CS989 : Big Data Fundamentals
Report : Term Deposit Subscription Status

Contents

List of Figures ii

1. Introduction to dataset 1
2. Key challenges and problems 3
3. General analysis and summary of the dataset 6
4. Unsupervised approach clustering 10

4.1 Attempt to cluster term deposit 10

4.2 Clustering timestamps 11

5. Supervised approach classification 13
6. Reflection of the chosen analysis method and conclusion 16

A Software version, data collection and packages used 17

Bibliography 18

List of figures

1.1 Bank dependency on big data 1
2.1 Outliers in data 3
2.1 Table label encoder 4
2.2 Distribution of data 4
2.3 Imbalance data 5
3.1 Outcome of previous marketing campaign 6
3.2 Education distribution 6
3.3 Job type for term deposit 7
3.4 Balance distribution 7
3.5 Age group distribution 8
3.6 Credit default, housing and loan 9
4.1 (a) Silhouette and elbow score 10
4.1 (b) Cluster in 2D 11
4.2 Homogeneity and completeness score 12
5.1 Correlation heat map 13
5.2 Feature importance 14
5.1 Table Decision tree working 14
5.2 Table Classification report 15
5.3 Area Under Curve 15

Chapter 1

Introduction to the dataset

Big data technologies are being widely used in various sectors from manufacturing to financial sectors for better decision making and problem solving. It is evident from figure 1.1 that use of big data tools and techniques is rising over the years. Practising advanced analytics methods like machine learning, deep learning and natural language processing, and targeting the specific customers so as to reduce the cost of resources and time.

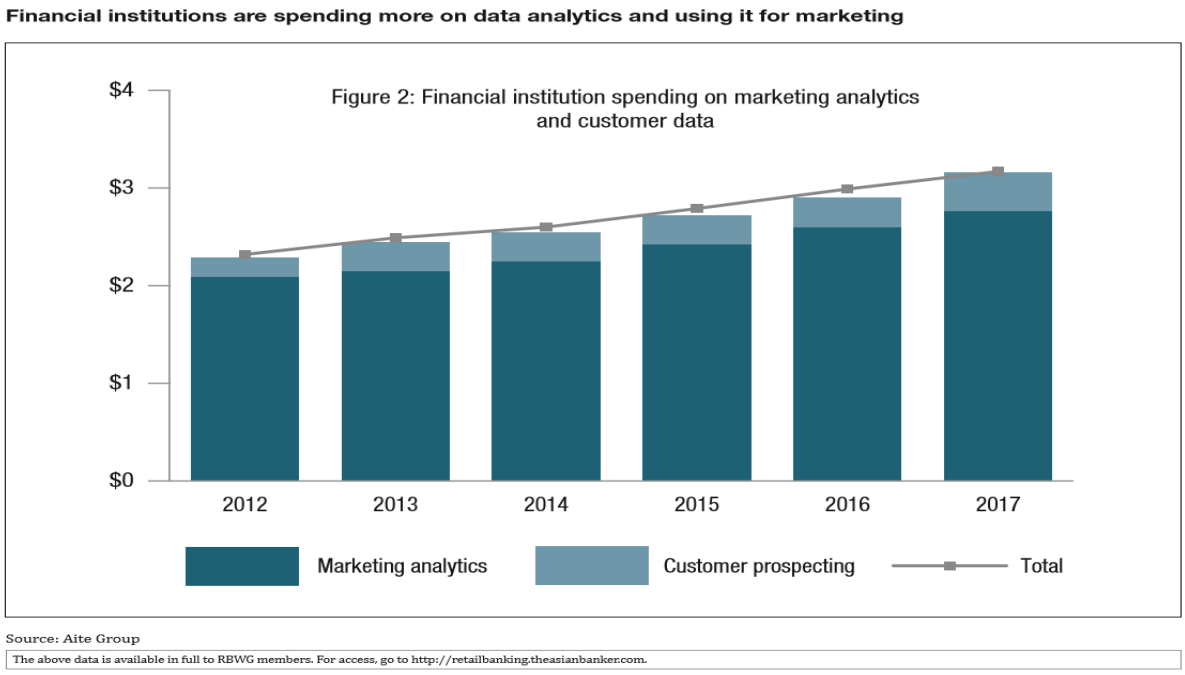


Figure 1.1 Banks dependency on big data over the years

Cost of marketing the products is huge unless banks are not aware which customer to target for that specific product like term deposit, mortgage loan or personal loan. A decade ago, in India banks were struggling to sell their schemes and products like fixed deposit, used banners to advertise which was expensive still results were devastating. In current scenario, enabling big data have improved their numbers significantly and helped in managing the resources efficiently (as per financial times India).

Chapter 1. Introduction to the dataset

The data this report is based on, to promote the term deposit among existing customers of a Portuguese bank from May 2008 to November 2010 having 45211 rows and 17 columns. This data has customers who have subscribed/not subscribed term deposit during this period, besides this data has others following features to categorise the specific customer.

