borrower Application Scordcard Classification Report on lendingclub data

Machine Learning Engineer Nanodegree li wang

2018.04.25

1.Definition

Project Overview

With the development of Peer-to-peer credit industry in recent years, more and more consumers choose to solve the problem of temporary funds shortage 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.

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.

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.

Metrics

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.

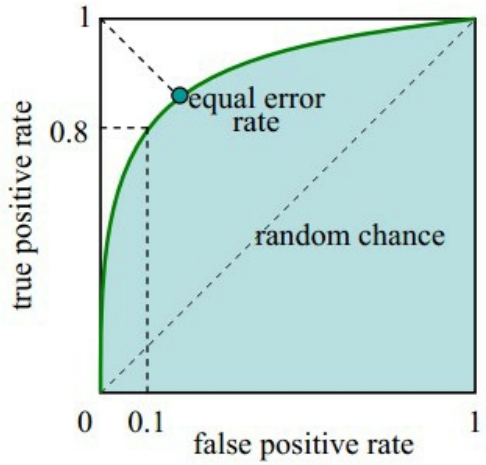
confusion_matrix:
two class as example, the row means real class of sample,the columns means predict class of sample. if predict class equal to real class, it means true or it means false.according to this thouht,confusion_matrix can shape four values,true postive ,ture negtive ,false postive,false negtive.

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

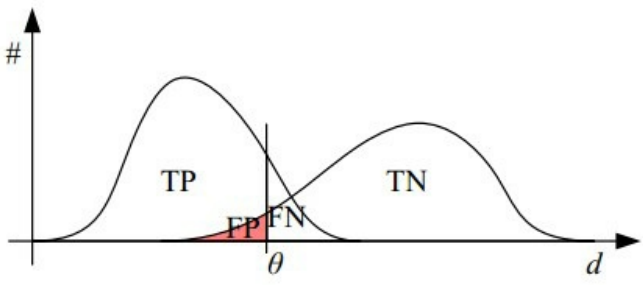
- ture postive : count of real positive and predict positve samples
- ture negtive : count of real negtive and predict negtive samples
- false positive : count of real negtive and predict postive samples
- false negtive : count of real positive and predict negtive samples

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)

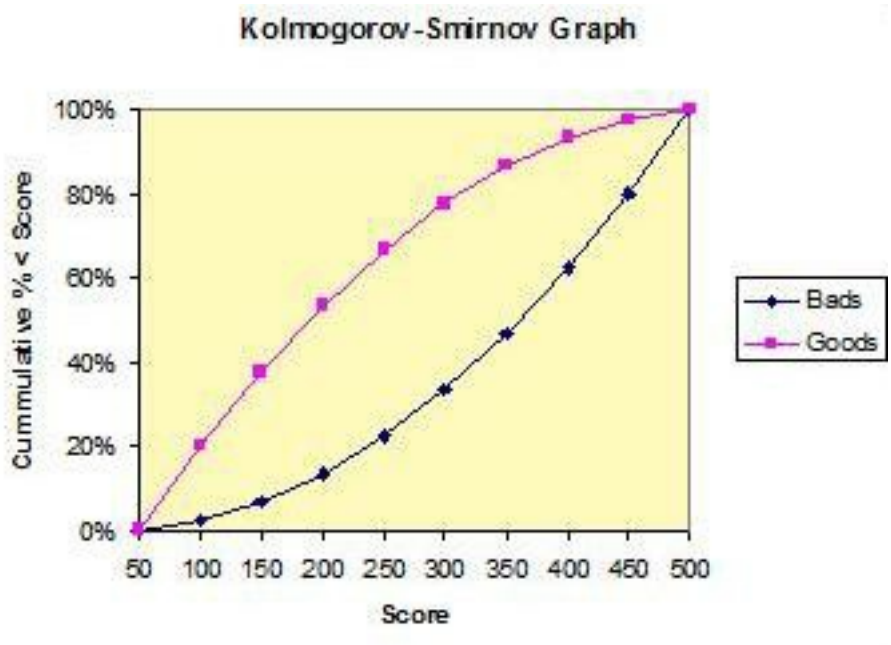
- (0,0),false positve rate = 0 and ture postive rate =0, it means all real positve sample have been wrong predicted,and no real negtive sample wrong predicted;
- (1,0),false positve rate = 1 and ture postive rate = 0, it means all real positve sample wrong predicted and all real negtive sample also been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

