Machine Learning Classification Algorithm Study on Loan Overdue Judgment

Wenjin Xu[1, 2], Yuan Feng[1, a, *], Di Zhou[3]

[1]Ocean University of China, Qingdao, Shandong, 266101, China

[2]Qingdao University of Science & Technology, Qingdao, Shandong, 266069, China

[3]The 91049 Army of the PLA, China

[a]15954263549@163.com

*Corresponding author

Keywords: Classification Algorithm, Stochastic Forest Algorithm, Anti-fraud, Loan Overdue

Abstract: With the development of Internet finance, in the field of financial anti-fraud, more and more accurate methods are needed to make users and enterprises have a two-way credit guarantee. This paper mainly studies the classification algorithm of machine learning, especially the stochastic forest algorithm. And applies it to the field of financial anti-fraud, and determines whether the user's credit in the overdue judgment of the user's loan.

1. Introduction

This paper mainly studies the classification algorithm of machine learning, especially the stochastic forest algorithm, and applies it to the field of financial anti-fraud, and determines the user's credit in the overdue judgment of the user's loan.

With the development of Internet finance, in the field of financial anti-fraud, more and more accurate methods are needed to make users and enterprises have a two-way credit guarantee; for enterprises, the user's credibility means whether to lend to the borrower; The machine learning algorithm can mine the user's historical borrowing information and getting the detection model. This is only one direction of anti-fraud research to achieve a forecast of whether users will be overdue.

This article mainly uses the random forest algorithm to carry out the application research. The full text can be divided into three parts, the first part is a small example to have a simple understanding to the machine study; the second is the definition of concrete explanation of machine learning algorithm meaning; and the third part is to understand the machine learning historical process and the basic concept.

This paper is organized as following: first we give some brief explanations for some algorithms commonly used in the industry at present. The second part turns to the classification algorithm, which is mainly introduced to understand the other algorithms contained in today's classification algorithm; then the paper briefly introduces the stochastic forest algorithm, and compare the loan overdue decision problem with other classification algorithms; the reason of choosing random forest algorithm is explained. The third part is the application of stochastic forest algorithm to the decision of overdue loans, including data set processing, feature selection, model training, etc.

2. MaCHiNE LEARNING in industry

Nowadays, the Internet covers the whole world and the output data of various products are massive. From pictures and words on the Internet to the big data from the industry, data appears in every corner of our lives. Now, the term of artificial intelligence is burning all over the world. In the future, data is still the theme of the Internet era, and machine learning allows us to find hidden rules from the large data. That means in future, data and artificial intelligence are the two keys to the science and technology.

In the past, people studied data and adapted it to the system that produced the data. However,

when the capacity of data is large enough that human themselves cannot perceive, calculate, and make specific plans for it. We have to think of ways to enable machines to learn, summarize, and study from massive amounts of data independently, especially from the intricacy of the connected network. According to the law, we must find a regular way of learning. However, there are so many machine learning shadows in the world; for example, we may realize that the pages that pop up automatically for you in a web page just happen to be what we need, which uses machine learning recommendation algorithms, or the popular fingerprint unlock image recognition.

Nowadays, there are many applications and researches based on machine learning, including bicycles, automobiles, anti-fraud in the financial field, automatic detection in the medical field, speech recognition and so on. Not long ago, people sneered at the ability of a product to apply machine learning. Now, more and more people are seeing the enormous role it can play, which should and is being looked forward to. As we have expected in the past, enterprises can make products that meet people's wishes and understand people's needs better.

Supervised learning and unsupervised learning are most commonly used in machine learning when applying the data of product output, and the anti-fraud behavior I want to study is using the supervised learning algorithm in machine learning. Therefore, the financial field obtains very comprehensive data such as user's personal information, reputation records, consumption records, behavior records, third-party credit records and so on. These are from the Internet, offline question and answer methods [1]. According to these data, we can rely on machine learning algorithm to build up the anti-fraud model of the data in the financial field. That is, to judge whether the user is able to bear the amount of the loan, and to monitor whether the user can repay the loan on time, which is not possible in the traditional financial field. Through accurate estimation, the bank decides whether to lend to the user, reducing the loss of the bank itself because the user may not be able to repay the loan on time.