- Age
- Job
- Marital
- Education
- Default
- Balance
- Housing
- Loan
- Contact
- Day
- Month
- Duration
- Campaign
- Pdays
- Previous
- Poutcome

By analysing this data using different libraries, it is possible to know the situation of subscription rate of term deposit and predict the future scenario for the same.

Chapter 2

Key challenges and problems

Firstly, main defiance was to clean the data and handle outliers, specifically in age, balance and duration columns in a way that it can be visualised and analysed.

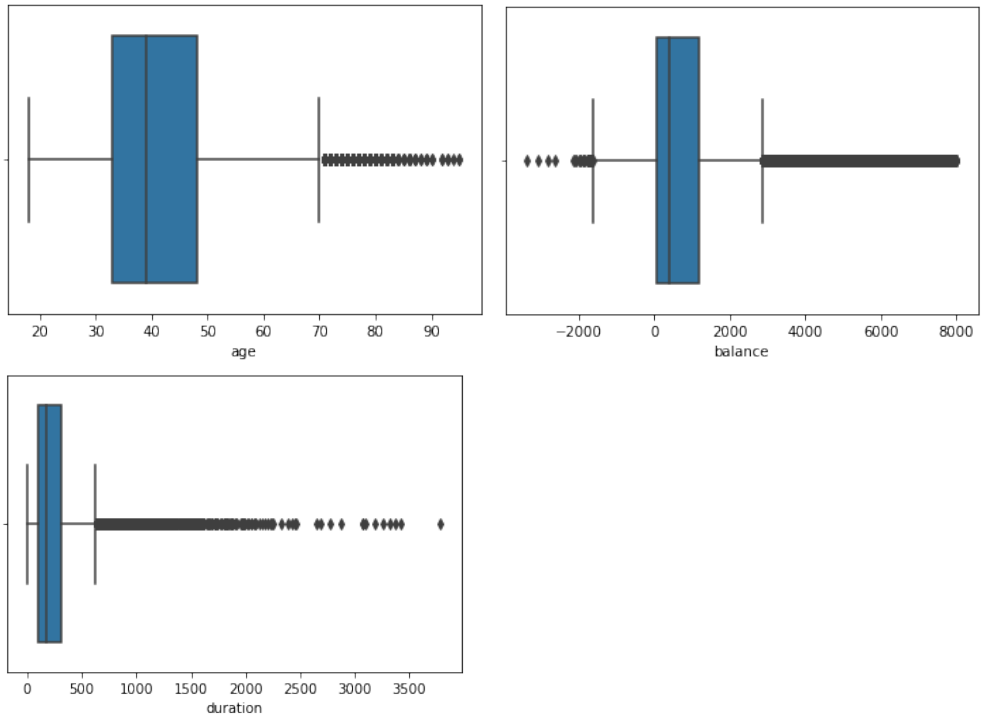


Figure 2.1 Outliers in Age, Balance, Duration

Performed outliers capping based on Inter Quantile Range (IQR) since data was right skewed, can not use z score method for capping because data is not normally distributed. After capping upper and lower wiskers, IQR is taken into account.

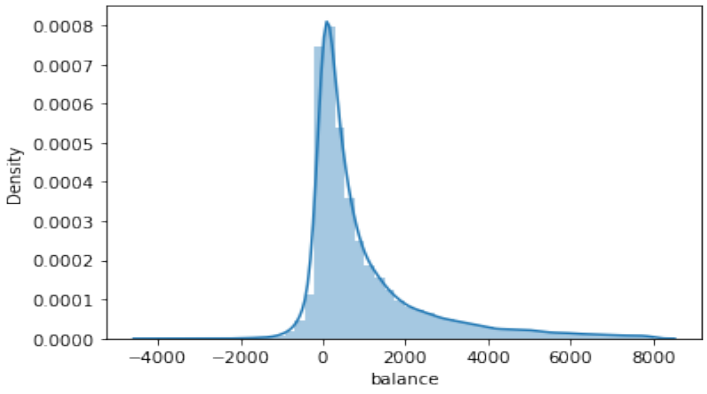
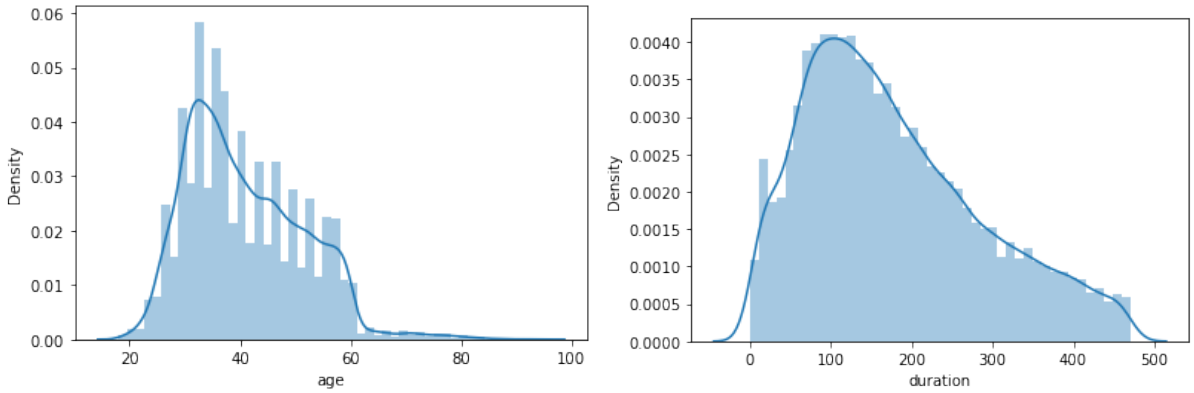


