China Family Loss Prediction Research Based on Machine Learning

Bin Qin[1], Zhili Wang[2*], Yufan Wu[3] and Haocheng Zheng[4]

[1]Shenzhen University Information Center, Shenzhen University, Shenzhen, Guangdong, 518060, China

[2]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[3]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[4]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, 518060, China

[*]Corresponding author's e-mail: 912427166@qq.com

Abstract. Social change has a direct influence on the family loss. To this end, we propose a novel approach to model the probability of family loss by learning the large-scale real-world survey data from China Family Panel Studies (CFPS). Especially, combining individual economic, health, identity, family profile and other information, we develop the scalable tree boosting model named Family Loss Prediction Model (FLPM) to model the family loss. Extensive experiments on real-world survey data validate the effectiveness of FLPM for measuring the probability of family loss in China. Furthermore, compared with the XGBoost and GBDT models, the experimental results show that the FLPM model is better and the precision is 98.48%, which provides an important basis for this study.

1. Introduction

China is experiencing a huge social change. Its wide scale, rapid speed, and great influence are unprecedented in human history, which performs particularly obvious in the following three aspects: economic growth, the spread of education and the demographic transition. On the economic front, China's GDP and per capita GDP has increased significantly since the start of China's economic reforms in 1978[1]: per capita GDP increased at an annual rate of 6.7% from 1978 to 2008, at the meantime, GDP increase from 0.37 trillion yuan in 1978 to 90.03 trillion yuan in 2018, with an increase of more than 240 times[2]. In terms of education, the number of Chinese university student has changed from about 1 million in 1978 to more than 20 million since the 21st century. In the aspect of the population, China has completed the demographic transition from high fertility rate and high death rate to low. China's total fertility rate (TFR) has fallen sharply since the end of the 1970s, which fell from 6 to 2[3], right at the replacement level. Additionally, Life expectancy has steadily increased since the 1950s, reaching 70 years old by the beginning of the 21st century on a level with developed countries around 1970 and far exceeds that of less developed countries.

Family change is affected by social changes, such as increasing social inequality, rising divorce rates, increased cohabitation before marriage, large-scale labor mobility. The starting point of the CFPS project is to allow scholars to have high quality, comprehensive and long term follow up data to study these social phenomena and change deeply. The factors above directly affect the disintegration and reconstruction of family. Specifically, the contribution of this paper can be summarized as follows:

- We explore to measure the loss of family with the hybrid model which can help to evaluate the its probability.
- We evaluate the proposed model with real-world data and extensive experiments. The experimental results clearly validate the effectiveness of FLPM.

2. Data Description

In this paper, we attempt to develop a data-driven approach for a comprehensive assessment of family loss in China. The data set used in the experiments is the survey data of CFPS project from 2010 to 2016, which contains a total of 178015 population records. Among them, 132695 records are selected for training and verification from 2010 to 2014, and 45320 data records for testing from 2016.

2.1. Feature selection

To be specific, the original data set consists of four years, including 2010,2012 and 2014, which 4 or 5 tables of data per year. We extracted the 31 potentially important features[2] from historical data. Among them, there are 19 nominal types and 12 numerical types, forming four new tables according to the year. Partial features description is summarized in Table 1.

Table 1. Partial feature description

Feature	*Meaning*
Pid	Personal ID
Fid	Family sample ID
Urban	Urban area
P_wage	Personal
Total_asset	Net family asset
Fproperty	Property income
fml2014num	Family size in 2014

2.2. Data processing

Generally, we need to construct a label to determine the probability of a person losing a family. To be specific, we use the *fid* intersection of adjacent years to mark the *pid*. If the *pid* has not been lost, it is marked as a positive class, otherwise, it is marked as a negative class. In order to reduce the complexity of the model, we first perform dimensionality reduction on the data. After converting these features into vectors. We use the PCA[2] algorithm for dimensionality reduction.

Define sample set as D, define samples as x_m, so

$$D = \{x_1, x_2, \dots x_m\} \tag{1}$$

Sample centralization for all samples

$$x_i = \leftarrow x_i - \frac{1}{m}\sum_{i=1}^{m} x_i \tag{2}$$

Let W represents projection matrix, x_i represents new sample after dimensionality reduction

$$z_i = W^T x_i \tag{3}$$

Define sample set as D' after dimensionality reduction, $z_m$is a new sample set of x_m mapping, so

$$D' = (z_1, z_2, \dots, z_m) \tag{4}$$

To ensure the quality of our analysis, we further extract features by using the *chi-squared goodness* of fit test. Finally, we use 45320 positive and 45320 negative data as test set; 132695 positive and 132695 negative data as the training set.