3. Random forest algorithm for financial

Random forest algorithm is based on decision tree algorithm through a given training sample, first according to the data set to establish a fixed number of decision trees, and then these decision trees together to form a random forest. The decision tree is constructed by self-help repeated sampling. In a given sample data set, N data are retrieved. The N data are integrated into a new sample data set, and then N decision trees are generated in the sample data set. This is the process of establishing the decision tree. Each of these N decision trees is formed by the samples collected above. Classification or regression of input data is based on the results of these N decision trees. Comparing these results, the one with the same result is the final result. Although the structure of N decision trees in random forests is the same, the different sample data also lead to different classification emphasis, and the correlation between trees is also different [9].

When building decision trees, we should pay attention to the following two aspects:

1) Sampling. The random forest algorithm takes sampling, row sampling and column sampling for the input sample data. When sampling, because of the use of a playback sampling method, the final sample data will have the same data. In order to prevent this method from affecting the results, we assume that there are M input samples, and the number of samples is also set to M, so that all the trees in the model do not have all the input samples during the training, avoiding the over-fitting phenomenon.

2) Sample data obtained by the above sampling method should be split completely when the decision tree is established. In order to ensure that the final classification results can be unified and get the same classification, it is necessary to avoid incomplete splitting of decision tree. In random forest algorithm, although there are many decision trees, some processing methods are not the same as decision trees. For example, pruning is sometimes required in decision trees to prevent over-fitting, which is not necessary in random forests.

Through the accuracy test of the above classification algorithm, I found that the accuracy of random forest is the best, followed by neural network and boost, random forest algorithm in the credit

risk prediction has more advantages than other algorithms, I choose to use random forest algorithm to study the application of loan overdue [10].

In the anti-fraud activities in the financial field, the signals are not monotonic and linear, but the random forest algorithm is composed of decision tree, which can process the two signals accurately at the same time. Compared with random forest, neural network is good at dealing with non-linear signals, but cannot deal with non-monotone signals. The worst logical regression is unable to deal with the two signals.

Random forest algorithm has the characteristics of no other algorithm, that is easy to use immediately. In addition, decision tree-based algorithms GBDT, XGBoost, have good results, but the parameter adjustment is more complex than random forest; because of the randomness of the two row variables in the random forest data set, compared with other algorithms, it is easier to avoid the occurrence of over-fitting; similarly, when dealing with the spatial value of data, at random. The impact of the algorithm on the prediction results is also smaller; and since the random forest algorithm does not need to adjust any parameters, its implementation is simpler [11].

3.1 Software preparation and process

The language I use is python, which downloads the latest python3.6 from the official website, and IDE uses pycharm.

Through the application research of random forest on loan overdue, this paper divides into the following steps [12]: data cleaning, feature engineering, model tuning, model training and overdue threshold determination:

The data used in this study include 32853 credit cards with transaction times from January 1, 2015 to January 30, 2017.

According to the two feature tables, data and features are integrated and clear.

One is the LC table, which represents the characteristics of the credit mark. Each row represents one transaction, and has 21 features, including the user name. All of these features are the user information that can be obtained when the transaction is going on. For specific field descriptions, see the data dictionary.

3.2 Data cleaning

First, the general package is required to call this chapter, and the extra package will be specified in every section after that.

3.3 Feature engineering

Feature selection is an indispensable part in machine learning. Choosing the appropriate feature can make the model train more accurately. Sometimes there are no suitable features in the original data, so we have to derive other features from the original features. The process of selecting features is repeated, and it is a step for every machine learning problem to determine the required features by analyzing the requirements of the problem.