Figure (a)



(b) (c)

Figure 2.2 (a), (b), (c) Distribution of data for age, balance, duration

Distribution can be referred from figure 2.2 that most of the data points are away from mean.

Secondly, deal with curse of dimensionality – since most of the columns in dataset are categorical particularly in job and month which have more than ten categories, if used pandas dummy variable library that will create more features which will increase the noise and redundancy during its analysis and machine learning will outperform. Another method used is label encoder that assigns the rank by calculating the frequency of that variable in a column. But if used for two categories, it will interchange the outcome values.

Job	Month	Month rank	Job rank
Blue collar	Jan	8	1
Technician	Feb	7	2
Management	Mar	11	3
Student	Apr	6	11
Entrepreneur	May	1	8
Admin	Jun	4	4
Retired	July	2	6
Services	Aug	3	5
Housemaid	Sept	10	10
Self employed	Oct	9	7
Unemployed	Nov	5	9

Table 2.1 label encoder

Also, while performing k-means clustering calculate the minimum distance between two data points and based on centroid assigns them into cluster but units of age, balance and duration have unique units which was difficult/inappropriate to compute the result. Scaling was necessary for age, balance and duration columns using Sklearn library minmaxscaler (range from 0 to 1) to bring these columns to the same scale.

Lastly, subscription status has imbalance data points required to perform oversampling, that virtually assigns values to lower frequency class to improve the model performance and accuracy.

Figure 2.3 Imbalance data : Subscription Status

Chapter 3

General analysis and summary of the dataset

So, in our current campaign only 11.6 % of the clients have subscribed the term deposit. Bank need to improve the run rate by better targeting, reducing the operational resources and cost by improving the prediction whom to call for subscription or not which can be clearly visible from figure 2.3. Comparing from past year campaign where most of the data is not available may be due poor data collection techniques but 3.3 % of those clients might be contacted as reflected in figure 3.1.

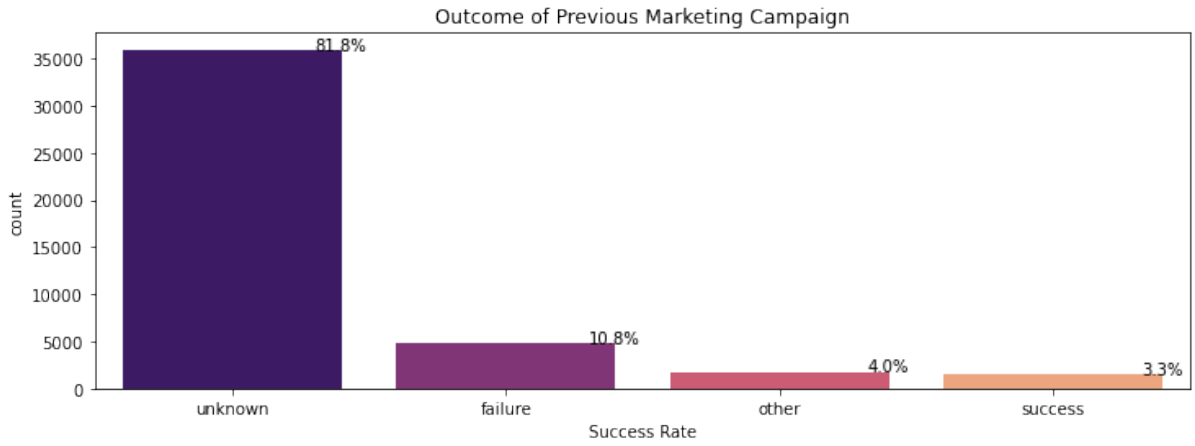


Figure 3.1 Outcome of Previous Marketing Campaign

Secondary and beyond education level accounts for 83 % of the total subscribed term deposit as shown in figure 3.2. Education is primary tool for country's economic growth and for one self to understand the banks legal and updated schemes to get better term plans.

