Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/54/20241411

Subscribing prediction of term deposit based on decision tree, random forest and support vector machine

Yifan Chen

School of Southampton, Southampton, United Kingdom

201900170215@mail.sdu.edu.cn

Abstract. To classify customers and predict their behaviour based on some of their features and what they do before is most people want to do. This study finds a data set about bank customers from Kaggle and use three different classification models to classify customers and predict whether they will subscribe a term deposit based on some of their features. The three classification models are decision tree model, random forest model and support vector machine model. Firstly, using these models to get the feature importance and accuracy rate to evaluate the result. In addition, this study changes the parameters about these models and find how can get better result. Based on these models to process data set and getting the result, this paper also compared their results and find advantages and disadvantages of three models. Finally, this paper also discusses how to improve the models so that they can get the better result and solve some other problems.

Keywords: Decision tree, random forest, support vector machine, subscribing prediction.

1. Introduction

In the modern society, people always want to predict other people's behaviours by some features. For example, hospital want to know what kind of people will need what kind of medicine so that they can be easier to make a prescription. Insurance company will explore whether people will buy insurance according to their characteristic to reduce the number of invalid calls. Similarly, bank will also do the same thing. For example, they will predict their customer churn [1]. Machine learn or other classification methods can help them predict accurately [2]. It is important for bank to predict customer behaviours that they can make different plans for different customers to make sure efficiency. When the bank gets every customer characteristic, Banks can combine customers previous actions and compare carefully so that not only they can predict what will they do next, but also can predict new customer's behaviours.

There are a lot of methods to classify data. For example, decision tree is a very classical method. Decision tree can choose several possible schemes and find the best scheme depend on calculate information gain rate. If decision tree has many decisions point in it, it will choose the root of the decision tree as final scheme., Decision tree has many advantages. Firstly, it is using white box model, so it is very easy to plain [3]. Second, it can handle multiple output problems. Decision tree can be used in many fields, such as it can be used to classify knee Injury Status [4]. Many people are trying to improve its function like predict cycle and definite the margin [5]. Decision tree is a basic method that depend on it can generate many new methods [6]. There is a method which is improve based on decision tree which called random forest. The principle of random forests is choosing several features as a

DOI: 10.54254/2755-2721/54/20241411

decision tree once and repeat operation. According to this, random forest is more accuracy than decision tree and for the same data set, the results of each experiment may be different because of each experiment choosing different decision trees. By generating many decision trees, random forest can get better result. However, if one repeats too few, the variable importance rankings will be different [7]. Identically, It will overfit when the trees too many. Besides, random forest is faster than detailed power flow studies and accurately [8]. Besides decision tree and random forests, support vector machine is also a good method to classify data. Support vector machine can map the data form low space to high space by kernel function so that the data can be separated in the new space. It can be divided into hard-margin and soft-margin [9]. When using these methods, one needs to be careful of losing information because it depends on support vector [10]. Support vector machine has many advantages like it can solve small sample problem and handle high dimensional data. Support vector machine has good robustness and interpretability as well.

In this paper will predict whether customer will subscribe a term deposit by some of their characteristics. The basis of the predict whether customer will subscribe a term deposit is based on their characteristics, such as age, job. Therefore, according to different characteristics should divide customers into two categories which is will subscribe a term deposit or will not subscribe a term deposit. In this paper will select decision tree, random forest and support vector machine as classification methods to classify and predict. Meanwhile, compare the result to understand the accuracy between different model is also necessary. In the second sections, this paper will present research data and three models and will also present the parameter and test indicators of the model. In the third section will show the result and compare the result from different models. In the fourth section will show the limitation and future outlooks. In the last section will make a conclusion of this paper.

2. Data and method

Before use different model to classifier customer and predict their behaviours, find a suitable data set is very important. This paper use bank customer data set which is found on Kaggle. The data set as follow conclude 42639 rows and 17 columns. In the data set, there are 16 columns of customers characteristic. In these columns, there are different contents in different columns. The types of jobs are unknown, technician, entrepreneur, blue-collar, management, retired, admin, housemaid, self-employed, services, student and unemployed. The types of marital are divorced, married and single. The types of education are primary, tertiary, and secondary. The type of default, housing and loan are divided into yes and no. The types of contact are unknown, cellular and telephone. The types of outcomes are success, failure, other and unknown. The other types are just number. The last column is term deposit. If the content is yes, it is mean that customer will subscribe a term deposit. On the contrary, if the content is no that means customer will not subscribe a term deposit. This paper will be based on the 16 columns information should classify and predict customers whether they will subscribe a term deposit.

