The study of methods of data mining in the field of bank marketing

Alissa Danko

Satbayev University, Kazakhstan, Almaty, Satpayev str., 22A

**Corresponding author's e-mail: Danko.alissa@gmail.com*

Abstract

The aim of the research is to study methods of data mining on the example of data set about results of bank marketing campaign. The data presented at Kaggle.com were used as the initial materials for the analysis. In the course of the study of the initial data set, analysis and correction were made by filling in empty values to avoid a large number of errors, and charts were built for clarity. During the study the following methods of data mining were used: Decision Tree, Random Forest, k Nearest Neighbor, and Adaptive Boosting algorithm was used to increase accuracy. The analyzed data set was studied in order to determine the probability of receiving a positive answer from a client to a term deposit offer. According to the results of the conducted analysis, the preferential dependence of the result on the duration of telephone calls with the client, as well as on his financial possibilities at the time of the research was revealed. The sphere of activity, level of education and age have less influence on the probability of positive answer. The presented data can serve for the marketing department at the bank as a tool for selecting a target audience for campaigns offering term deposits.

Keywords: Data mining, Decision Tree, Random Forest, K Nearest Neighbor, Adaptive Boosting

1 Introduction

Nowadays, the banking sector has a huge number of representatives, which certainly requires development of competitiveness. Today it is not enough to have bright and loud advertising headers, it needs for tools and methods to turn the focus on services. Defining the needs of customers in one or another period of time, the identification of the target segment, forecasting and coordination of possible ways of development - the solution of these issues can be achieved by using data mining.

The purpose of the study is to study methods of data mining. The main task is to identify regularities and main dependencies affecting the result of the advertising campaign of the bank. The work is useful at the beginning of study by most frequently used methods and algorithms. The data set was presented by Kaggle.com. Anaconda 3, Jupiter and Python programming language were used as application software for the study.

2 Theoretical background

Data mining is a set of algorithms and methods to extract useful and practically applicable information and knowledge from data sets. It is based on data preparation - processing, cleaning, addition based on existing data, transformation, as well as the use of statistics, optimization, pattern recognition, visual representation, etc.

The tasks of data mining can be conventionally divided into two types - descriptive and predictive. The tasks of descriptive type include grouping, cluster analysis, sequence search, etc. The predictive tasks include classification, regression analysis, etc.

As noted earlier, data mining is a set of algorithms, including learning algorithms. They are divided into supervised learning and unsupervised learning. The main difference between these two types of algorithms is determined by the need to select input and output vectors. For supervised learning the model is built using predefined parameters. In turn, in unsupervised learning the value of parameters is selected automatically during the process of detecting internal dependencies and regularities between the data.

To solve the problem by applying data mining, you should follow the next steps:

1. task definition;
2. data collection;
3. data preparation;
4. selection of algorithms for data analysis;
5. definition of parameters and training algorithms;
6. model training;
7. analysis of certain regularities.

The subject area of the research is the bank's marketing campaign aimed at attracting customers to sign a term deposit agreement. A marketing campaign is a series of activities carried out by a company for informing, reminding or persuading its target audience about its product or service [1].

A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution. Term deposit investments usually carry short-term maturities ranging from one month to a few years and will have varying levels of required minimum deposits [2]. The term deposit agreement provides for the return of the deposit at the end of the term established in the agreement. However, the client has the right to take out the money

before the end of the contract term, considering the prescribed penalties.

In order to study the data set the following methods of data mining were used: Decision Tree, Random Forest, k Nearest Neighbor. They represent supervised learning, i.e. they require the definition of input and output parameters. Below is a summary of the methods.

Decision Tree is a classifier built on the basis of the deciding rules of the "if, then" type, ordered in a hierarchical tree structure [3]. The idea of this method is to form queries that are directed to the data. Decision Tree forms nodes, containing samples from the original data set, belonging to the same class. Its task is to detect parameters with similar values.

Random Forest is a composition of a set of Decision Trees, which makes it possible to increase accuracy in comparing to a single tree [4]. The prediction is the result of aggregation of the responses of the set of trees. Random Forest works on the basis of two concepts - sampling is random, random sets of parameters are selected when divided into nodes. The trees are trained independently of each other. The result is a class for which most trees have voted.

The k Nearest Neighbor method (k-NN) is a metric algorithm for automatic classification of objects or regression. It is used to solve the classification problem. It classifies objects into a class that owns most of k closest neighbors in the multidimensional attribute space [5]. It is one of the simplest algorithms for teaching classification models. This algorithm is applied to large multidimensional samples.