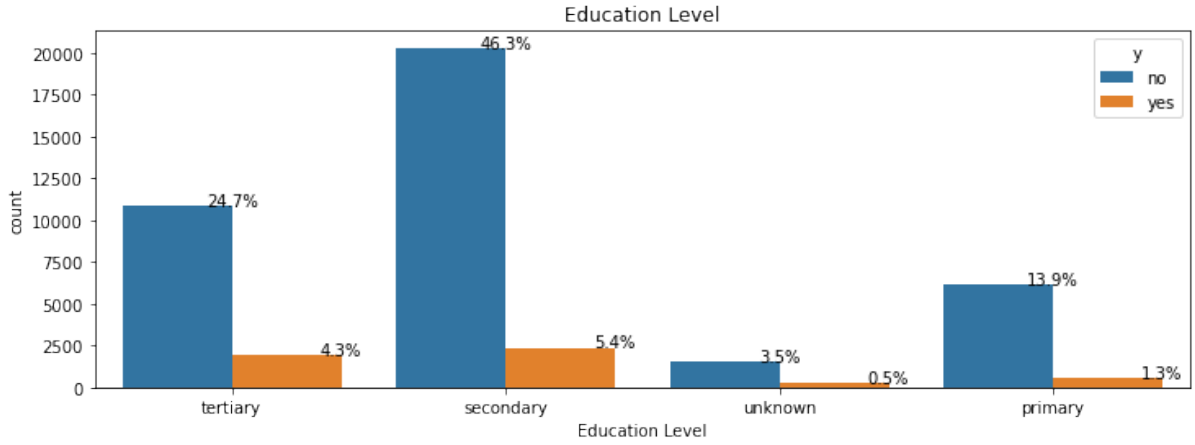


Figure 3.2 Education Distribution

Chapter 3. General analysis and summary of the dataset

Profession accounts for balance in your bank statements, only clients having enough funds can opt for term deposit. Management, technician, blue-collar, admin and retired professionals are responsible for 76.7 % of the subscription rate as represented in figure 3.3.

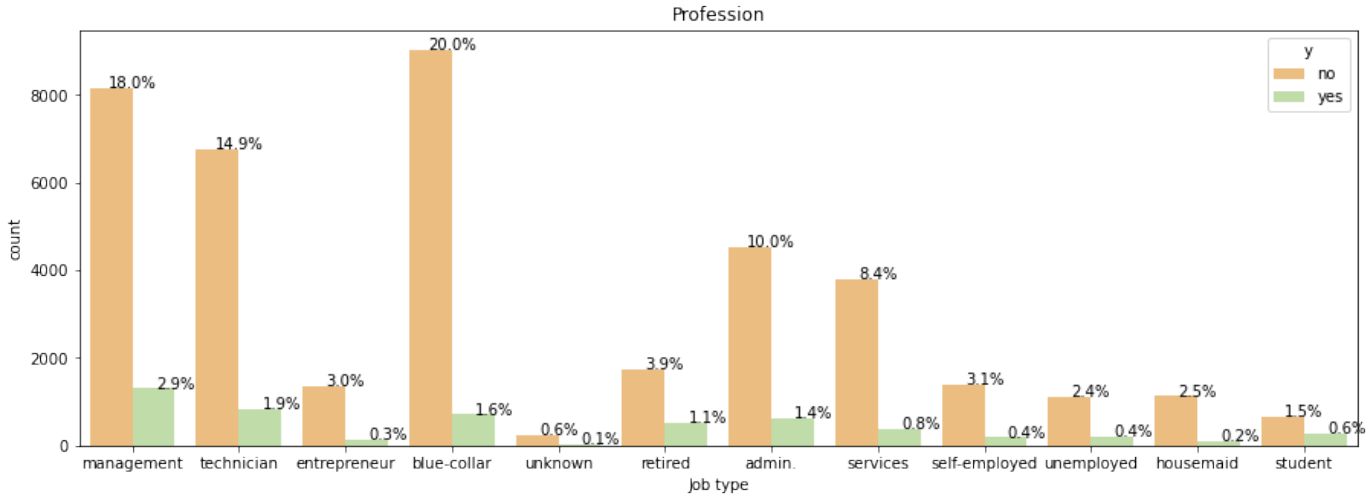


Figure 3.3 Job Type in Term Deposit Status

Banks should more often contact to customers having low and medium account balance holders between £1000 to £4000 as shown in figure 3.4. Avoid calling negative and high account balance holders to save its capital and human resource. Since low and medium balance holders have family responsibilities like children education, rents, etc by getting term deposit that will help them in long run.