The feature engineering of this project is divided into 3 parts: 1, feature selection 2, feature derivation 3, feature abstraction.

1) First of all, the features we captured from this data are video validate, failed count, lender count, first sucessborrowtime, owing amount, credit code. Credit validate, amounttoreceive, success count, overduelesscount, remain funding, graduate school, listing id, lastbidti I (the last transaction time, do not know whether to lend or to borrow, estimated is to borrow), normal count (normal repayment), certificate validation (diploma certification), fistbidtime, borrow name (borrower name), overduemorecount (more than 15 days overdue), study style (education, adults, ordinary), o Wingprincipal (now owed principal), current rate (interest rate), waste count (flow number) Gender, age, months, register time, cancel count, amount, phone validate, educate validate, ncincidentitycheck, auditing time, education and egree.

Through feature comparison, the final selected features are: loan amount, loan term (months), loan interest rate (current rate), loan success date, initial rating (credit code), loan type, whether the

first tender, age (age), gender (gender), mobile phone certificate (phone validate), household registration certification, visual. Video validate, education certificate or certificate validate, credit validate, Taobao certificate, historical success count, historical success loan amount, total owing amount, historical normal repayment period, history Overdue repayment period (this may be less count + more count), overdue

2) When the existing features cannot meet the needs, through the combination of features, generating features that meet the requirements is feature derivation.

3.4 Random forest model adjustment

There are many performance indicators involved in evaluating classification models, such as Confusion Matrix, ROC, AUC, Recall, Performance, lift, Gini, K-S and so on. Here, we use AUC as a judgement index.

Next is the process of adjusting parameters:

First, run according to the default parameters of the model.

```
Rf0 = RandomForestClassifier (oob_score=True, random state=10)
Rf0.fit (x_train, y_train)
Print (rf0.oob_score_)
Y_predprob = rf0.predict_proba (x_train) [: 1]
Print ("AUC Score (Train):%f"% metrics. roc_auc_score (y_train, y_predprob)
OUT: 0.904743272032
AUC Score (Train): 0.985649
```

First of all, the more adjustment is n-estimator, the better.

```
Param_test1= {'n_estimators': list (range (1001001100))}
Gsearch1= GridSearchCV (estimator = RandomForestClassifier (min_samples_split=100).
Min_samples_leaf=20, max_depth=none, max_features='sqrt', random_state=10).
Param_grid =param_test1, scoring='roc_auc', CV=5)
Gsearch1.fit (x_train, y_train)
Print (gsearch1.grid_scores_, gsearch1.best_params_, gsearch1.best_score_)
```

After adjusting three parameters, see if the score of the present model is rising or not.

```
Rf1= RandomForestClassifier (n_estimators= 1000, max_depth=9, min_samples_split=70,
Min_samples_leaf=20, max_features='sqrt', oob_score=True, random_state=10)
Rf1.fit (x_train, y_train)
Print (rf1.oob_score_)
OUT:
```

Zero point nine two four zero nine zero two five six nine four five

Thus, the fraction is higher than that before the parameter adjustment, and then the minimum sample number min_samples_split and the leaf node minimum sample number min_samples_leaf are optimized.

3.5 Predict overdue rate and determine overdue value

First, the model of parameter tuning is trained, and the model is saved. Because the training samples are not small, it takes a lot of time to train the random forest model every time, and the trained models are the same. In machine learning, sklearn algorithm has a special package for model preservation.

In python, there is a joblib that can save the model and take the saved model out for different test suites:

```
From sklearn.ensemble import RandomForestClassifier
RFC = RandomForestClassifier (n_estimators=1000, max_depth=9, min_samples_split=70).
Min_samples_leaf=30, max_features=5, oob_score=True, random_state=10)
Rfc.fit (x_train, y_train)
#y_test_pred = rfc.predict_proba (x_test)
```

Conservation model

From sklearn.externals import joblib

Joblib.dump (RFC,'randomforest.pkl', compress=3)

Remove the trained model, input the test set which is divided into the front, and predict the overdue rate.

