Capstone Project (Week 2)

Bank Marketing Dataset

Introduction:

Now-a-days the marketing of bank products are done by calling customers personally. This kind of personal calling marketing is called as tele-marketing, as it is done through telephone. It is done remotely and there are dedicated call centres for calling up the customers for asking whether they want to buy a product or not. This marketing sometimes done without analyzing the customers. But it is a very good idea to analyze the customers and group the customers who have tendency to subscribe a product.

After grouping the customers, banks contact them directly who are likely to buy a product. This kind of target marketing is done for meeting the bank's goals. The customers also have a central contact centre for any inquiries.

After segmentation, the data about customers' data is recorded and it can be used for predicting whether a customer buy the product or not. Based on this data a model can be built for predicting the customers who are going to buy a product. This model is useful for banks as they can predict and recommend for their customers. This process enables banks to focus on customers who are more likely to buy the product. The collected information gives bank the metrics about their customers and buying needs.

Data:

The data used for this analysis is a telemarketing dataset of a portuguese bank. The dataset is related to direct marketig campaign through telephone. The data set contains a total of 17 variables including the target variable. Here the target variable is of binary class which tells whether a customer will buy a product or not.

There are four datasets related to this problem. I considered bank-full.csv dataset for the analysis. Bank-full dataset consists of 45211 instances and 17 attributes.

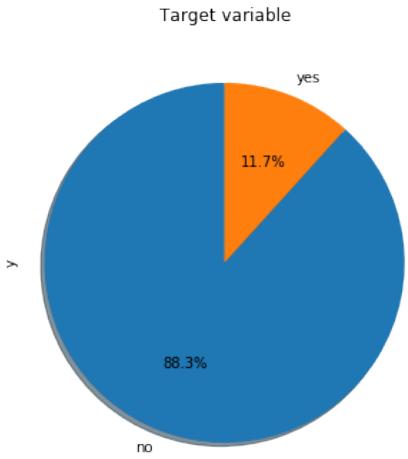
1. Age – Numeric attribute
2. job – categorical attribute having type of job they have
3. marital – Marital status of customer (Married, single, divorced)
4. education – categorical attribute (Primary, secondary, tertiary, unknown)
5. default – categorical attribute (has credit in default or not)
6. balance – Numeric attribute (average yearly balance in euros)
7. housing – categorical attribute (has housing loan or not)
8. loan – categorical attribute (has personal loan or not)
9. contact – mode of contact (cellular or telephone)
10. day – numeric (last contact day of the month)
11. month – categorical (last contact month of the year)
12. duration – Numeric (contact duration)
13. campaign – numeric (number of times contacted during the campaign)
14. pdays – numeric (number of days passed by after contacting)
15. previous – numeric (number of contacts performed before the campaign)
16. poutcome – categorical (out come of previous campaign)
17. y – categorical - customer subscribed to product or not (target variable)

This is a publicly available dataset and is found at UCI Machine Learning repository "https://archive.ics.uci.edu/ml/datasets/bank+marketing".
It is a bank marketing dataset and ofter more than one call is needed for accesing whether a customer has subscribed to the product or not.

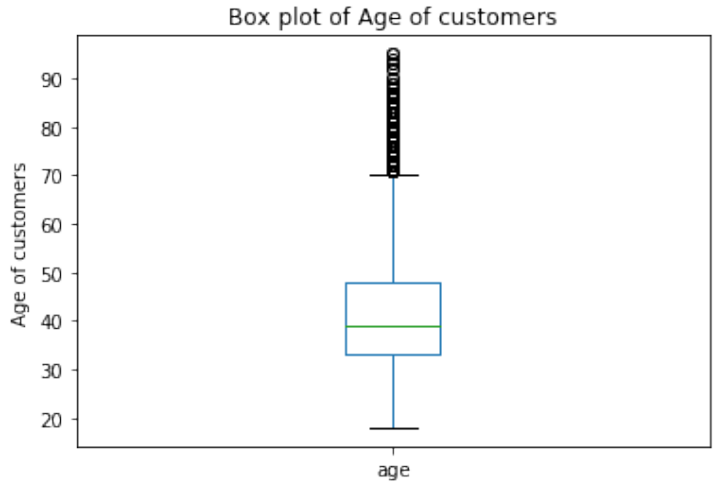
Methodology:

Exploratory data analysis:

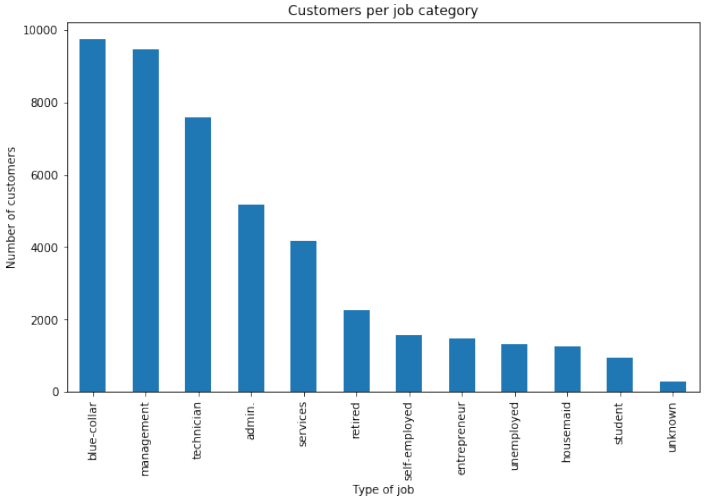
The target variable in the dataset is ‘y’. It contains ‘yes’ / ‘no’ i.e. the customer subscribed to the product or not.



As we can see in the above figure, the customers who subscribed to the product are only 12% of the total records. Whereas 88% have not subscribed. There is a huge gap between the proportion of two classes. By building a good model out of this data, we can predict the customers who are most probable to buy and target those customers. Which hopefully increase the proportion of ‘yes’ class in future.



The above box plot show us the range of age we have in the dataset. The minimum age is 18, maximum age is 95. The average age of a customer in our dataset is 41.



Here is a bar plot which shows the job categories of all the customers in the dataset. It is sorted in descending order of number of customers per category. As we can see there are blue-collered customers highest in number followed by management, technician and so on.

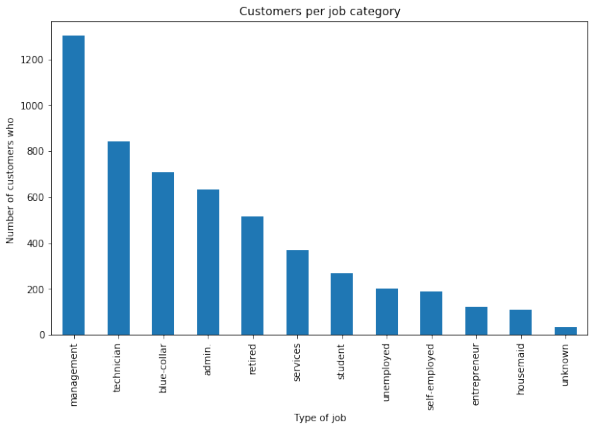


Fig: Customers of class ‘yes’

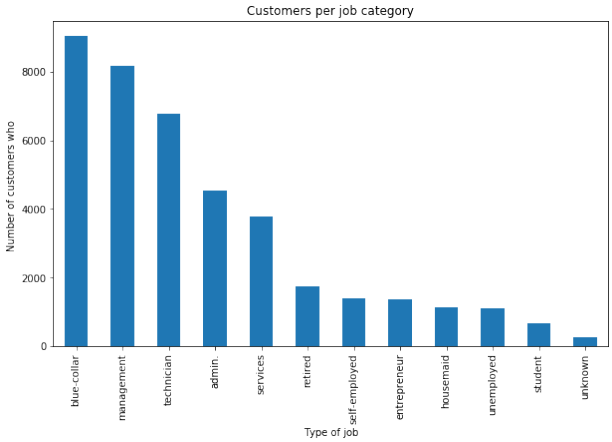
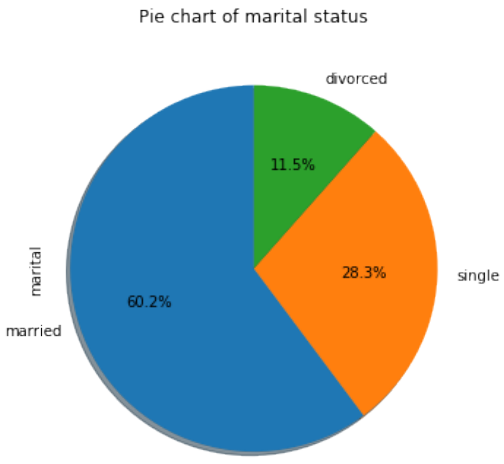


Fig: Customers of class ‘no’

In the above two figures, we can infer that the customers who subscribed to the product are ‘management’ type job are highest in number. Whereas the customers who are not subscribed are ‘blue-coller’ type job. So we can focus more on the customers who are having ‘management’ and ‘technician’ jobs.



The distribution of marital status of customers is 60% of the customers are married, 28% are single and 12% are divorced.

The values of other variables are explained in the Data section.

Preprocessing of dataset is a crucial step and usually it takes large amount of time in the model building process. But here in the dataset used pre-processing of data is not needed as the data is in well structured format and there are no missing values in the data.

