Application Classification Report on lendingclub data

Machine Learning Engineer Nanodegree li wang

Background

With the development of Peer-to-peer credit industry in recent years, more and more consumers choose to solve the problem of temporary shortage of funds through Peer-to-peer loan. Unfortunately, although this is conducive to the allocation of funds, but the loan applicants for various reasons overdue by the phenomenon of many times, this gives Peer-to-peer company a certain degree of loss of bad debts.

The machine learning algorithm can be used to discriminate the applicants, to deny the users with great overdue risk before the loan is issued, it could control the financial risk and reduce the risk of the enterprise's potential bad debts effectively.

Because I am now working for a P2P company,and i hope that my studies can be better serve my work, so i choose this project as my graduation project.In this project, i will use the well-known Peer-to-peer platform LendingClub Open loan Applicant data for the pre-loan applicants overdue classification model.

reference:

- https://is.cuni.cz/webapps/zzp/download/120269679
- https://www.sciencedirect.com/science/article/pii/S095741741101342X
- https://www.diva-portal.org/smash/get/diva2:824593/FULLTEXT01.pdf
- http://cs229.stanford.edu/proj2014/Kevin%20Tsai,Sivagami%20Ramiah,Sudhanshu%20Singh,Peer%20Lending%20Risk%20Predictor.pdf
- https://www.mwsug.org/proceedings/2016/AA/MWSUG-2016-AA02.pdf

Problem Statement

This project is a two-class supervised learning model designed to use existing data to build a machine learning model that divides all loan applicants into ' good users ' and ' bad users '.

I will expore the basic statistic information of data,and transforme the discrete features to numerical variables through dummy variable coding and Labelcode technique. After that, the whole dataset will be divided into training set and test set, and the machine learning classification algorithm will be used to learning on training set, including parameter tuning,after that the model will by evaluated by evaluation metrics like auc,ks,etc.

Data

In this project, the goal is to use the applicant's basic information, credit card transactions, information and other data to build a machine learning model to identify users. The data used is the internationally renowned Peer-to-peer credit platform LendingClub official website opened in the third quarter of 2017 borrower data, involving 420,000 applicant samples, 145 feature information,including a good customer sample 350,199, and bad Customer sample 66175.It belongs to the sample distribution imbalanced data.

data source:https://www.lendingclub.com/info/download-data.action

Solve Problem

The tools I used were mainly Python and R,python is to complete data preprocessing, feature engineering, modele training and tuning work, R-Packet complete data visualization, and statistical computing work. The data set contains continuous feature and discrete character, the continuous feature needs to deal with outliers, the discrete character needs to be transformed to numerical data, and the continuous and discrete features are missing and need to be treated.

The main pretreatment technologies used include: normalization technology, dummy variable coding technology, missing filling technology, abnormal data detection technology; The feature engineering part includes: Variance threshold technology, lasso regression L1 truncation technique, vif technique, recursive feature deletion technique.

in the model phase, This project chooses the logical regression classification model and the randomforest classification model; the techniques used in the optimization of the model are: Grid search technology, cross validation technology, the main techniques used in model evaluation are: Confusion Matrix, ROC curve, learning curve, KS curve.

Since the data used is a classification imbalance data, this will be adjusted when the model is built, and its classification ratio can be as balanced as possible.the technique i mainly use is give less class more weight.

In the process,i split randomly all data into 70% training set and 30% testing set,all the training process,i only use the training set. tesing set has been used to testing my model. In training process,i will use five folds cross-validation to training model.

Basic Model

In this project, i plan to use LogisticRegressionclassifier and RandomForestClassifier algorithm for training and tuning, and the advantages and disadvantages of the two algorithms are as follows:

- LogisticRegressionclassifier

The predicted result of logistic regression is between 0 and 1, it can be applied to continuous features and classification features, very easy to use and explain, they are good characteristics for consumer financial loan users risk control, because in the business scene, we want to know what factors to affect the borrower's repayment ability.