To classify the customers, there are many classification models can be used. This paper will use decision tree model, random forest model and support vector machine model to complete. In data set, the last column will be set as label and the other columns will be set as feature. When use the classification model based on python, set parameters should be put firstly. In decision tree model, there are many parameters one can set and change. For example, the criterion can be changed which means different criteria for classifying features. Classification strategy can choose one of best or random. Besides these, there are many parameters can be chosen in decision tree model. When use random forest model and support vector machine model, they are the same as decision tree model that should be set the appropriate parameters. After using classification models and getting the result, test indicators of the model is necessary. Accuracy rate is the most typical test indicators in classification task. The principle of accuracy is using the same part of predict label as the test label to divided by the test label. The accuracy rate higher the predict result better.

DOI: 10.54254/2755-2721/54/20241411

3. Results and discussion

3.1. Decision tree

The first classification model is decision tree model. Before classify the customers, determine the degree of importance of each feature is also important. The result is shown in Fig. 1. According to this picture can understand the degree of importance that duration is the most important feature and default is the least important feature. The reason why the degree of importance of each feature show like this is mainly depend on the type of content of each feature. For example, the type of content of duration is the most so that its feature importance is the highest. Besides, the more same type of content of one feature correspond to the same result means this feature is more important. After getting the result of degree of importance of different features, it is time to set different parameters to get the predict result. In Fig.3 show the accuracy rate of different max depth.

