PROJECT PROPOSAL ON

CREDIT SCORING PREDICTION MODEL

July 10, 2022

Introduction

Credit scoring model is simply a mathematical machine learning model used to estimate the probability of default. The probability of default here refers to the choice outcomes of the customers that may somehow trigger a credit event, for example, failure to pay, bankruptcy, obligations default and cross-event defaults. Credit scoring model is widely and greatly known to be an essential tool assisting economic growth. It is however a good tool that is valuable and essential on improving financial inclusion, efficiency and credit access for individuals and micro, small, and medium businesses. The use of Credit scoring and various scoring outcomes has therefore led to greater access to a lot of companies and organization data. This has led to great need of efficiency and computing power and economic growth as well.

The credit scoring model has rose up from traditional methods of decision making of accepting or else rejecting the credit requests of consumers. The credit scoring model has however included other faces of crediting such as pricing of the financial services to determine or come up with the risk profile of the creditors, consumers, business and the settings of the limits of crediting.

Credit scoring model has several methods which gets complex each an every day with the ballooning of the computing industry. The methods have evolved from traditional methods which involves the manual decision making of crediting to the modern methods which includes computing power usage such as artificial intelligence, that includes, machine learning and incorporation algorithms which combined builds what we call models. The models help in determining the scorecard of a creditors.

The use of such modern methods (Innovative methods) has several benefits and opportunities which include, financial inclusion, accuracy gains of the used models, access to credit, efficiency due to the automation the models have provision for and last but not least an improved customer experience in the crediting industry.

In this project, the model will be based on HMEQ data collection taken from applicants granted credit recorded and given via the current existing crediting system, we shall use several prediction models which will assist us in choosing the best model to have an effective credit scoring model for a business and even maybe an organization. We shall gain knowledge much about data science as the main core subject in this project.

Statement of the Problem

This simply refers to a concise description of the issues or problem or an improvement the project addresses. It identifies the gaps that exists in mentioned sector (Crediting sector). The gaps must however be solved or fixed by the project.

The following gaps are to be fixed by this project:

1. How is the accuracy of the model?
2. How can the model be improved and in what means?
3. Is the model reliable in crediting decision making?
4. How to fix thresholding problem for our model to work efficiently?

The above gaps are explained in the Research and Methodology below.

Publications

This part details the published information in conjunction to credit scoring prediction model. This helps a reader to understand more about the project and give comparisons on the output of the project. Several publications to read in conjunction to this project exists out there. We have journals, past papers, magazines, text books and articles written by individual peoples. We have existing publication one can enjoy reading to name a few;

1. Paper title - Scoring Models of Bank Credit Policy Management.

   Written by:

   Aida Hanic of Institute of Economic Sciences (Serbia), Emina and Ednan of University of Sarajevo

2. Journal title - Journal of Advanced Research in Business and Management Studies

   Witten by:

   Yosi Lizar Eddy and Engku Muhammad Nazri Engku Abu Baka

Objectives of the Study

The objectives of the project are as follows:

1. To Determine the accuracy of the credit scoring prediction model.
2. To study ways to use to improve the efficiency of the model.
3. To determine reliability of the model by doing model prediction comparisons.
4. To determine how threshold impacts credit scoring prediction model.

Research Methodology and Design

Methodology refers to the research methods simply known as contextual research framework, coherent and logical scheme-based views, beliefs, and values that guides the researcher's choices.