However, the essence of logistic regression is based on the linear regression model, the premise is that there is no correlation between the features used for modeling, so it is sensitive to the collinearity problem, so it is necessary to consider the relationship between the variables in the model building, in order to achieve a reliable result. The prediction result of logistic regression is ' S ' type, which is transformed by log function, this ' s ' curve at both ends of the probability changes with log value, the probability change is very small, the marginal value is too little, and the change of intermediate probability is very sensitive, which leads to the unsensitivity of the difference between the two ends.

- RandomForestClassifier

Random forest algorithm is an enambeling algorithm based on decision tree, in the actual application scene has the very good effect performance, it can handle the high latitude characteristic and has the very strong anti-interference ability, simultaneously also may process data with missing data, can also output the feature important degree score. it has fast computation speed, and can balance the error of unbalanced data.

However, the random forest algorithm still inherits the local optimal feature of the decision tree, and the feature attribute division is more likely to have a significant impact on the results of the random forest, so a special attention should be paid to the attribute classification in practical application.

In this project, I mainly use the above two algorithms, the logic regression to eliminate the characteristics of the collinearity problem, mainly based on the feature correlation coefficient and vif judgment;

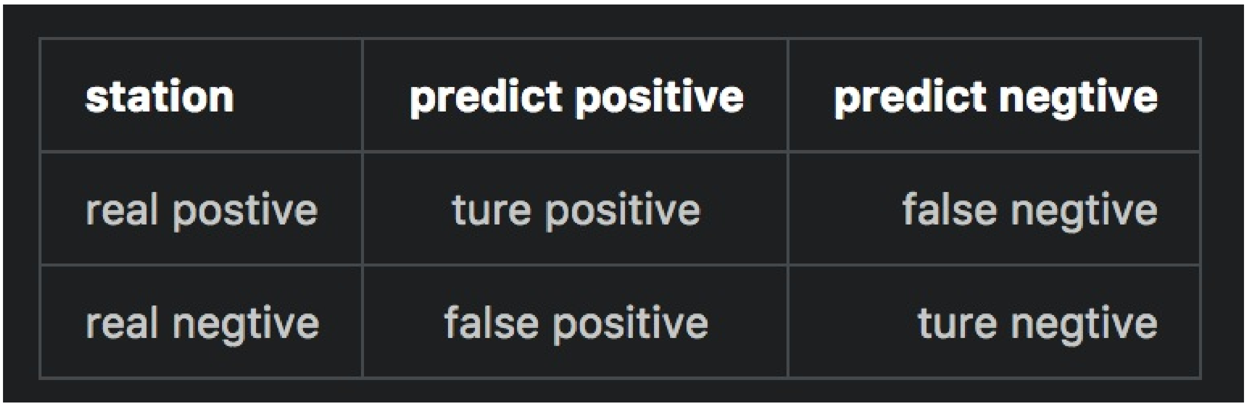
in the random forest algorithm, I mainly use the feature of the chi box, the maximum number of boxes per feature is not more than 5, in order to avoid the effect of excessive attribute value on RandomForestClassifier prediction.

> please attention: in this project i will choose RandomForestClassifier algorithm without any parameter adjusted as my benchmark model

Evaluation Indicators

The model is two classification model, I use the index of confusion_matrix,roc_curve,auc,ks to evaluate the model, because these indicators is much more effction for imbalance data.

confusion_matrix:
two class as example, the row means real class of sample,the columns means predict class of sample. if predict class equal to real class, it means true or it means false.according to this thouht,confusion_matrix can shape four values,true postive ,ture negtive ,false postive,false negtive.

