American Express - Default Prediction

Tianyi Cheng

Problem Statement:

The topic is mainly about using machine learning to predict credit default.

The target binary variable is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.[1] And it provides five categories of features: delinquency variables, spend variables, payment variables, balance variables, and risk variables. The goal is to predict the probability of a future payment default.

The metric to evaluate is $M = 0.5.(G + D)$. G is the normalized Gini coefficient and D is the default rate captured at 4%.

After checking the data and requirements, there are some specific difficulties. First, the data is very large, so the computation complexity has to be reasonable and it is important to choose the right data type to process. Second, some data is missing in the data set, leading to extra disturbance during training and predicting. Last, researchers already have some prior knowledge about this topic. But it still takes some work to be employed in the model.

On the other hand, large data makes some special methods practicable.

Applications

First and foremost, it will help America Express to better predict the future default, paying less cost at the same time. There is no doubt that data-driven prediction is more reasonable than sensorial prediction.

Meanwhile, data-driven prediction can be used in many other aspects of daily life, for example, the prediction of the remaining useful life of lion batteries based on former charging data.

Literature review and Open Source research

After reviewing the description, I have found that the task is mainly about processing the original data and building the classification model.

The most distinguishing feature of this dataset is its number of samples. There are variables of over 5.53 million customers. It has a large effect on both the demand for processing the data and enables lots of unusual classification models. Meanwhile, the data is very large, so the computation complexity has to be reasonable and it is important to choose the right data type to process.

Data processing

Description: The target binary variable is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.[1]

Change the data type. The figures are mostly stored as float. Considering that all of them have been normalized, we don't need that high precious. Determine the final data type according to the requirements of the model

Dealing with empty data. Some samples miss some features. For some certain features, over 20% of the samples are missing. There are mainly three ways to solve this. First of all, we can simply put away these features because there are 190 features in total. And some data is even irrelevant to the result. But it depends on the result after we have the test results. On the other hand, we can combine some of the features to train a special model, which is can be used for the samples missing lots of features, and use the remaining data to train a model with better feasibility.

Feature Selection. Some features have little relevance to the result. In most cases, they would not help classify the samples, and their abnormal value may even disturb the outcome. So, before starting to train the model, we have to select the most prevalent features. They might need some process to be useful, such as taking the logarithm.

Priori knowledge. Different from other black-box problems, some researchers have already built up some theories about it. If we can bring these theories into our model,

our accuracy would undoubtedly increase a lot. What's more, a big disadvantage of machine learning is that it lacks interpretability. And prior knowledge is quite useful on this issue.

Dealing with outliers. To be honest, because the data has been standardized, it is usually the case that we cannot tell outliers and distinguished features. it would be solved only after we have already built our model and got a better understanding of the features.

Models to process data

I have searched for some models which may be useful for processing data.

1. SVM.

Support vector machine maps training examples to points in space to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.[2]It can help avoid overfitting; but considering that the data set is so huge, the amount of calculation may be too large\

2. Boosting

In bagging, a random sample of data in a training set is selected with replacement. These data are used to train a small classification model, we call them weak learners. After we have trained enough models in parallel, we will these models to classify together.

For boosting, and bootstrap aggregation we also have weak learners. And they learn sequentially. During the process, the weights of the misclassified data in the previous model are increased to improve the performance.

3. ANN

Considering that the data is so huge, we can also use some artificial neural networks.

Literature the be reviewed and open source

[1] Thomas L C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers[J]. International journal of forecasting, 2000, 16(2): 149-172.

[2] Pandey A, Shukla S, Mohbey K K. Comparative Analysis of a Deep Learning Approach with Various Classification Techniques for Credit Score Computation[J]. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 2021, 14(9): 2785-2799.

[3] Mushava J, Murray M. A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function[J]. Expert Systems with Applications, 2022, 202: 117233.

Citation

[1]https://www.kaggle.com/competitions/amex-default-prediction/overview/evaluation

[2] https://www.wikipedia.org/