SATBAYEV UNIVERSITY

Computing & Engineering

Volume 1 (2023), Issue 1, 36-40

https://doi.org/10.51301/ce.2023.i1.07

Applying machine learning methods for analysis socio-economic survey data

G.S. Rysmendeyeva*

Satbayev University, Almaty, Kazakhstan

**Corresponding author: g.rysmendeyeva@satbayev.university*

Abstract. To ensure the content of decision-making information systems in the individuals' assets management process requires the development of mathematical models of complex social systems. Studying the expectations of young people on socio-economic issues is of great importance for understanding the future development of the state for developing social policy strategies. A priority throughout the life cycle of an individual is happy and stable marriage, for the stability of which material and moral well-being is important. The next important factor of growing up is closely related to solving the housing problem. The strategic goal of most universities is to train highly paid specialists who are capable to develop the country and support the well-being of own family. Planning expected income is one of the steps of the family welfare planning algorithm. The purpose of this work is to study factors that are important for maturation and well-being. Using machine learning methods, the work explores socio-economic problems from the point of view of first-year university students. The influence of various factors for making decisions regarding the expected age of marriage, solving the housing problem, and expected job income is considered in the research work. Pre-processing of survey data applies data mining techniques. A comparative analysis of the forecast accuracy of classification methods is carried out: logistic regression, neural networks, support vector machines. Students are clustered using the K-means method.

Keywords: *machine learning methods, mathematical model, forecasting, behavior patterns, youth problems*.

1. Introduction

The long-term priority in the field of social policy in the direction of ensuring social guarantees and increasing the personal responsibility of the country's citizens is the development strategy of our state. Having a conscious management trajectory, a person can plan personal finances at a young age without resorting to the help of a credit institution. Information decision-making systems in the field of asset management for individuals require the development of mathematical models of complex functional social systems and algorithms for their solutions.

Research into issues of investment and human capital during the life cycle of an individual is relevant. Thus, in [1] it is noted that transferring subsidies to an earlier age increases aggregate welfare and human capital. Determining the optimal portfolio for an investor with increased risk aversion in the stock market suggests that older investors should reduce their allocations to risky assets, which is consistent with the empirical relationships between age, wealth and portfolios [2]. Work [3] notes the importance of conducting research on the life cycle of individuals in families in urban conditions. Optimal strategies for financing and investing a defined contribution pension plan when changing consumption, attitude to risk and the level of human capital. Assessing intergenerational social mobility is important for understanding the role of family in explaining income inequality. Issues of modeling optimal investment behavior considering personality factors such as abilities, human capital, strength, etc. are studied within the framework of the dynamic conflict model.

This paper aims to describe issues of family financial well-being using mathematical modeling methods for further use in the development of an information system. To achieve this goal, it is first necessary to describe the processes of creating family financial well-being and determine the limitations of the mathematical model. A study of the literature allows for the application of existing institutional investment management solutions to family finances. By family we mean three generations: children, parents and the older generation of grandparents. The welfare of the family is considered as family savings, the income of family members is considered as private cash flows. Earnings can be attributed to individuals, while family savings are distributed among family members. Savings also serve as reserves and insurance for rainy days. Since the well-being of an individual is interdependent on the well-being of the family, the totality of family assets can be considered as a long-term fund, which implies the financing of certain long-term goals. Examples of goals could be financing the education of children, financial assistance in creating a young family, purchasing real estate, and financial assistance for elderly parents. Planning expected income is one of the steps of the family welfare planning algorithm [4-6]. The purpose of this work is to study factors that are important for maturation and well-being.

