Case-Based Teaching for Python Language Under the Background of Big Data

FENG LI, YUJUN HU, LINGLING WANG*
School of Management Science and Engineering
Anhui University of Finance and Economics
Bengbu 233030, CHINA

Abstract: - This paper proposes a new teaching method for the Python language programming course, which can better enable students to understand and use in the background of big data. Python is an open-source programming language with a community-based model. In this paper, firstly, various functions are described in Python Language. Additionally, different application areas are presented in this paper, such as transportation logistics, urban management, biomedical field, smart power grid, energy field, and commercial field. Finally, bank customer churn as case-based teaching is introduced can improve the students' confidence in their future studies.

Key-Words: Case-Based Teaching; Python Programming; Big Data; Bank Customer Churn.

Received: April 16, 2021. Revised: April 19, 2022. Accepted: May 15, 2022. Published: July 6, 2022.

1 Introduction

In recent years, information technology has developed rapidly with the advent of the Internet era. Data processing and mobile Internet have become hot topics for people's application and research, accounting for an increasing proportion of people's lives and becoming an irreplaceable part of life. In this information age, almost every minute, every corner of the world, is generating data, and the volume of data is growing faster and faster [1]. In the face of the rapid growth of massive information, the original data processing technology can no longer meet the needs of current data process intelligent it cannot effectively deal with a complicated and large amount of data, the amount of data is increasing, and people's exploration of it is further deepened, big data technology arises at the historic moment. Big data technology can quickly and effectively process a large number of complex types of data and screen out valid data [2,3].

After the release of Hadoop in 2006, Yahoo first applied it, and then more and more large companies began to use Hadoop for big data storage and computing [4]. In 2008, Hadoop officially became Apache's top project, and many big data commercial companies began to appear. At the same time, the programming model of MapReduce is complicated. Yahoo internally developed Pig as a scripting language, which provides SQL-like syntax. Developers can use Pig scripts to describe operations on data sets, and after Pig is compiled, MapReduce programs are generated and run into Hadoop clusters [5].

With the in-depth application of big data and other high-tech technologies in the field of transportation, big data technology plays an increasingly important role in the construction of convenient, efficient, economical, and green urban transportation systems, as well as in the planning and making scientific and accurate decisions of urban transportation departments [6].

At present, the application technology of big data information technology has developed rapidly with the advent of the Internet era supporting the development of the industry. The basic technical framvariousework of the big dasystemstem has reached a relatively mature and stable degree. In the constant pursuit of efficiency of the social deinformation ageent direction of big data

technology has also begun to change to improve efficiency.

2 Overview of Python Language

Python language is a free, open-source, cross-platform advanced dynamic programming language [7]. It supports various programming methods and has a large number of powerful built-in objects, standard libraries, and extension libraries. Powerful programming functions can be realized by directly calling the built-in functions or standard libraries. By its nature, Python is both an "object-oriented" language and an "interpreted" language [8]. Python is relatively easy to get started. Its syntax is similar to That of English, and programs can be executed directly through the interpreter, but it consumes a lot of hardware resources.

On the application side, the Python Language is particularly well suited for data analysis and processing. Matplotlib is a 2D drawing tool that is often used to chart data with a few simple lines of code [9]. Pandas is an open-source tool for manipulating complex two-dimensional and three-dimensional arrays and for manipulating data in relational databases [10]. Python's powerful and rich libraries and data analysis capabilities make it well suited for the field of artificial intelligence. In neural networks and deep learning, Python can find mature packages to call. And Python is an object-oriented, dynamic language for scientific computing, which makes Python a favorite for artificial intelligence. The power of scientific computation remains Python's strongest competitiveness in the field of AI and Big Data. Python aims to train students to make use of Python language in the application of their major and plays an important role and position in the curriculum system and major construction of machine learning, pattern recognition, computer vision, and so on. The language is an interpretive, object-oriented computer programming language for data statistics, analysis, visualization, and other tasks, as well as machine learning, artificial intelligence, and other fields.