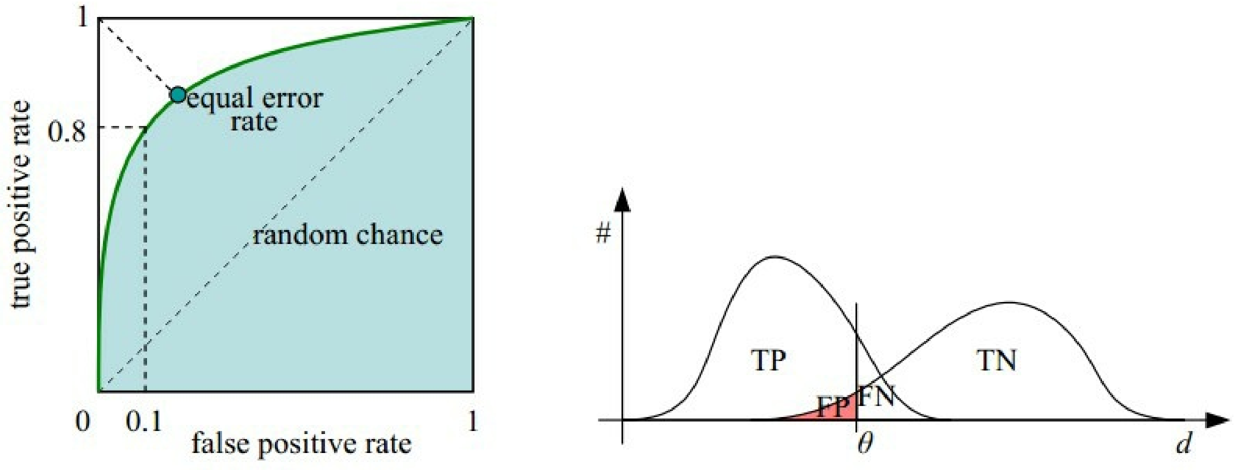


station	predict positive	predict negtive
real postive	ture positive	false negtive
real negtive	false positive	ture negtive

- ture postive : count of real positive and predict positve samples
- ture negtive : count of real negtive and predict negtive samples
- false positive : count of real negtive and predict postive samples
- false negtive : count of real positive and predict negtive samples

roc_curve:

roc_curve describe the predict accuracy,especially in imbalace situation,it will much better than Calculate accuracy rate. please see the following graph:



(a)

(b)

The x axis of roc_curve is false positve rate,the y axis of roc_curve is true postive rate.

false positve rate = false positive/ real negtive;

ture positive rate = true positive / real positive

there are four key points in roc_curve:(0,0),(1,0),(0,1),(1,1)

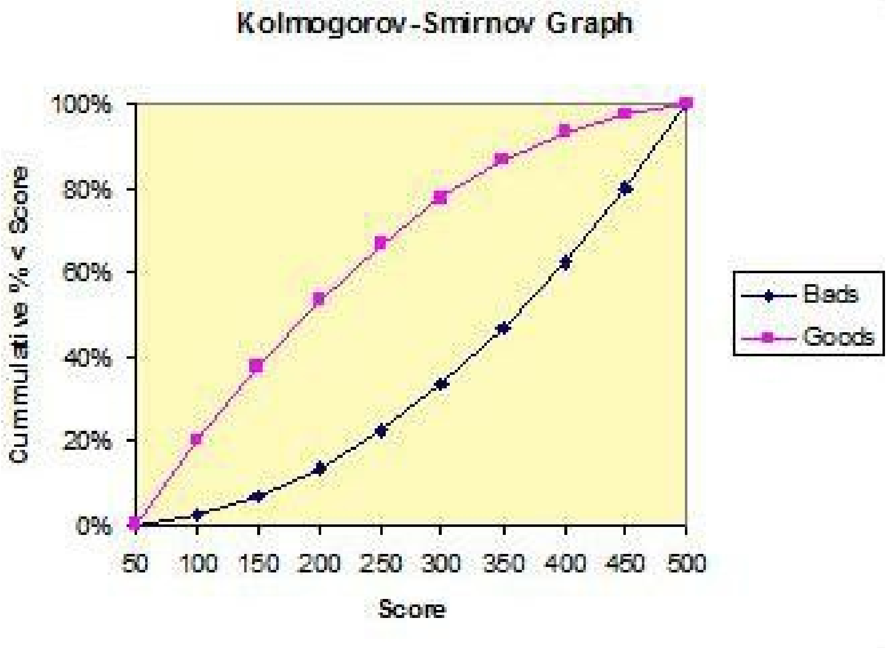
- (0,0),false positve rate = 0 and ture postive rate =0, it means all real positve sample have been wrong predicted,and no real negtive sample wrong predicted;
- (1,0),false positve rate = 1 and ture postive rate = 0, it means all real positve sample wrong predicted and all real negtive sample also been wrong predicted
- (0,1), false positve rate = 0 and ture positve rate = 1, it means all real negtive sample have been predicted right and all real positve sample also been predicted right.
- (1,1),false positive rate =1 and ture positve rate = 1, it means all sample have been predicted right.