Prediction of overdue rate

RFC = joblib.load ('randomforest.pkl')

Y_test_pred = rfc.predict_proba (x_test)

Comparison

Y_pred_df = pd.DataFrame (y_test_pred, columns= ['not overdue,' overdue ')

Y_test_df = pd.DataFrame (np.array (y_test))

Compare_df = pd.merge (y_pred_df, y_test_df, left_index=True, right_index=True)

Look at the comparison between forecast and actual results.

Print (compare_df.columns)

Print (compare_df.describe ())

OUT:

Index ("not overdue", "overdue", 0], dtype='object')

No overdue 0

Count 15335 15335 15335

Mean 0.922646 0.077354 0.074405

STD 0.045143 0.045143 0.262437

Min 0.375076 0.014665 0

25% 0.898167 0.045909 0

50% 0.932045 0.067955 0

75% 0.954091 0.101833 0

Max 0.985335 0.624924 1

Next, determine the overdue threshold.

From the point of view of products, the number of real overdue but predicted to be overdue should be reduced, and whether a more accurate overdue probability prediction value should be given to customers should be considered.

Measure = compare_df [compare_df[0] = = 1]

Print (measure.shape)

H = measure [measure. overdue > 0.08], when the actual overdue period is overdue, that is, the prediction is correct.

Print (h.shape)

Mea = compare_df [compare_df[0] = = 0].

Print (mea.shape)

P = mea [mea. overdue > 0.08] is not overdue.

Print (p.shape)

We set the overdue threshold to 0.07, 0.08 and 0.09 respectively, and get the following results.

Table 1. The Result with Overdue Threshold 0.07

threshold	0.1	truth	…….
……	……	overdue	not overdue
Prediction	overdue	823	6523
……	not overdue	318	7671

Table 2. The Result with Overdue Threshold 0.08

threshold	0.1	truth	…….
……	……	overdue	not overdue
Prediction	overdue	731	5586
……	not overdue	410	8608

Table 3. The Result with Overdue Threshold 0.09

threshold	0.1	truth	…….
……	……	overdue	not overdue
Prediction	overdue	652	4653
……	not overdue	489	9541

From the above results, in order to reduce the real overdue and forecast the number of non-overdue targets, but also take into account the real number of non-overdue prediction of non-overdue targets, I finally choose the overdue threshold of 0.08. Under the condition that the overdue value is 0.08, the number of real overdue forecast is 731, the number of real overdue forecast is 8680, and the correct rate is above 60%. From the above forecast results, the correct rate of overdue forecast is still relatively high.

Acknowledgments

This work is supported by Shandong Critical ProjectcNo.2007GG10004018, and No.2018GGX105005, Key Auditing Project No. 1516SDSJ0102, and Qingdao National Labor for Marine Science and Technology Open Research Project QNLM2016ORP0405.

References

[1] Andrew. Ng. Marchine Learning [M]. Stanford University. 2012.

[2] Poll. Mechine Learning & Algorithm. 2015.06. https://www.cnblogs.com/maybe2030/p/4585705.html

[3] Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China, LingxiaoTang, FeiCai, YaoOuyang, Technological Forecasting and Social Change, Available online 17 March 2018.

[4] A hybrid KMV model, random forests and rough set theory approach for credit rating Knowledge-Based Systems, Ching-Chiang Yeh, Fengyi Lin, Chih-Yu Hsu, Volume 33, September 2012, Pages 166 - 172.

[5] Relationship between Capital Operation and Market Value Management of Listed Companies Based on Random Forest Algorithm, WenLong, LinqiuSong, LingxiaoCui, Procedia Computer Science, Volume 108, 2017, Pages 1271 - 1280.