Additionally, it can meet almost all the functional requirements of data processing, statistical model, and graph drawing under data mining. A large number of third-party modules support content ranging from statistical computing to machine learning, from financial analysis to biological information, from social network analysis to natural language processing, and from various databases, and various language interfaces to high-performance computing models.

To sum up, the introduction of the pre-class targeted preview method in the course of big data analysis and application effectively promotes the development of classroom teaching; In the course of teaching system theory, small cases of Python are integrated to deepen students' understanding and mastery of relevant theoretical knowledge. Students are required to conduct data mining and analysis for applications in aviation, e-commerce, public service, power, and other industries. In the process of project development, students learned to use octopus and other tools for data collection; Completed data cleaning, attribute reduction, data transformation, and other data processing work; Programming to realize the process of data visualization; Completed the model building and analysis of group tasks, discussed common algorithms such as k-means clustering optimal K value selection scheme, and conducted comparative experimental analysis; For the optimization and application expansion of the model, the author also puts forward his views and makes a preliminary attempt. Compared with the traditional single theoretical teaching mode, the mixed teaching mode gives full play to the students' main role and awareness of classroom participation, forms a good classroom interaction, greatly stimulates students' interest in learning, improves their hands-on ability, enhances the teaching effect, and achieves the expected teaching objectives.

3 Application of big data technology

Big data technology includes data collection, data access, infrastructure, data processing, statistical analysis, data mining, model prediction, and result presentation. Its main applications are as follows:

3.1 Application of big data technology in the field of transportation logistics

At present, the following problems are common in urban traffic in China, such as the inadequate and unreasonable establishment of traffic facilities and lax traffic management. The urban traffic construction speed will be difficult to match the growth rate of vehicle ownership in the current year and the following years, resulting in traffic problems. A variety of reasons, resulting in a high traffic accident rate. For example, in terms of intelligent transportation, big data technology is used to study the relationship between vehicle traffic efficiency and traffic light segmentation time, speed, and road congestion, and establish a traffic light management model to improve vehicle traffic efficiency and alleviate traffic congestion [11].

The AI intelligent camera based on big data technology can monitor and record the speed, quantity, and road conditions of vehicles on the road in real-time, and send them to the comprehensive management platform for analysis and processing through a high-speed information transmission network, to help the traffic management department make the current judgment and decision. Using the advantages of big data technology, through the rapid analysis and feedback of a large number of detailed traffic data in real-time, it can judge and predict the traffic events and accident risks that may exist on the road and can be linked with hardware products for early warning, to effectively prevent traffic accidents and avoid causing road congestion [12].

3.2 Application of big data technology in urban management

Through the use of big data technology, each part of urban management can be converted into accurate and detailed data, which can provide scientific and effective solutions for urban administration, traffic management, ecological management, and so on. Reasonable planning can promote the development of industry and trade, while unreasonable development will lead to the waste of government investment and economic loss of investors, and inhibit economic development to a certain extent [13].

Big data technology can analyze macro geographical spatial distribution and promote urban construction and development under scientific and reasonable planning. Big data is also widely used in healthcare and education, energy, manufacturing, finance, and cultural media. Realize intelligent transportation, environmental monitoring, urban planning, and intelligent security.

3.3 Application of big data technology in the biomedical field

With the application of big data technology, it can efficiently collect, manage, query, and analyze the data with continuous and rapid growth. So that the medical staff can know the side effects of the drug, the situation and people of the drug are not easy to apply before prescribing drugs to patients, which greatly reduces the probability of medical accidents; Before prescribing painkillers, doctors can learn whether patients are at risk of drug addiction from big data on their medical records. If so, doctors can choose different treatment methods in advance [14]. With big data technology, prescriptions, treatment plans, and other medical data can be efficiently analyzed, helping to discover the optimal treatment plan, confirm the patient's disease development, and identify chronic diseases. Intelligent management of big data technology has realized personalized diagnosis methods, played an important role in disease prediction, disease analysis, and patient control, greatly promoted the development of medical technology, and enabled human beings to explore deeper mysteries of life.

