithm was ran to identify the best number of clusters that reduces the most the Davies-Bouldin index. For this case was the model with 7 clusters reducing this value up to 1.87. In

the following image there is an example of the evolution of the Davies-Bouldin index.



References

Min Li, Amy Mickel, and Stanley Taylor. Should this loan be approved or denied?: A large dataset with class assignment guidelines. *Journal of Statistics Education*, 26(1):55–66, 2018. doi: 10.1080/10691898.2018.1434342.