> if the max KS value is greater than 0.3, it means the model is better; if the max KS value is greater than 0.2, it means this model can be use but not very good;if the max KS value is less than 0.2,it means the model very poor and can not be used; If the KS value is negative, it shows that the score is inconsistent with the good and bad, and the model is wrong.

> The disadvantage of the KS index is that it can only represent the distinction of the best score and cannot measure other points.

2.Analysis

Data Exploration

This project is use the internationally renowned Peer-to-peer credit platform LendingClub official website opened in the third quarter of 2017 to analyze the borrower data, involving 420,000 applicant samples, 145 features information, including a good customer sample 350,199, and bad Customer sample 66175.It belongs to the sample distribution imbalanced data.

According to the Basel Agreement, the repayment status will stabilize after 12 months of repayment, and the loan term of the customer data used in the development model of the

project is 36 or 60, and the repayment period is much more than 12 months,so the applicant repayment status is stable.

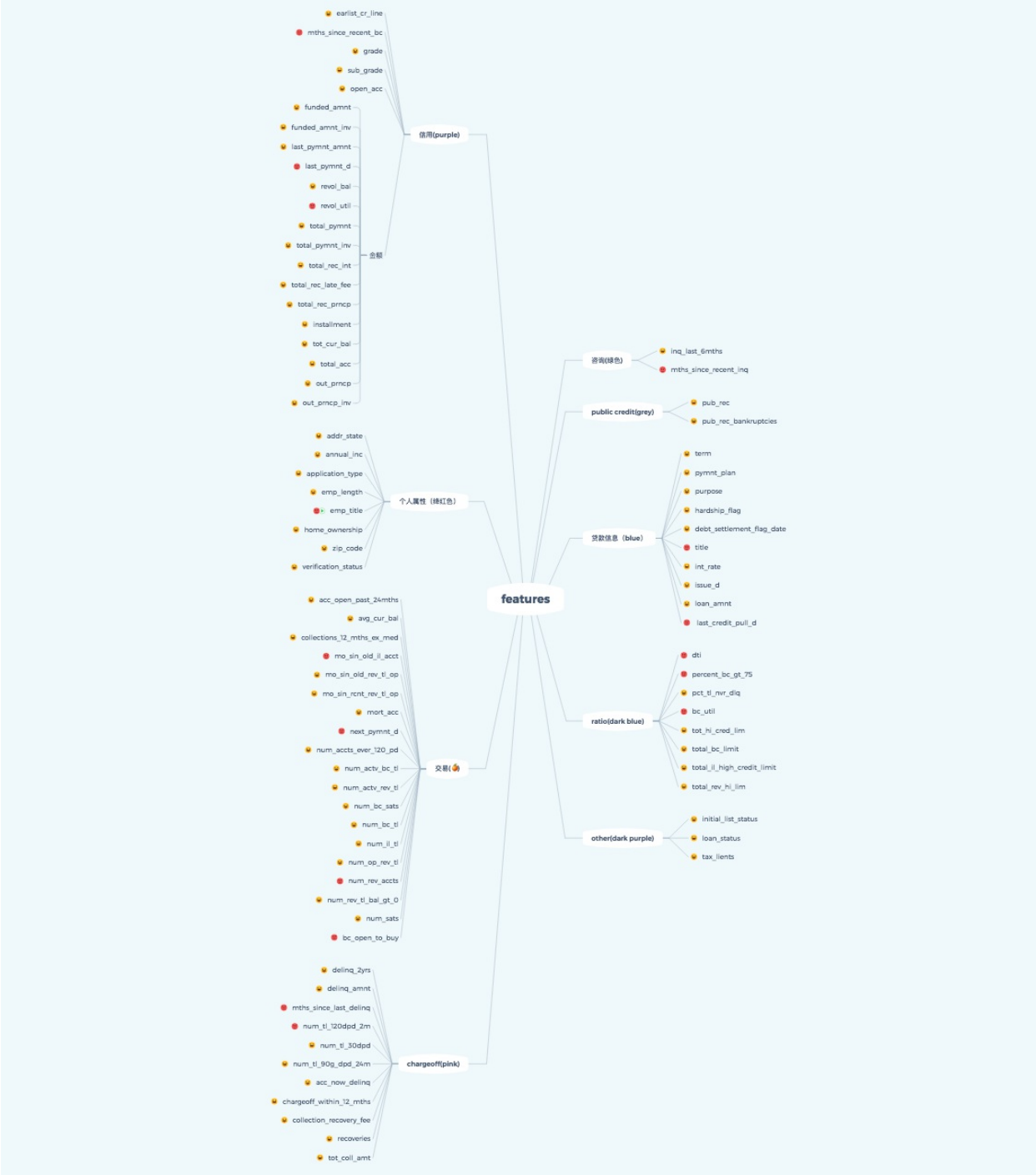
In this project i will only use data with 36 term,and i will use the apply date between Jan-2015 to Oct-2015 as my training set,the apply date is Nov-2015~Dec-2015 as my test data testing my model.