3.4 Application of big data technology in the smart power grid

The era of big data has brought new development opportunities and new challenges to the power industry. With the deepening of information construction in electric power enterprises, the amount of data generated by business systems is increasing explosively. At present, massive storage brought by big data and some business systems are facing challenges such as high storage upgrade costs and slow system response speed. A smart grid refers to the

integration of big data technology into the traditional energy network to form a new power grid, which optimizes the production, supply, and consumption of electric energy through information such as users' consumption habits. For example, by analyzing data in the power grid, we can know which areas have excessive power loads and the frequency of outages, or predict which lines are likely to fail. These results are helpful for power grid upgrading and maintenance [15].

For power equipment, big data can be used for real-time monitoring and early warning diagnosis of the environmental status information, mechanical status information, and operating status information of the equipment, to do a good job of fault prediction and equipment maintenance in advance, to improve the level of equipment maintenance, automatic diagnosis, and safe operation. Big data technology can improve the perception ability of primary equipment, and well combine it with secondary equipment to realize joint processing, data transmission, comprehensive judgment, and other functions, and improve the technical level and intelligence degree of the power grid.

3.5 Application of big data technology in the energy field

At present, promoting a carbon-neutral environment, new energy vehicles in just a few years to get the market share of the rapid increase, thereby charging pile demand is also increasing, combined with the current market research, and even in some areas, the number of new energy automobile ownership and charging pile than serious, charging pile is in short supply, unreasonable setting of many problems. The application of big data technology has played a great role in promoting the intelligent operation of new energy vehicles. Relying on big data technology, a cloud platform for the operation of new energy vehicles is formed. By using the data analysis and mining capabilities of big data technology, new energy vehicle service providers and charging pile suppliers can integrate vehicle information and charging information on the cloud platform. Through the analysis of big data technology, the scientific and reasonable construction of charging stations can not only provide users with a better experience but also maximize the use of charging stations [16-18].

3.6 Application of big data technology in the commercial field

In the current era, the competition between enterprises in the market is very fierce. An enterprise, to obtain a foothold in the rapidly changing market environment, needs to distinguish the truth and fallibility of information, peep out business opportunities, make reasonable business decisions, grasp business opportunities and create enterprise value. Big data makes concepts like consumer behavior more comprehensive and measurable. It promotes the advancement of inductive reasoning by creating a more dialectical data-first scientific approach [19]. The creation of big data has also enabled many companies to fully integrate business analytics, giving more non-technical employees access to data and data-driven insights. Although this ability, the leader's vision and strategy, the strength of the enterprise staff and pay, all are closely linked but need more accurate data information as a basis, scientific and reasonable business decisions as to the guidance, to get a business plan, will play to the strength of the enterprise, seize business opportunities, create business value, to obtain a larger market. The application of big data can search for the latest data, screen useful information, and conduct scientific analysis of data [20].

In addition, the enterprise data, including capital, order quantity, department personnel, and customer information, can be scientifically analyzed in all aspects to obtain the results after data calculation, rather than the subjective judgment of decision-makers, which can effectively make up for decision-making errors caused by decision-makers lack of strength or insufficient information control. With the application of big data technology, data analysis, scientific decision-making by leaders, and efficient work of employees enable enterprises to achieve scientific development, greatly enhance the

core competitiveness of enterprises, and maximize profits for the enterprise [21].

4 Case-based teaching for Python Programming

With the vigorous development of the economy, the deepening of economic globalization and diversification, the banking industry has been impacted, and the loss of customers has increased. Therefore, it is very important to forecast the loss of customers of banks and discuss its influencing factors. To analyze the key factors affecting bank customer churn, and then formulate corresponding customer retention strategies to solve the problem of bank customer churn, the bank customer churn prediction system came into being. However, in the face of a large amount of customer information, it is difficult for the traditional bank customer churn prediction system to have a high prediction accuracy to solve this problem.

This example collects anonymous data from foreign banks, including credit scores, deposits and loans, gender, age, and a series of customer information. Data attributes can be divided into 14 columns.

