Data Modelling to Predict Customer Subscription

Practical Data Science (COSC2670)

Assignment 2: Data Modelling and Presentation

Anooja Mathew
S3767921
Masters in Data Science (093313B)
RMIT
Email: s3767921@student.rmit.edu.au

Jewel James
S3763905
Masters in Data Science (093313B)
RMIT
Email: s3763905@student.rmit.edu.au

Date of Report: 2 June 2019

Table of Contents

1. Executive Summary.....1

2. Introduction.....1

3. Methodology.....1

1. Data Collection, Cleaning 3
2. Data Exploration.....3
3. Data Modelling 6

1. KNN.....6

2. Decision.....7

4. Results.....**7**

5. Discussion.....8

6. Conclusion.....10

7. References.....11

Executive summary:

A Portuguese banking institute conducted marketing campaigns through phone calls. The data collected through these campaigns were used to determine if the customer will subscribe to a term deposit. This kind of prediction will enable the bank in accessing their profit forecast and analyse if their marketing strategies are effective. The data collected consist of details of customers such as age, education, job, marital status, credit status etc. Based on these details, the analytical model was trained to predict if the customer would agree to start a term deposit in the bank. Analysing these results of the model, the bank can provide budget and profit calculations and also contact the expected customers in regard to the deposit plans. The model obtained after training the available dataset predicts the customers who will not subscribe 90% accurately. However, it does not predict the customers who can subscribe to the deposit. The reason for this false prediction is the less support data related to customers who have subscribed. A more balanced dataset would have solved this issue and provided more appropriate results.

Introduction:

With the increasing number of banks and financial institutions each day, drawing clients towards their products like term deposit play a crucial role in withholding the bank's position in the market and achieving their yearly targets. Thus, banks rely highly on software which can help them identify the potential customers and thus strategize their resources in obtaining them. This report will discuss the models which can analyse the customer details in the best possible way and provide a list of suitable candidates who can be further contacted with the term deposit plans.

Methodology:

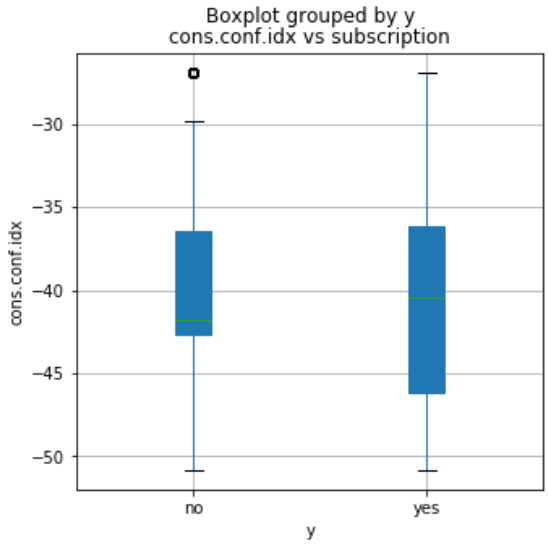
The data was collected through marketing campaigns conducted via phone calls to a wide range of population. The dataset obtained out of this campaign contained 41188 observations with 20 attribu tes containing the personal, social and employment details that can be useful in predicting if the per son would enrol for the term deposit. Below are the variables and their description.

Variable	Description	Expected Values
Age		
Job	Type of job	'admin.','blue-collar','entrepreneur','housema id','management','retired','self-employed','ser vices','student','technician','unemployed','unk nown'
Marital	Marital status	'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed

Education		'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown'
Default	Has credit in default?	'no', 'yes', 'unknown'
Housing	Has housing loan	'no', 'yes', 'unknown'
Loan	Has personal loan	'no', 'yes', 'unknown'
Contact	Last communication type	'cellular', 'telephone'
Month	last contact month of year	'jan', 'feb', 'mar', ..., 'nov', 'dec'
day_of_week	last contact day of the week	
duration	last contact duration, in seconds	
campaign	number of contacts performed during this campaign and for this client	(numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign	(numeric; 999 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client	(numeric)
Poutcome	outcome of the previous marketing campaign	(categorical: 'failure', 'nonexistent', 'success')
emp.var.rate	employment variation rate - quarterly indicator	(numeric)
cons.price.idx	consumer price index - monthly indicator	(numeric)
cons.conf.idx	consumer confidence index - monthly indicator	(numeric)
euribor3m	euribor 3 month rate - daily indicator	(numeric)
nr.employed	number of employees - quarterly indicator	(numeric)

Data Cleaning:
White spaces - The white spaces were stripped from all the observation.
Typos - Typos were validated by checking for the unique values for each attribute.
No other data cleaning was done as there were no confirmed outliers for the numerical data and no Null values for the attributes.

