Predicative Analytics to Determine Credit Worthiness

An approach to analyze how different machine learning algorithms handles an individuals' credit risk

MATH2191 – Applied Research Project

Saurabh Mallik (s3623575) | Dilip Chandra (s3574580)
Shiyuan Lou (s3639669) | Mohammad (3650497)

Introduction

Loans are one of the most ancient instruments used by banks and lending institution to generate revenues. Individuals usually approach banks and lenders to get assistance in accomplishing their investment and personal goals. In return, the lending institution charges the individual a certain rate of interest. As per the article by Romero Avila & Diego (2011), Loans play a significant role in the economic development of a nation. Lending practices rise capital for investment decisions. Capital leveraging maximize the potential to produce or develop a product which is made possible by approving loans.

When individuals have an existing credit history, it is easy for the banks to predict if an individual would default on loans. However, banks today want to untap the potential of the unbanked population who do not have any existing track record nor sufficient credit history. To achieve of this objective lending institutions are trying to gain better insights into big data to make more accurate decisions which would reduce the credit risk.

The process of availing loans has been evolving over the recent years. Earlier individuals needed to visit the bank to avail loans, however today we have an online lending platform where individuals are able to obtain loans at a click. Banking institutions are utilizing big data to accurately predict if a borrower will repay a loan or default in the near future.

The probability of loss due to individual's failure to make repayments on debt is referred as credit risk (Lan et al, 2012). Banks and lending institutions apply credit risk management techniques to mitigate uncertainty.

Lending institutions use comprehensive credit prediction models to determine whether to approve a loan request. A good prediction model can provide the lender with a leverage to advance the maximum borrowing potential to an individual without exceeding the risk threshold. Our main aim in this project is to test out a variety of machine learning algorithms which would predict whether or not an individual will be able to repay a loan.

Today, machine learning algorithms are being used across various industries, to provide complex solutions. These algorithms have the ability to self-teach and automatically apply complex mathematical calculations to data and have been detailed in the methodology section that follows.

The purpose of this project is to identify accuracy scores of the each of the machine learning algorithms and compare which approach handles the data better, limitations of the approaches and further studies requires.

Machine learning algorithms in helping understand consumer credit risk

It has been widely studied that the most influential drivers of macro-economic conditions are probably consumer spending. In their article Amir, Adler & Andrew (2010), state that, in the studies conducted on one of the worst financial crises in modern history, consumer behavior has been seen to have a central role at every stage. In their study Amir, Adler & Andrew (2010), produced a machine-learning model for consumer credit default which is surprisingly accurate in forecasting credit events 3–12 months in advance. Although the sample is only a small percentage of the total customer base. This supports the project research question that machine learning techniques, specifically regression trees help answer consumer credit risk and behavior.

The task of consumer credit scoring has been considered a classification task where each individual end up getting tagged with either a good or a bad credit status. The probability that an individual will default, provides more robust information about the credit worthiness of consumers, and they are usually estimated by logistic regression (Jochen et al 2013). In their study, Jochen et al (2013) during logistic regression, regressed the customers' characteristics on the logistic transformation of the default probability. Here x_i are the covariables of subject i. The aim is to estimate the default probability $P(y_i = 1|x_i)$ of a credit given the variables x

$$ln\frac{p_i}{1-\ p_i} = ln\ ln\ \frac{P(y_i = 1|x_i)}{1-\ P(y_i = 1|x_i)}\ =\ x_i^{'}\beta$$

This again supports our research question of a predictive analysis algorithm approach to understand whether an individual is a credit risk for an institution or not.

In a more recent study, Guegan & Hassani (2018), using 181 features, created models using approaches such as logistic regression, random forest, gradient boosting and other deep learning techniques. The RMSE results of their analysis show that random forest and gradient boosting had the best scores out of 7 specific models. In fact, the AUC scores were good too. The authors understand that while predictive analytics has a decent outcome, there are various factors that come into play which are hard for an algorithm to fathom – quality of data, sample size and more. The project aims to uncover these challenges faced by using different modelling techniques to find a good fit for credit risk data, and hence this article greatly supports the research question, and will help act as a guidance. Guegan & Hassani (2018), also show that partial selection of top 10 variables doesn't necessarily yield stable results, hence an understanding of feature variable before selection is of key importance. This will be a major part of the pre-processing step in the project.

Data

The datasets for this project have been sourced from Home Credit Default Risk on Kaggle (https://www.kaggle.com/c/home-credit-default-risk). Brief details of the individual datasets in this project have been outlined below.

- Application Train – 122 variables and approximately 308k observations (including contract type, gender, amount credit etc.). The main table which has been broken down into train and test.
- Application Test – 121 variables and 48.7k observations (same variables as train set minus the target variable)
- Bureau – 17 variables and 1.72 million observations which highlight key bureau data of individuals banking at the institution (credit_active, credit_currency, amt_credit_sum_debt etc.). This dataset talks about a client's previous credit history.
- Bureau_Balance – 3 variable and 27.3 million observations. This dataset shows monthly balances of previous credits of client.

- POS_Cash_balance – 8 Variables and 10 million observations. This dataset shows the monthly balance of previous point of sale and cash loans of applicant.
- Credit_card_balance – 23 variables and 3.84 million observations. Showing monthly credit card balances of the applicants.
- previous_application – 37 Variables and 1.67 million observations. An entire record of all previous pplications for loans.
- installment_payments – 8 Variables and 13.6 million observations, which talk about repayment history for previous credits of the institution.

For the purpose of this project, it is intended to join the application train and test dataset initially, so that a better randomization of train and test can be conducted during model fitting and tuning stages. This will also help in understanding the features better and help with the selection task.