In addition to the classification algorithms described earlier, the AdaBoost algorithm (hidden from adaptive boosting) was used. This is an algorithm that builds a composition from the basic training algorithms to increase their effectiveness during training [6]. Its function is to have each next classifier built on objects that are poorly classified by previous classifiers.

For practical realization of the research Python programming language was chosen. Python is a high-level general-purpose programming language focused on improving developer's performance and code reading [7]. The main advantages of the language are that it is quite simple, easy to understand and learn, its standard library includes a large number of useful functions.

3 Practical realization

3.1 INPUT DATA

For application of data mining methods, it is necessary to follow algorithm of actions. At the first stage the purpose of studying is set. The main question is whether the client will sign a term deposit agreement during the bank's marketing campaign. The offers were made through telephone calls with the bank's clients.

The second stage - data collection, in this case - uploading of the data provided on Kaggle.com.

The data set contains the main characteristics describing the customers and information on interaction with them, which is shown in Figure 1. The data set contains 31647 records, on the basis of which it is possible to obtain results with high accuracy.

```
bank.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31647 entries, 0 to 31646
Data columns (total 18 columns):
ID            31647 non-null int64
age           31647 non-null int64
job           31647 non-null object
marital       31647 non-null object
education     31647 non-null object
default       31647 non-null object
balance       31647 non-null int64
housing       31647 non-null object
loan          31647 non-null object
contact       31647 non-null object
day           31647 non-null int64
month         31647 non-null object
duration      31647 non-null int64
campaign      31647 non-null int64
pdays         31647 non-null int64
previous      31647 non-null int64
poutcome      31647 non-null object
subscribed    31647 non-null object
dtypes: int64(8), object(10)
memory usage: 4.3+ MB
```

Figure 1 Data set

Table 1 describes the values in the data array.

Table 1 Data set information

Name	Info
ID	Unique customer ID
age	Customer's age
job	Type of activity
marital	Marital status
education	Education
default	Are there any debts, default on the loan?
balance	Balance of funds
housing	Housing loan
loan	Consumer loan
contact	Type of communication with the customer
day	Monthday of contact
month	Month of contact
duration	Contact duration in seconds
campaign	Number of customer contacts in this company
pdays	Number of days since last contact
previous	Number of customer contacts in previous company
poutcome	Was the previous company successful?
subscribed	Has the customer signed a term deposit agreement?

3.2 THE MODIFICATION OF DATA

There are no null values in the present dataset, but when each column is analyzed individually, ‘unknown’ values were identified. There are four columns where ‘unknown’ values are found:

1. job (206 out of 31647);
2. education (1314 out of 31647);
3. contact (9177 out of 31647);
4. poutcome (25929 out of 31647).

So, by finding the mean and median values a part of null values for the ‘education’ column was filled in. Columns ‘poutcome’ and ‘contact’ have too many empty values, they should be removed. The column ‘job’ has a small number of unknown values, which can not essentially influence the research results.

3.3 THE RESEARCH OF DATA SET

For visual representation of dependencies, correlation, the heatmap chart was used - Figure 2.

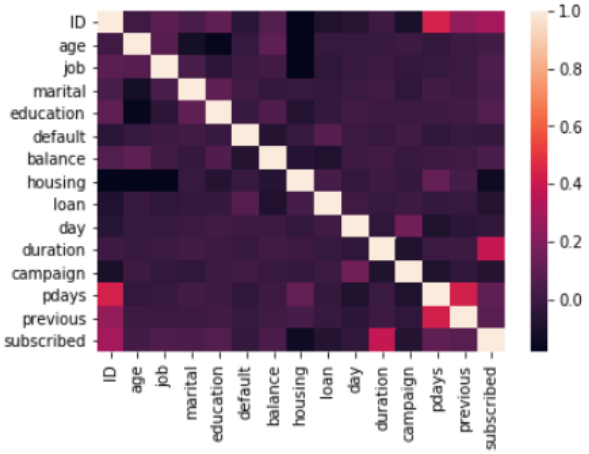

Figure 2 Heatmap

According to the chart, the parameter 'subscribed' depends most of all on the duration of contact with the client. Education, job, marital status, balance and age have less influence.

Based on the assumptions about the parameters that influence the client's decision and the correlation graph, charts have been built. In Figure 3, you can see that the number of signed contracts is greater if the telephone calls with the client was longer. This can be due to the personal qualities and communicative abilities of the manager, his client focus.

