Mobile User Credit Prediction Based on LightGBM

Qiangqiang Guo, Zhenfang Zhu*, Hongli Pei, Fuyong Xu, Qiang Lu, Dianyuan Zhang and Wenqing Wu
Shandong Jiao tong University, China
*Corresponding author

***Abstract*—LightGBM algorithm is used to build an effective credit score prediction model for mobile users and improve the prediction system of personal credit score. Firstly, linear correlation is analyzed to build feature set, then k-means algorithm is used to analyze feature set clustering, and finally, credit scoring model is built by LightGBM. Experiments on real data provided by the digital China innovation competition show that this method has higher accuracy than GBDT, XGBoost and other algorithms. By clustering the data feature set based on linear correlation analysis and applying it to LightGBM credit scoring model, mobile users' credit scores can be predicted more accurately.**

Keywords—score prediction; LightGBM algorithm; K-means; feature data

I. INTRODUCTION

With the deepening of the construction of social credit system, the social credit standard construction develops rapidly, relevant standards have been published one after another. But a multi-level standard system which is including credit service standards, credit data acquisition and service standards, credit repair standard, city credit standard, and industry credit standards, urgently needs to be promulgated, and social credit standard system is expected to advance rapidly. The construction of social credit system is a systematic project, and improving the credit scoring system will help promote the upgrading of the credit system of the whole society. The constitution of personal credit evaluation is the basis of social credit evaluation system, Building a scientific personal credit evaluation system is the basis of building a scientific social credit evaluation system, while the mobile user credit evaluation is one of the most important components of personal credit evaluation. With the progress of science and technology and the development of society, personal credit score is becoming more and more important to individuals, while traditional credit scores are mainly measured by a few dimensions such as personal spending power, which is difficult to reflect personal credit comprehensively, objectively and timely. Nowadays, with the vigorous development of e-commerce and Internet finance, personal credit evaluation needs to meet the requirements of the times and change to the direction of big data under the background of big data.

This algorithm aims to solve the problem of credit score prediction in the environment of large sample and high-dimensional data. We propose a mobile user credit score model based on LightGBM algorithm: K-LGB model, to achieve mobile user credit score. The algorithm can effectively improve the accuracy of credit score prediction, and at the same time improve the algorithm execution efficiency.

II. RELATED RESEARCH

A. Selecting a Template

Score prediction problem[1] belongs to a branch of recommendation system, the performance of recommendation system is largely affected by the accuracy of scoring prediction. With the in-depth study of scholars at home and abroad, two kinds of statistical and non-statistical methods have been developed in credit evaluation[2]. Non-statistical methods include neural networks, genetic algorithms, expert systems, etc. Statistical methods include logistic regression, linear regression, non-linear regression, nearest neighbor estimation, etc. Many scholars have used user history scoring behavior and item attributes to model early[3] to solve the scoring prediction problem, In previous studies, Maher Alarajden et al.[4] combines neural networks, support vector machines, random forests, decision trees, logistic regression and naive Bayes with LR, and achieved good results. So far, the credit evaluation system proposed by Maher Alaraj is still regarded as the industry standard model of credit scoring model. Maysam F. Abbod et al.[5] proposed an algorithm that integrates Gabriel near-field graph editing and multivariate adaptive regression spline method in data preprocessing to achieve predictive credit score. In addition, a new classifier combination rule based on consensus method of different classification algorithms in the stage of set modeling was proposed. Cuicui Luo et al.[6] comparing deep learning algorithms such as belief network and restricted Boltzmann machine with popular machine learning algorithms such as logistic regression, support vector machine and multi-layer perceptron, we find that DBN performs best in area evaluation performance using classification accuracy and receiver performance curve. CK Leong et al.[7] proposed a Bayesian network model to solve the problems of truncated samples, sample imbalance and real-time implementation in credit risk scoring. Compared with the competition model (logical regression and neural network), it performs better in accuracy, sensitivity and other dimensions.

With the rapid development of machine learning technology, domestic scholars pay more attention to the combination and application of these models. Comprehensive application of multiple machine learning methods for credit scoring is gradually becoming the main means to solve the problem of insufficient accuracy of single algorithm results and obtain better prediction results. For example, Jiang Minghui[8], Wang Lei and others[9] have achieved good results by improving Logistic model and establishing credit scoring model.

