

PRACTICAL DATA SCIENCE

Assignment 2: Data Modelling and Presentation

Lecturer: Dr. Yongli Ren

Student : Tu Nguyen - [s3608114@rmit.edu.vn](mailto:s3608114@rmit.edu.vn)

Bui Anh Tu - [s3634162@rmit.edu.vn](mailto:s3634162@rmit.edu.vn)

Date: 9/12/2019

TABLE OF CONTENT

[**Abstract**](#_n2cudyuut6jv) **3**

[**Introduction**](#_2qz54r9feprc) **3**

[**Methodology**](#_jpnz6mrpl1hr) **3**

[Data](#_su8207ni18kk) 3

[Data Analysis](#_7uz0rp2tiwq7) 3

[Data Visualization](#_7idbywy92ue2) 3

[Machine Learning](#_scikbcfy6was) 4

[k-Nearest Neighbors Classifier](#_v53kpetu041m) 4

[Decision Tree Classifier](#_4h7ct4e51yv7) 4

[**Results**](#_h6wx3e6q2wqr) **4**

[Data Pre-processing](#_h9bthvyqj41a) 4

[Data modeling](#_hcjede759btx) 8

[Feature Scaling](#_w4zx7j8bff8n) 8

[kNN Classifier and Decision Tree](#_vtnls83tz5hf) 9

[Overall](#_aejskfygxtv) 9

[**Discussion**](#_c99uycmc0p90) **9**

[**Conclusion**](#_4obx7s28bilo) **10**

[**References**](#_bzvw9m1lc5wl) **10**

[**Appendix**](#_2798coabk2ud) **10**

# **Abstract**

Bank marketing campaign is researching in Portuguese banking institutions. This campaign was collected by a phone call to the client and convince them to open a term deposit. The phone call would be recorded by the customer’s acceptance or not subscribe to the deposit. After the recording, we will use these statistics to create the dataset. The assignment includes data preparation, exploration, modeling and compares the result of two modelings. Overall, we use two predictive classification models included Decision Tree, K- nearest neighbor. Considering this, we can predict the target group for the next campaign, how to improve the percentage of clients to open the term deposit.

# **Introduction**

The term deposit is one of the contributions to bank development. The bank can use it to invest and pay for the customer a part of the income from the investment. But a lot of people just want to invest in the housing or other investments instead of opening term deposit in the bank. Moreover, the term deposit is also new in Portuguese so not all the clients trusted and subscribe to it. In this report, we have recorded the statistics about if the client will subscribe a term deposit (variable y) based on their age, job, marital status, etc. In addition, we also predict whether the customer does not subscribe to the deposit and recommend to convenience the client.

# **Methodology**

## Data

This dataset is recorded the client’s opinion about the term deposit from May 2008 to November 2010 in Portuguese. It is published on <http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>. The dataset is included the 17 columns and 45211 rows. It contained the opinion based on Job, age, marital status, education, the credit in default, average yearly balance (in euros), housing loan or personal loan, contact communication type, last contact day of the month,last contact month of year, last contact duration,number of contacts performed during this campaign and for this client, number of days that passed by after the client was last contacted from a previous campaign, number of contacts performed before this campaign and for this client, outcome of the previous marketing campaign and the customer’s opinion to subscribe the term deposit or not ( Pie Chart 1 ).

## Data Analysis

In the data analysis, we imported python packages pandas (https://pandas.pydata.org/) and Numpy library (<http://www.numpy.org/>). The previous dataset was checked and corrected for the errors, null values, missing values. We also describe the count, mean, std, min, mean, max to have a better point of view in general dataset per column.

## Data Visualization

In this case, we plot package Matplotlib (<https://matplotlib.org/>), Seaborn (<https://seaborn.pydata.org/>) for data visualization.

## Machine Learning

### k-Nearest Neighbors Classifier

K Nearest Neighbors algorithm is one of the most popular supervised machine learning algorithms used for classification and regression. The model accuracy score indicates the ability of kNN to correctly predict the target label. In this data, the target is to predict if the client will sign for a deposit in the next visit.

However, the high number of features can result in the problems of overfitting. Therefore, to avoid that problem, Hill Climbing method is applied to cut off the unnecessary features.

To have a better understanding of the performance of the model, it is recommended to observe the confusion matrix and performance scores (precision, recall, f-1 score, classification error rate) for the model with different sets of parameters. In this study, the comparison is between the default parameter and the optimized parameter.

The tool used to select the best-suited parameter is GridSearchCV. The procedure is that GridSearchCV will perform a test with different sets of parameters for the kNN Classifier and choose the best parameters in which the model gives the best accuracy score.