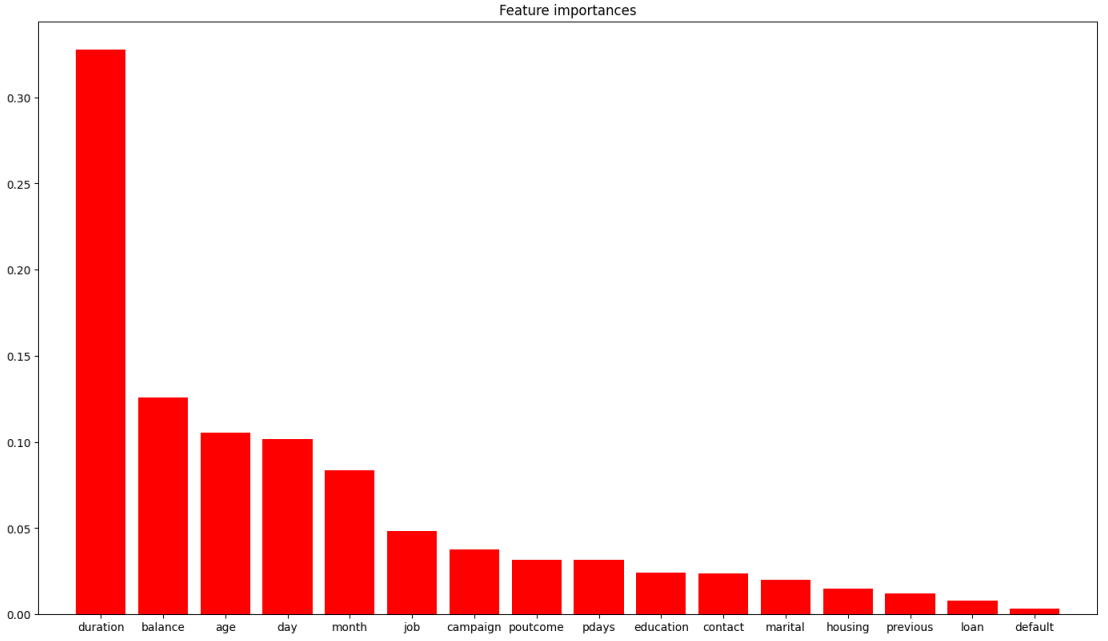

Figure 1. Feature importance of decision tree

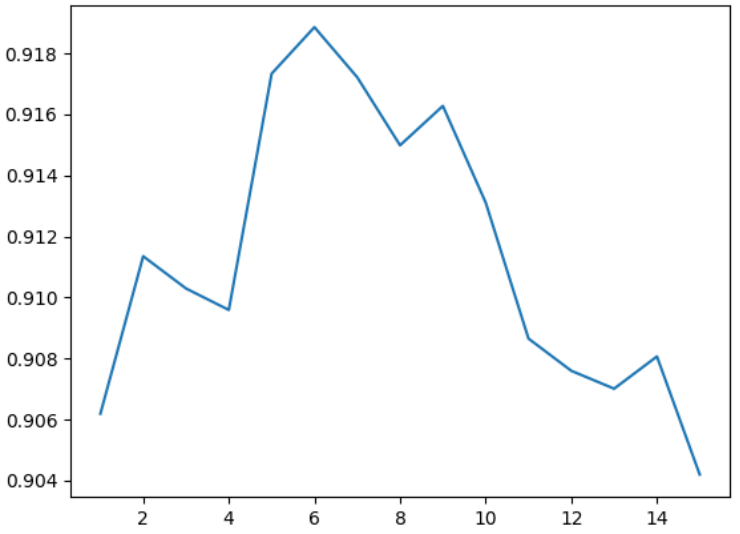

Figure 2. Accuracy result of decision tree

DOI: 10.54254/2755-2721/54/20241411

According to Fig. 2, one can understand when the number of max depth change from one to six, the accuracy rate is keeping rising. However, if the max depth keeps rising, the accuracy rate will start to fall. The reason for this is because when the number is not too big, the more max depth, it means the decision tree will consider more features so that the accuracy rate will be higher. On the contrary, if the max depth becomes too big, the decision tree will consider too many features to be influenced by some insignificance features. In this case, the accuracy rate will start fall.

3.2. Random forest

The second classification model is random forest model. The result of determine the degree of importance of each feature is presented in Fig. 3. The result is almost as the same as the degree of importance of each feature in decision tree models. The important of some features has changed is because that random forest choose different features to make a decision tree and there are many decision trees in a random forest model. In random forest model, this paper tries to change the number of decision trees from one to sixteen when the max depth is six or default value which means no limit. The results are given in Fig. 4.

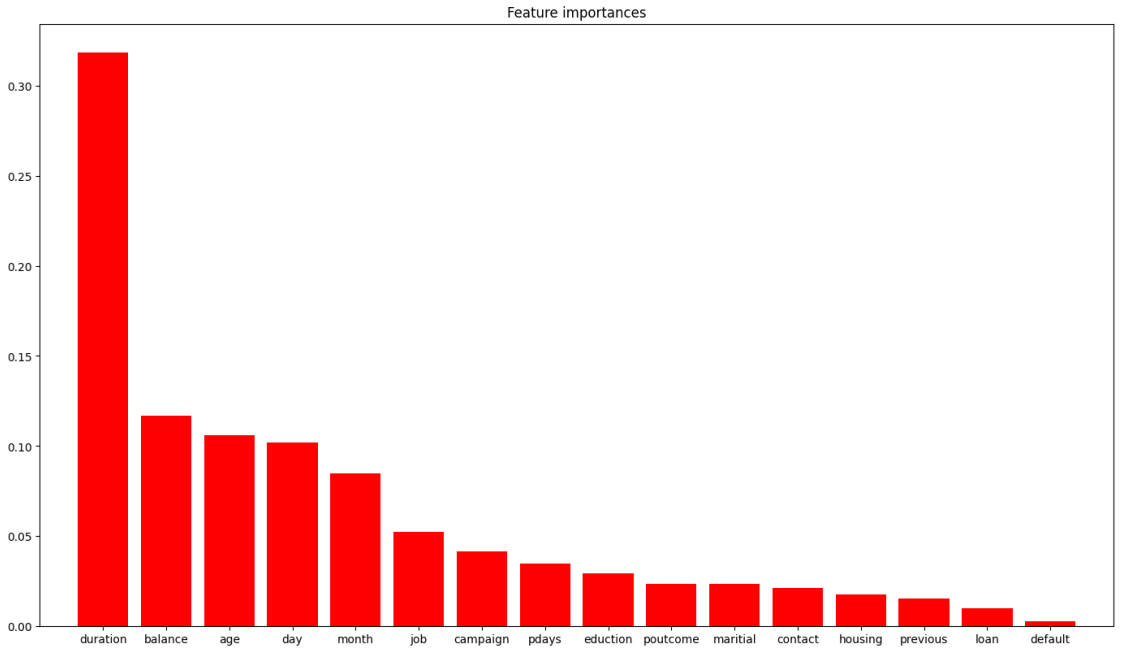

Figure 3. Feature importance of random forest.