1.1. Data collection and preparation

This work uses supervised learning models for labeled data for classification problem and an unsupervised learning model

for student clustering. The methods help to identify factors that divide students into clusters and classes. Required data set for learning models is collected by anonymously surveying students using online surveys via Google forms. One of the main tasks is systematization and processing of survey data. It requires preliminary data preparation, data analysis, selection of features for training models, and assessment of forecasting efficiency. Pre-processing of survey data Excel spreadsheets is carried out using intelligent and exploratory data analysis methods using the Python programming language. The volume of data is 90 questionnaires of first year students of various specialties who voluntarily filled out questionnaires. A variety of data were obtained: structured, semi-structured, unstructured regarding values and attitudes of students on different aspects of future life. Traditionally, forecasting is based on numbers, indicators, coefficients based on statistics and mathematical modeling. However currently one uses machine learning methods (MLM), since the use of unstructured data allows making predictions deeper and information technology - doing calculations faster [7-10].

Data preparation and ranking involves cleaning and organizing the source data into a consolidated format so that the resulting data set is suitable for further analysis and training of the model. The quality of the data influences the training of the selected machine learning models. Since there is relatively small data set, deleting incorrectly completed questionnaires is not suitable. All missing, blank responses were replaced with minimum, maximum, or average values. To make the data suitable for analysis, text and unstructured data were digitized. Aggregation of similar data into one variable was not produced, although some of the questions were for assessment of value factors of family well-being, for example, attitude towards future work, spouse, and income. To aggregate or isolate factors that do not have an impact and are not associated with output data, additional expert knowledge is required. No external expertise was involved, so all collected answers left without aggregation or removal. Some data relates to material and time factors like the age of marriage, the age of purchasing own home, income. In general, resulting variables refer to different types: ordered, continuous, categorical. For example, income is continuous, the choice to live with or not with parents is categorical, the assessment of the degree of importance of real estate is discrete and ordered. After converting text data, it was received up to 90 observations and 56 features of digital data set. The data obtained are not only of different types, but also lie in different ranges of values. Normalization is sensitive to outliers and is used when the data distribution unknown. Standardization is less sensitive to outliers, since it depends on the average value and standard deviation. In this work, the data were standardized according to formula:

$$x_{norm} = \left((x - \mu)\right) / \sigma$$

2. Materials and methods

2.1. Preliminary data analysis

Visualization of data in the form of histograms allows you to make a preliminary analysis of students' attitudes to issues of interest. Data analysis shows that the majority of modern young people prefer to join into marriage after they buy a house, a car and settle down. They prefer to conclude marriage at 23 or 28 years. A minority would like to get married before the age of 21 or after 33. It also shows that most students want to earn no less than 400 thousand tenge per month. Few people expect to maintain an income of 150-300 thousand tenge. There are students who want to earn more than 1 million tenge. The overwhelming majority believe that an apartment is the most important property that can be acquired by spouses during marriage. The survey of students about what does a family spend most of its income on gives the results that most students think that this is buying a real estate.

2.2. Nearest neighbors' method

Scatterplot and nearest neighbor method are used to understand the relationships between two features. So, there is no clear connection between the following characteristics: how much do parents spend on average in month for student support and at what age does the student plan to decide his own family's housing issue. In most cases the boundaries of separation between students take complex forms. However, there is decision boundary between students whose expected age of marriage before and after 27 years old. Some division occurs between students regarding opinion, that saving on a deposit is ineffective or the costs of purchasing real estate is not the largest family expenses. One can interpret the division as different attitude towards a saving and real estate among different groups of students. There is a division of opinions regarding the point that the most important thing is to pass on healthy genes to the heirs.

2.3. Clustering

The clustering method of k-means is used to determine k-groups of students with similar answers to questionnaires. In this case, the structure of the division is unknown in advance, in contrast to the classification based on a predetermined feature. The elbow method was used, which helps to select the optimal number of clusters for the clustering task. Based on the distance metric between the clusters it follows that the students are optimally divided into three clusters considering all 56 characteristics. Fig. 2 shows a two-dimensional visualization of three clusters based on the first two features, since it is impossible to display 56 features on the graph. Fig. 2 shows the centroid of three clusters according to coordinates x= feature "real estate purchase" and y= "feature financing education of children". It is possible to conclude that students who consider the financing of children's education as insignificant expenses fall into the first cluster. The rest students, who do not agree with this, are divided into two groups according to attitude to the cost of buying real estate. Thus, the second cluster estimates the cost of buying real estate on the scale of 3 and below (middle priority), the third cluster - on the scale 4 – 5 (high priority).