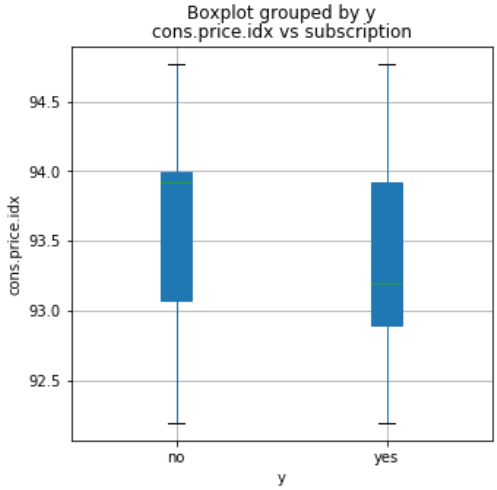
Data Exploration:
Different variables were explored with each other and with the target label to analyse the relationship among the attributes as well as to understand if they are a contributing factor to the target. The exploration also assisted in deciding upon the features which would be useful for the prediction.
Below are the variables that were explored for analysing the different hypothesis and research questions.



Is consumer confidence index a factor affecting the client subscription?

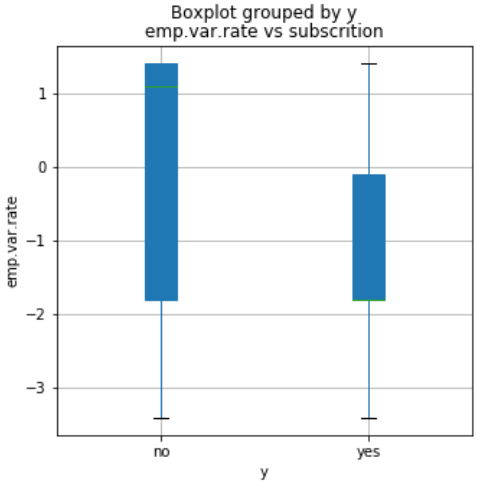
The graph depicts that people with a lower cons.conf.idx can subscribe for the deposit.

The cons.conf.idx measures how optimistic or pessimistic consumers are with respect to the economy in the near future. Thus, concluding that those with a lower index have more chances of subscribing for the term deposit.



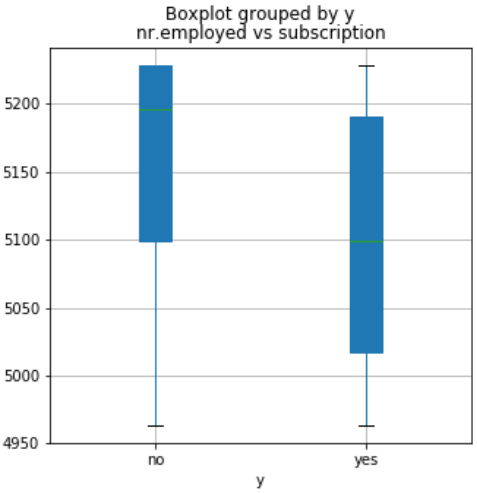
Is consumer price index a factor affecting the client subscription?

The graph depicts that customers with a lower cons.price.ind tend to subscribe to the term deposit while people with a slightly higher index would not prefer to continue with the term deposit.

