Scientific Research Publishing

Applied Mathematics, 2020, 11, 8-22
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Identification Model for Needy Undergraduates Based on FFM

Luwen Hu[1], Xiaoyong Zhao[1], Shihao Fan[2], Yufeng Gui[1*]

[1]College of Science, Wuhan University of Technology, Wuhan, China
[2]Shanghai Xuhui High School, Shanghai, China
Email: *guiyufeng@whut.edu.cn

How to cite this paper: Hu, L.W., Zhao, X.Y., Fan, S.H. and Gui, Y.F. (2020) Identification Model for Needy Undergraduates Based on FFM. *Applied Mathematics*, **11**, 8-22.
https://doi.org/10.4236/am.2020.111002

Received: December 6, 2019
Accepted: December 28, 2019
Published: December 31 2019

Abstract

In recent years, as the enrollment rate of Chinese colleges has increased year by year, the identification of needy undergraduates has become increasingly important. However, the traditional way to identify college students with financial difficulties mainly relies on manual review and collective voting, which easily causes subjectivity and randomness. To alleviate the problem above, this paper establishes an automatic identification model for needy undergraduates based on the 1842 questionnaires collected from undergraduates in WHUT. Firstly, this paper filters the questionnaire preliminary using the local outlier factor algorithm. Secondly, this paper combines mutual information, Spearman rank correlation coefficient and distance correlation coefficient by rank-sum ratio to select features for eliminating noise from irrelevant features. Thirdly, this paper trains filed-aware factor machine model and compares it with other models, such as Logistic Regression, SVM, etc. Eventually, this paper finds that filed-aware factor machine performers much better than other models in the identification of needy undergraduates, and prominent features affecting the identification of needy undergraduates are the year of the family income, cost of living provided parents, etc.

Keywords

Local Outlier Factor, Rank-Sum Ratio, Field-Aware Factor Machine, Identification Model for Needy Undergraduates

1. Introduction

The number of undergraduates in higher education institutions in China has been on the rise since 2000, when the country began to expand enrollment in higher education. At the same time, the tuition fees of various universities and colleges are also rising, which has caused certain economic pressure for many students

admitted to universities from rural areas. Therefore, it is essential to identify and fund needy undergraduates, which is not only to train country's qualified talents, but also to increase the vitality to the world of knowledge. However, due to the large number of undergraduates, the identification process of needy undergraduates is not stable, so an accurate and effective way is needed to provide relevant support for the financial aid of needy undergraduates.

At present, all countries in the world take the family economic survey as the main means to identify needy undergraduates, and the standards of each country are also based on the restrictions on the supporting objects. Among them, the United States takes family income as the only criterion for identifying needy undergraduates, because the perfect income verification and tax collection system in the United States can effectively report and supervise residents' non-earned income. In Japan, household income and assets indicators are combined with various classification indicators to determine undergraduates' family economic status. Uganda in Africa relies on proxy variables, such as the class of father's job and vehicle, to measure family income. The Nigerian Student Loan Board uses a four-factor property test that measures a family's financial status by its parents' occupation, income, household size and the number of children in education. In the small African country of Malawi, the family which wants to receive a student loan must meet one of the following conditions: the parents or guardians are unable to provide financial assistance to him, and the parents or guardians do not have a clear and fixed source of income and other economic reasons approved by the loan committee. In some Latin American countries, the household economic survey is quite rigorous and detailed. For example, in Peru, parents of undergraduates, applying for student loans, are even interviewed on property, such as houses, cars, land, parents' jobs, employers and wage earners [1].

Pathman, DE (Pathman, DE), Konrad, TR (Konrad, TR) *et al.*, exploited the data of 723 undergraduates from 69 states collected by statistical methods to analyze the impact of receiving state-funded scholarships and repaying loans on needy undergraduates in 2004 [2]. Jiyun Kim, Stephen L. DesJardin, Brian P. McCall used a random utility model to explore the effects of student expectations about financial aid on postsecondary choice focusing on income and racial/ethnic differences in 2009 [3]. Luo Suo and Jian Gong, in 2015, used BP neural network to create a nonlinear mapping between the economic conditions of college students and the needy undergraduates identifying [4]. Aifeng Li, Zhineng Xiao, Biyun Liang, in 2017, collected 36,546 data concerning dining consumption of students in three months, used Datist, a big data analysis software to build a model, acquired concerning dining habits, consuming behaviors, situations in school and consumption indicators of the students, and then selected needy undergraduates [5]. Tao, BR (Tao, Bairui), Liu, KD (Liu, Kaida) *et al.*, based on GA-SVM [6], established a targeted poverty reduction model for the needy undergraduates in 2018 based on the information of freshmen's admission and undergraduates' daily life consumption. Yao Bei, in 2019, extracted five categories of feature clusters

through data cleaning, based on the campus big data platform and data mining technology before using logistic regression, random forest and other algorithms for data mining and analysis, and established a model on the identification of needy undergraduates eventually [7].