Proposed Methodology

To achieve our goal of understanding which predictive analysis technique is most suitable for big data pertaining to credit risk of an individual, our project will be divided into the following stages.

- Stage I – Data Pre-Processing and Understanding
- Stage II – Feature Selection - Data Visualization of Univariate and Multivariate features to understand significance and relation of features with target (whether good credit risk or bad)
- Stage III – Model Specification and Training – Decision Tree Approach, Logistic Regression, Random Forest, Nearest Neighbor.
- Stage IV – Model Tuning, Limitations and Further Study Requirements.

Stage I - Data Pre-Processing and Understanding

1. In the first step of this stage of the project, the 7 datasets along with the application train and test dataset will first need to be compiled in order to have a full dataset consisting of all variables.
2. Identification and populating missing values with aggregated data to create homogeneity.
3. Tidying of variable names, variable factors and levels to simplify tasks for further stages.

Stage 2 - Feature Selection

1. In this process we will get to know the significance of each variable and its correlation with the datasets results.
2. We will reduce the variables and visualise the variables and its effect on the target variable.
3. Demonstrate data exploration technique which will allows us to locate the outliers.
4. Variables with a strong correlation with the target would be chosen for the modelling.

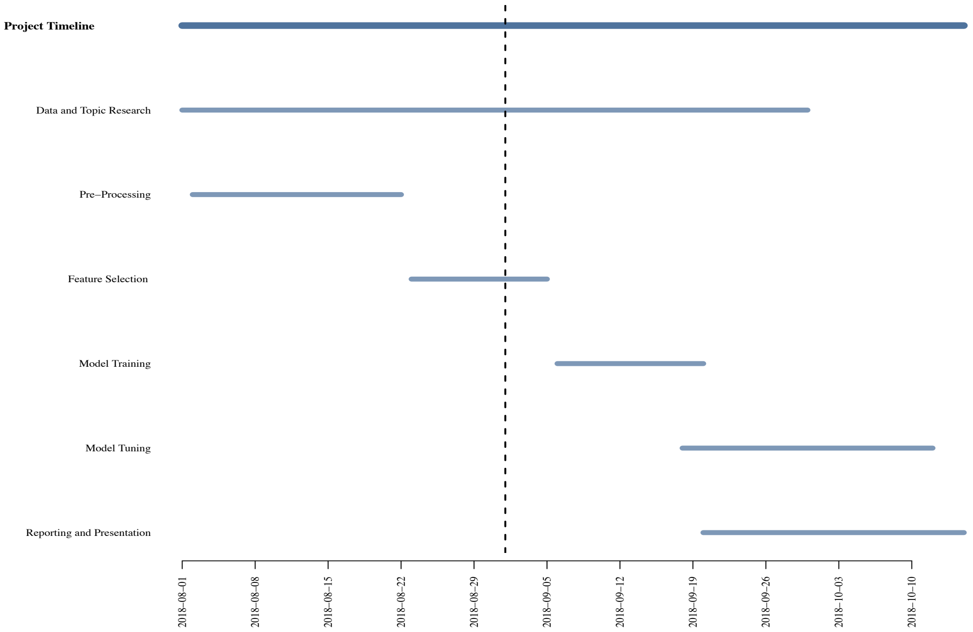
Stage 3 - Model Specification and Training

1. We will be using the classifying approach which will assist us supervised on machine learning.
2. We will be using four different approaches for model specification and training, It would be based on Logistic regression which is the baseline classifier, with other classifiers such as Decision Tree Approach, Random Forest, Nearest Neighbour.
3. Logistic regression would be applied on the binary variables to estimate the parameters of logistic model.
4. K nearest neighbours will allow classification and regression through a non-parametric algorithm.
5. Decision tree will allow us to model the decisions and their possible consequences.
6. Random forest is the constructs the multitude of decision trees.

Stage 4 - Model Tuning, Limitations and Further Study Requirements

1. We will adjust fine tuning thresholds by training each classifier.
2. Adjust prediction thresholds to make probability predictions.
3. The data will be split into training and test.
4. Then lastly, models will be applied, and accuracy of the models would be found.

Project Management: Gantt chart showing tasks and time line



Expected Outcomes:

The ideal model is expected to achieve accuracy of approximately 70% in classifying new test data set. In other words, the ideal-trained model can provide the lender with a leverage to advance the maximum borrowing potential to an individual without exceeding the risk threshold.

Based on our research and pre-processing of the data, the following have been outlined as key expected outcomes for the project.

a. Identifying a machine learning model that works best with this data and provides a decent accuracy score of about 70%.
b. Understanding which variables make the most impact on the accuracy of prediction.
c. Ascertaining whether missing values make too much of an impact on the results without having to populate them based on means and deviations.
d. Comparing results and findings from each model to ascertain the best approach to leverage this kind of data.
e. Understanding limitations of our approach and further studies required to improve the result.

References

1. Romero Avila & Diego, 2011. Information disclosure, banking development and knowledge driven growth. *Economic Modelling, 2011, Vol.28(3), pp.980-990*
2. Iscoe Ian, Kreinin Alexander, Mausser Helmut & Romanko Oleksandr, 2012. Portfolio credit-risk optimization. Journal of Banking and Finance, June 2012, Vol.36(6), pp.1604-1615
3. Khandani, Kim, & Lo, 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking and Finance,34*(11), 2767-2787.
4. Kruppa, Schwarz, Arminger, Ziegler, & Kruppa, J. 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*(13), 5125-5131.
5. Guegan, D., & Hassani, B., 2018. Regulatory learning: How to supervise machine learning models? An application to credit scoring. *IDEAS Working Paper Series from RePEc, 6*(2), IDEAS Working Paper Series from RePEc, 2018.