Machine Learning Engineer Nanodegree

Capstone Proposal

Shubhendra Kumar
November 26th, 2018

Proposal

Domain Background

The project background is based on analysis performed on potential customers or candidates in financial industry especially by investing firms and banks. Since banks are those institutions which generate most of their revenue by giving out loans hence it is very important to analyze potential customers which can contribute to generating targeted revenue in the banking industry. Since the fundamental background of customers differ from each other, hence it becomes of utmost importance to leverage machine learning to analyze financial history and background of the customers to generate meaningful and predictive insights from their data.

Using machine learning, we can create Models that could predict or classify whether the customer will be interested in utilizing services and products offered by banks and hence efforts could be directed to engaging those customers and on the other hand predicted, less interested customers can be targeted with much better schemes or service rates that they couldn't refuse to the offers made to them. This article https://www.stoodnt.com/blog/scopes-of-big-data-data-science-in-the-banking-finance-fintech-sector/ illustrates the impact of data science in the functionality of financial services and institutions such as corporate banks.

Also this research paper（ https://www.researchgate.net/publication/327011881_Predicting_Credit_Worthiness_of_Bank_Customer_with_Machine_Learning_Over_Cloud) on predicting credit worthiness of a bank customer user data science and machine learning to generate insights and profits for a bank.

Problem Statement

The problem statement is to classify customers of a financial institution so as to find that they would be interested in signing E-loan services offered or not by a financial bank or a firm, keeping in mind the financial history of customers. To process each of the customer's financial background manually, would not simply suffice, hence this problem requires a supervised machine learning model that can learn from the existing data and make predictions for customers. The financial parameters such as their age, years of employment, debt etc. could be utilized as inputs of the learning model.

Datasets and Inputs

The dataset has been obtained/downloaded from Super Data Science Machine Learning Practical (https://www.superdatascience.com/machine-learning-practical/) Module 5. The dataset is a real time dataset but the identity of the users has been anonymized. The dataset contains the financial history of the customers that will be used to build up a classification model to find whether the user will be interested in signing an E-Loan or not when offered by a financial institution.

The parameters of the datasets are:

1.) entry_id = User's entry id/ unique identification

2.) age = user's age

3.) pay_schedule = how often the applicant gets payed

4.) home_owner = owns a home or not

5.) income = monthly income

6.) years_employed = years passed since users started to do job

7.) months_employed = months employed after the previous job year completion

8.) current_address_year = years living in the same house till today

9.) personal_account_m = months for which user has personal account after the latest year completion

10.) personal_account_y = years for which user has had account

11.) has_debt = has pending debt

12.) amount_requested = amount requested by user from the financial institution

13.) risk_score, risk_score_2, risk_score_3, risk_score_4, risk_score_5 = risk score attached with the customer which signifies the risk percentage that the user shall be able to return money or not within the time allotted.

14.) ext_quality_score, ext_quality_score_2 = external quality score of a customer

15.) inquiries_last_month = inquiries made in the last month by the user

16.) e_signed = signed an e-loan when offered by the financial institution

These shall be the input parameters for the classification model.

The dataset contains a total of 17908 rows.

The no of customers that didn’t sign the E-loan = 8270 and those who signed = 9640. Therefore the dataset is pretty much balanced and this is also beneficial for our model as it contains sufficient classes to learn from where in this case it will be a classification model.

Solution Statement

The solution is implemented by developing a classification model for the problem statement. At most two classification algorithms: Support Vector Machines and Logistic Regression shall be selected and fed to the input features and the model shall be trained with the training data. The model that gives the best metrics (accuracy) on the test data, shall be selected and will be hyper-tuned to enhance its performance. This shall be implemented using k-fold cross validation and grid search algorithm available under scikit learn library in python

Benchmark Model

For benchmarking, the simplest of the classification model i.e. Logistic Regression shall be use. As Logistic Regression model is well built for binary classification, hence we chose it as a benchmarking model. Training time and predicting time of this model shall be evaluated and compared to the actual implemented model and finally will make predictions on the initial test dataset once again to determine the test accuracy. Also other metrics like precision, f1-score and recall shall be evaluated using classification report and confusion matrix.

Evaluation Metrics

The model shall be evaluated on the basis of accuracy metric where accuracy = (true positives+true negatives)/(true positives+true negatives + false positives +false negatives). More the evaluated accuracy, more is the success of the classification model which means that the model more accurately classifies the customers into the categories of signing or not signing the E-loan. The classification report for the model shall also generate the evaluation of precision and recall scores too. Also we shall evaluate f1 score because f1-score mathematically represents the balance between precision and recall scores as f1-score = (2*precision*recall)/(precision+recall) i.e. harmonic mean of precision and recall scores.

Project Design

The project shall start by incorporating the dataset into the IPython notebook. At the first stage, the data shall be tried for cleaning where presence of any possible outliers or Null values shall be removed or rectified. Then Exploratory analysis shall be performed which would include several statistical visualization plots to understand the data and features more clearly. This phase shall be followed by selecting appropriate features for our models and then developing two of the popular classification models i.e. logistic regression and SVM for evaluation on the test set. The best performing model shall undergo hypertuning using grid search technique and will perform again on the test set to determine the final performance and the hyper-tuned model’s accuracy on the test set.