The policy, however, still needs to be improved in terms of simplicity, effectiveness and accuracy, on grounds of that the current efficient subsidy for needy undergraduates in China has been widely used in the past decade. For example, the school's identification of needy undergraduates mainly focuses on manual audit and class voting, which easily causes subjectivity and randomness.

Field-aware Factorization Machines was first proposed in 2016, which has not been used to identify needy undergraduates and it is suitable for sparse data. Therefore, this paper adopts the three-classification Field-aware Factorization Machines method to establish an identification model for needy undergraduates, so as to solve the problem of the lack of uniformity in the identification standards and the subjectivity of the identification process and obtains a better performance than the predecessors. All the work above is supported by Wuhan University of Technology.

2. Data Collection and Processing

2.1. Data Collection

The subject of the questionnaire survey is the undergraduate of Wuhan University of Technology. The design of the questionnaire is mainly based on the questionnaire of Chinese college students' family economic situation and the intermediate process of the identification of needy undergraduates. There are totally 19 questions in the questionnaire. Some questionnaires are issued through the Internet, while others are issued in paper form. Finally, the questionnaires were screened by two staff members and 1842 questionnaires were left.

2.2. Data Encoding

In order to train the model, we transformed all discrete features, such as "Nation", "Locality", "City", etc., into one hot encoding, while coding of other continuous variables remains the same. Final outcomes of the data encoding are listed in **Table 1**.

2.3. Anomaly Detection

On grounds of invalid questionnaires affecting the accuracy of the model, it is necessary to clean these data. And this paper chooses the LOF (Local Outlier Factor) algorithm to solve it.

The definitions of some terms and symbols in LOF are as follows [8]:

1) *K*-distance

$$k\text{-}dist(p) = d(p,o) \tag{1}$$

where o is the kth point closest to point p (does not include p).

Table 1. Data encoding.

Field	Feature	Assignment
Nation	Han	Han = 1, others = 0
	Ethnic	Minority = 1, others = 0
Locality	Northeast	Northeast = 1, others = 0
	Northern Coast	Northern Coast = 1, others = 0
	Eastern Coast	Eastern Coast = 1, others = 0
	Southern Coast	Southern Coast = 1, others = 0
	Middle reaches of Yellow River	Middle reaches of Yellow River = 1, others = 0
	Middle reaches of Yangtze River	Middle reaches of Yangtze River = 1, others = 0
	Southwest	Southwest = 1, others = 0
	Northwest	Northwest = 1, others = 0
City	Metropolitan	Metropolitan = 1, others = 0
	Large city	Large city = 1, others = 0
	Medium-sized city	Medium-sized city = 1, others = 0
	Small city	Small city = 1, others = 0
	Countryside	Countryside = 1, others = 0
	Poverty-stricken countryside	Poverty-stricken countryside = 1, others = 0
Family	Insured residents on record	Insured residents on record = 1, others = 0
	Insured residents	Insured residents, others = 0
	Divorced family	Divorced family = 1, others = 0
Laborers	The number of laborers	Range over positive integer
Elderly people	The number of the elderly people	Range over positive integer
Family background	The number of kids	Range over positive integer
	The number of preschoolers	
	The number of kids in middle school	
	The number of members in High school	
	The number of members in University or college	
	The number of the salary family member	
	Others	
Health	Personal condition in health	Fitness = 1, minor ailment = 2, critical illness = 3
Health of parents	Healthy	Healthy = 1, others = 0
	Either seriously ill	Either seriously ill = 1, others = 0
	Both seriously ill	Both seriously ill = 1, others = 0
	Single family	Single family = 1, others = 0

Continued

Disability	The number of the disable family member	Range over positive integer
Finance	Subsidized	Subsidized = 1, others = 0
	Farmland	Farmland, others = 0
	Civil servants	Civil servants, others = 0
	Migrant workers	Migrant workers, others = 0
	Individual business	Individual business, others = 0
	Cooperation	Cooperation, others = 0
	Financially supported by relatives	Financially supported by relatives, others = 0
Financial support	Financial Support from family	Range over positive integer
Annual income	Annual income of family	Range over positive integer
Department or house	House in countryside	House in countryside = 1, others = 0
	Small house in countryside	Small house in countryside = 1, others = 0
	Department in city: 1	Department in city: 1 = 1, others = 0
	Department in city with 1 as well as house in countryside with 1	Department in city with 1 as well as house in countryside with 1 = 1, others = 0
	Department in city: 2 or more	Department in city: 2 or more = 1, others = 0
	House in countryside: 2 or more	House in countryside: 2 or more = 1, others = 0
	Zero	Zero = 1, others = 0
House's size	The size of the house or department	Range over positive integer
Accidents	Not any	Not any = 1, others = 0
	Infertility caused by drought or flood	Infertility caused by drought or flood = 1, others = 0
	Property Damage	Property Damage = 1, others = 0
	Earthquake or fire	Earthquake or fire = 1, others = 0
	Incurable disease	Incurable disease = 1, others = 0
Debt	Debt	Range over positive integer
Academic performance	Failed before	Failed before = 1, others = 0
	Always passed	Always passed, others = 0