table1:data view

table2:feature description

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the pas
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations,
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, ex
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one ye
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.

The specific features as follows:



> Discrete features{'id', 'term', 'int_rate', 'grade', 'sub_grade', 'emp_title', 'emp_length', 'home_ownership', 'verification_status', 'issue_d', 'pymnt_plan', 'desc', 'purpose', 'title', 'zip_code', 'addr_state', 'earliest_cr_line', 'revol_util', 'initial_list_status', 'last_pymnt_d', 'next_pymnt_d', 'last_credit_pull_d', 'application_type', 'verification_status_joint', 'hardship_flag', 'hardship_type', 'hardship_reason', 'hardship_status', 'hardship_start_date', 'hardship_end_date', 'payment_plan_start_date', 'hardship_loan_status', 'disbursement_method', 'debt_settlement_flag', 'debt_settlement_flag_date', 'settlement_status', 'settlement_date'}

> Continuous features{'member_id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv',
> 'installment', 'annual_inc', 'url', 'dti', 'delinq_2yrs',
> 'inq_last_6mths',
> …
> 'deferral_term', 'hardship_amount', 'hardship_length', 'hardship_dpd',
> 'orig_projected_additional_accrued_interest',
> 'hardship_payoff_balance_amount', 'hardship_last_payment_amount',
> 'settlement_amount', 'settlement_percentage', 'settlement_term'}
> label：{0,1} 0:normal，1:overdue

table 1: feature example

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	hardship_last_payment_amount
0	NaN	NaN	8800.0	8800.0	8800.0	36 months	9.80%	283.13	B	B3	...	NaN
1	NaN	NaN	23100.0	23100.0	23100.0	60 months	20.50%	618.46	E	E4	...	NaN
2	NaN	NaN	24250.0	24250.0	24250.0	60 months	24.24%	701.01	F	F3	...	NaN
3	NaN	NaN	17925.0	17925.0	17925.0	60 months	17.27%	448.09	D	D3	...	NaN
4	NaN	NaN	15850.0	15850.0	15850.0	60 months	23.13%	448.01	F	F2	...	NaN