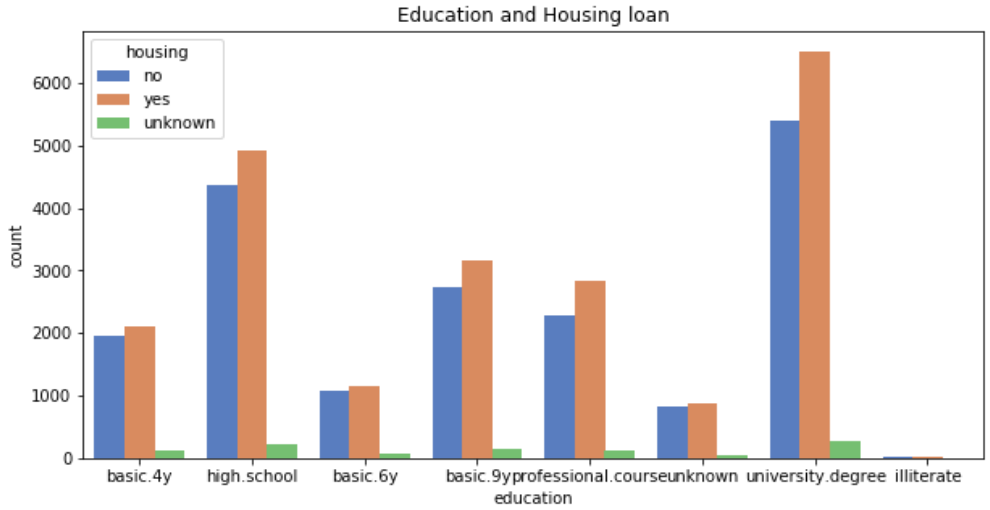


Is employment variability factor affecting the client subscription?

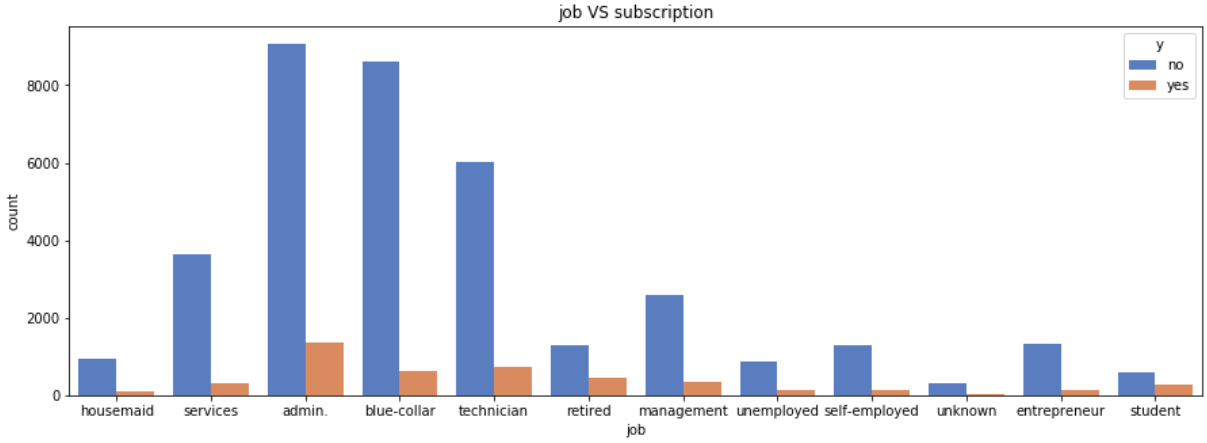
The graph depicts that people with lower employment variation rate subscribe to the term deposit thus indicting that more stability in a job environment encourages them to plan for a term deposit.



The graph depicts that people with a lower no_of_employees index tend to subscribe to the term deposit. This would indicate that lower of number of employees a person is working under during the quarter, more are the chances that he would think about taking the deposit.



The graph above illustrates the education level for people who have opted for housing loan. It concludes that among people opting for housing loans, majority of them have university degree or have cleared high school



The graph illustrates that the people with jobs in the field of admin have more chances of subscribing to the product.

Feature selection:

After data exploration, the below features were shortlisted to be provided to the data modelling tec hniques used for the prediction purpose.
Below are the selected features:

Age	Job
Education	Default (already in credit)
nr.employed	Housing (already taken a housing loan)
Loan (already taken a personal loan)	Poutcome (outcome of the previous marketing ca mpaign)
Emp.var.rate (employment variation rate)	Cons.conf.idx (consumer price index)

The dataset was trained based on the corresponding target from column 'y'.