2) Reachable distance

The kth reachable distance from point o to point p is defined as:

$$reach\text{-}dist_k(o,p) = \max\left\{k\text{-}dist(p), d(o,p)\right\} \tag{2}$$

that is, if point o is within the k nearest neighbors of point p, the kth reachable distance will be the k-distance of p. Otherwise, it will be the real distance between o and p.

3) *K*-domain

K-domain of point *p*, marked as $N_k(p)$, is a collection containing all points within the *k*-distance radius of point *p*, including the points on the circle. Number of elements in the $N_k(p)$ is marked as $\|N_k(p)\|$.

4) Local reachability density

The local reachable density of point *p* is expressed as:

$$lrd_k(p) = 1 \Big/ \left(\frac{\sum_{o \in N_k(p)} reach\text{-}dist_k(o,p)}{\|N_k(p)\|} \right) \tag{3}$$

represents the reciprocal of the average reachable distance of points in the *k*-domain of point *p*. The higher the density value is, the more likely it is to belong to the same cluster. The lower the density, the more likely it is to be an outlier.

4) Local outlier factor

The local outlier factor of point *p* is expressed as:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} = \frac{\sum_{o \in N_k(p)} lrd_k(o)}{|N_k(p)|} \Big/ lrd_k(p) \tag{4}$$

represents the average ratio of the local reachable density of points in the *k*-domain of point *p* to the local reachable density of point *p*. If the ratio is close to 1, it means that the density of *p* is similar to the density of its domain point, and *p* may belong to the same cluster as the domain point. If this ratio is less than 1, it means that the density of *p* is higher than the density of its domain point, and *p* is the dense point. If this ratio is greater than 1, it means that the density of *p* is less than the density of its domain point, and *p* is more likely to be an outlier [9].

The main flow of the LOF is as follows:

1) For each data object *p* in the overall data set, find its k-domain and calculate their reachable distance;

2) Calculate local reachability density of data object *p*;

3) Calculate local outlier factor of data object *p*;

4) Repeat the above steps, calculate local outlier factor for all data object. Sort them and select outliers based on the preset threshold.

The algorithm cleans 1842 questionnaires collected from the survey of needy undergraduates and selects 1756 valid questionnaires finally.

2.4. Feature Selection

As some of the features in this questionnaire are excessive, feature selection is carried out in order to make better use of prior knowledge and avoid or alleviate the problem of overfitting [10]. The following will use mutual information, Spearman rank correlation coefficient and distance correlation coefficient to carry out feature selection.

2.4.1. Mutual Information

Mutual information measures the amounts of information that one random variable contains about another, and the reduction in the uncertainty of one random variable due to the knowledge of the other variable. By the consideration of two random variables, feature s and true label t, with a joint probability mass function $P\left(s=s_i,t=t_j\right)$ and marginal probability mass function $P\left(s=s_i\right)$ and $P\left(t=t_j\right)$. Mutual information $R1$ is the relative entropy between the joint distribution and the product distribution $P\left(s=s_i\right)P\left(t=t_j\right)$ [11]. The specific calculation formula is as follows:

$$R1=\sum_{i=1}^{N}\sum_{j=1}^{N}P\left(s=s_i,t=t_j\right)\log_2\frac{P\left(s=s_i,t=t_j\right)}{P\left(s=s_i\right)P\left(t=t_j\right)} \tag{5}$$

When $R1$ is 0, there is no correlation between the two; When $R1$ is positive, it means that the probability of both occurrences is relatively high; when $R1$ is negative, it means that the two are negatively correlated, that is, they are mutually exclusive.

The result is shown in Figure 1. It can be seen that the two most representative features are monthly average living expenses and annual family income, followed by the number of elderly support; the less significant features were number of workers, failing status, ethnicity, family type, etc.

2.4.2. Spearman Rank Correlation Coefficient

Spearman rank correlation coefficient $R2$ is a nonparametric measure of statistical dependence and assesses the monotonic relation between two variables. For Spearman rank correlation coefficient, the variable's rank is used instead of the value itself, which is the average of their positions in the ascending order of the values. A perfect monotone function occurs a value of or 1 for Spearman coefficient, and 0 occurred to no correlation [12].

