Machine Learning Engineer Nanodegree

Capstone Proposal

Roberto Pinto Martins Junior
February 14th, 2018

Domain Background

Access to credit is important for market economies to develop and flourish, having a big impact on companies and people lives. Banks and financial institutions play an important role in the market, defining who is going to have access to credit.

Money is a limited resource and financial institutions try to evaluate the risks using methods to guess the probability of default and then decide whether or not a loan should be granted. One of these methods is credit scoring which assigns a score to the borrower - it helps the bank evaluate the risk of default and compute interest rates. Using this score, a bank may also assign some credit score category or band, such as bad, poor, fair, good or excellent.

There has been an increase in the interest for machine learning models in credit risk and scoring, as observed in this paper[1] from Moody's, the risk analysis agency. It has also become a business where you can build models: GiniMachine.com.

This proposal is based on Kaggle competition "Give Me Some Credit"[i].

Problem Statement

The goal is to create a model that predicts borrowers' credit score so the bank will be able to determine the likelihood of a default and then compute the interest rate for the loan. In this binary classification problem, the model will classify if the potential borrower will experience a 90+ day past due delinquency, but will also output the probability of the classification. This model may also be used by borrowers to help them make the best financial decisions.

---

[1] "Machine Learning: Challenges, Lessons and Opportunities in Credit Risk Modeling": https://www.moodysanalytics.com/risk-perspectives-magazine/managing-disruption/spotlight/machine-learning-challenges-lessons-and-opportunities-in-credit-risk-modeling

Datasets and Inputs

This project dataset contains financial and credit information from 251,503 anonymous individuals, already divided in 150,000 records for training and 101.503 for testing, although this test set is meant to be used when submitting the results in the competition.

While performing a quick exploration of the dataset, it was observed to be definitively an unbalanced one. Our target class is sparse - individuals who experienced credit distress account for only 6.68% (10,026) of our samples. This has some implications to the model, because if it considers that all samples fall into the delinquency category (naïve approach), it will still have a 93.32% accuracy.

For our analysis, the training dataset will be split in training and test sets while also accounting for the sparse characteristic of the target class. Information such as age, monthly income, number of credit lines or loans, debt ratio and number of times the borrowers has been past due are included in the dataset, depicting each person's financial "life-style" or condition.

This dataset will be used to predict the probability of experiencing 90 days past due delinquency in the next two years and therefore will help not only financial institutions defining credit score but will also give a tool to borrowers understand how everyday decisions may affect their financial life as well.

Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer

NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

This dataset was obtained from the Kaggle competition "Give Me Some Credit".

Solution Statement

The solution will predict if the borrower will be in financial distress within two years and also provide the probability of being in that category. First, some aspects of the dataset will be visualized to improve understanding over the dataset and then prepared for being processed in the training and testing stages.

In this binary classification solution, some models will be tested, such as Gradient Boosting, KNeighbors and SGD. These models will be evaluated by metrics such as f-beta score and area under curve.

Benchmark Model

As observed before, our dataset is unbalanced – samples of credit delinquency account only for 6.68% of the samples - and therefor one possible benchmark model is the naïve predictor, which will consider that all borrowers experienced 90+ days past due delinquency.

Additionally a linear logistic regression (with no parameter tuning) will be used to benchmark.

Evaluation Metrics

Evaluation of the models will be based on the requirements of the Kaggle competition: area under curve (AUC); after computing the receiving operating characteristic (ROC) and therefore, the closer the classifier gets to 1.0, the better.

Additionally, it's also going to consider the F-beta score for model evaluation, because it's the kind of problem that recall is important. This importance is crucial to the bank or credit institution, which need to correctly classify bad payers and thus avoiding giving credit to someone who is going to be late or not

paying at all. Meaning that the model must show a high rate of correctly classified bad payers over all the borrowers who previously experienced credit delinquency.

Project Design

The solution tries to predict if someone will be in the SeriousDlqin2yrs category, which indicates whether or not someone had experienced financial distress in the past two years, in the form of 90 days or more of past due delinquency. The credit score will be the probability of being the SeriousDlqin2yrs category.

Firstly, a better understanding of the dataset is needed and therefore an exploration of the data set, followed by visualization and data preparation will be performed. In this preparation, the best approach to missing information will be evaluated and features encoded, if needed. The dataset will also be split in training and test sets.

Next, to predict if someone will be in financial distress, some training models will be used and compared, such as Gradient Boosting classifier, KNeighbors classifier and SGD classifier. To evaluate the best model, two methods will be used: area under curve (a competition requirement) ad f-beta score

And finally, the most promising model will be selected and improved, fine tuning its hyperparameters using Grid Search or other suitable method of evaluating the best parameters combination.

Additionally, if the model implementation provides, the most significant features will be selected and the model tested again, but this time looking for improvements. Also a PCA analysis may be used to understand the dataset variance and if the new components improve the model.

References

---

[i] *"Give Me Some Credit" https://www.kaggle.com/c/GiveMeSomeCredit*