- *RowNumber*
- *CustomerID*
- *Surname*
- *CreditScore*
- *Geography*
- *Gender*
- *Age*
- *Tenure*
- *Balance*
- *NumOfProducts*
- *HasCrCard*
- *IsActiveMember*
- *EstimatedSalary*
- *Exited*

4.1 Data Processing

Generally, the obtained dataset has redundant attributes, noise, or non-numerical attributes, which cannot be used directly. Therefore, it is necessary to process the data in advance, and then train the data set with high quality. Geography and Gender are listed as non-numerical features in the original data set, which need to be converted into numerical features. For other continuous variables, they need to be discretized.

There are some non-numerical features in the original data set, such as Geography (France, Spain, Germany) and Gender (female, male). These non-numerical features may play a large role in classification. Therefore, to enable the model to process these non-numerical features, we need to neutralize these two features.

The factorize function in Python's Pandas library numerals non-numerical features by mapping the same nominal types to the same numbers.

```
import pandas as pd
def quantification(dataPath,outputPath):
    df=pd.read_csv(dataPath)
    x=pd.factorize(df['Geography'])
    y=pd.factorize(df['Gender'])
    df['Geography']=x[0]
    df['Gender']=y[0]
    df.to_csv(outputPath)
```

The decision tree algorithm needs to deal with discrete data. As there are continuous variables such as credit score, age, deposit and loan, and estimated income in the original data set, these continuous variables need to be transformed into discrete variables.

	CreditScore	Age	Balance	EstimatedSalary
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	76485.889288	100090.239881
std	96.653299	10.487806	62397.405202	57510.492818
min	350.000000	18.000000	0.000000	11.580000
25%	584.000000	32.000000	0.000000	51002.110000
50%	652.000000	37.000000	97198.540000	100193.915000
75%	718.000000	44.000000	127644.240000	149388.247500
max	850.000000	92.000000	250898.090000	199992.480000

Fig 1. Statistics of variables

For the CreditScore attribute, 25% of the data is less than 584, 50% of the data is less than 652, and 75% of the data is less than 718, so it is divided into four categories based on the quartile. For credit scores less than 584 data is divided into first gear, the credit score is 584 ~ 652 data is divided into the second leg, credit scores of 584 ~ 718 data is divided into third, credit scores greater than 718 data is divided into the fourth gear, and so on, the

```
    import pandas as pd
    def filtering(dataPath, outputPath):
        df = pd.read_csv(dataPath)
        df_new = pd.DataFrame(
            columns=['Geography', 'Age', 'EstimatedSalary', 'NumOfProducts', 'CreditScore', 'Tenure',
'HasCrCard','IsActiveMember', 'Exited', 'Gender'])
        ones = sum(df["Exited"])
        length = len(df["Exited"])
        zeros = length - ones
        i = 0; flag_0 = 0; flag_1 = 0
        while i != length:
            if df["Exited"][i] == 0 and flag_1 < 1 * ones:
                df_new = df_new.append(pd.DataFrame(
                    {'Gender': df["Gender"][i], 'Geography': df["Geography"][i], 'Age': df["Age"][i],'EstimatedSalary':
df["EstimatedSalary"][i], 'NumOfProducts': df["NumOfProducts"][i],'CreditScore': df["CreditScore"][i], 'Tenure':
df["Tenure"][i], 'HasCrCard': df["HasCrCard"][i],'IsActiveMember': df["IsActiveMember"][i], 'Exited': df["Exited"][i]},
index=[i]))
                flag_1 = flag_1 + 1
            if df["Exited"][i] == 1 and flag_0 < 1 * zeros:
                df_new = df_new.append(pd.DataFrame(
                    {'Gender': df["Gender"][i], 'Geography': df["Geography"][i], 'Age': df["Age"][i],'EstimatedSalary':
df["EstimatedSalary"][i], 'NumOfProducts': df["NumOfProducts"][i],'CreditScore': df["CreditScore"][i], 'Tenure':
df["Tenure"][i], 'HasCrCard': df["HasCrCard"][i],'IsActiveMember': df["IsActiveMember"][i], 'Exited': df["Exited"][i]},
index=[i]))
                flag_0 = flag_0 + 1
            i = i + 1
        df_new.to_csv(outputPath)
```

age, the deposit, and lending situation, and discretization estimated income in this way. By observing the original data set, it was found that in the deposit and loan column, there were a large number of users whose data were 0, that is, there was no deposit and loan. Therefore, it was divided into a single level, and the other non-zero values were divided according to their uniqueness.