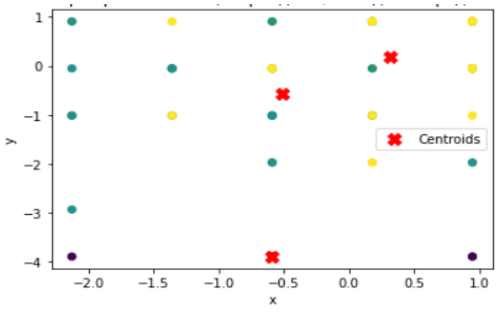

Figure 1. Scatter diagram of three clusters on x= "real estate purchase" and y= "financing education of children"

2.4. Classification

The choice of training method depends on the classification task, type of data and data set size. Most machine learning methods involve training on structured data, so the unstructured data was initially processed and organized into a structured format. In our case there were 90 observations and 56 features in the data set, so it was decided to use several methods and compare the results to study the advantages of each method. In this research work, the most important factor in family well-being is the expected income of a student. Therefore, the forecast of an expected student income was taken as the output variable of the classification model. The model for predicting student income can be used in the process of education and training of highly paid specialists. Collected from the questionnaires, the expected income data set was marked and divided into two classes: label 1, if the expected income is at least 500 thousand tenge and label 0 if it below 500 thousand tenge. As seen in Figure 2 such a division divides student into two almost equal classes.

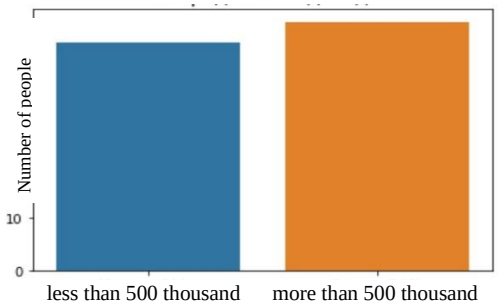

Figure 2. Output data of income

3. Results and discussion

Three classification methods were used in the research work: logistic regression, neural networks (NN) and the SVM (support vector machine method). To apply classification methods, 80% of the data was used as training data, 20% as test data. The neural network architecture consists of three layers. There are 64 neurons in the hidden layers and one neuron in the output layer. The activation function was the relu function in the hidden layers and the sigmoid function at the output layer. Forecast results for test data are presented in the error matrix (Figures 3-5). For test data, error analysis showed that logistic regression works better at reducing errors Type I, and the SVM method is better in reducing the Type II error. The NN method is optimal.