### Decision Tree Classifier

Like kNN Classifier, Decision Tree Classifier is one of the top-used supervised algorithms. The algorithms aim to split the dataset into local partitions and detect the homogeneity of the set. The default method to perform that task is using the Gini index. The Gini index is calculated by calculating the sum of squares of probability for success and failure (+ ).

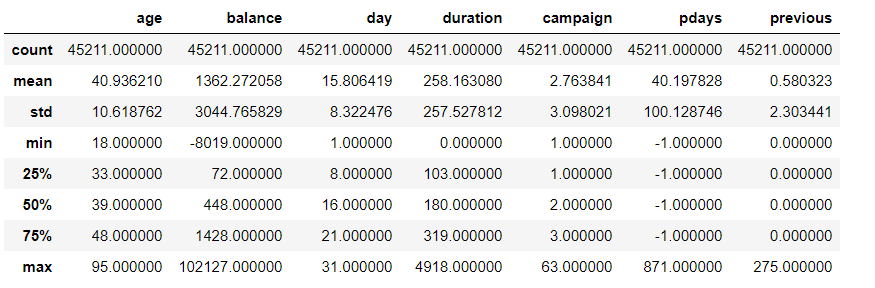
For this Classifier, an optimized set of parameters also need to be determined to prevent overfitting of the dataset. Similar to kNN, GridSearchCV is used to find parameters in which the algorithm results in the best performance score.

# **Results**

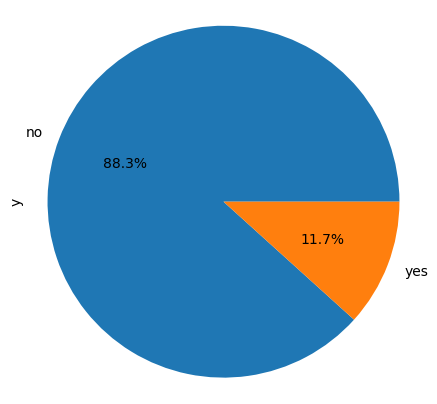
## Data Pre-processing

Data collected the 17 columns and 45211 rows included: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y ( term deposit ).

The statistic for the numerical column in the dataset shows the count, mean, std, min, max, 25%, 50%, 75%.



With the label column, I use Pie charts to plot because we can easily see the percentage and the difference. I consider the main column is ‘y’ to observe the percentage of people who say ‘yes’ to a term deposit ( Pie chart 1 ). From 2008 to 2010, the deposit is still new so the percentage of clients accept to open deposit is low.



#### 

**Figure**: The distribution of client who say ‘yes’ and ‘no’ in the campaign

After exploring the general dataset, I find the relationship between columns. I compare the marital column, education column, month column, default column, housing column, and loan column to the y (deposit) column.

In marital with y, we see that the highest percentage of people who say ‘yes’ and also ‘no’ to the deposit is ‘married’. But most people less likely to have the term deposit. In the next campaign, we should focus on the two other ‘single’ and especially ‘divorced’ because they are also potential customers. They do not marry so they do not spend money on the family’s life, they can have money to open the term deposit and maybe they interest in investing.

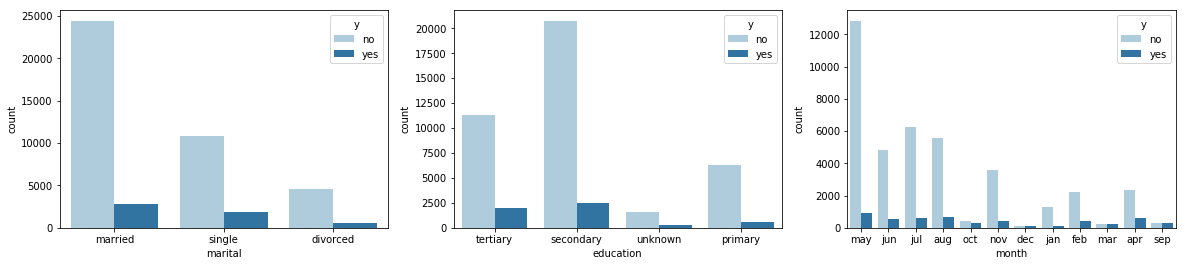
With education and y, we also see the same things as in the first chart, the part of customer ‘secondary’ has the highest and lowest percentage to open the term deposit. In the next campaign, I support that we need to focus on people in ‘tertiary’ because they are the second-highest percentage in the graph and they

In the third chart, we should target the customer that contacted in the month with lowest percentage accept the term deposit. In the next campaign, we recommend to focus on the part of customers who are less likely to subscribe the the term deposit contacted during the months of March, September, October and December because there are the months with the lowest percentage to have term deposit.