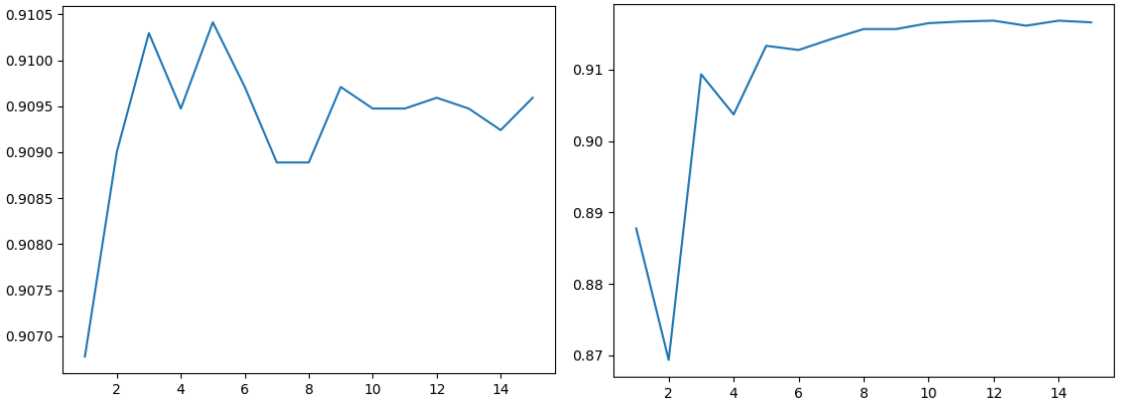

Figure 4. Result of random forest.

DOI: 10.54254/2755-2721/54/20241411

According to the result, when the max depth is default value and the numbers of decision trees is one, the accuracy rate is the same as decision tree with no limit max depth. As the number of trees rising, the accuracy rate is also rise as well. When the max depth is six, the accuracy rate rises firstly and tends to a stable value. In general, the accuracy rate increases with the number of the decision trees number. However, if the max depth is definite, only increase the number of trees is useless to improve the accuracy rate.

3.3. SVM

The third classification model is supported vector machine model. In this model, one can also change its parameters to get the different result. In this paper changed the penalty parameter to get the different result. When the penalty parameter rising, it means it will increase the penalties for misclassification. In this case, the accuracy rate will improve because the model tends to train the whole results correctly. The results of different penalty parameter are illustrated in Fig. 5. According to the picture, the accuracy rate is truly increasing with the penalty parameter increasing. However, the growth rate is not obvious, the accuracy rate just increases about 0.3 percentage. The reason is because the number of different labels in data set which used in this test have a big difference in quantity and the support vector machine model maybe classify the two labels in one easily.

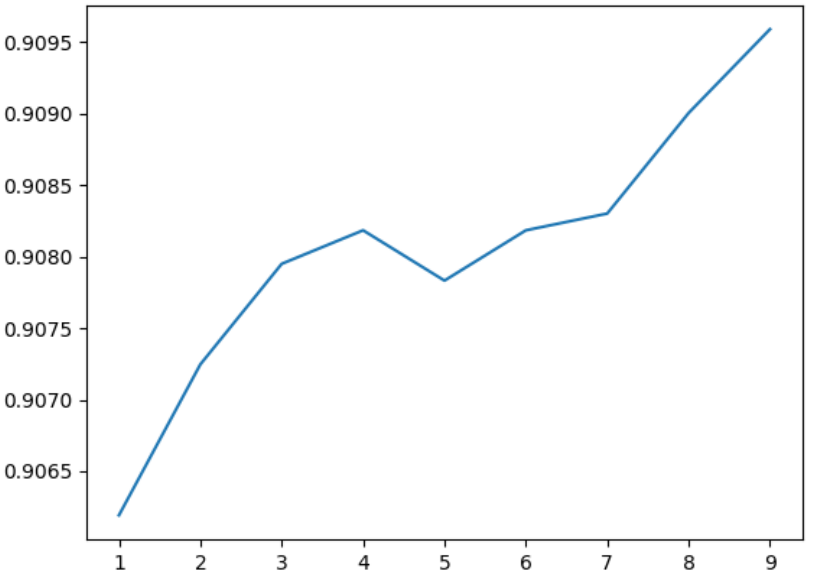

Figure 5. Result of SVM.