In recent years, with the deepening of credit evaluation research, non-statistical methods such as artificial intelligence have been introduced. The focus of scholars' research has shifted to integrated learning algorithm, neural network (NNs), support vector machine (VSM) and other algorithms. The existing research results show that a group of individual learners are constructed according to training data, and a learning method of integrating multiple learners is adopted with some strategy. Compared with single classifier such as logical regression, decision tree and evaluation model of neural network[11] and fuzzy analysis and evaluation model, the integrated learning method has higher accuracy and better robustness[10]. In these ensemble learning methods, LightGBM has the advantages of faster training speed, lower memory consumption, better model accuracy, support for parallel learning, and fast processing of massive data. In view of this, this paper builds a credit scoring model based on LightGBM algorithm to predict the credit score of China Mobile users.

III. Research on Mobile User Credit Score Based on Light GBM Algorithm

The existing credit scoring models often only use the Bagging method (such as RF algorithm) or Boosting method (such as LigthGBM) in integrated learning, which has great limitations in multi-dimensional feature extraction and linear relationship mining. In view of this, in the face of large sample and multi-dimensional data environment, in order to solve the model over-fitting problem, construct effective feature information, and improve the accuracy of the credit score of this model, this paper proposes a K-LGB model to obtain credit scores of mobile users. We first construct a feature set by analyzing linear correlations. Then the K-means algorithm is used to cluster the feature sets, and the feature set clustering analysis results are added to the data set as effective feature information. Finally, the data set with valid feature information is added as input to the LightGBM model, and the credit score is generated by the LightGBM model.

A. Linear Correlation Analysis

It is found that the analysis of linear correlation can not only solve the problem of model over-fitting, but also solve the problem of multi-dimensional feature extraction and linear relationship mining. There, this paper uses the Pearson correlation coefficient for linear correlation analysis. Pearson correlation coefficient, also known as Pearson product moment correlation coefficient, is commonly used in statistics to measure the correlation between two sets of data. The value of the Pearson correlation coefficient is between -1 and 1. The larger the absolute value, the stronger the linear correlation; the closer the absolute value is to 0, the weaker the linear correlation. Suppose that given the data set $X=\{x_1,x_2,...,x_i\}$ and $Y=\{y_1,y_2,...,y_i\}$ containing *i* items, the Pearson correlation coefficient formula is as shown in 1:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum X^2{}_i - (\sum X_i)^2} - \sqrt{n\sum y^2{}_i - (\sum y_i)^2}} \tag{1}$$

Where n is the number of values of the variable, r_{xy} is the Pearson correlation coefficient value of the data set $X \quad Y$.

We calculate the Pearson correlation coefficient between features, features and credit scores, determine linear correlation, and select feature sets with strong linear correlation with credit scores as linear correlation analysis results. The linear correlation between some data features and credit scores is shown in TABLE I.

TABLE I. LINEAR CORRELATION BETWEEN PARTIAL DATA FEATURES AND CREDIT SCORES

Feature data	Linear correlation value
User network age (month)	0.55
Average consumption value in recent June (¥)	0.49
Number of people in the telephone circle during the month	0.48
…	...
Is it the tourist attraction of the month	0.27
Whether 4G unhealthy users	-0.15
Sensitivity of monthly telephone charges	-0.24

After linear correlation analysis, we found 7 characteristics and credits such as “user network age (month)”, “user average consumption value in recent June (¥)”, “number of people in the current month”, and “what is the attraction in the month”. The points have a strong linear correlation. Therefore, we choose this part of the feature set for further analysis.

At present, the credit evaluation presents a large sample and high dimensional characteristics. The irrelevant and redundant variables will adversely affect the accuracy of the model prediction. Selecting valid feature information directly determines the accuracy of the credit evaluation model. In view of this, we first fully exploit the effective feature information in the data through linear correlation analysis method, then cluster the linear correlation analysis results to construct effective feature information, and finally the constructed effective feature information is manually added to the data set as a feature column as an input to the LightGBM scoring prediction model. Although the effective feature information is added to the data dimension, the experiment proves that the method improves the accuracy of the score prediction model.