3. Probability Model of Family loss

After the experimental data is ready, we turn to the details of our novel FLPM approach for effectively measuring the likelihood of family loss. FLPM is the hybrid model based on XGBoost and GBDT. To the end, the models are integrated with weights of 0.85 and 0.15, achieving the optimal. XGBoost, an efficient system implementation of Gradient Boosting, it is an optimized and improved version of GBDT. XGBoost can parallel computing, effective processing of sparse data and optimize the use of CPU and memory [6]. For GBDT[5] [7], the model could be given by

$$F_m(x) = \sum_{m=1}^{M} T(x; \varphi_m) \tag{5}$$

Where $T(x; \varphi_m)$ is a decision tree, φ_m is the parameters of the tree, and M is the number of decision trees. Then, in order to train the model, we should adjust the parameters according to the objective function, which could be defined by the following formula

$$\xi(\varphi) = \sum_{i=1}^{h} l(\hat{y}_i, y_i) + \sum_{k=1}^{g} \Omega(f_k) \tag{6}$$

Where $\hat{y}_i$ is predicted value, y_i is a real value, l is loss function. Here $\Omega(f_k)$ stands for regularization term and f_k represents a decision tree. The algorithm of the model uses a forward step-by-step algorithm, let L be the loss function and solve the loss function with a square error. The loss function of the model can be expressed as follows

$$L[y, F_{m-1}(x) + T(x; \varphi_m)] = [y - F_{m-1}(x) - T(x; \varphi_m)]^2 \tag{7}$$

we could then estimate the family loss rate by

$$P\_fid\_loss_i = \sum_{j=1}^{n} \frac{P\_pid\_loss_j}{n} \tag{8}$$

where subscript i represents *i-th* family, the subscript j in $P\_pid\_loss_j$ represents *j-th* person in family i, is the number of the family members.

4. Experiments

In this section, we will evaluate the effectiveness of our purposed model. Especially, our FLPM will be validated on real-world CFPS (China Family Panel Survey) data sets compared with two typical baselines.

4.1. Experimental Setup

Firstly, we summarize the details of the experimental setup in this subsection, including data pre-processing, experimental results and model evaluation.

4.2. Data Pre-processing

As introduced in Section 2, we conducted our experiments on real data from the 2010 to 2016 CFPS survey project. In addition to dimension reduction, we also process missing values[8]. To ensure the effectiveness and avoid missing values, we process the missing values in the data, the numeric features are filled with the median, and the nominal features are binned and one-hot encoded.

4.3. Experimental Results

To validate the performance of FLPM for predicting family loss, we use the 2016 data as the verification set, and the comparison between predicted result and the actual situation is shown in Figure 1. As shown in Figure 1, the predicted result is in line compliance with the actual situation. The predicted probability distribution of the number of Pid is shown in Figure 2.

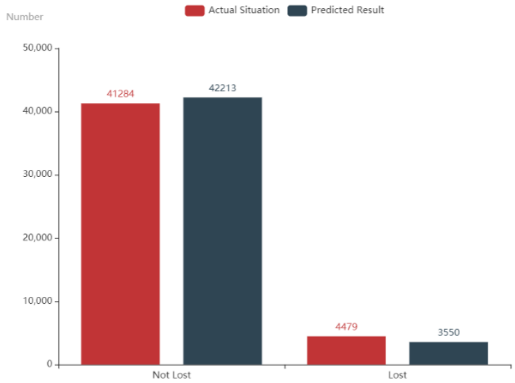

Figure. 1. Comparison result

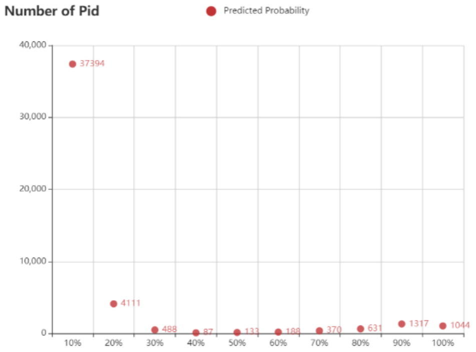

Figure. 2. Predicted probability distribution

The model fusion and corresponding evaluation parameters are shown in Table 2.

Table 2. Model Fusion and comparison of relevant evaluation parameters

XGBOOST weight	*GBDT weight*	*Precision*	*Accuracy*	*Recall rate*	F_1 *Score*	*AUC*
0.1	0.9	0.91838	0.91662	0.54169	0.68145	0.91662
0.25	0.75	0.92401	0.91983	0.54397	0.68480	0.91983
0.45	0.55	0.93900	0.92307	0.54852	0.69251	0.92307
0.65	0.35	0.95247	0.92547	0.55458	0.70100	0.92547
0.85	0.15	0.98482	0.97747	0.7814	0.87140	0.97949
0.9	0.1	0.95542	0.92765	0.56065	0.70664	0.92765

4.4. Model evaluation

For the comprehensive verification of FLPM, we use three methods to evaluate metrics including ***accuracy***, ***precision*** and ***recall rate*** [2], $\boldsymbol{F_1}$***Score*** and ***ROC*** curve. Since the CFPS project in this paper attempts to find the 1000 most likely loss families, which focuses on the model's precision. Among them, the precision of the FLPM model has reached 0.956. The results are shown in Table 3, which indicates that our FLPM approach performs better with the ability to predict family loss compared with baselines.