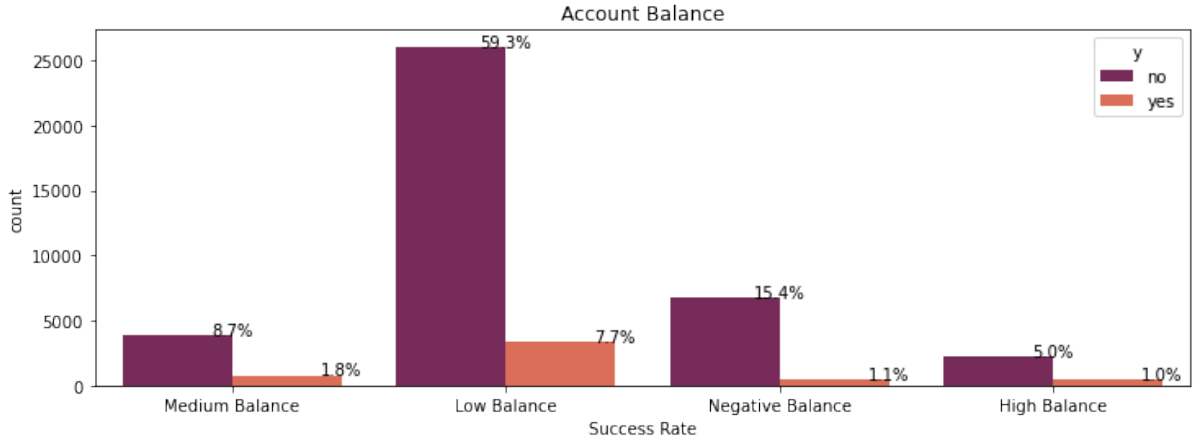


Figure 3.4 Account Balance Distribution

Chapter 3. General analysis and summary of the dataset

Teenagers grouped at 22 years old may be does not have enough income source, so they meant not to be contacted. As depicted in figure 3.5, middle age citizens accounts majorly for the subscription rate of term deposit since they are earning higher as per their profession which can be concluded from above figures as well.
Banks can also particularly target senior citizens since less contacts were made to them compared to others still they account for 2.2 % of term deposit maybe they are getting pension and savings to supports.

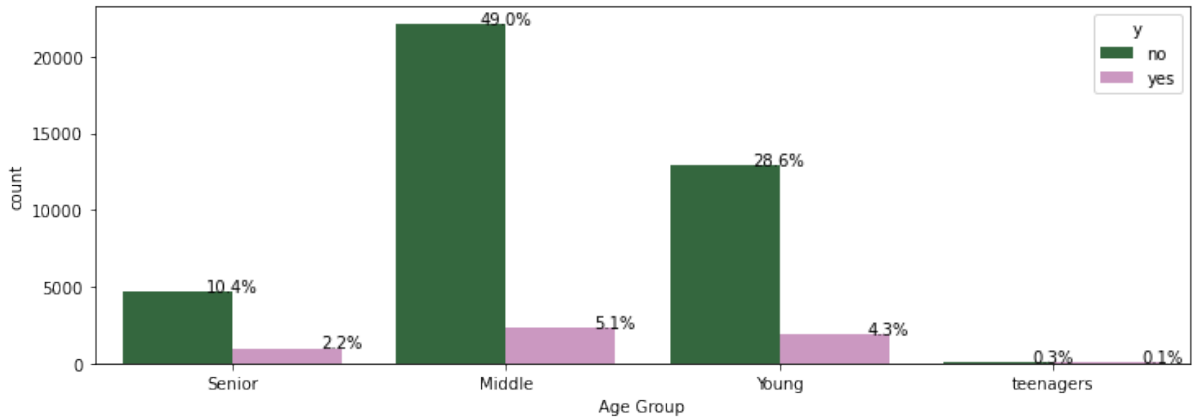
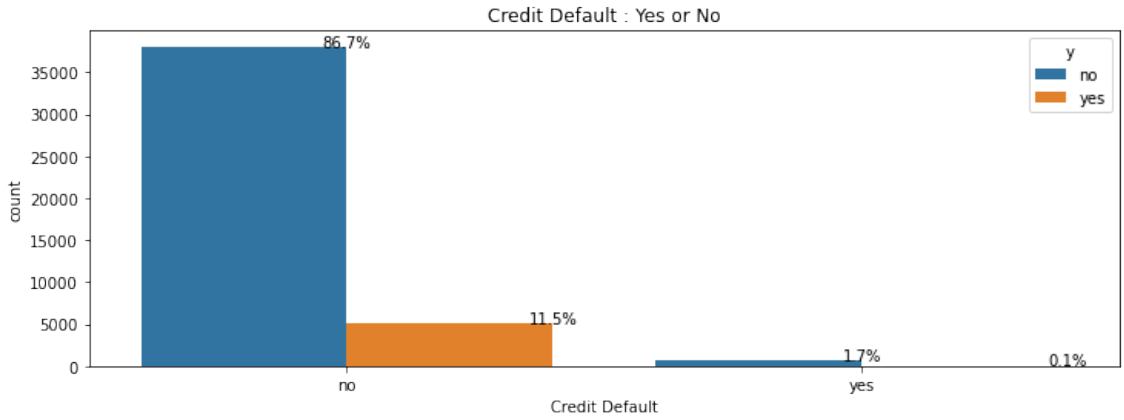
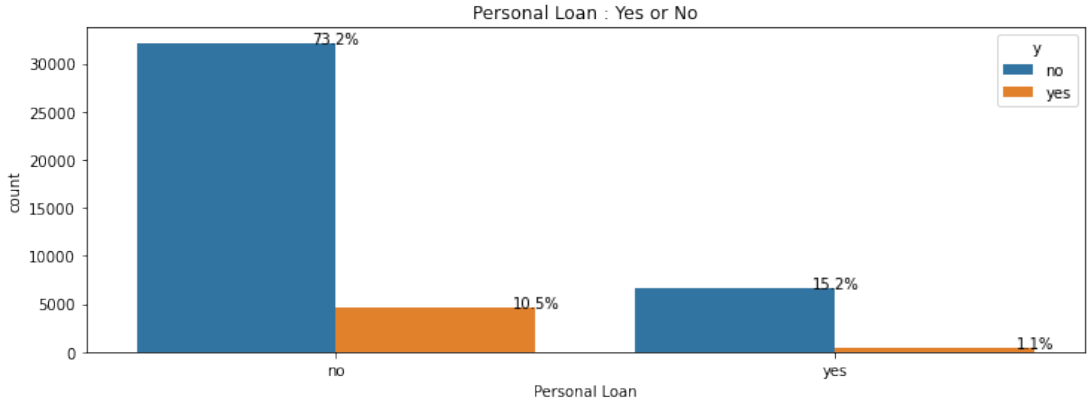