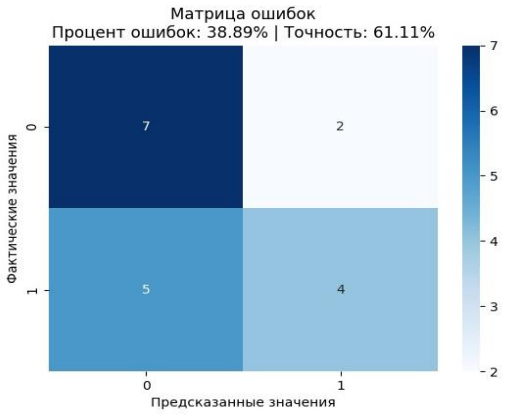

Figure 3. Logistic Regression Error Matrix

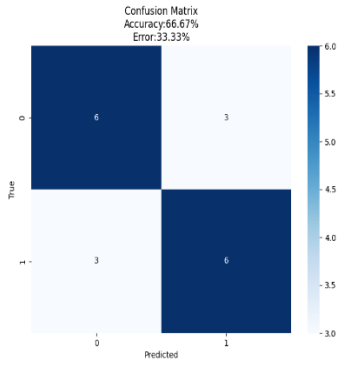

Figure 4. Error matrix of the neural network method

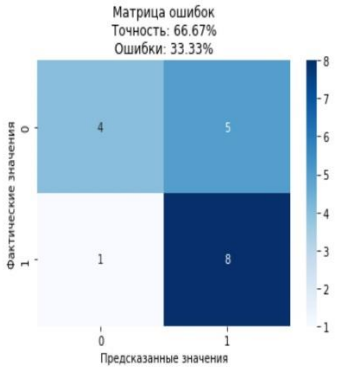

Figure 5. Error matrix, SVM method

Table 2 shows the classification performance indicators of each method, of which one can see that considering the sample size, heterogeneous data types, the ratio of the number of observations and features, the NN and SVM methods give better results than logistic regression. Moreover, the NN method gives balanced performance measures in the sense of type I and type II errors, which can be explained by the fact that in the data set the number of observations exceeded the number of features. In addition, most of the features had low variability, since it took values from a limited set {1; 2; 3; 4; 5}. Thus, the forecast accuracy of the learning model for forecasting expected income according to the survey data was 67%. This is higher than just guessing 50%, but the result is not efficient enough.

4. Conclusions

Thus, the research work collected and analyzed data from a sociological survey of university students. Structured and unstructured survey data were digitized and preprocessed by exploratory data analysis methods for further processing by machine learning methods. The clustering method of k-means was used to determine three groups of students with similar answers to questionnaires. Several machine learning methods were applied for classification students by expected future income. The research work provides a comparative analysis of the effectiveness of classification methods: logistic regression, neural networks, support vector machines – for specific task of classifying students, where the output attribute is the expected income, and the input features are answers to the socio-economic questions of the questionnaire. Received models allow studying behavior patterns and classifying and clustering students by expected income depending on the expected age of entry into adulthood, the

financial capabilities of students' parents and other factors. One can conclude that the students are optimally divided into three clusters considering 56 characteristics. However further detailed analysis is needed to interpret the data and apply for long-term forecasting and planning of family wealth.

References

[1] Caucutt, E.M. & Lochner, L. (2020). Early and Late Human Capital Investments, Borrowing Constraints, and the Family. *Journal of Political Economy, 128(3), 1065-1147*

[2] Back, K., Liu, R. & Teguia, A. (2019). Increasing risk aversion and life-cycle investing. *Mathematics and Financial Economics, (13), 287–302.* https://doi.org/10.1007/s11579-018-0228-1

[3] Caplin, A. (2018). Introduction to symposium on «Engineering Data on Individual and Family Decisions Over the Life Cycle». *Economic inquiry, 56(1), 9-12.* https://doi.org/10.1111/ecin.12468

[4] Rysmendeyeva, G.S. (2023). Modeling the age of adulthood depending on economic factors using machine learning method. *International Satbayev Conference, (4), 148-160*

[5] Rysmendeyeva, G.S. (2021). Development of mathematical models for the information system of decision-making in the process of asset management. *Vestnik KazNITU, (3), 65-75*

[6] Rysmendeyeva, G.S. (2020). Development of visual models of the information system of decision-making in asset management process of physical persons. *Vestnik KazNITU, (5), 434-439*

[7] Zakharova, I.G. (2018). Machine learning methods of providing informational management support for students' professional development. *Obrazovanie I Nauka-Education and Science, 20(9), 91-114*

[8] Dhruvil, S., Devarsh, P., Jainish, A., Pruthvi, H. & Manan, S. (2021). Integrating machine learning and blockchain to develop a system to veto the forgeries and provide efficient results in education sector. *Visual Computing for Industry Biomedicine and Art, 4(1), 18-28*

[9] Shuo-Chang Tsai, Cheng-Huan Chen, Yi-Tzone Shiao, Jin-Shuei Ciou, Trong-Neng Wu. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education,17(12), 12-25*

[10] Zubarev, A., Bekirova, O. (2020). Analysis of Bank Default Factors in 2013-2019. *Ekonomicheskaya politika, 15(3), 106-133*

Әлеуметтік-экономикалық сауалнамалардың мәліметтерін талдау үшін машиналық оқыту әдістерін қолдану