3.4. Summary of the models

This section discusses three models and get their results which showed by accuracy rate. Although three models are classification models, their results have a little different. Because random forest model is based on decision tree model, when the parameter of decision tree model and the trees in random forest model are the same, the result of random forest will be better than the result of decision tree. However, if the parameters in random forest model are not better than decision tree model, the result of random forest maybe worse than decision tree model, though there are many decision trees in random forest model. In support vector machine model, the principle is unlike the other two models so that the result will be more susceptible to a larger number of labels. The result will be good for the data set which the number of different labels is about equal.

4. Limitations and prospects

This study uses three classification model to classify data set. However, there are still some limitations. Firstly, there are still a lot of parameters do not change in this paper. There are too many parameters of

DOI: 10.54254/2755-2721/54/20241411

these models so that this paper chooses the most useful and classical parameter to change instead of changing all of them. Secondly, in decision tree model, it is a weak classifier and its information gain is biased towards the features which are more numerical so that it will be influenced by some features and classify the data set based on them. Thirdly, although random forest model can prevent overfitting, both decision tree model and random forest model are easy to overfit when there has loud noise. Because the data set in this paper do not have any noise so that the results are still good. Also, random forest model can't solve the continuous data [11]. Finally, support vector machine costs a lot of machine memory and computing time so that if the data set is too large, this method is not good.

In the future should improve the models. When people face different problem, they need a targeted approach so that create some new models which based on basic model is very important. For example, decision tree model can't estimate the forecast result so that when it is used in self-training, the result will be bad. In this case, use a method of node split can help decision tree model to solve the problem [12]. In random forest model, because it cannot solve the continuous data so that add additional spatial discretization for data set can discretize the data [11]. Besides these, support vector machine model can also be improved. There is a new method based on support vector machine and it has high accuracy rate than traditional model [13].

5. Conclusion

To sum up, this study chooses the data set which from Kaggle about bank customers whether will subscribe a term deposit. After determining the data set, this paper use three classification models to classify customer and predict them whether will subscribe a deposit term. The result shows in chapter three that can see their feature importance and accuracy rate. It also shows their advantages and disadvantage. After getting the results, this paper also has some limitations and gives some advice to improve the models. In a word, this paper use three different classification models to accomplished the goal and hope to do better in the future.

References

[1] Alizadeh M, Zadeh D S, Moshiri B and Montazeri A 2023 IEEE Access vol 11 pp 29759-29768.
[2] Dawood E A E, Elfakhrany E and Maghraby F A 2019 IEEE Access vol 7 pp 109320-109327.
[3] Girard C I 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada pp 5494-5497.
[4] Fu Y, Yin Z, Su M, Wu Y and Liu G 2020 IEEE Access vol 8 pp 138046-138057.
[5] Ma Y, Zhang Q, Li D and Tian Y 2019 IEEE Access vol 7 pp 70319-70331.
[6] Liu Z, Wen T, Sun W and Zhang Q 2020 IEEE Access vol 8 pp 128337-128348.
[7] Hsu C H 2021 IEEE Access vol 9 pp 41334-41343.
[8] Cañete-Sifuentes L, Monroy R and M A Medina-Pérez M A 2019 IEEE Access vol 9 pp 110451-110479.
[9] Zhu X, Xiong J and Liang Q 2018 IEEE Access vol 6 pp 33583-33588.
[10] Lučin I, Čarija Z, Družeta S and Lučin B 2021 IEEE Access vol 9 pp 155113-155122.
[11] Wang W, Keen J, Bank J, Giraldez J and Montano-Martinez K 2023 IEEE Open Access Journal of Power and Energy vol 10 pp 327-334.
[12] Kammoun A and AlouiniFellow M S 2021 IEEE Open Journal of Signal Processing vol 2 pp 99-118.
[13] Behnamian A, Millard K, Banks S N, White L, Richardson M and Pasher J 2017 IEEE Geoscience and Remote Sensing Letters vol 14(11) pp 1988-1992.