The data is splitted into 80-20 partition for building model and its evaluation. An out-of sampling technique used for splitting. 80% of data is used for training the model and then tested on the remaining 20% of data. A random sampling without replacement is performed for the analysis. The test data is not used in model building.

As the target variable is binary class. It is a classification problem. I used classification algorithms for classifying the classes.

Firstly I decided to go with the decision tree algorithm as it gives us if-then rules. The rules are very intuitive and can easily be interpreted. The decision tree algorithm is also used widely in banking domain because of its simplicity in understanding and useful in decision making process.

Then I tried out using Logistic Regression and followed by K-Nearest Neibors algorithm. Logistic regression algorithm gives us probabilities of each customer for subscribing to the product. The probabilities are very useful for bank. By analysing the probabilities bank can decide on customizing the product according to the needs of customer.

K-NN algorithm is useful as the dataset is relatively large. KNN is effective when we are analysing the dataset of large number of records. But choosing the best K is sometimes tricky and troblesome. For determining the best K in KNN multiple iterations are conducted and decided to go with K value as 4.

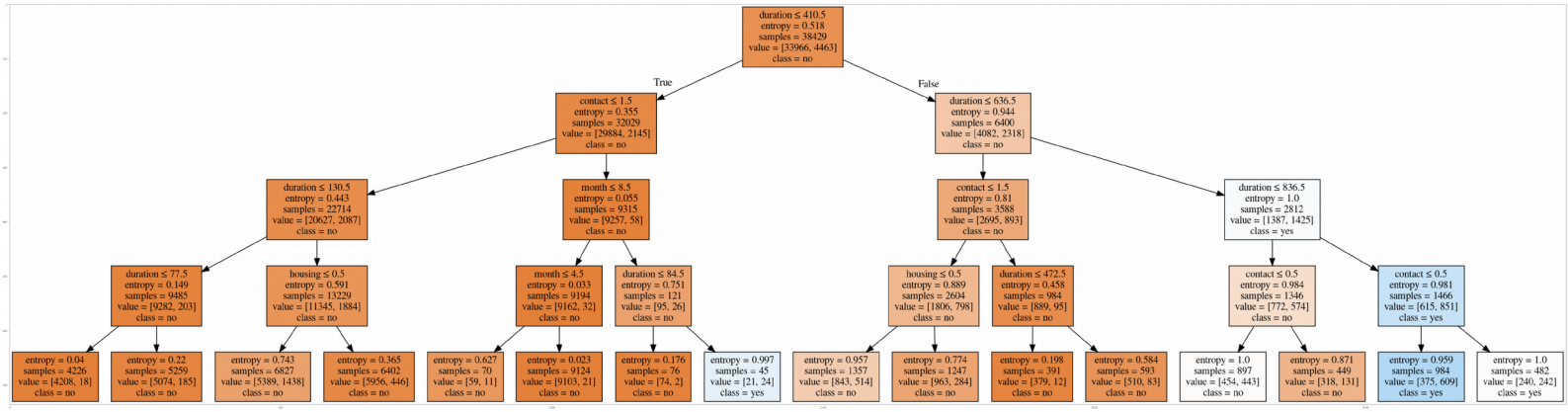
As the dataset is of 45211 instances, SVM is not so useful as it took very long time to converge. So I decided to go with only decision tree, logistic regression and KNN for comparative analysis.

After trying out these algorithms, the models are evaluated and tested on the test dataset. Various classification metrics are used for checking the accuracy of models. I have used Log loss, Accuracy, F1-score and jaccard similarity score for reporting

Results and discussion:

Algorithm	Accuracy	Jaccard similarity	F1 score	Log loss
Decision Tree	0.89	0.89	0.86	NA
Logistic Regression	1	1	1	0.01
KNN	0.99	0.99	1	NA

The results obtained are very promising and there is a very much improvement in the accuracy that is reported in the base paper. As we can see above, Logistic regression out performed all the classifiers. Eventhough decision tree gave the lowest accuary of all the classifiers, it can be considered as we can see the rules and form the ruleset. The ruleset is easy in understanding and can be applied in the domain easily.



As we can see in the decision tree plot, duration of the call is identified as the important column and the whole tree is splitted based on it. By traversing the tree we can also identify the important variables for our analysis and drop the other variables. In this way it is also useful in feature selection.

Conclusion:

By analysing the above results and other attributes we can conclude that Logistic regression is helpful in predicting the customers who are very likely to subscribe to the product. The logistic regression also gives probabilities of each customer for buying the product.

However there can be a debate why one can deploy decision tree algorith. The decision tree as we can see above is very helpful in traversing through the important variables. It helps the bank to identify and focus on the group of customers. Eventhough it has less accuracy, the accuracy obtained in the modeling can not be ignored as it is 89%.

References:

1) [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014
2) UCI ML repository (https://archive.ics.uci.edu/ml/datasets/bank+marketing)