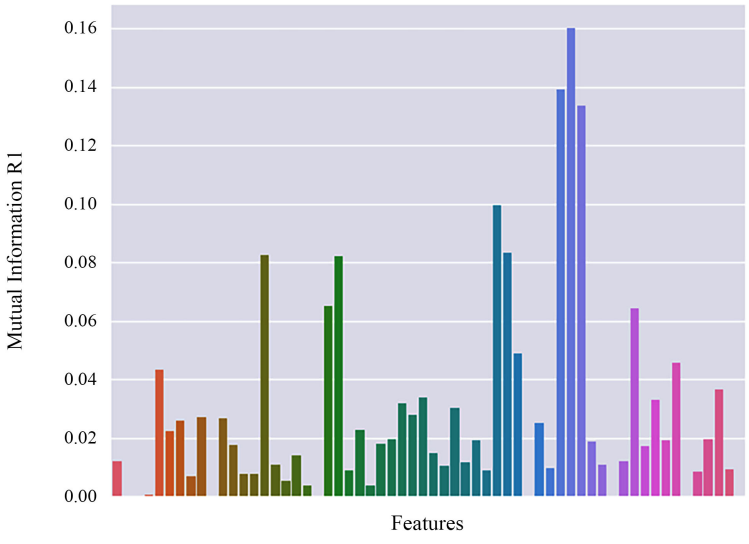

Figure 1. The final scores of the features using mutual information. Refer to Table 1 for the features in Figure 1.

The specific calculation formula is as follows:

$$R2 = \frac{\sum_{i=1}^{N}\left(R_{s_i}^{*} - \bar{R}_{s_i}^{*}\right)\left(R_{t_i}^{*} - \bar{R}_{t_i}^{*}\right)}{\sqrt{\sum_{i=1}^{N}\left(R_{s_i}^{*} - \bar{R}_{s_i}^{*}\right)^2 \sum_{i=1}^{N}\left(R_{t_i}^{*} - \bar{R}_{t_i}^{*}\right)^2}} = 1 - \frac{6\sum_{i=1}^{N}\left(R_{s_i}^{*} - R_{t_i}^{*}\right)^2}{N\left(N^2 - 1\right)} \tag{6}$$

where $R_{s_i}^{*}$ is the rank of s_i, and $R_{t_i}^{*}$ is the rank of t_i, $\bar{R}_{s_i}^{*} = \sum_{i=1}^{N} R_{s_i}^{*}$, $\bar{R}_{t_i}^{*} = \sum_{i=1}^{N} R_{t_i}^{*}$, and N corresponds to the number of samples, $\left(s_i, t_i\right)$ is the observed value of sample points.

The value of $R2$ is between −1 and 1. When the value is 1, it means that the two random variables s and t are positively correlated. When the value is −1, it means that there is a completely negative correlation between s and t. When the value is 0, it means that s and t are linearly independent [13].

The result is shown in **Figure 2**. It can be seen that the two most representative features are monthly average living expenses and annual family income, followed by the number of old people to support, housing area, household debt, household work and city size. The less significant features are failing status, ethnicity, etc.

2.4.3. Distance Correlation Coefficient

The distance correlation coefficient $R3$ is used for the independence of the two variables s and t. When $R3 = 0$, it means that s and t are independent of each other. And the larger the R3, the greater the correlation between s and t.

The correlation coefficients of s and t are expressed as follows [14]:

$$R3 = r_{st} = \frac{\sum_{i=1}^{N}\left(s_i - \bar{s}\right)\left(t_i - \bar{t}\right)}{\sqrt{\sum_{i=1}^{N}\left(s_i - \bar{t}\right)^2}\sqrt{\sum_{i=1}^{N}\left(s_i - \bar{t}\right)^2}} \tag{7}$$

The results are shown in **Figure 3**. It can be seen that the two most representative features are monthly average living expenses and annual family income. And the contribution of the other variables is relatively insignificant.

2.4.4. Rank-Sum Ratio

In order to integrate the above three methods for feature selection, we applied rank-sum ratio comprehensive evaluation method to mutual information $R1$, Spearman rank correlation coefficient $R2$ and distance correlation coefficient $R3$ to obtain the contribution ranking of 60 features. Equation (8) is used to calculate the rank sum ratio, where n is the number of indices.

$$\mathrm{RSR} = \sum_{i=1}^{n} \frac{Ri}{3n} \tag{8}$$

The ranking results are shown in **Figure 4**. Remove features ranked 50th to 60th, including Northeast, Divorced Family, Insured Residents on Record, Ethnic, Small City, Property Damage, Earthquake or Fire, Always Passed, the Number of Preschoolers, Northern Coast. Finally, 50 features are left.