Due to the unbalanced classification of training data in the original data set, to achieve a better model effect, there are generally two methods of over-sampling and under-sampling to solve the problem of unbalanced classification. Here, the simplest under-sampling method is adopted to delete the redundant classification data.

4.2 Data modeling

After preprocessing the data set, we first use the decision tree to analyze the bank customer churn prediction and take the churn prediction results as a benchmark. Then, based on the decision tree model, we compare the results with the optimization results of several common classification algorithms.

The meaning of the decision tree is intuitive and easy to explain. For practical applications, the decision tree has the speed advantage that other algorithms are difficult to compare. Therefore, for the decision tree, on the one hand, it effectively processes and learns large-scale data, on the other hand, it can meet the real-time or higher speed requirements in the test or prediction phase.

Sklearn provides the training model of the decision tree, which uses cart algorithm. Cart algorithm only generates a binary tree, that is, a non-leaf node generates two child nodes each time, indicating whether it meets or does not meet the conditions of the node. In this example, the problem to be solved is to determine whether a customer is a vulnerable customer, which is a classification problem. Therefore, using the decision tree classifier in the skeleton library to solve this problem.

4.3 Experiment results

The decision tree model with the maximum depth of 5 and the minimum number of samples required for node splitting of 100 is used to verify and generate an evalua-tion report. It is concluded that the overall accuracy of the model reaches 76.93%. In view of the unbalanced characteristics of bank customer churn data, we should not only refer to the accuracy, but also make a comprehensive evaluation in combina-tion with indicators such as

```
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(criterion="gini",max_depth=5,min_samples_split=100)
dt_model.fit(feature_train,target_train)
#Resusts
predict_results = dt_model.predict(feature_test)
#Scores
scores = dt_model.score(feature_test,target_test)
```

the area under the curve and the confusion matrix.

Through the classification effect evaluation, it can be found that the performance of the bank customer churn prediction model obtained by the decision tree algorithm is good, which reflects the good adaptability of the decision tree algorithm.

Firstly, the anonymous user data of foreign banks are collected, the data is pre-processed, the non-numerical attributes are numeralized, the continuous variables are discretized, and the undersampling method is used to balance the training data cate-gories. Then the decision tree algorithm is used to analyze the prediction of customer churn and calculate the prediction results. Furthermore, this result can be used as a benchmark for rewriting to increase the depth of decision tree and the number of samples required for maximum node splitting, and the classification effect before and after improvement can be compared. Finally, the overall accuracy can be ob-tained, and the result is 76.93% in Fig 2.

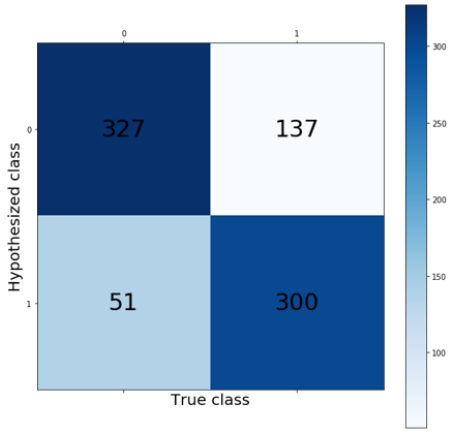

Fig 2. Experiment results

6 Conclusions

Big data technology has been applied in various fields of contemporary people's life. Meanwhile, it has brought an undoubted impact on everyone. At the same time, the application of big data technology has also become a person's core competitive-ness in the new era. Therefore, in the industrial Internet stage, big data will be gradually introduced, but inevitably. Based on the above situation, we can come to the conclusion that big data is applicable to every aspect of life and big data behavior is everywhere. In the future, the development space of big data will be bigger and bigger, and the demand for human resources will also be bigger and bigger.