B. K-means Clustering Based on Feature Set

1) K-means Cluster Analysis

The core idea of K-means algorithm is to divide n data into k clusters, so that the similarity is lower in different clusters, the similarity in clusters is higher, and the degree of similarity is measured by the distance between objects. It is the European distance. Assuming a given data set, $X=\{x_m \mid m=1,2,...,h,h \in R\}$ Y the sample has n attributes (dimensions) $A_{1,} A_2 ... A_n$, then the Euclidean distance formula is as follows:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \qquad (2)$$

The smaller distance of $d(x_i, x_j)$, the higher the sample x_i and x_j and similarity, and the smaller the difference; the sample and similarity are low, and the difference is large. The K-means clustering algorithm generally uses the sum of squared errors as a standard measure function, which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2 \qquad (3)$$

P is a point representing the space of the object, m_i is the mean of the clusters C_i (m_i and C_i both are multidimensional). Where E is the squared error sum of all objects in the dataset, and the size of different clusters E will be different. Therefore, the algorithm needs to adjust E to the minimum to make the cluster optimal.

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

K-means is a clustering algorithm belonging to the partitioning method and is a classical clustering algorithm. Due to the simplicity and speed of the algorithm process, it is widely used in industry. The main advantages are as follows:

- The algorithm tries to minimize the squared sum error of the determined K divisions.
- When the clustering data is dense (convex), and the data between the cluster and the cluster is large, the K-Means algorithm has a better clustering effect.
- When dealing with large data sets, the algorithm is efficient and relatively scalable.

2) Cluster analysis based on linear correlation analysis results

As mentioned earlier, the method flow for constructing effective feature information is as follows:

- Selection of clustering algorithm: Different clustering algorithms have different advantages and disadvantages. We use the attributes of data (whether the algorithm is independent of data input order; data dimension) and algorithm processing ability (algorithm complexity) as the basis for clustering algorithm selection. We compare the Hierarchical methods, the partition-based methods (K-means), the support vector machine (SVM) and other algorithms in the clustering algorithm, and finally choose the partition-based method (K-means) as the model of this paper. Clustering Algorithm.
- Input of K-means clustering algorithm: linear correlation analysis result (N-dimensional feature set with strong linear correlation with credit score), number of cluster clusters k (K value is 4, K value is determined as this article) Section 4. 3. 1)

K-means clustering algorithm output: effective feature information (1D). K-means clustering algorithm results are shown in TABLE II.

TABLE II. K-MEANS CLUSTER ANALYSIS RESULTS

User network age (month)	**Average consumption value in recent June (yuan)**	**Number of people in the telephone circle during the month**	**…**	**K-means Cluster analysis result**
186	163. 86	83	…	3
145	109. 64	70	…	3
62	162. 98	77	…	2
78	98. 33	45	…	1
10	78. 86	29	…	0

The LightGBM algorithm belongs to the Boosting method in the integrated learning method. The method combines the weak learning algorithm into the LightGBM of the strong learning algorithm. It is a lightweight GB framework that supports high-efficiency parallel training, which improves the training speed of the model and reduces the memory. It has the advantage of improving model training speed, reducing memory consumption and improving accuracy. LightGBM provides users with a variety of parameter settings, and adjusts the different parameters to get the optimal model.

We assume that the training set sample is $T=\{(x_1,y_1),(x_2,y_2)...(x_m,y_m)\}$. Where $x_i \in T$ is the i-th sample in the training set, $\hat{y}_i$ is the predicted value, y_i is the true value， l is the loss function. In the t-step model, the x_i prediction is as shown in equation (4), Its objective function is shown in equation (5):

$$\hat{y}^{t}{}_{i} = \hat{y}^{t-1}{}_{i} + f_t(x_i) \qquad (4)$$

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_{i,} \hat{y}_i^{t}) + \sum_{i=1}^{t} \Omega(f_i) \qquad (5)$$