3. Factorization Machine

3.1. Principles of Factorization Machine

Factorization Machine (FM) was first proposed by Rendle in 2010 [15]. The factorization method is essentially applied to solve the problem of feature combination under sparse data. It is a general model that can be applied to all real data sets.

For a given vector $X = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$, The expression of the second-order FM model is as follows:

$$\hat{y}_{FM}(x) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \langle V_i, V_j \rangle x_i x_j \tag{9}$$

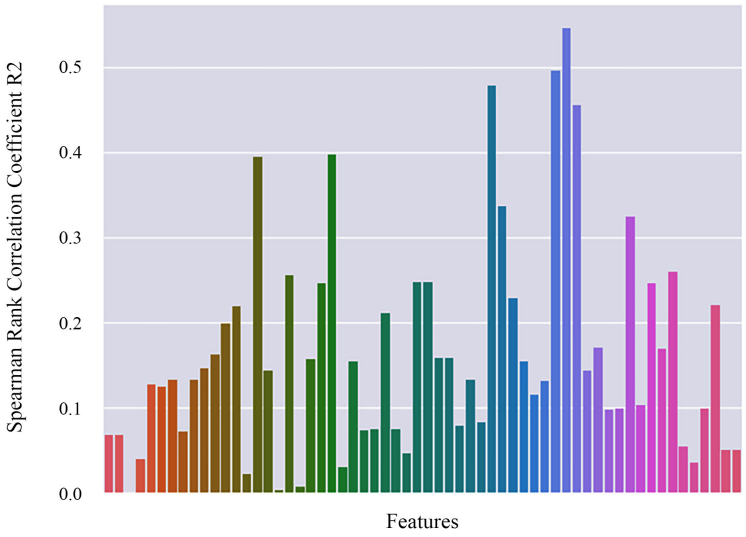

Figure 2. The final scores of the features using the Spearman rank correlation coefficient. Refer to **Table 1** for the features in **Figure 2**.

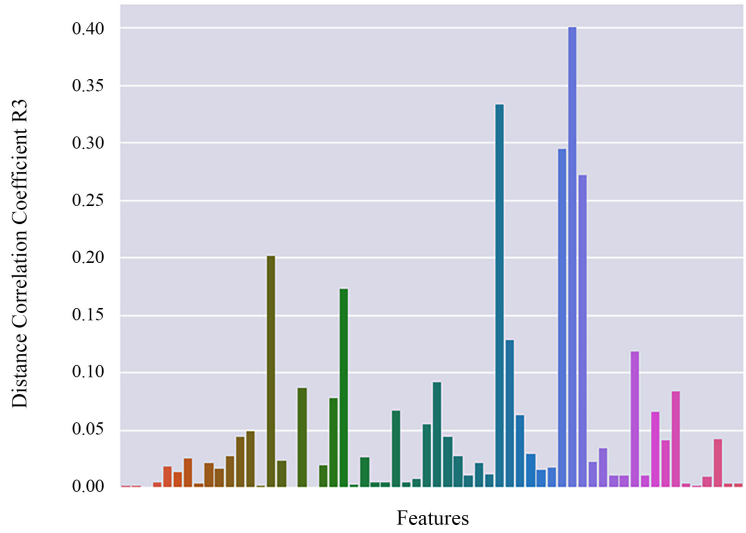

Figure 3. The final scores of the features using distance correlation coefficient. Refer to **Table 1** for the features in **Figure 3**.

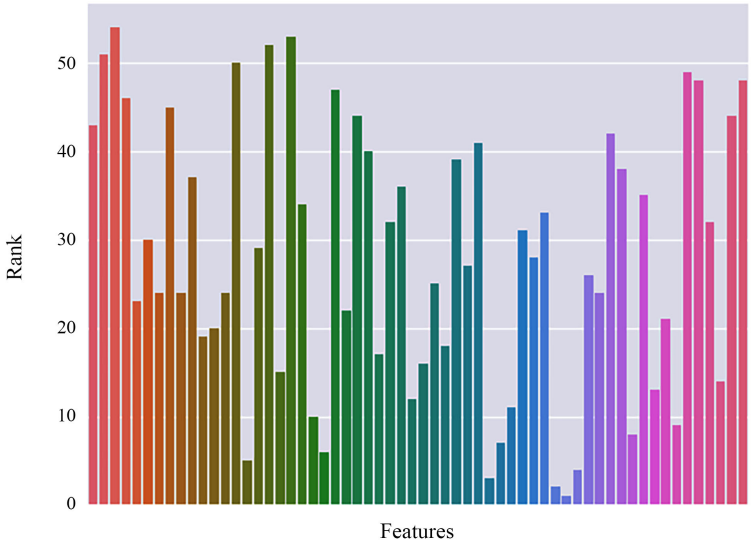

Figure 4. The outcomes of the rank-sum ratio of the 60 features. Refer to **Table 1** for the features in **Figure 4**.