Figure 3.5 Age Group Distribution

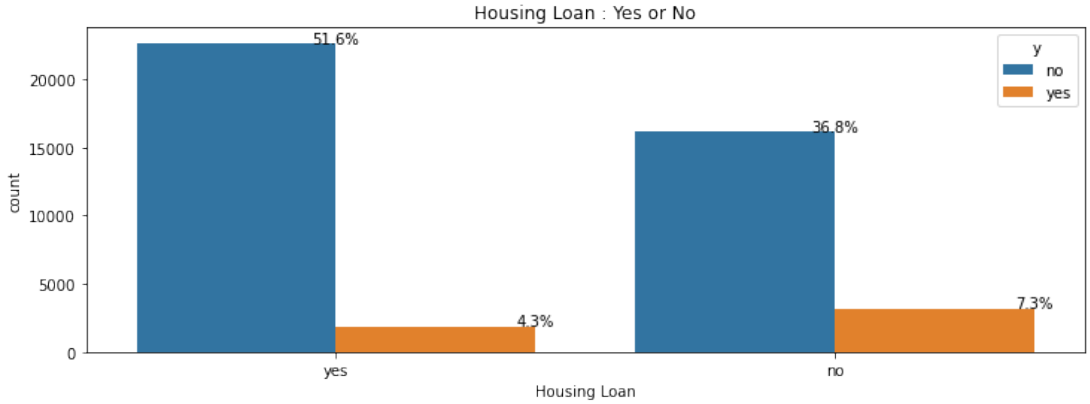
Moreover, it is interesting to know the fact that customers not having personal loan are accountable for 10.5 % to term deposit. On the contrary, clients having not housing loan contributes 7.3 % as shown in figure 3.6(a, b, c), this may be due to banks usually lend huge money on mortgage to its clients after credit check, which in turn clients have burden to pay instalments that hesitate them of getting benefits from such schemes.



(a)



(b)



(c)

Figure 3.6 (a, b, c) Credit Default, Housing Loan and Personal Loan

However, after analysing different features, bank should contact clients based on their job type, account balance, age group and type of loan customer is having. Furthermore, previous year subscription status data is not available. Then, calculations would be more precise and accurate but without that this could only be used for observations.

Chapter 4 Unsupervised Approach

Clustering Analysis using K-means Algorithm

The objective of clustering is to find grouping of various features in dataset based on term deposit status, so as to get clear input about importance of those features which make it easy and efficient for lender to target specific category customers. Later on, those features can also be beneficial for classification machine learning algorithms. For the same, data was firstly scaled to adjust it to the same level for better results and clusters.

4.1 Attempt to cluster term deposit status

Clustering term deposit status will be essential for the bank enabling them to effective decision making and using resources efficiently, in such a away, which customer and his behaviour pattern to target not only for term deposit/marketing but can also for suggesting other loan schemes, bonds and other banking products that will help it grow exponentially. Also, quite useful in for it in policy making and restructuring schemes.

To find optimal number of clusters is quite a task used both the methods silhouette score and elbow method. For maximum silhouette score > 0.8, number of clusters are 2 or 3 and for elbow method, it starts to attain a saturation point after wcss(within clusters sum of squares) = 0.25 at 3 clusters refer to figure 4.1(a).

