Home Credit Default Risk

Capstone Proposal

Omar Alqasem
September 29th, 2018

Proposal

Domain Background

Credit risk is the risk of not meeting the legal obligations from a borrower not making the payments required. Banks uses strict credit risk management to minimize lending borrowers who fail to repay the obligated payments upon agreed terms based on applicant credit history. Therefore, many applicants with insufficient or non-existing credit history struggle to get loans. As a result, these applicants are taken advantage of by untrustworthy lenders.

Problem Statement

Home Credit Group's (HCG) goal is to open their doors for applicants with insufficient or non-existing credit history to apply for loans and receive a positive loan experience. At the same time, HCG wants to eliminate applicants who fail to repay their obligated payments. Therefore, in order to help eliminating rejections for capable borrowers of repaying their loans, a solution will be implemented that uses machine learning to predict applicants' ability to repay the loan through a set of data including telco and transactional information.

https://www.kaggle.com/c/home-credit-default-risk

Datasets and Inputs

The dataset is taken from Kaggle competition for Home Credit Default Risk https://www.kaggle.com/c/home-credit-default-risk/data.
The dataset contains 221 columns and around 300k records which include information related to loan applicants such as gender, number of children, total income, whether applicant owns a car, whether applicant owns a reality, family status, income source, education level, days of birth …etc. The dataset also includes information related to the applicant's loan such as credit amount of the loan, loan annuity, application day of the week, application hour of the day, purpose of the cash loan, contract status of previous application, and most importantly whether the client has fully paid the loan. The data is already split for training and testing data. Therefore, there's no need to do any further splits for this project.

The dataset provided will help distinguishing clients whom had successfully made full loan payments and those who failed to do so. As a result, this shall help solving the classification problem of predicting new applicants' ability to repay the obligated loan based on previous applicants' information.

Solution Statement

Home Credit Default Risk is a classic supervised learning problem. The aim for solving this problem is to train a classification model to predict whether a new applicant is capable of repaying the obligated loan installments.

Benchmark Model

There are many existing researches that predict whether a loan will default using different datasets [1]. These studies used confusion matrix results as a benchmark model to compare with. The aim for this project is to use the same approach by benchmarking the final results with a confusion matrix results.

[1] http://rstudio-pubs-static.s3.amazonaws.com/203258_d20c1a34bc094151a0a1e4f4180c5f6f.html

Evaluation Metrics

The aim for this project is to use Accuracy, Sensitivity, Specificity, as evaluation metrics. These can be computed as follows:

Accuracy = TP + TN / P + N
Precision = TP / TP + FP
Sensitivity = TP / P = TP / TP + FP
Specificity = TN / N = TN / FP + TN

Project Design

The project will follow a series of steps that are required for each and every machine learning project. These steps can be further explained as follows:

1- Getting to Know Data
This step involves getting a sense of how data looks like and try to deeply understand and make sense of it. It includes visualizations and calculating few statistics that gives valuable insights about the dataset.

2- Data cleansing and preprocessing
This step includes removing columns with missing data, converting categorical variables into binary numerical, removing outliers, scaling and normalizing features …etc. The data has already been split into training and testing datasets, therefore, no need for further splitting.

3- Model Selection

This step includes trying out different classification models to find best performant one. There are many classification models that can be used for this project including Linear Classifiers, Support Vector Machines, Decision Trees, Random Forest …etc.

4- Testing and Benchmarking
This step ensures that the model selected and the final results are competing the results given from the benchmarking model and that final result is not under-fitting nor over-fitting.