$\Omega(f_i)$ is a regular term, f_i is a decision tree. Set the loss function to square loss. Then the objective function is:

$$O\mathrm{bj}^{(t)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{t-1} + f_t(x_i)))^2 + \Omega(f_t) + cons\tan t \qquad (6)$$

As mentioned earlier, the irrelevant and redundant variables can adversely affect the accuracy of model predictions. Selecting valid feature information, directly determines the accuracy of the credit scoring model. Therefore, we manually input the K-means clustering algorithm (the constructed effective feature information) into the data set as a new feature column. Integrate the dataset of the new feature column as input to our LightGBM model. The specific LightGBM credit scoring model training process is as follows:

Input: The output of the K-means clustering algorithm is used as the effective feature information, and is manually added to the data set in the form of a new feature column as the LightGBM model input.

Output: Mobile users predict credit scores.

Algorithm steps:

- The algorithm determines the objective function, Set the loss function to square loss, and generate each node of the decision tree through a greedy strategy, find the best tree structure.
- The algorithm calculates the first derivative and the second derivative of the sample point of the loss function before each iteration, generates a new decision tree and calculates the predicted value for each node.
- Iteratively generated N decision tree iterations are added to the model. Initialize N decision trees, and distribute the weight of the training sample evenly.
- Train the weak classifier, update the weight to get the final classifier, and output the mobile users' predicted credit score.

IV. Experiment and Evaluation

A. Experimental Details

This experiment uses the data set of the 2019 Digital China Innovation Competition titled "Consumer Crowd Portrait - Credit Intelligence Score". The data set is the sample data provided by China Mobile Fujian Corporation for the 2018 x month (desensitization). It includes a variety of multi-dimensional data such as customer's various communication expenses, arrears, travel, consumption places, social, personal interests and so on. There are 50,000 training sets and 50,000 test sets. The data mainly includes information on user identity, consumption ability, personal relationship, location trajectory, application behavior preference, etc. It contains 30 fields of user code, credit score and other features. A sample data set is shown in TABLE III.

TABLE III. SAMPLE DATA SETS

User age	Average consumption value in recent June (yuan)	User Fee Sensitivity	……	Credit score
44	163. 86	3	……	664
23	22. 81	2	……	547
……	……	……	……	……
73	40	4	……	576
108	70	3	……	601

B. Data Analysis Preprocessing

In the data set, features of different dimensions have different metrics, but the eigenvalues should be correct and effective. Through statistical analysis of the data set, it is found that there are data missing and outliers in the dataset. which leads to the loss of validity and correctness of the feature values. Therefore, it is necessary to perform missing data restoration and outlier processing on the data set.

C. Evaluation Index

There are many indicators for evaluating the performance of the user credit scoring model. such as Accuracy, Recall, F score, MAE, ROC and Precision. In order to verify the performance of the model, we choose MAE and ROC and AUC(Area Under Curve)as the evaluation indicators for this model. We converted MAE into a Score indicator. The specific formula is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| pred_i - y_i | \tag{7}$$

$$Score = \frac{1}{1+MAE} \tag{8}$$

Where $pred_i$ is the forecast sample, y_i is a real sample. The smaller the value of MAE, the closer the forecast data is to the real data. The higher the value of all Scores, the better the evaluation.

D. Test Results and Analysis

1) Selection of k value

The clustering result depends on the setting of the initial value. However, the selection of values often takes many experiments to find the optimal number of clusters. This experiment uses different K values for comparison of evaluation results. Through experimental results, when the K value is 4, the evaluation results of this model are optimal.

2) LightGBM parameter adjustment

Although the LightGBM model parameters contain many types of parameters, the structure is relatively simple. The parameter settings are directly proportional to the model effect. The better the parameter adjustment, the better the effect. The LightGBM model provides users with multiple types of parameters. And provides convenient CV functions for users to adjust. In the process of adjusting the model parameters, this paper splits the training set by 80% as a new training set. and the remaining 20% of the data is used as a new test set. By fine-tuning the parameters based on the predicted results of the new test set and the actual results, and using the CV function to. Get the optimal parameters of the LightGBM model, such as learning_rate is 0.01, Objective is regression_l1, metric is mae, feature_fraction is 0.6, bagging_fraction is 0.8, bagging_freq is 2, num_leaves is 31, verbose is -1, max_depth 5, lambda_l2 is 5, and lambda_l1 is 0.