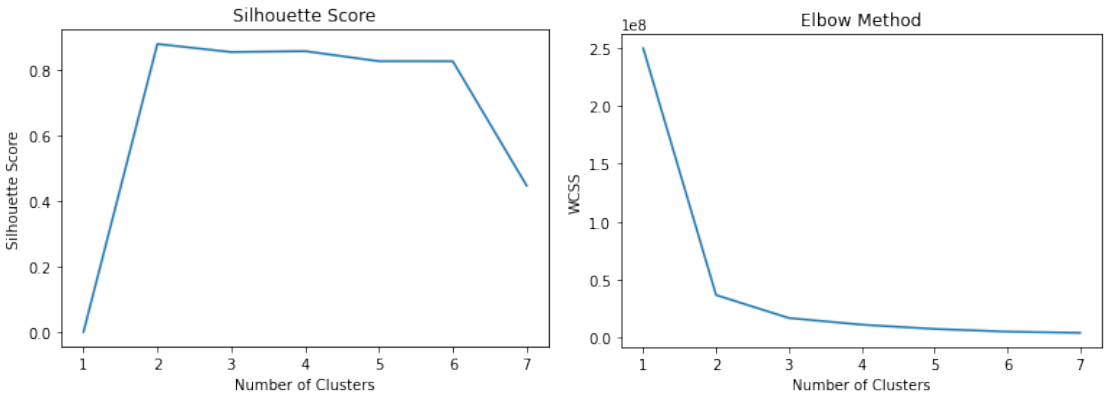


Figure 4.1 (a) Silhouette and Elbow Score

Chapter 4. Clustering analysis using k-means algorithm

As k-means algorithm clusters the data points which was having too many features/dimensions, for visualising data use principle component analysis(PCA) that reduces the dimensions as per requirement for analysis but could not be effective for machine learning algorithms, while reducing dimensions most of the important features gets deleted. As observed in figure 4.1(b), most of the data points are collapsing over each other.

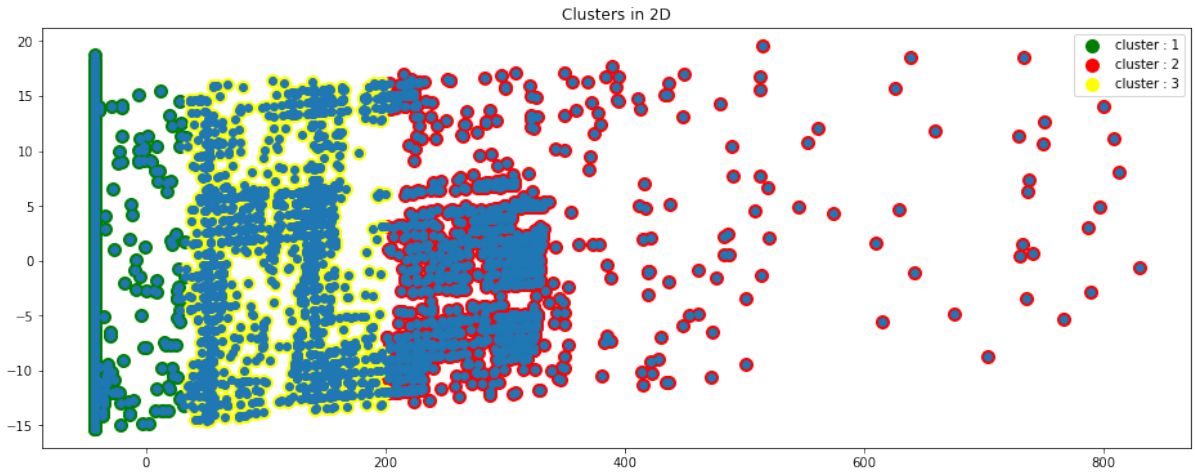


Figure 4.1 (b) Clusters in 2 Dimensions

After concluding from the k-means clusters, it is clear that k-means failed in identifying features. This is the reason to drop the clustering analysis, has to use another technique for the same which will be discussed in the coming chapter.

4.2 Clustering Timestamps

As the prior clustering did not perform quite well, so thought of increasing the clusters for other scenarios. Bank can identify credit defaulters for example bank usually verifies credit score which basically get improve after every six months if you are having any type of loan assuming you are paying instalments on time, clustering helps in detecting previous defaulters who are having current loan which in turn bank can restructure its interest rate and minimize the risks associated with that customer. To deal with such situation, analysis was performed with multiple clusters using k-means algorithm, increasing one more cluster at time and the result has been plotted using homogeneity score(clustering of members of single class in the single cluster) and completeness score(clustering of members within a cluster of single class). Range of score lies between 0.0 to 1.0.

As the figure 4.2 (a) shows homogeneity score has been increasing at a negligible rate or can be treated as constant which is trying to cluster every feature independently while completeness score has drastically fall after 2 clusters. This means that it is not possible to classify using k-means algorithm. So, applying decision tree algorithm to classify the target variable which will be explained in the next chapter of this report.