In the ‘default’ with ‘y column, the number of people who have credit, did not open the term deposit. Because of this, we should focus on people who have less credit in default.

The graph between ‘y’ and ‘housing’, we can see that people who do not have house say ‘yes’ to the term deposit more than people who have a house. May be they do not have enough money now to buy a house, so the term deposit is one of the ways for them to have more money for the house. The next campaign should list them in a target.

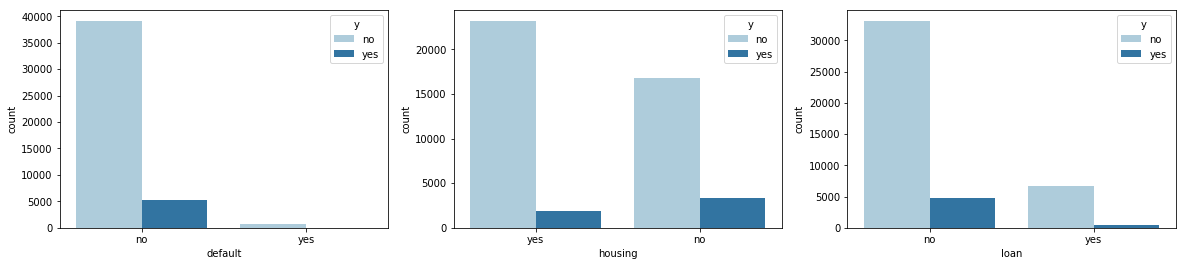
The last graph is ‘y’ and ‘loan’, most of the people who have loan, say ‘no’ to the term deposit. Loan is also the burden for them so it is hard for them to have the term deposit so we should focus on people who do not have loan.



**Graph 1**: The relationship between marital status and the term deposit

**Graph 2**: The relationship between the education and the term deposit

**Graph 3**: The relationship between the month of the last contact and the term deposit

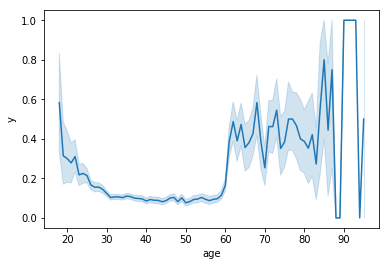


**Graph 1**: The relationship between credit in default and the term deposit

**Graph 2**: The relationship between housing and the term deposit

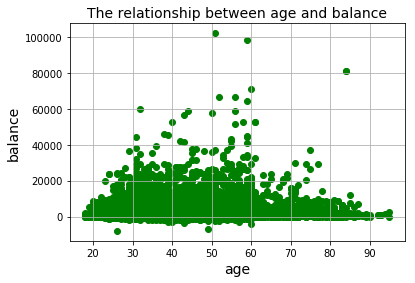
**Graph 3**: The relationship between loan with the term deposit

In the statistics below, we can see the relationship between ‘y’ and ‘age’ column. The highest percentage of clients who say ‘yes’ to the term deposit is older people from 60 years old. We should focus more on the people who have the lowest percentage to say ‘yes’ like people from 50 to 60 years old. Because in this age, the graph shows that this group tended to increase to open the term deposit.