where, n is the feature dimension, w_0 is the global offset, w_i is the strength of the ith variable $\left(w_0, w_1, \cdots, w_n \in \mathbb{R}\right)$, and $V_i = \left(v_{i1}, v_{i2}, \cdots, v_{ik}\right)^{\mathrm{T}} \in \mathbb{R}^k, i = 1, 2, \cdots, n$ is a latent vector introduced for the feature x_i with the hyperparameter $k\left(k \in \mathbb{N}^+\right)$ [16].

Let's say $\left\langle V_i, V_j \right\rangle = \sum_{l=1}^{k} v_{il} v_{jl}$ is the interaction between the ith and jth variables [17]. Simple parameters w_{ij} for each interaction are not adopted, because the parameters for each cross terms of parameters are no longer independent of each other. For example, the coefficient of $x_h x_i$ and $x_i x_j$ ($\left\langle V_h, V_i \right\rangle$ and $\left\langle V_i, V_j \right\rangle$ respectively) both have a common item V_i. That is, all samples containing non-zero combination of x_i can be used to learn the latent vector V_i. Even under the condition of sparse data, FM is able to learn the parameters of cross terms well.

3.2. Principles of FFM

Field-aware Factorization Machine (FFM) was first proposed by Yuchin Juan in 2016 [18]. On the basis of FM, FFM groups features of the same properties into the same field. To take the above classification of "Health of parents" as an example, "Health of parents = Healthy", "Health of parents = Either seriously ill", "Health of parents = Both seriously ill" and "Health of parents = Single family" all represent parents' conditions in health and can be put into the same field. The same categorical feature generated by one hot encoding can all be placed in the same field.

In FFM, a latent vector V_{i,f_j} is learned for each feature x_i and for each field f_j. Therefore, the latent vector is not only related to the feature, but also to the field. In other words, different latent vectors are used when the feature "Health of parents = Healthy" is associated with the feature "annual household income"

and "household housing situation", which is consistent with the intrinsic difference between feature "annual household income" and "household housing situation".

If there are n features of the sample belonging to f fields, the quadratic term of FFM has nf latent vectors. In the FM model, there is only one latent vector for each feature. Actually, FM can be regarded as the special case of FFM, which is the FFM model when all features are attributed to a field.

According to the field sensitivity of FFM, Equation (10) can be obtained below.

$$\begin{aligned}\hat{y}_{FFM}(x) &= w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \left\langle V_{i,f_j}, V_{j,f_i} \right\rangle x_i x_j \\ &= w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{w}_{ij} x_i x_j\end{aligned} \tag{10}$$

where, f_j is the field to which the jth feature belongs.

If the dimension of the latent vector is k, then the number of quadratic term of FFM is nfk, far greater than FM model's [19] [20] [21].

4. Establishment of Identification Model for Needy Undergraduates

4.1. Application of FFM to the Identification of Needy Undergraduates

In order to fit our data set, we changed traditional FFM model to three-class classification FFM. The final model is described as follows:

$$\hat{y}(x) = w_0' + \sum_{i=1}^{n} w_i' x_i + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \hat{w}_{ij}' x_i x_j \tag{11}$$

where $w_0' = (w_{01}, w_{02}, w_{03})^{\mathrm{T}}$, $w_i' = (w_{i1}, w_{i2}, w_{i3})^{\mathrm{T}}, i = 1, 2, \cdots, n$, $\hat{w}_{ij}' = (\hat{w}_{ij}, \hat{w}_{ij}, \hat{w}_{ij})^{\mathrm{T}}$, $\hat{y}(x) = (\hat{y}_1(x), \hat{y}_2(x), \hat{y}_3(x))^{\mathrm{T}}$.

The input of the model is the data set consisting of the remaining 50 features after feature selection.

The output of the model is the probability of non-needy undergraduates, needy undergraduates and extremely needy undergraduates. According to the principle of maximum probability, we choose the best classification. Use softmax function as activation function:

$$f_i(x) = \frac{\mathrm{e}^{\hat{y}_i(x)}}{\mathrm{e}^{\hat{y}_1(x)} + \mathrm{e}^{\hat{y}_2(x)} + \mathrm{e}^{\hat{y}_3(x)}}, i = 1, 2, 3 \tag{12}$$

Use cross entropy loss as loss function:

$$H_Y(\hat{Y}) = -\sum_{p} Y_p \log(\hat{Y}_p) = -\sum_{p=1}^{N} Y_p \log(\hat{Y}_p) \tag{13}$$

where Y_p is the true label of sample p expressed in the form of one hot encoding. Specifically, if sample p is a non-needy undergraduates, then $Y_p = (1, 0, 0)$, $\hat{Y}_p$ is the output vector obtained after the sample p is fed to the model, namely $\hat{Y}_p = \hat{Y}(x^{(p)}) = (\hat{f}_1(x^{(p)}), \hat{f}_2(x^{(p)}), \hat{f}_3(x^{(p)}))^{\mathrm{T}}$ [22].