Developing big data technology has become a national development strategy, but the industry still faces many challenges. In the future, the development trend of big data technology will be to establish comprehensive database, diversified fusion anal-ysis will gradually replace single analysis, and data mining technology will be more mature.

Acknowledgment

We thank the anonymous reviewers and editors for their very constructive comments. This work was supported in part by the Natural Science Foundation of the Higher Education Institutions of Anhui Province under Grant No. KJ2020A0011, Innovation Support Program for Returned Overseas Students in Anhui Province under Grant No. 2021LCX032. the Science Research Project of Anhui University of Finance and Economics under Grant No. ACKYC20085, Undergraduate teaching quality and teaching reform project of Anhui University of Finance and Economics under Grant No. acszjyyb2021035.

References:

[1] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques, and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

[2] Cheng, Ying, et al. "Data and knowledge mining with big data towards smart production." Jour-nal of Industrial Information Integration 9 (2018): 1-13.

[3] Hu, Han, et al. "Toward scalable systems for big data analytics: A technology tutorial." IEEE access 2 (2014): 652-687.

[4] Khetrapal, Ankur, and Vinay Ganesh. "HBase and Hypertable for large scale distributed storage systems." Dept. of Computer Science, Purdue University 10.1376616.1376726 (2006).

[5] Jena, Bibhudutta, et al. "A survey work on optimization techniques utilizing map reduce frame-work in hadoop cluster." International Journal of Intelligent Systems and Applications 9.4 (2017): 61.

[6] Hu, Han, et al. "Toward scalable systems for big data analytics: A technology tutorial." IEEE access 2 (2014): 652-687.

[7] Sanner, Michel F. "Python: a programming language for software integration and development." J Mol Graph Model 17.1 (1999): 57-61.

[8] Srinath, K. R. "Python–the fastest growing programming language." International Research Journal of Engineering and Technology (IRJET) 4.12 (2017): 354-357.

[9] Ari, Niyazi, and Makhamadsulton Ustazhanov. "Matplotlib in python." 2014 11th International Conference on Electronics, Computer and Computation (ICECCO). IEEE, 2014.

[10] McKinney, Wes. "pandas: a foundational Python library for data analysis and statistics." Python for high performance and scientific computing 14.9 (2011): 1-9.

[11] Ayed, Abdelkarim Ben, Mohamed Ben Halima, and Adel M. Alimi. "Big data analytics for logis-tics and transportation." 2015 4th international conference on advanced logistics and transport (ICALT). IEEE, 2015.

[12] Yan, Zengwen, et al. "The application of big data analytics in optimizing logistics: a developmen-tal perspective review." Journal of Data, Information and Management 1.1 (2019): 33-43.

[13] Liu, Yang. "Big data technology and its analysis of application in urban intelligent transportation system." 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS). IEEE, 2018.

[14] Luo, Jake, et al. "Big data application in biomedical research and health care: a literature review." Biomedical informatics insights 8 (2016): BII-S31559.

[15] Jaradat, Manar, et al. "The internet of energy: smart sensor networks and big data management for smart grid." Procedia Computer Science 56 (2015): 592-597.

[16] Wang, Jin, et al. "Big data service architecture: a survey." Journal of Internet Technology 21.2 (2020): 393-405.

[17] Marlen, Azamat, et al. "Application of big data in smart grids: Energy analytics." 2019 21st Inter-national Conference on Advanced Communication Technology (ICACT). IEEE, 2019.

[18] Mohammadpoor, Mehdi, and Farshid Torabi. "Big Data analytics in oil and gas industry: An emerging trend." Petroleum 6.4 (2020): 321-328.

[19] Al Nuaimi, Eiman, et al. "Applications of big data to smart cities." Journal of Internet Services and Applications 6.1 (2015): 1-15.

[20] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analyt-ics." International journal of information management 35.2 (2015): 137-144.

[21] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." Information sciences 275 (2014): 314-347.

Sources of funding for research presented in a scientific article or scientific article itself