3) Comparative analysis of model effects

In order to verify the superiority of the method in this paper, we used the evaluation index Score. Use LightGBM, XGBoost[12], K-LGB, K-XGB four models, and use the evaluation index (Score, effectiveness, Accuracy) to compare experimental results. The evaluation index Score results are shown in TABLE IV

TABLE IV. MODEL EVALUATION SCORE RESULTS AND EFFICIENCY

Model	**Score**	**Running time /minute**	**Average accuracy**
K-LGB	6. 412	8	69.58%
XGBoost	6. 340	15	67.78%
LightGBM	6. 276	6	66.97%
K-XGB	6. 391	20	68.61%

The experimental results shown in TABLE IV show, the score of the algorithm in this paper is 6. 412, and the model runs for 8 minutes. It is 5. 412 percentage points higher than the Score of the LightGBM model.

V. CONCLUSIONS

This paper onducts cluster nalysis based on the results of linear correlation analysis. and fully tap into data characteristics, The LightGBM algorithm is used as a typical big data technology to predict the credit scores of China Mobile users. In terms of data preprocessing, the data loss problem uses the NaN method, and the method of setting the upper and lower limits is used for the problem of the first and last outliers of the data. In a large sample, high-dimensional environment of data sets, compared with algorithms such as GBDT and XGBoost, experimental results show that the proposed algorithm has better prediction accuracy and computational efficiency, suitable for handling large-scale data. At the same time, this article also has shortcomings, the feature sets with weak linear correlation with credit scores are not used in model training. Therefore, the next step will be to use a feature set that has a weaker linear correlation with the credit score in the model, to more fully mining data features, At the same time, improve the prediction accuracy of the deep learning model, which allows the model to achieve a more accurate credit score in a large sample, high dimensional data environment.

ACKNOWLEDGMENT

This work was supported National Social Science Foundation (19BYY076); Science Foundation of Ministry of Education of China (14YJC860042), Shandong Provincial Social Science Planning Project (19BJCJ51, 18CXWJ01, 18BJYJ04, 17CHLJ07C).

REFERENCES

[1] Yang,G ; Xu,X ; Zhao,F. Predicting User Ratings with XGBoost Algorithm. Data Analysis and Knowledge Discovery, 2019, vol.3(01), pp.118-126.

[2] Jin ,X. Research on the Index and Model of Personal Credit Scoring. Master's degree, Zhejiang University of Finance & Economics, Hangzhou, China ,2018

[3] Deng,X; Jin C, Han Q ,at. al. Improved collaborative filtering model based on context clustering and user ranking. Systems Engineering —Theory & Practice, 2013, vol.33(11) , pp.2945-2953.

[4] Ala"Raj M ,;Abbod M F . A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Systems with Applications, 2016, pp. 64:36-55.

[5] Ala'Raj M; Abbod M. Classifiers consensus system approach for credit scoring . Knowledge-Based Systems, 2016, vol.104 ,pp.89-105.

[6] Luo C; Wu D;Wu D . A deep learning approach for credit scoring using credit default swaps. Engineering Applications of Artificial Intelligence, 2016:S0952197616302299.

[7] Leong C K. Credit Risk Scoring with Bayesian Network Models . Computational Economics, 2016, vol.47(3) , pp.423-446.

[8] Jiang M; Xu P; Han Y. Optimized CBR for Personal Credit Scoring. China Soft Science,2014, vol.12, pp.148-156.

[9] Wang L; et al. Application of Data Mining Models in Credit Scoring for Small Business Owners. Statistical Research ,2014, vol.31(10), pp.89-98.)

[10] Brown I; Mues C . An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 2012, vol.39(3), pp.3446-3453.

[11] Feng J. Research on personal credit risk assessment model based on BP neural network . Master's degree,Dalian University of Technology, Dalian, China,2017.

[12] Chen T; Guestrin C. XGBoost :A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA; August 13 - 17, 2016;pp:785-794.