4.2. Experiments

4.2.1. Coefficient Setting

According to the passage above, an identification model for needy undergraduates based on three-class classification Field-aware Factorization Machines was established.

In the experiment, the feature dimension(n) is 50, the latent vector parameter(k) is 30 and the number of fields is 18. Adagrad algorithm is used for gradient updates, with a learning rate of 0.01; the maximum number of iterations is 30.

4.2.2. Bootstrap Method

In order to minimize the randomness in the experiment, we repeat the experiment 10 times using bootstrap. The following is the specific process:

1) Randomly select 100 sets of data as the test set;

2) In the remaining data set, randomly select 1000 sets of data as the training set;

3) Train FFM model using the training set;

4) Calculate the accuracy $acc_i\left(i=1,2,\cdots,10\right)$ of the FFM model in the test set;

5) Return to step 2, repeat the experiment for ten times;

6) Calculate the average accuracy of *acc* as the final performance of FFM model.

4.3. Result

In order to show the excellent performance of our model, we took many models, such as Logistic Regression, SVM, Bayesian Network, Decision Tree and FM for comparative experiments. The specific experiment process is consistent with the passage above. Final results are shown in Table 2 below.

It can be seen from Table 2 that the FFM model has the best performance.

5. Conclusions

From the results of feature selection, it is found that prominent features affecting the identification of needy undergraduates are the year of the family income, cost of living provided parents, family farming, family houses, and total number of children. So when it comes to making artificial evaluation of needy undergraduates, more considerations should be taken into the factors above.

Table 2. The outcomes of the six different methods.

Model	Accuracy
Logisitc Regression	82.8%
SVM	87.4%
Bayesian Network	87.7%
Decision Tree	86.5%
FM	89.5%
FFM	91.2%

In addition, through comparative experiments, it can be found that FM and FFM have excellent performance in solving the classification problem under sparse data, because they can effectively learn the combination of features. Compared with FM, the concept of field is introduced in FFM, the features of the same property are attributed to the same field, different latent vectors are learned for different fields, so as to further improve the precision of the model.

6. Future Work

In the future, we will continue our work in two directions. One is the data, the other is the model.

When collecting data, we only collected data of needy students from Wuhan University of Technology. However, in reality, there are still some differences in the scale of identification of undergraduates in different schools. Therefore, we will collect data of undergraduates from other schools to further improve the generalization ability of the model.

Recently, more complex Factorize Machine models have been proposed, such as DeepFM [23] and xDeepFM [24]. We will try to apply these models to the identification of needy undergraduates in the future.

Acknowledgements

This paper is financially supported by the National Undergraduates Innovation and Entrepreneurship training Program, Wuhan University of Technology, China (No. S201910497069).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Bi, H.-X. (2009) An Introduction and Analyses of Determination of the Needy Undergraduates in University or Colleges in China and Abroad. *Comparative Education Review*, **31**, 62-66.

[2] Pathman, D.E., Konrad, T.R., King, T.S., Taylor, D.H. and Koch, G.G. (2004) Outcomes of States' Scholarship, Loan Repayment, and Related Programs for Physicians. *Medical Care*, **42**, 560-568. https://doi.org/10.1097/01.mlr.0000128003.81622.ef

[3] Kim, J., Desjardins, S.L. and Mccall, B.P. (2009) Exploring the Effects of Student Expectations about Financial Aid on Postsecondary Choice: A Focus on Income and Racial/Ethnic Differences. *Research in Higher Education*, **50**, 741-774. https://doi.org/10.1007/s11162-009-9143-x

[4] Suo, L. and Gong, J. (2015) Identification of University Poor Students Based on Data Mining. 8*th International Conference on Intelligent Computation Technology and Automation*, 14-15 June 2015, 462-465. https://doi.org/10.1109/ICICTA.2015.121

[5] Li, A.F., Xiao, Z.N. and Liang, B.Y. (2017) Recognition and Analysis of Poor Stu-

dents on College Students Campus Card Consumption Data Based on Big Data. 2*nd International Conference on Mechatronics Engineering and Information Technology*, Dalian, 13-14 May 2017, 111-114. https://doi.org/10.2991/icmeit-17.2017.21

[6] Tao, B.R., Liu, K.D., Miao, F.J., Sun, T.R. and Miao, R. (2018) Targeted Poverty Reduction Model for the Needy Undergraduates Based on GA-SVM. 4*th International Conference on Social Sciences*, *Modern Management and Economics*, Chengdu, 22-23 Jun 2018, 263-266.

