Machine Learning Engineer Nanodegree

Capstone Proposal

Jaime Ignacio Rovirosa
June 2nd, 2018

Domain Background

For the capstone project I want to use the Kaggle competition “Home Credit Default Risk” [1]. The main objective of the competition is to determine how capable is a person to repay a loan based on non-traditional information.

The lending industry is a mature one with a long history. In the US most lenders use a FICO-based model to determine who to lend to and at what interest rate. There are three main credit bureaus: Equifax, Experian and TransUnion. The FICO was first introduced in 1989 and is made up of the following components [2]:

1. 35% for payment history
2. 30% debt burden
3. 15% length of credit history
4. 10% types of credit used
5. 10% recent searches for credit or hard credit inquiries

Among the many shortfalls of the FICO model is that the FICO score may vary between the bureaus depending on the information they have collected. Another problem is that consumers without sufficient credit history and information have a hard time to secure a loan or do so at a very high interest rate, or even worse, they need to resort to terrible terms with predatory lenders.

This a clear example where Machine Learning and Artificial Intelligence can leverage more information about consumers to determine more efficiently their creditworthiness. Providing transparency and opportunities to underserved populations provides purpose to the work that companies like *Home Credit* or *Upstart* are doing and I’m inspired by this novel use of technology to improve our society.

Problem Statement

The main objective of the competition is to determine whether a person will be able to repay a loan. The challenge is that the traditional information used in the FICO-based model to answer such question might not be available or complete. In its place a broader, more varied information set is presented to assess an individual's creditworthiness.

Besides making sure a person capable of repaying the loan is not rejected, a secondary objective is to be able to determine the loans' principal, maturity and repayment schedule [3].

Datasets and Inputs

All the information for this competition is proprietary and provided by *Home Credit* via *Kaggle* [4]. Most of this information has been gathered by Home Credit directly and the rest comes from the credit bureaus. There is no ethical conflict with the customer data as the data is properly anonymized.

The data can be broken down into three main categories. The main one is the information about the applications with Home Credit. The second one is information regarding previous applications, repayment history, monthly balance snapshots of loans and credit cards applicants had with Home Credit. The final category contains information coming from credit bureaus regarding previous credits with other financial institutions and their monthly balances.

There are approximately 200 attributes to consider about the applicants and their previous behavior. This includes plenty information about previous loans and credit card payments. Some attributes in the data set have already been normalized and most attributes have sparse data.

Solution Statement

As stated in the competition evaluation section "Submissions are evaluated on area under the ROC curve [5] between the predicted probability and the observed target" [6]. The solution that needs to be submitted for this kaggle competition is very simple, for every loan application in the data set a predicted probability between 0 and 1 needs to be predicted.

Benchmark Model

As part of the Kaggle competition, the data set includes a sample submission. This submission has the same value of 0.5 assigned for all loan applications in the dataset. This is equivalent to assigning a 50% chance to all the loans without considering any data and would represent a good naive benchmark. When looking at the Leaderboard this sample submission has a score of 0.5. Another higher benchmark is given for the an estimation using a Random Forest regressor, and it has a score of 0.688. Two weeks into the competition and the current high score in the leaderboard is 0.800 [7].

Evaluation Metrics

The submissions for the competition are evaluated on area under the **ROC curve** between the *predicted probability* and the *observed target*. To draw a ROC curve we need to know the *true positive rate* (**TPR**) and the *false positive rate* (**FPR**) from the confusion matrix. The **TPR** defines how many correct positive results occur among all positive samples available during the test.

FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test [8].

The **TPR**, also known as *recall* or *sensitivity*, is calculated as the ratio between the sum of the true positive and the sum of the condition positive values. Similarly, the **FPR**, also known as *fall-out* or (1 - *specificity*), is calculated as the ratio between the sum of the false positive and the sum of the condition negative values [9].

A good evaluation metric would be one that measures both of this values. The *positive likelihood ratio* or ***LR+***, is one such value and is calculated as the ratio between the *sensitivity* or the **TPR** and (1 - *specificity*) or the **FPR** [10].

Project Design

To solve the *Home Credit Default Risk Kaggle* competition I will take the following steps. First I will start by **exploring the data** and getting familiar with it. To do so I will create multiple visualizations, observe the distributions of the data and analyze the feature set. The result of this step will be used to determine what actions are need to preprocess the data, which will be the second step.

To **preprocess the data**, I will prepare it. I will check to see if there are any invalid or missing entries, normally called cleaning; then I will check the format and structure of the data. After preparing the data I will convert any *categorical variables* (non-numerical) into *numerical features* using the *one-hot encoding* scheme. Finally, I will check if there are any highly skewed continuous features and perform a transformation; and *normalize* the numerical features by applying a scaling to the data. This scaling does not change the shape of each feature distribution and it ensures that each feature is treated equally when applying supervised learners.

The third step will be to shuffle and **split the data** (both features and their labels) into training and validation sets. Normally 80% percent of the the available data is used for training and the remaining 20% for validation.

The fourth step will be to determine and codify the **evaluation metrics** to measure the model performance. I proposed using the *positive likelihood ratio* to evaluate the model's performance but I would also consider other alternatives like the *F1 score* and the *R squared* (or the *coefficient of determination*) score.

The fifth step will be to do **model selection**. The benchmark model provided for the competition is the random forest. I will start with this model and see how close I can get to the benchmark score of 0.688. This is the list of the models I will try:

1. Gradient Tree Boosting or Gradient Boosted Regression Trees (GBRT) [11]
2. AdaBoost Ensemble model [12, 13]
3. Stochastic Gradient Descent Classifier (SGDC) model [14]
4. Light Gradient Boosting Machine model (or LGBM) [15]

For each model I will:

1. Fit the model with the train data.
2. Use the trained model to perform predictions on the validation data.
3. Use the *evaluation metrics* (*LR+, f1_score, r_square*) to measure the performance.
4. Tune the model hyperparameters using *grid search* [16] and repeat steps 1 through 3 above.
5. Determine *feature importance* (which features provide the most predictive power)
6. Submit the prediction results and assess the results versus the naive and benchmark predictors and the other submissions.

After doing all the above in step 5, I would consider which other models could be used, find other ways to tune the hyperparameters and review if the data needs to be process further to obtain a higher submission score. With all this information I will elaborate the capstone project report.

Sources:

[1] https://www.kaggle.com/c/home-credit-default-risk
[2] https://en.wikipedia.org/wiki/Credit_score_in_the_United_States#FICO_score
[3] https://www.kaggle.com/c/home-credit-default-risk
[4] https://www.kaggle.com/c/home-credit-default-risk/data
[5] http://en.wikipedia.org/wiki/Receiver_operating_characteristic
[6] https://www.kaggle.com/c/home-credit-default-risk#evaluation
[7] https://www.kaggle.com/c/home-credit-default-risk/leaderboard
[8] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#ROC_space
[9] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Basic_concept
[10] https://en.wikipedia.org/wiki/Likelihood_ratios_in_diagnostic_testing#positive_likelihood_ratio
[11] http://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting
[12] https://en.wikipedia.org/wiki/AdaBoost
[13] http://scikit-learn.org/stable/modules/ensemble.html#adaboost
[14] http://scikit-learn.org/stable/modules/linear_model.html#stochastic-gradient-descent-sgd
[15] https://github.com/Microsoft/LightGBM
[16] http://scikit-learn.org/0.17/modules/generated/sklearn.grid_search.GridSearchCV.html