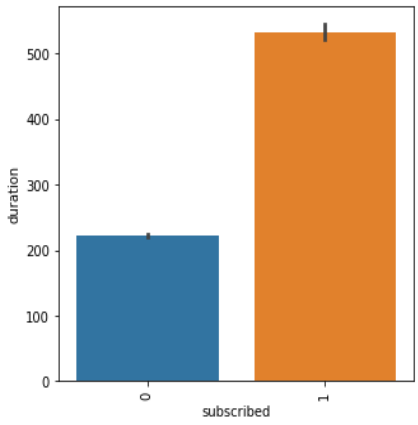

Figure 3 The dependency 'subscribed – duration'

The next parameter that affects the decision of the client is his current balance of the bank account - the more he is, the higher the probability of agreement. The chart of the dependency of the parameters is shown in Figure 4.

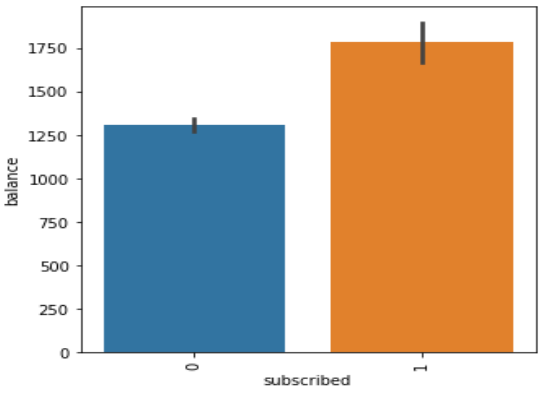

Figure 4 The dependency of 'subscribed – balance'

As the chart in Figure 5 shows, clients with higher education are more likely to sign a term deposit agreement.

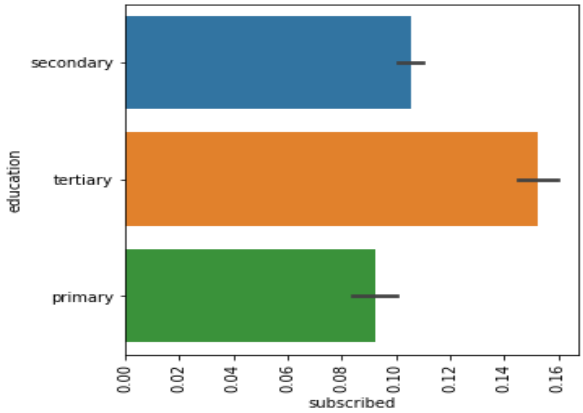

Figure 5 The dependency of 'subscribed – education'

Figure 6 shows a graph of the relationship between the age of the client and his decision in response to the bank's offer. This graph shows a pattern that in most cases a term deposit agreement is signed by young people between 18 and 25 years old as well as older people over 60.

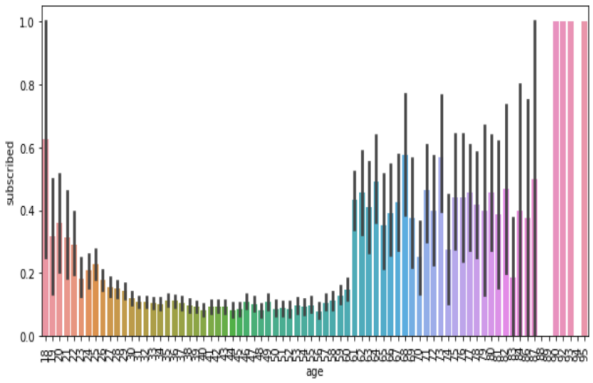

Figure 6 The dependency 'subscribed – age'

Figure 7 shows the relationship to the previous chart. It shows that students (in most cases young people under 25) and retired people over 60 are more likely to invest in a term deposit.

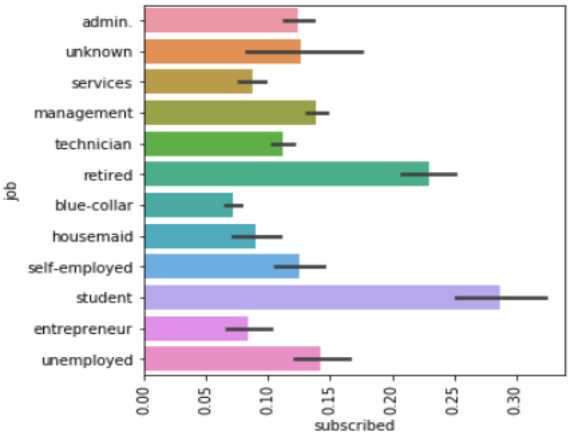

Figure 7 The dependency 'subscribed – job'

The relationship between marital status and signing a deposit agreement is shown in Figure 8. Single clients are more likely to use the term deposit.