[7] Yao, B. (2019) Research on the Application of Data Mining Technology in University Wisdom Aid Financially. PhD, Anhui University, Hefei.

[8] Breunig, M., Kriegel, H.-P. and Sander, J. (2000) Fast Hierarchical Clustering Based on Compressed Data and Optics. *European Conference on Principles of Data Mining and Knowledge Discovery*, Lyon, 13-16 September 2000, Lecture Notes in Computer Science, Vol. 1910, 232-242. https://doi.org/10.1007/3-540-45372-5_23

[9] Yang, H., Li, D.-N. and Wang, Y.-J. (2019) K-Means Algorithm Based on LOF. *Communications Technology*, **52**, 1884-1888.

[10] Andrew, H.A. (2018) Preliminary Study on Feature Selection and Feature Generation. https://www.cnblogs.com/LittleHann/p/9384698.html

[11] Dominique, N., Ma, Z., Yang, C.N., Peng, S.L. and Jain, L.C. (2020) Enhancing Network Intrusion Detection System Method (NIDS) Using Mutual Information (RF-CIFE). 2*nd International Conference on Security with Intelligent Computing and Big-Data Services*, 14-16 December 2018, 329-342. https://doi.org/10.1007/978-3-030-16946-6_26

[12] Zhang, W.-Y., Wei, Z.-W., Wang, B.-H. and Han, X.-P. (2016) Measuring Mixing Patterns in Complex Networks by Spearman Rank Correlation Coefficient. *Physica A*: *Statistical Mechanics and Its Applications*, **451**, 440-450. https://doi.org/10.1016/j.physa.2016.01.056

[13] Xu, H. and Deng, Y. (2018) Dependent Evidence Combination Based on Shearman Coefficient and Pearson Coefficient. *IEEE Access*, **6**, 11634-11640. https://doi.org/10.1109/ACCESS.2017.2783320

[14] Johannes, D., Dominic, E. and Donald, R. (2016) Distance Correlation Coefficients for Lancaster Distributions. *Journal of Multivariate Analysis*, **154**, 19-39. https://doi.org/10.1016/j.jmva.2016.10.012

[15] Rendle, S. (2010) Factorization Machines. 10*th IEEE International Conference on Data Mining*, Sydney, 995-1000. https://doi.org/10.1109/ICDM.2010.127

[16] Guo, S.C., Chen, S.C. and Tian, Q. (2019) Ordinal Factorization Machine with Hierarchical Sparsity. *Frontiers of Computer Science*, **14**, 67-83. https://doi.org/10.1007/s11704-019-7290-6

[17] Wang, X.P. (2018) A Research on CTR Prediction Based on Ensemble of RF, XGBoost and FFM. PhD, Zhejiang University, Hangzhou.

[18] Juan, Y.C., Zhuang, Y., Chin, W.-S. and Lin, C.-J. (2016) Field-Aware Factorization Machines for CTR Prediction. *The* 10*th ACM Conference*, Boston, 15-19 September 2016, 43-50.

[19] Zhang, L., Shen, W.C., Huang, J.H., Li, S.J. and Pan, G. (2019) Field-Aware Neural Factorization Machine for Click-Through Rate Prediction. *IEEE Access*, **7**, 75032-75040. https://doi.org/10.1109/ACCESS.2019.2921026

[20] lijingru1 (2019) The Illustration of the Factorization. https://blog.csdn.net/lijingru1/article/details/88623136

[21] XtyscutI (2019) In-Depth Understanding of FFM Principles and Practices. https://blog.csdn.net/xty5057212/article/details/89202934

[22] Dopami (2018) The Use of the tf. Nn. Softmax_cross_entropy_with_logits. https://www.jianshu.com/p/648d791b55b0

[23] Guo, H.F., Tang, R.M. and Ye, Y.M. (2017) DeepFM: A Factorization-Machine Based Neural Network for CTR Prediction. *Twenty-Sixth International Joint Conference on Artificial Intelligence*, 1 August 2017, 1725-1731. https://doi.org/10.24963/ijcai.2017/239

[24] Lian, J.X., Zhou, X.H., Zhang, F.Z., Chen, Z.X., Xie, X. and Sun, G.Z. (2018) xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. 24*th ACM SIGKDD International Conference*, 14 March 2018, 1754-1763. https://doi.org/10.1145/3219819.3220023