Machine Learning Engineer Nanodegree

Capstone Proposal

Omar Villa

November 2nd 2018

Proposal

Abstract

The project proposal is to create a ML model to accurately predict Home Credit Defaults Risks in order to do this we will have to merge several tables and analyze the data using a feature engineering technique with Featuretools module and utilize chart tools as seaborn to present the data and use scikit-learn or keras to get predictions.

Domain Background

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of

repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful. [1]

Supervised learning is one of the most successfully utilized methods in the machine learning field when it comes to predict future values of many different assets in this case our customer (Home Credit) needs help to predict Home Credit Defaults and with this reduce risk, this will not only help the lender to avoid high risks but also will help their applicant to get unnecessary debt. With the use of Machine Learning tools our customer will reduce the human error at the moment their employees evaluate potential clients.

For this project we will refer to the top 2 kernels under the kaggle competition:

Start Here: A gentle Introduction [2]

Home Credit: Complete EDA + Feature Importance [3]

Problem Statement

The Home Credit Default Risk competition is a supervised classification machine learning task. The objective is to use historical financial and socioeconomic data to predict whether or not an applicant or client will be able to repay the loan. This is a standard supervised classification task:

- Supervised: The labels are included in the training data and the goal is to train a model to learn and predict the labels from the features
- Classification: The label is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying the loan)

Dataset

The data is provided by Home Credit, a service dedicated to provide lines of credit to the unbanked population. [4]

There are 7 different data files:

- ✓ Application_train/application_test: the main training and testing data with information about each loan application at Home Credit. Every loan has its

own row and is dentified by the SK_ID_CURR. The training application data comes from the TARGET with indicating 0: the load was repaid and 1: the load was not paid.

The Application_train dataset haves a 122 columns and 307,512 rows, which is big enough to split it and create a test dataset considering sometimes kaggle testing sets don't come with the correct answers.

To help with a potential inbalance problem after we run the first part of the analysis and clean up the data we will use performance metrics as Confustion Metrix and calculate Precision, Recall and F1-Score

- ✓ Bureau: data concerning client's previous credits from other financial institutions. Each previous credit as its own row in the bereau and is identified by the SK_ID_BUREAU, each loan in the application data can have multiple previous credits.
- ✓ Bureau_balance: monthly data about the previous credits in the bureau. Each row is one month of the previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
- ✓ Previous_application: previous application for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK_ID_PREV
- ✓ POS_CASH_BALANCE: monthly data about previous points of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
- ✓ Credit_card_balance: monthly data about previous cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
- ✓ Installments_payments: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

Solution Statement

We want to understand the relationship between the 7 datasets and its features, the outcome may sound very straight forward but because we have so many features we can fall in an overfitting problem because of this we may want to analyze the information to see if we need to do some cleanup or even reduce the number of features with tools as PCA. At the moment we consider to implement a classification algorithm and a Neural Network algorithm for the prediction but during the progress of the investigation this might change based on the output.

Benchmark Model

As this is a kaggle competition a benchmark model would be SK_ID_CURR in the test set, who must predict a probability for the TARGET variable. The submission will be evaluated on area under the ROC curve between the predicted probability and the observed target.

We will Logistic Regression and KNN algorithms as our benchmark model because of their simplicity to create a baseline score

Evaluation Metrics

Our evaluation metrics will be based in a ROC Curve because is what the competition specification requests but for learning purpose we will also include a Confusion Metrix to compare results.

Project Design

Because we are predicting a category based on the scikit-learn map [5] we will follow the classification method first and at least implement one of the different algorithms under the classification umbrella as SGD Classifier or KNN considering that most of data can be numeric but Naïve Bayes is in the picture if there is more text in the data of what we expect.

References

[1] H. Credit, "home credit default risk," September 2018. [Online]. Available: https://www.kaggle.com/c/home-credit-default-risk.

[2] W. Koehrsen, "start here a gentle introduction," September 2018. [Online]. Available: https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction.

[3] Lathwal, "kaggle.com," September 2018. [Online]. Available: https://www.kaggle.com/codename007/home-credit-complete-eda-feature-importance.

[4] H. Credit, "kaggle.com," Home Credit, September 2018. [Online]. Available: https://www.kaggle.com/c/home-credit-default-risk/data.

[5] Scikit-Learn, "machine_learning_map," 2017. [Online]. Available: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html.