Machine Learning Engineer Nanodegree

Loan Default Prediction

Ricardo Szczerbacki
October 6, 2019

Proposal

Introduction

In the credit loan business, excessive uncertainties about credit applications default risk are dangerous to credit companies and borrowers alike.

Greater uncertainties lead to more restrictions to credit applications and too many restrictions can impede potential capable creditors' access to the credit they need and reduce the income of credit companies, while insufficient restrictions can impose unpayable debts to clients and losses to loan companies.

Looking forward to reduce this uncertainty and broaden the financial inclusion for the unbanked population, the competition “Home Credit Default Risk” was created on **Kaggle ®,** by the company **Home Credit ®,** focused on the search of algorithms to better predict the default credit applications based on data available of the applicants, their assets and their previous credit applications and payments history. The competition can be accessed in https://www.kaggle.com/c/home-credit-default-risk.

The main idea is that, provided that we have enough previous loans information, including the asset data and credit history of the applicants, and the “outcome” of the loan, i.e., if the clients were indeed able to pay the loan, a Machine Learning algorithm could be used to create a model that, to a certain extent, predicts if an applicant will have problems paying his loan.

We can see this as a binary classification problem, where we want to predict, with a certain probability, if a borrower will have payment difficulties or not.

About the Data

The dataset used is the one originally provided for the competition and contains a train set with 307,511 previous loan applications, with information about the application and the applicant, including data on previous credits (and payments) from the credit company and from other financial institutions and repayment history for the previous disbursed credits.

The data available includes relevant information about the applicant’s car, house, income, education, occupation, marital status, children, documents presented

and other information that should, in different degrees, correlate to the capacity of the applicants to honor their debts.

Also, the data on previous loans and payments can, at some extent, help to identify behavior patterns that could have influence on the probability of a credit default loan.

Solution Statement

As the first step I propose to generate more features combining the existing ones with previous knowledge. Then I will use three classifiers: a Logistic Regression, that is one of the most used classifiers; a Random Forest, that has outperformed other Machine Learning methods in previous studies on Loan Default Prediction [1][2]; and a Gradient Boosting strategy using the XGBoost [3], as one of the most popular classifiers in Machine Learning hackathons and competitions.

At the end the models will be evaluated individually and combined trying to maximize the quality of our predictions.

Benchmark Model

Although there are previous studies for this domain, as the data chosen as predictors vary widely in these studies, the direct comparison of the results obtained is not adequate.

Therefore, I chose to benchmark the proposed solution with two models: a random prediction algorithm, as the baseline model to be surpassed, and the winning contest solution, as the state-of-the-art classifier for this specific data, but that is too much complex to be adopted as a general and practical solution.

Evaluation Metrics

For the evaluation metric I will use, as proposed by the competition, the area under the ROC (Receiver Operating Characteristic) curve, between the predicted probability of an application being default and the observed target.

The area under the curve (AUC) relates to how well the model is capable of distinguishing between classes and is often used for model comparison in Machine Learning [4].

Project Design

The first step is to perform some exploratory data analysis (EDA) with the following objectives:

- Find missing values and for each situation give the proper solution (like row deletion, mean/median replacement, constant replacement, backward/forward filling, etc).
- Transform categorical features in numerical, using the proper encoding.

- Check for skewed data and transform the data if necessary, to obtain better results.
- Check for strong correlations between features and use this knowledge to help to decide upon features removal.

The next step is to create new features based on the combination of existing features and prior knowledge. Also, new features will be created based on the information on payments and installments for each applicant.

Then a tree classifier will be used to estimate the features importance and create a ranking of the features. This information will be used to perform the feature selection, reducing the number of features, but keeping the most important ones. After the full implementation of the methods, different scenarios varying the number of features should be evaluated.

To evaluate the models that will be created, a k-fold cross-validation will be used with the AUC-ROC metric. Initially 10 folds will be used, but this number may change if it seems necessary.

The first model to be implemented and evaluated is a simple random prediction algorithm, that chooses at random a classification in each prediction. This will serve as a baseline model.

Then will be implemented a Logistic Regression followed by a Random Forest and the XGBoost model.

For the Random Forest and the XGBoost a Grid Search or a Random Search strategy will be used for the hyperparameters tuning.

At last, the three models and also each two models will be combined using a simple ensemble averaging strategy. The combinations generated will also be evaluated as candidates.

The model with the best final score will be the final proposed solution.

References

[1] Alomari, Zakaria. (2017). Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications. New Zealand Journal of Computer-Human Interaction.

[2] Lifeng Zhou, Hong Wang. (2012). Loan Default Prediction on Large Imbalanced Data Using Random Forests. TELKOMNIKA (e-ISSN: 2087-278X), Vol.10, No.6, October 2012, pp. 1519~1525.

[3] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. KDD.

[4] Wikipedia contributors. (2019). Receiver operating characteristic. In Wikipedia, The Free Encyclopedia. Retrieved 24, October 5, 2019, from: https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=917442289