Г.С. Рысмендеева*

Satbayev University, Алматы, Қазақстан

**Корреспонденция үшін автор: g.rysmendeyeva@satbayev.university*

Аңдатпа. Жеке активтерді басқару процесінде шешім қабылдаудың ақпараттық жүйелерінің мазмұнын қамтамасыз ету үшін күрделі әлеуметтік жүйелердің математикалық үлгілерін жасау қажет. Жастардың әлеуметтік-экономикалық мәселелер бойынша үміттерін зерделеу мемлекеттің даму болашағын түсіну және әлеуметтік саясаттың стратегияларын әзірлеу үшін үлкен маңызға ие. Адамның бүкіл өмірлік циклiндегі басымдық - бұл бақытты және тұрақты неке, оның тұрақтылығы үшін материалдық және моральдық әл-ауқат маңызды. Бұл өсудің маңызды факторы тұрғын үй мәселесін шешумен тығыз байланысты. Көптеген университеттердің стратегиялық мақсаты – елді дамытуға және отбасының әл-ауқатына қолдау көрсетуге қабілетті, жалақысы жоғары мамандарды дайындау. Күтілетін табысты жоспарлау отбасының байлығын жоспарлау алгоритмінің кезеңдерінің бірі болып табылады. Бұл жұмыстың мақсаты - жетілу және әл-ауқат үшін маңызды факторларды зерттеу. Машиналық оқыту әдістерін қолдана отырып, жұмыс университеттің бірінші курс студенттері тұрғысынан әлеуметтік-экономикалық проблемаларды қарастырады. Зерттеу жұмысы күтілетін неке жасына, тұрғын үй шешімдеріне және жұмыстан күтілетін табысқа қатысты шешім қабылдауға әртүрлі факторлардың әсерін зерттейді. Сауалнама деректерін алдын ала өңдеу деректерді іздеу әдістерін қолдану арқылы жүзеге асырылады. Жіктеу әдісінің болжам дәлдігіне салыстырмалы талдау жүргізіледі: логистикалық регрессия, нейрондық желілер, тірек векторлық машиналар. Оқушылар «К-орталары» әдісі арқылы топтастырылады.

Негізгі сөздер: *машиналық оқыту, математикалық модель, болжау, мінез-құлық үлгілері, жастар мәселелері.*

Применение методов машинного обучения для анализа данных социально-экономических опросов

Г.С. Рысмендеева*

Satbayev University, Алматы, Казахстан

**Автор для корреспонденции: g.rysmendeyeva@satbayev.university*

Аннотация. Для обеспечения содержания информационных систем принятия решений в процессе управления активами личности необходима разработка математических моделей сложных социальных систем. Изучение ожиданий молодежи по социально-экономическим вопросам имеет большое значение для понимания перспектив

развития государства и разработки стратегий социальной политики. Приоритетом на протяжении всего жизненного цикла личности является счастливый и стабильный брак, для стабильности которого важно материальное и моральное благополучие. Данный важный фактор взросления тесно связан с решением жилищного вопроса. Стратегической целью большинства университетов является подготовка высокооплачиваемых специалистов, способных развивать страну и поддерживать благополучие своей семьи. Планирование ожидаемого дохода является одним из этапов алгоритма планирования благосостояния семьи. Целью данной работы является изучение факторов, важных для взросления и благополучия. С помощью методов машинного обучения в работе исследуются социально-экономические проблемы с точки зрения студентов-первокурсников вуза. В исследовательской работе рассматривается влияние различных факторов на принятие решений относительно предполагаемого возраста вступления в брак, решения жилищного вопроса, ожидаемого дохода от работы. Предварительная обработка данных опроса осуществляется методами интеллектуального анализа данных. Проводится сравнительный анализ точности прогноза методов классификации: логистической регрессии, нейронных сетей, опорных векторов. Проводится кластеризация студентов методом К-средних.

Ключевые слова: *машинное обучение, математическая модель, прогнозирование, паттерны поведения, проблемы молодежи.*

Received: 28 November 2022
Accepted: 16 March 2023
Available online: 31 March 2023