Data Modelling

In lines with the goal of the research, a model that could correctly identify the outcome of a new observation, i.e. if the customer would subscribe to a term deposit (yes) or if the customer would not subscribe to a term deposit (no) has to be used. The dataset consisted of a target column named 'y' with values as 'yes' and 'no'. This column had to be used to train the observations and could be tested on a new set of observations and compared with the available target label to analyse the accuracy of the model. This method of modelling can be accomplished by the classification modelling technique.

Out of the different classification techniques, K Nearest Neighbour and Decision Tree techniques were used for the analysis.

K Nearest Neighbour Technique:

- Different combinations of test- train split was used. The dataset was divided into the below test training sets to train the data and test if the predictions are accurate based on the available results.

Training Size	Testing Size
50	50
60	40
80	20

- This technique was applied and the results for various values of k along with different parameters for weight and p were analysed.
- Ranges of k from 30 to 50 were applied for each of the combinations to analyse the confusion matrix and classification report.
- Observations from the results:
  1. Most of the models had high accuracy.

2. Weighted recall ranges from 88% to 90%.
3. Weighted precisions were ranging from 87% to 90%.

- A model with ideal values for recall and precision was chosen from each set of combination.
- Hill climbing - Feature selection was performed using Hill climbing on each of these models and further analysis was done to understand the variations in recall, f1-score and precision.
- Hill climbing technique further shortlisted Features which were most appropriate for the analysis.
- K nearest was again tried with these short listed features to get accurate values for the deciding parameters.
- The ideal model from K nearest was then concluded as described in results.

Decision Tree Technique:

- This technique was applied with different values of criterion like 'gini' and 'entropy' along with tuning the model using parameters for max_depth and max_features.
- Observations from the results were similar to those from k nearest neighbour as below:
    1. Most of the models had high accuracy.
    2. Weighted recall ranges from 88% to 90%.
    3. Weighted precisions were ranging from 87% to 90%.
- A model with ideal values was chosen from each set of combination.

The results from both the modelling techniques were analysed considering the research question i.e. predicting potential clients who will subscribe to the term deposit and the most accurate and ideal model for suggested for the prediction.

Results:

Below are the results obtained from K Nearest Neighbour and Decision Tree.

<u>K Nearest Neighbour:</u>

- The recall value for 'Yes' is low overall due to the availability bias of the data. After feature engineering, there was a slight improvement in the recall value.
- The precision for 'Yes' target was improved by 10% after feature selection.
- The matrix value for False Positive was higher for both the sets after feature engineering. However, for the third combination of 80-20, the values were better than the first 2 scenarios. Thus, the model obtained after feature engineering for dataset combination of 80-20 and K value of 50 would be considered as the ideal k – nearest model.

K Nearest Neighbour								
			Model Recall		Model Precision		Model Confusion Matrix Values for False Positive	
K Value	Train	Test	Before Feature Engineering	After Feature Engineering	Before Feature Engineering	After Feature Engineering	Before Feature Engineering	After Feature Engineering
			Yes	**Yes**	**Yes**	**Yes**		
46	50	50	4	14	56	67	72	160
50	60	40	4	17	56	69	52	123
50	80	20	5	15	59	69	34	64

Decision Tree:

- Similar to the recall value of K Nearest Neighbour, the recall value for target 'Yes' in decision tree is low overall. Thus the best among them was taken into consideration.
- The precision for 'Yes' target was also in the range of 34% to 37%.
- The matrix value for False Positive was higher when modelled through 'Gini' instead of 'entropy'. Though the values were in the range of 20-40, these values were considered while choosing the ideal model.