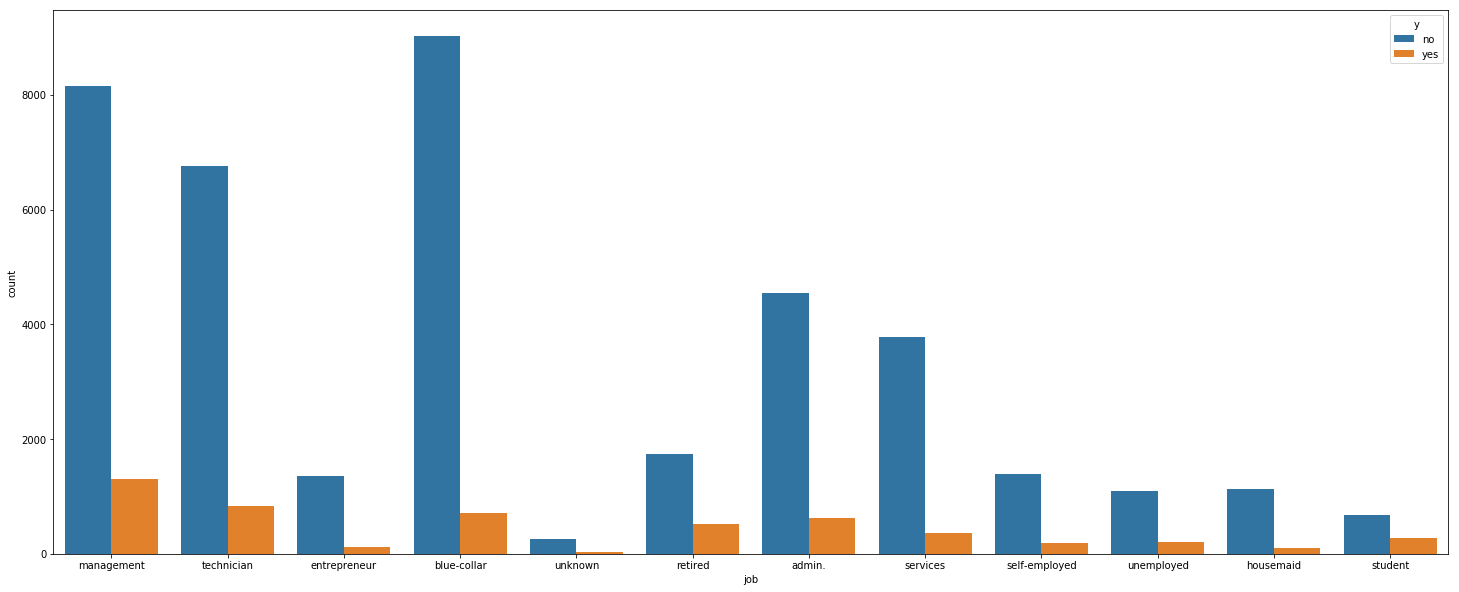


**Figure**: The relationship between age of client and the term deposit

In the relationship between age and balance, people who are from 50 to 60 have the highest balance. This is the target group we should work for the next campaign. People who in the high balance were more likely to open a term deposit.

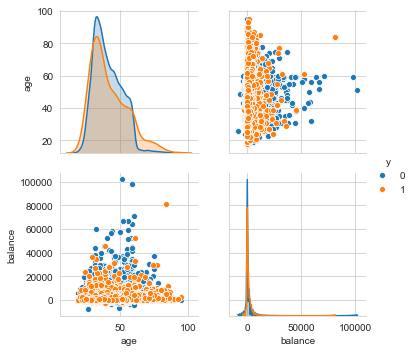


Customers with 'unknown' jobs are less likely to subscribe for term deposit. Moreover, customers in 'management' and 'services' likely to have deposit. We should focus on customer who have the fine job such as ‘management’, ‘technician’, or ‘blue-collar’. Moreover, ‘retire’ is also the target because at this age, they want to save money without working.



**Figure:** The statistic about the ‘job’ and the term deposit

The graph below is the relationship between ‘age’ and ‘balance’ and ‘y’ (term deposit). People who are older and have the high balance to subscribe the term deposit.



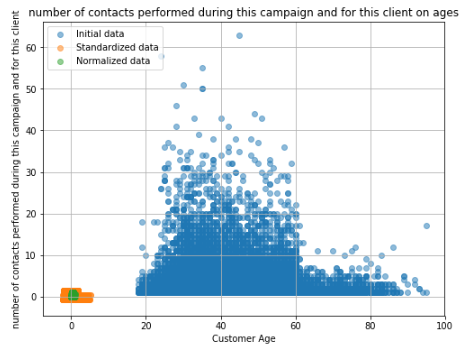
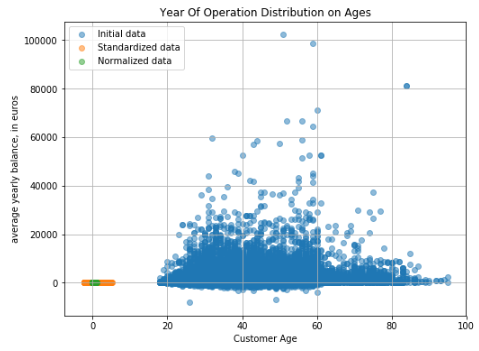
**Figure**: The relationship between ‘age’, ‘balance’ and ‘y’ (term deposit)

## Data modeling

### Feature Scaling

While working with data set in terms of Machine Learning, rescaling data is a good approach. Since the range of values of the untrained, raw data are not in the same range, the variance will be increasing. High variance leads the data model to create more random noise while performing prediction, the model then becomes less stable that will result in a low prediction accuracy rate.

By applying Feature Scaling, the range of the features will be centered around zero so that the data model can achieve a unit variance. This practice helps reduce computational expenses and improve the performance of the model. The most popular Feature Scaling practices are Standardization and Normalization.