as describe on above, near to (1,0) as soon as possible will be the best choice.

auc area:
it can be discribed as the area of roc,most of time it between 0.5 and 1.the greater the value,the best of the model.

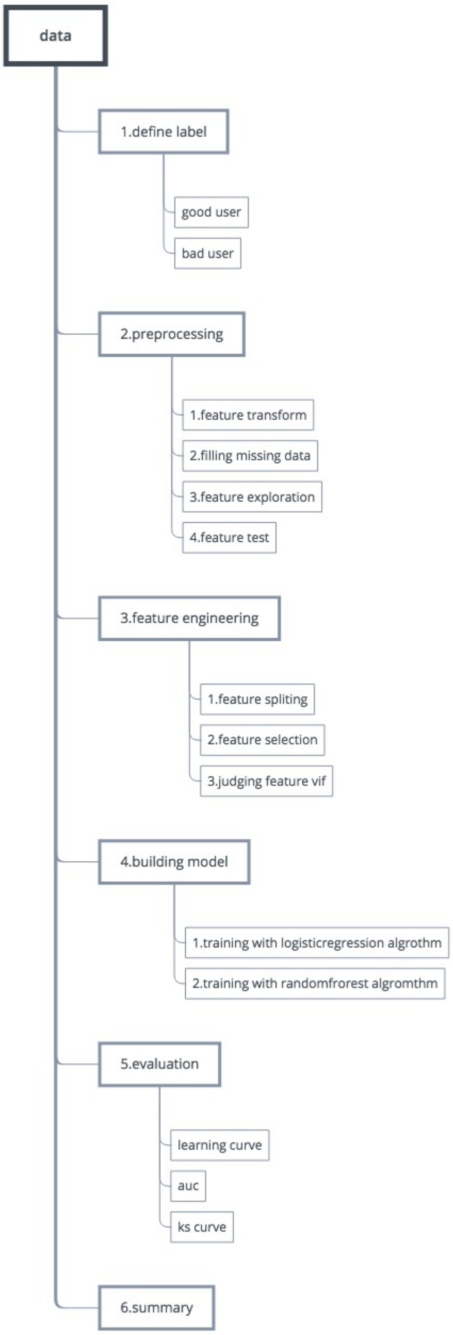
> The common threshold value of AUC is 0.7, if AUC value is greater than 0.7,it means the model has a strong degree of discrimination;if the auc value between 0.6 and 0.7, it means the model has a certain degree of discrimination; if the auc value is between 0.5 and 0.6, it means the model has a weak degree of differentiation; if auc value is Lower than 0.5, it means the model is less differentiated than random guesses.

ks_curve:
it can be discribed as the metrics of divergence,the higher ks value means a good model to saperate different class.



> transfering the model result into probability,sort the probability ascendingly,the x axis is total sample comulative probability and the y axis is good sample comulative probability and bad sample comulative probability.
> ks = max{good/total good - bad/total bad}

Design Frame


data
1.define label
good user
bad user
2.preprocessing
1.feature transform
2.filling missing data
3.feature exploration
4.feature test
3.feature engineering
1.feature spliting
2.feature selection
3.judging feature vif
4.building model
1.training with logisticregression algrothm
2.training with randomfrorest algromthm
5.evaluation
learning curve
auc
ks curve
6.summary