Decision Tree							
		Model Recall		Model Precision		Model Confusion Matrix Values for False Positive	
		Gini	Entropy	Gini	Entropy	Gini	Entropy
		Yes	**Yes**	**Yes**	**Yes**		
50	50	27	27	34	35	1179	1136
60	40	26	26	36	35	880	842
80	20	27	26	37	37	422	408

Discussion

Out of the different model variations provided for K and Decision Tree, best variation was chosen based on the below factors:

- **Imbalance Dataset-** The data set available for the prediction of customers who would take a term deposit is a bias dataset i.e. the observations for target 'no' (customers who wouldn't subscribe) are greater than the target 'yes'. Thus, the predictions of 'yes' scenarios are very less as compared to the 'no' scenario. Below is an example of the confusion matrix generated to better understand the analysis.

```
[[18131   160]
 [ 1979   324]]
```

(Confusion matrix)

		Predicted	
		No	**Yes**
Actual	**No**	Not Predicted and Not Subscribed **(True Negative)**	Predicted and Not Subscribed **(False Positive)**
	Yes	Not Predicted and Subscribed **(False Negative)**	Predicted and Subscribed **(True Positive)**

(Description of the confusion matrix)

As illustrated in the table above, there are 4 categories that need to be analysed to decide which model would be the better one for predicting the probability of customers taking up the term deposit.(Atwa, T. ,2017).

- **Not predicted and not subscribed –** All the models predicted this category with the highest accuracy. The major reason for this would be large number of observations that belonged to the 'Not subscribed' category.
  However, in terms of useful data for the bank, it won't be of much importance as they wouldn't be interested in analysing the details of customers who wouldn't subscribe to the product.
- **Not predicted and subscribed –** The value for this category was relatively high as compared to predicted and subscribed category. This indicates that the model predicts the target values with 'Yes' falsely. The major factor for this irregularity is the low availability of observation with target 'Yes' which does not allow the model to be trained as per expectation.
  However, the bank would not be so concerned with this false prediction because it would not damage its profit calculations or impact its stability in any way.
- **Predicted and Subscribed (True Positive) –** The value for this category should ideally be high to indicate that the model predicts the target as expected. But in this case, the actual values are relatively low due to the unbalanced data.
  The bank would want this category to predict accurately so that it can estimate a better financial forecast for the financial year.
- **Predicted and Not Subscribed (False Positive) –** The value of this category should ideally be low. But the models predict a high count of observations that were predicted to apply for the term deposit, but actually do not apply.
  This can be a major setback for the bank as there might be chances that they would have included these figures in their financial forecast. A positive prediction and false outcome would mean fall in the actual profit gains as compared to what was projected. This is not a situation which the bank would want to deal with.

Taking into considerations the above mentioned factors, the below attributes of a classification report were taken into account for providing an accurate model. (Shung, 2018)

- **Recall –** The weighted recall value was considered as it takes into account the fraction of relevant instances that are successfully predicted. Also, for multiple models having similar

weighted recall, the recall for 'Yes' target was given more emphasis considering the research goal.

- **Precision -** The weighted precision value was considered as it takes into account the fraction of correctly predicted instances. Also, for multiple models having similar weighted precision, the precision for 'Yes' target was considered.
- **Predicted and Not Subscribed (False Positive) –** Model having a lower value for this category was considered, as a higher value for this category would not be desirable for the bank.

Note: Accuracy was not considered as an important factor due to availability bias of data. The accuracy would always be on the higher side as the dataset has about 90% of observations for 'No' category as compared to only 10% observations for 'Yes' category.

Conclusion:

The Decision Tree for 80 training and 20 testing combination was the ideal model recommended according to the values obtained for the deciding parameters like re-call, precision and confusion matrix. Although the results provided are ideal according to the dataset provided, it can be further enhanced if the biasing of the dataset is reduced i.e. if more support data related to 'yes' target (customers who agree to subscribe to the term deposit) are added.

Thus, for the research goal of predicting customer who will take up the term deposit, the model would not be effective and reliable. There would be high chances of false predictions for subscribers which can impact the budget and financial forecast of the bank. However, it can be used if the only consideration is prediction without much emphasis on the customers who will subscribe to the deposit.

References:

UCI. (2019). [online] Available at: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing [Accessed 2 Jun. 2019].

(Shung, 2018) - Shung, K. (2018). Accuracy, Precision, Recall or F1?. [online] Towards Data Science. Available at: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9 [Accessed 2 Jun. 2019].

Atwa, T. (2017). How To Plot A Confusion Matrix In Python. [online] Available at: https://tatwan.github.io/How-To-Plot-A-Confusion-Matrix-In-Python/ [Accessed 1 Jun. 2019].