**Figure:** the distribution of clients before (blue dots) and after standardization (red dots) or normalization (yellow dots)

As observed in the figure above, the distribution of dots is centralized around zero, it means the variance decreases. It is also clear that the shape of the plot remains similar indicating that the feature scaling does not change the shape of the dataset.

### kNN Classifier and Decision Tree

The two models of Classification are performed in different test\_size (20%, 40%, and 50%). For each case of test\_size, the accuracy score of the models using hyper-tunned parameters is optimized to at least 2%, accept for test\_size = 40% of kNN, the score is slightly optimized.(**Appendix Table 1)**

The results in **Appendix Table 2** also indicates that while using optimized parameters, the value of accuracy scores vary really slight when the test\_size change. Also, precision, f-1 score and recall remain almost the same.

### Overall

Overall, both kNN and Decision Tree Classifiers both obtain the best performance scores using optimized parameters. However, test\_size where kNN performs the best is 50%. Whereas, for Decision Tree, it is 20%. (**Appendix Table 1)**

# **Discussion**

The kNN Classifier makes predictions based on the nearest numbers. So it is natural that the higher n\_neighbors parameter, the better the performance score is obtained. That is the reason why the optimized parameters for kNN always have higher n\_neighbors value than the default ones (**Appendix**).

The Decision Tree Classifier makes predictions by generating and giving a visualization of the tree which is really helpful to observe the performance of the algorithm. One more benefit of this learning model is that it has more parameters to be tuned. It means that there are more options to look for the optimized parameter for this algorithm.

Overall, both kNN Classifier and Decision Tree Classifier have both proven their high accuracy in classifying the clients based on the feature attributes. The performance scores of kNN and Decision Tree are both over 89% which is a reasonable score. The kNN acquire the best score (89.39%) at test\_size = 40% and the Decision Tree performs the best at test\_size = 20% with higher accuracy(89.65%). However, the difference is not important (89.39 vs. 89.65), the results show that both these two classification models have good ability to make predictions with this dataset.

# **Conclusion**

The goal that we want to towards to find for future strategies in order to target future marketing campaigns for the bank. Moreover, we use Decision tree, K-nearest neighbor to predict and have the specific data. According to the study, we can get the target client is older people, has the highest balance, focus on which months, people who are divorced and single. Moreover, people who have credit is also need to be listed in the target group. Potential client is also who less likely to loan and housing loan. Based on these results, we can have the better bank marketing campaign and have more clients who will subscribe the term deposit.

# **References**

Archive.ics.uci.edu. (2019). *UCI Machine Learning Repository: Bank Marketing Data Set*. [online] Available at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing [Accessed 13 Sep. 2019].

Www2.1010data.com. (2019). *Bank Marketing Data Set*. [online] Available at: https://www2.1010data.com/documentationcenter/beta/Tutorials/MachineLearningExamples/BankMarketingDataSet.html [Accessed 13 Sep. 2019].

Talend Real-Time Open Source Data Integration Software. (2019). *What is Data Preparation? (+ How to Make It Easier) - Talend*. [online] Available at: https://www.talend.com/resources/what-is-data-preparation/ [Accessed 13 Sep. 2019].

Khandelwal, R. (2019). *Bias and Variance in Machine Learning*. [online] Medium. Available at: https://medium.com/datadriveninvestor/bias-and-variance-in-machine-learning-51fdd38d1f86 [Accessed 13 Sep. 2019].

# **Appendix**

**Table 1:** The results of the accuracy scores of kNN model and Decision Tree model using default parameters and optimized parameters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | kNN | | | Decision Tree | | |
|  | 50% | 40% | 20% | 50% | 40% | 20% |
| default param | 85.15% | 89.28% | 87.44% | 87.33% | 86.88% | 87.06% |
| optimized param | 87.65% | 89.39% | 89.35% | 89.33% | 89.07% | 89.65% |

**Table 2:** The performances scores of kNN Classifier and Decision Tree using optimized parameters

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | test\_size | precision | recall | f-1 score | accuracy score | error rate |
| kNN | 50% | 0.85 | 0.88 | 0.86 | 87.65% | 12.35% |
| 40% | 0.87 | 0.89 | 0.86 | 89.39% | 10.61% |
| 20% | 0.87 | 0.89 | 0.86 | 89.35% | 10.65% |
| Decision Tree | 50% | 0.89 | 0.89 | 0.89 | 89.33% | 10.67% |
| 40% | 0.89 | 0.89 | 0.89 | 89.07% | 10.93% |
| 20% | 0.88 | 0.9 | 0.89 | 89.65% | 10.35% |