table 2: show description

	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	url	dti	delinq_2yrs	inq_last_6mths	...	hard
count	0.0	286606.000000	286606.000000	286606.000000	286606.000000	2.866060e+05	0.0	286604.000000	286606.000000	286606.000000	...	
mean	NaN	15184.672861	15184.672861	15177.383216	439.707307	7.752131e+04	NaN	19.139224	0.346448	0.566345	...	
std	NaN	8609.142780	8609.142780	8605.037156	245.918986	7.677776e+04	NaN	9.008141	0.924757	0.863738	...	
min	NaN	1000.000000	1000.000000	900.000000	30.120000	0.000000e+00	NaN	0.000000	0.000000	0.000000	...	
25%	NaN	8400.000000	8400.000000	8400.000000	261.830000	4.700000e+04	NaN	12.540000	0.000000	0.000000	...	
50%	NaN	14000.000000	14000.000000	14000.000000	382.550000	6.500000e+04	NaN	18.580000	0.000000	0.000000	...	
75%	NaN	20000.000000	20000.000000	20000.000000	577.790000	9.200000e+04	NaN	25.350000	0.000000	1.000000	...	
max	NaN	35000.000000	35000.000000	35000.000000	1445.460000	9.000000e+06	NaN	999.000000	30.000000	5.000000	...	

Data visualization

Data Explot summary

- label: from label columns 'y' distribution,i find in this project it belongs to two classes imbalanced sample classification problems
- cat_col: we can see from the following graph,most of the borroweres have house but in mortgage station and people who rent aparterment have much higher overdue rate than others;the main purpose of apply loan is debt_consolidation.
- word_cloud: tearcher is the more job for borrower filling and plenty of people didn't fill this information;'945XX'and '750XX' is the most zip_code for borrower filling and the 'TX' and 'CA' for 'addr_state'.
- youxu_col:in this part,i find most borroweres's FISO score in grade C to grade E,and most people have work at least ten years.many feature have serious outlier.

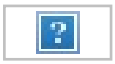
- lianxu_col: some feature have beatiful normal distribution but some not,like youxu_col,some features have serious outlier.

1.histogram of target y

It can be seen from the histogram that the number of overdue users is about 14%, it means this project belongs to imbalanced sample classification.

2.catergery features histogram

3.text features word cloud graph

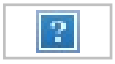


4.ordered features histogram

5.continue features histogram

6.Correlation coefficient matrix

Figure 5 shows：The darker the red is, the higher the correlation coefficient is,part of continuous features have very high correlation coefficient, this problem will be solve in the following process.



3.Algorithms and Techniques

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.

1.Logistic Regression classifier

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.

2.Random Forest Classifier

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.

==========

4.Basic Model

In this project, i plan to use RandomForestClassifier algorithm to train the top 10 important feature with default paremeter as my basic model.The result is:

> the features used are:'bc_open_to_buy_Merge', 'emp_title', 'int_rate_Merge', 'sub_grade', 'grade', 'acc_open_past_24mths_Merge', 'num_tl_op_past_12m_Merge', 'total_bc_limit', 'tot_hi_cred_lim', 'avg_cur_bal'

5.Method

Data preprocessing

The statistical analysis and visualization of the data show that the data sets are missing seriously, and some of the features are all missing, and part of the discrete feature dtypes is been marked as float type, and there is a obviously collinearity between some features.

Processing details strategy as follows:

- Remove features that are missing more than 60% and every feature should have at least two different values;
- Feature transformation: Converts character features that are supposed to be numeric types to numeric values;
- outliers treatment: There are obvious outliers, delete and adjusted the abnormal values;
- Missing processing: using proper strategy to filling the Missing data;
- select important features based on feature selection technique (e.g. IV values, L1, feature importance, etc.)

According to the above preprocessing, the algorithm will be iterated and the model evaluation results are obtained.

Implementation

1.In the preprocessing phase, delete all the columns values missing more than 60% columns, and row values or column values are all missing data.