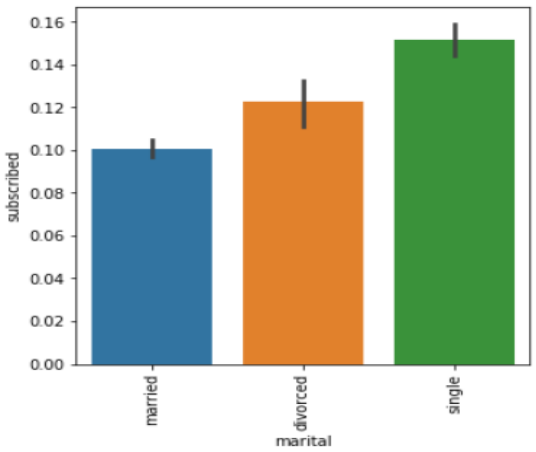

Figure 8 The dependency 'Subscribed – marital'

The next stage in the research of the data set was the application of supervised learning algorithms. Taking into account the identified dependencies and relationships, the input and output parameters for using the algorithms were selected. The dependent parameter is "subscribed". The parameters that have an influence were selected as follows: 'age', 'job', 'duration', 'education', 'balance'. With the help of these algorithms, the accuracy of the predicted result - positive answer to the bank offer - is determined.

When using Decision Tree, the following accuracy indicator was obtained - Figure 9.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn import metrics
clf = DecisionTreeClassifier()
clf = clf.fit(X_train,y_train)
y_pred = clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.833070036861506
```

Figure 9 Decision Tree

Figure 10 shows the result of determining accuracy using Random Forest.

```
rdmf = RandomForestClassifier(n_estimators=20,
                              criterion='entropy')
rdmf.fit(X_train, y_train)
rdmf_score = rdmf.score(X_test, y_test)
rdmf_score_tr = rdmf.score(X_train, y_train)
print(rdmf_score)
```

```
0.8769878883622959
```

Figure 10 Random Forest

Using the k-Nearest Neighbors algorithm, the following accuracy was determined, as shown in Figure 11.

```
knn = KNeighborsClassifier(p=2,
                           n_neighbors=10)
knn.fit(X_train, y_train)
knn_score = knn.score(X_test,
                      y_test)
print(knn_score)
```

```
0.8814112690889943
```

Figure 11 k-Nearest Neighbors algorithm

The Adaptive Boosting algorithm was used to increase the accuracy of calculations. When Decision Tree was applied to the following result was received, which is shown in Figure 12.

```
ada = AdaBoostClassifier(DecisionTreeClassifier(max_depth=10),
                         random_state=42).fit(X_train, y_train)

print("Decision tree accuracy: %.2f" % clf.score(X_test, y_test))
print("AdaBoost accuracy: %.2f" % ada.score(X_test, y_test))
```

```
Decision tree accuracy: 0.83
AdaBoost accuracy: 0.87
```

Figure 12 Adaptive Boosting

4 Conclusions

As the result of the research, information about the factors that influence customers decisions was obtained. A correctly selected target segment increases the effectiveness of the campaign and increases the profit from it. Term deposits are more likely to be of interest to customers with the following parameters ordered by the correlation value:

1. Age groups under 25 and over 60 years of age;
2. Has a tertiary education;
3. Has a large bank account balance;
4. The marital status - single.

Besides the considered personal characteristics of customers, the result of contact with customers is significantly influenced by the duration of the contact. This may be the result of certain communication abilities of the bank manager, as well as his focus on the needs of the client. This factor can be used as a recommendation for development and training of managers in the field of relationships with clients.

References

[1] *Marketing campaign definition* **E-source:** https://www.mbaskool.com/business-concepts/marketing-and-strategy-terms/14292-marketing-campaign.html 15 March 2020
[2] *Term deposit definition* **E-source:** https://www.investopedia.com/terms/t/termdeposit.asp 15 March 2020
[3] *Decision trees* **E-source:** https://wiki.loginom.ru/articles/decision-trees.html 21 March 2020
[4] *The implement and parsing the random forest algorithm in Python* **E-source:** https://tproger.ru/translations/python-random-forest-implementation/ 23 March 2020
[5] *Method K-nearest neighbor* ***E-source:*** https://wiki.loginom.ru/articles/k-nearest-neighbor.html 23 March 2020
[6] *AdaBoost algorithm* **E-source:** http://www.machinelearning.ru/wiki/index.php?title=AdaBoost 28 March 2020
[7] *Python programming language* **E-source:** https://web-creator.ru/articles/python 28 March 2020