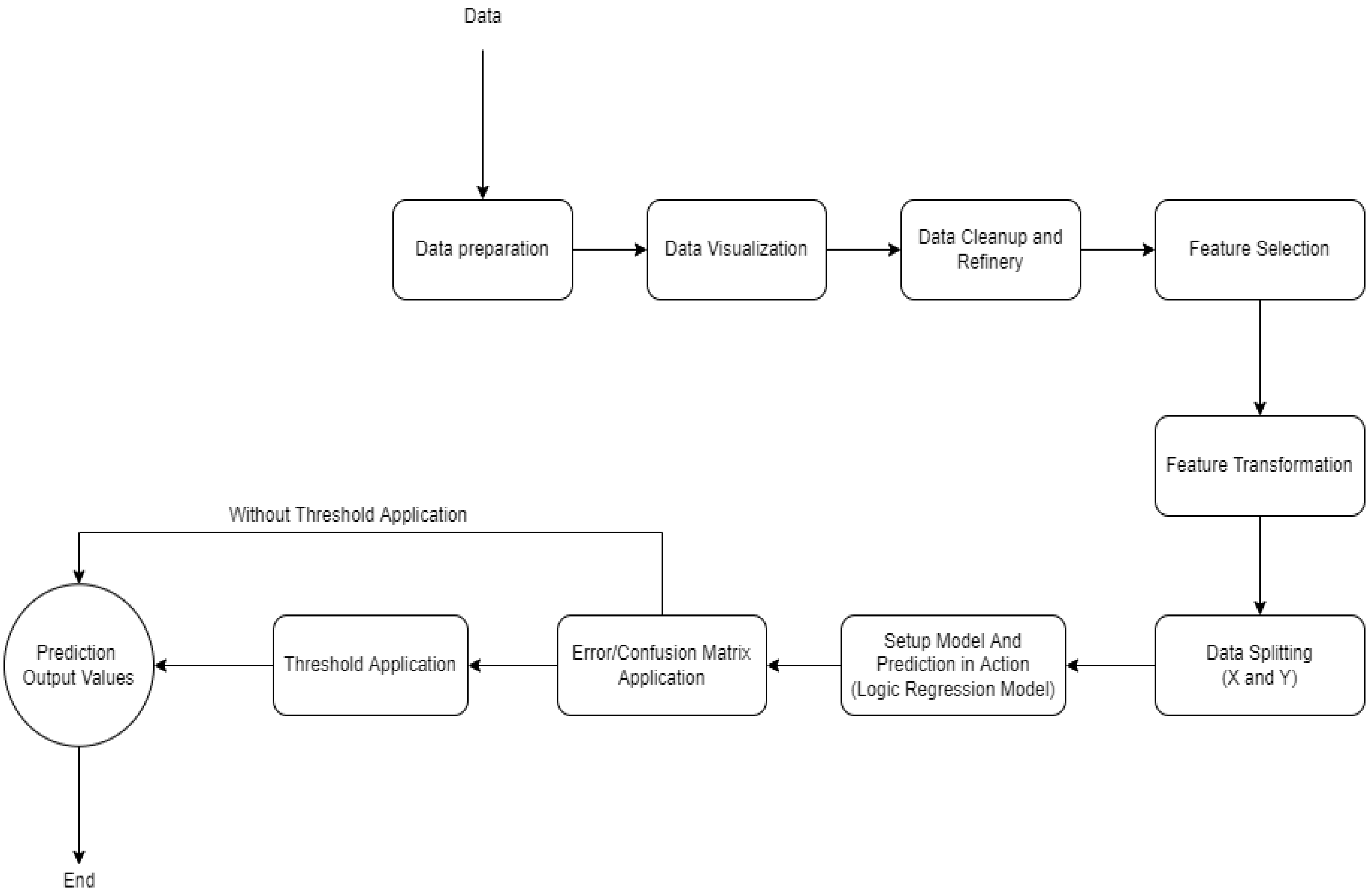
In research design, details a description of how the design was carried out. That is, knowing what kind of input will generate to output. Research design provides the glue that holds the research project together. A design is used to structure the research, to show how all the major parts of the research project, the samples or groups, measures, treatments or programs and methods of assignment. This is the most significant element of the research process where the whole research is designed, all the decisions that were made and details of the research scripted down.

Preferred to use the following methods to perform the research:

1. Logistic Regression Method
2. Decision Tree Regression method
3. Random Boost Regression Method

Using the above three methods, We shall have the comparison of the performance at its best.

The Design of the Model was designed using the flowchart design below:


Data
Data preparation
Data Visualization
Data Cleanup and Refinery
Feature Selection
Feature Transformation
Data Splitting (X and Y)
Setup Model And Prediction in Action (Logic Regression Model)
Error/Confusion Matrix Application
Threshold Application
Prediction Output Values
Without Threshold Application
End

Dataset

The dataset used in this project is a collection of HMEQ publication. HMEQ dataset is a set of Home equity loans consisting of baseline and performance data information for around 5960 recent given out equity loans.

The target variable, BAD, simply shows that an applicant just either defaulted the loan or the loan was delinquent. For each applicant in the list, had 12 input target variables noted down which denotes something vital in conjunction to giving out loans.

Conclusion

In conclusion, the following might be used to enhance efficiency and high performance of the model;

Threshold – It is advisable to use the default threshold value, 0.5, for most general cases. You are again advised to choose a threshold value that works at its best.

Resampling - This simply refers to data resampling, that is, data division. This helps in improving the performance of the algorithms one uses. We always have two types of sampling, Under-sampling and Over-sampling. One is required to use either, which fits.

Bibliography

Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". JAMA. 316 (5): 533–4

YuryWallet (July 3, 2020). "https://medium.com/@yurywallet/python-credit-scoring-modeling-under-a-hood-c0cecdeec1d7"