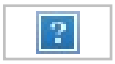
2.Delete after_loan feature:'pymnt_plan', 'collection_recovery_fee', 'recoveries', 'hardship_flag', 'out_prncp_inv', 'out_prncp';

3. delete 90% value same in one column

4.calculating IV values of missing features, remove the feature with the IV value less than 0.01, and filling missing features with IV value greater than 0.01.

5. missing filling
   Classification feature: Emp_title use the mode filling
   Continuous features :
   'dti', 'mo_sin_old_il_acct', 'mths_since_recent_inq','percent_bc_gt_75''bc_util'
   'bc_open_to_buy', 'mths_since_recent_bc'

Missing features Distribution Example:



Now,The data preprocessing has been completed, in this stage ,i have divided the variables into classification features and continuous features.

first, i have encoding the catergory features;second,i have used the ChiMerge algorithm to split the features into several box(details: if the feature only have less than five unique value or equal to five, i will not split them but if some value have only one label, i will use merging strategy merged them;if the feature have more than five unique values, i will use chiMerge algorithm to split them);

second,after finishing the chiMerge process,calculate the woe and iv value of every feature. delete the features that iv value is less than 0.01;

third,implement feature selection strategy, keep high iv value feature and if two features have strong relationship,i choose higher iv value and drop low iv value feature;

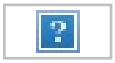
finally,use the featues selected to build LogisticRegressionclassifier and RandomForestClassifier model.

chiMerge result show：

mths_since_recent_bc_Bin	mo_sin_old_il_acct_Bin	bc_open_to_buy_Bin	emp_title_br_encoding_Bin	mths_since_recent_inq_Bin
Bin 1	Bin 1	Bin 0	Bin 0	Bin 2
Bin 1	Bin 2	Bin 1	Bin 3	Bin 0
Bin 3	Bin 2	Bin 1	Bin 0	Bin 2
Bin 2	Bin 1	Bin 2	Bin 0	Bin 2
Bin 4	Bin 1	Bin 1	Bin 2	Bin 2

been kept features correlation coefficient matrix heatmap：

model result show：



perfect

I mainly use grid research to get the best parameters for LogisticRegressionclassifier and RandomForestClassifier algorithm.

- LogisticRegression algorithm

> the default parameters model 1: is:{'C':1.0,'class_weight':None,'penalty':'l2'};the best parameters model 2:{'C': 0.01, 'class_weight': {0: 1, 1: 1}, 'penalty': 'l2'}

- RandomForestClassifier algorithm

> the default paramters model 3: {'class_weight':None,'criterion':'gini','max_depth':None,'nn_estimators':10};the best parameters model 4:{'class_weight': {0: 1, 1: 1}, 'criterion': 'gini', 'max_depth': 6, 'n_estimators': 20}

Justification

Based on the above evaluation, the LogisticRegression algorithm have much better result. i will choose model 2 as the final result of the model, because the paramters of model 2 have been search,so may it much stable than others.

Result Visualization

featues coefficient

Reflection

This project is a supervised binary classification problem. The dataset is downloaded from lendingclub official website. There are 145 total features and 220 throusands clients. It is an imbalanced dataset. I applied Logistic Regression and Random Forest Classifier to the dataset. After tuning parameters, I found that Logistic Regression performed best. The AUC is 0.90. After that, I applied tuned Logisitc Regression model to dataset of different testing sizes. AUC scores were relatively the same.

Improvement

For further improvements, i will try to use neural Network technique to apply to the dataset. because neural Network is now very popular technique and have achieved some impressive success.

References

- https://zhuanlan.zhihu.com/p/21550547
- http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://blog.csdn.net/kevin7658/article/details/50780391
- http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html#sklearn.model_selection.GridSearchCV
- http://www.360doc.com/content/16/0726/23/20558639_578629644.shtml
- https://www.lendingclub.com/info/download-data.action
- 信用评分模型技术与应用
- 统计学习方法
- 机器学习与R语言实战
- feature engineering for machine learning