Table 3. Model performance index comparison

Model	*Precision*	*Accuracy*	*Recall rate*	F_1 *Score*	*AUC*
XGBoost	0.9548	0.9537	0.5610	0.8373	0.9282
GBDT	0.9088	0.9489	0.5402	0.7998	0.9136
FLPM	0.9848	0.9774	0.7814	0.8714	0.9795

4.4.1. ROC curve, Confusion matrix

To more comprehensively evaluate the model, we introduce the *ROC*[2] *curves* and Confusion matrix, which are to evaluate the model’s performance. Especially, the ROC curve is used to evaluate the generalization performance of the model. The larger the area under the line (***AUC***) of the ***ROC*** curve, the better the ubiquitous performance of the model.

According to the confusion matrix, the experiment has achieved the expected result for family loss prediction. In 24993 positive samples, 1660 samples classified to negative class by mistake, while in 1546 negative samples, only 68 samples classified to positive class by mistake.

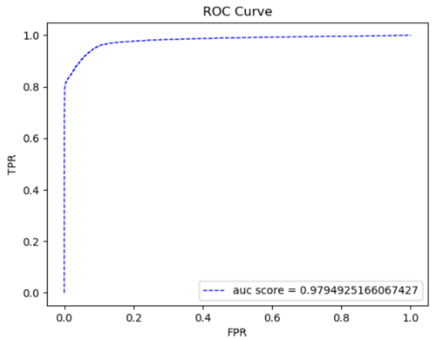

Figure. 2. ROC Curve of FLPM model

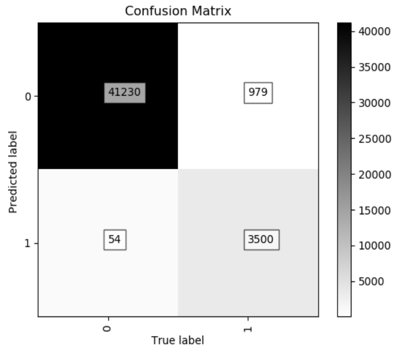

Figure. 3. Confusion matrix of FLPM model

5. Conclusion

In this paper, we proposed a data-driven approach to predict the probability of family loss based on the analysis of data provided by CFPS. Specifically, we first extract 31 important features from the data and clean it up. Then, we developed the GBDT, XGBoost and hybrid model to calculate the probability of family loss. And we found that the hybrid model works best. Therefore, with the help of the hybrid model, we can effectively rank the family loss probability based on their information.

In the future, we will consider more "personalized" factors for this research, i.e., to reveal the phenomenon about increasing social inequality, rising divorce rates, increased cohabitation before marriage, as well as large-scale labor mobility. Besides, more comprehensive studies will be conducted.

Statement: This project was funded by the Ministry of Education and the "Shenzhen University-Sangfor Cloud Computing Innovation Teaching Laboratory" (No. 201801277010).

[1] Xie Y. (2016) Why China needs empirical research. In: Empirical Research of Social Sciences. Shanghai. pp. 18-24.

[2] China Family Panel Studies. (2018) CFPS User's Manual (3rd Edition)(CHN). http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/45LCSO.

[3] Pan Y. (2017) Future trends and common challenges of globalization. Tsinghua Financial Review, 9:36-40.

[4] M. Dash, H. Liu. (1997) Feature selection for classification. Intelligent Data Analysis, Volume 1, Issues 1–4, 1997, Pages 131-156.

[5] Pingan Sun，Beizhan Wang. (2017) Comparative analysis of dimensionality reduction algorithms, case study: PCA. International Conference on Intelligent Systems, 10.1109/ISCO.2017.7855992

[6] Chen T, Guestrin C. (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining. New York. pp. 785-794.

[7] Friedman J. (2001) Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29:1189-1232.

[8] Hongxiao Hu, Jia Xie, Bing Han. (2007) Comparative study on missing value processing methods. Market Modernization, 10.3969/j.issn.1006-3102.2007.15.235.

[9] Goutte C., Gaussier E. (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In: Losada D.E., Fernández-Luna J.M. (eds) Advances in Information Retrieval. Berlin. pp. 345-359.

[10] Rao G. (2018) What is an ROC curve? The Journal of family practice, Volumes 52, Issues 9, Pages 695.