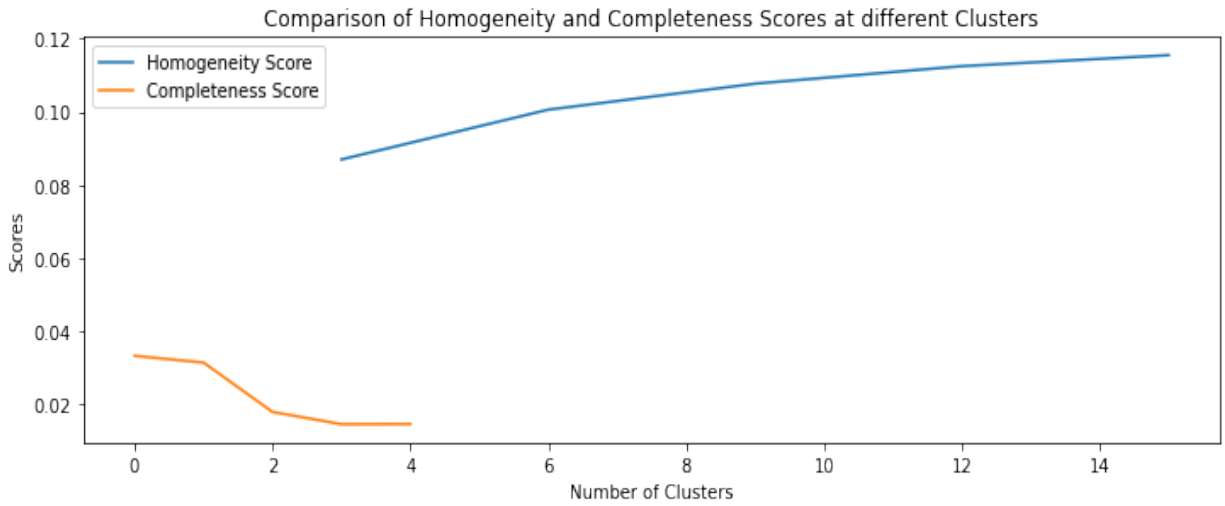


Figure 4.2 Homogeneity and Completeness Score

Chapter 5 Supervised Approach

Classification using Decision Tree

Classification is being preformed to classify the class into probabilities of occurrences in the term of 0 and 1. For this target variable term deposit had to be manipulated in the correct manner, so it could be used properly for analysis. Term deposit had categorical variables in the form of yes and no which was handled by assigning numeric values 0 for not subscribing and 1 for subscribing term deposit. This has been achieved using the below code:

```
def y_status(y):
    if y == 'yes':
        return 1
    else:
        return 0

df['y'] = df['y'].apply(y_status)
```

Since dataset had various columns feeding all columns to machine learning algorithm, columns which has no correlation with target just affect the model's time complexity and accuracy. This had been done using plotting correlation heat map. Higher the value with target variable those dependant variables will be consider for classification.

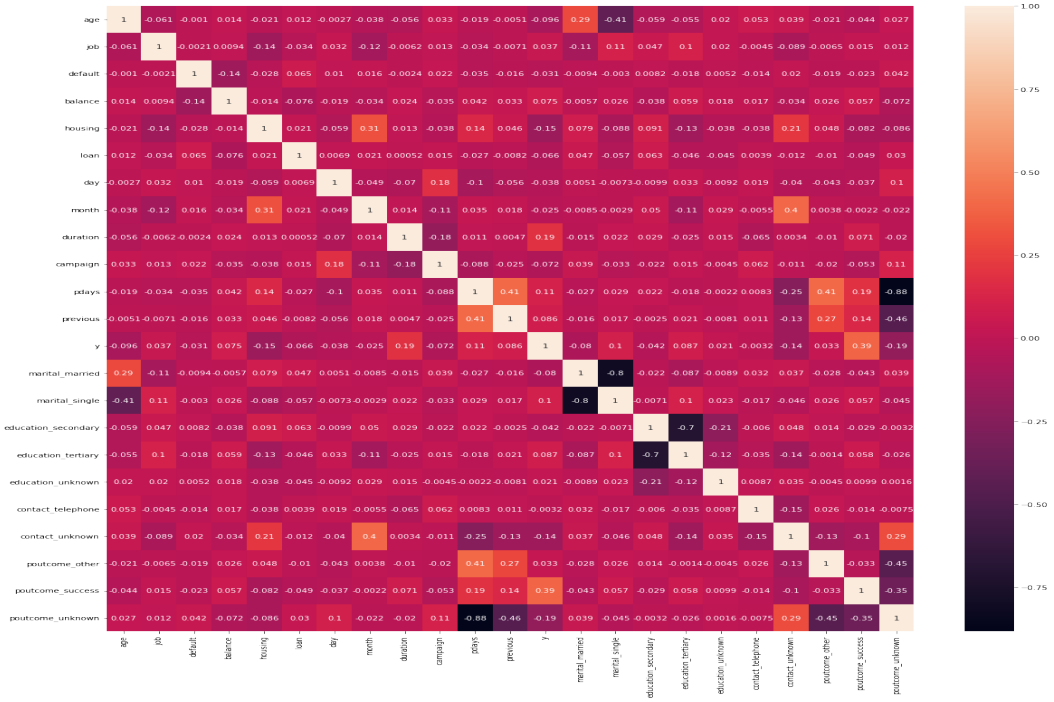


Figure 5.1 Correlation Heat Map

Another technique used to get clear picture about feature selection was feature importance under sklearn library named ExtraTreesRegressor which basically plot important features related to target variable. Technically both techniques were taken into consideration while selecting corelated features with independent variable. (refer to figure 5.1 and figure 5.2)

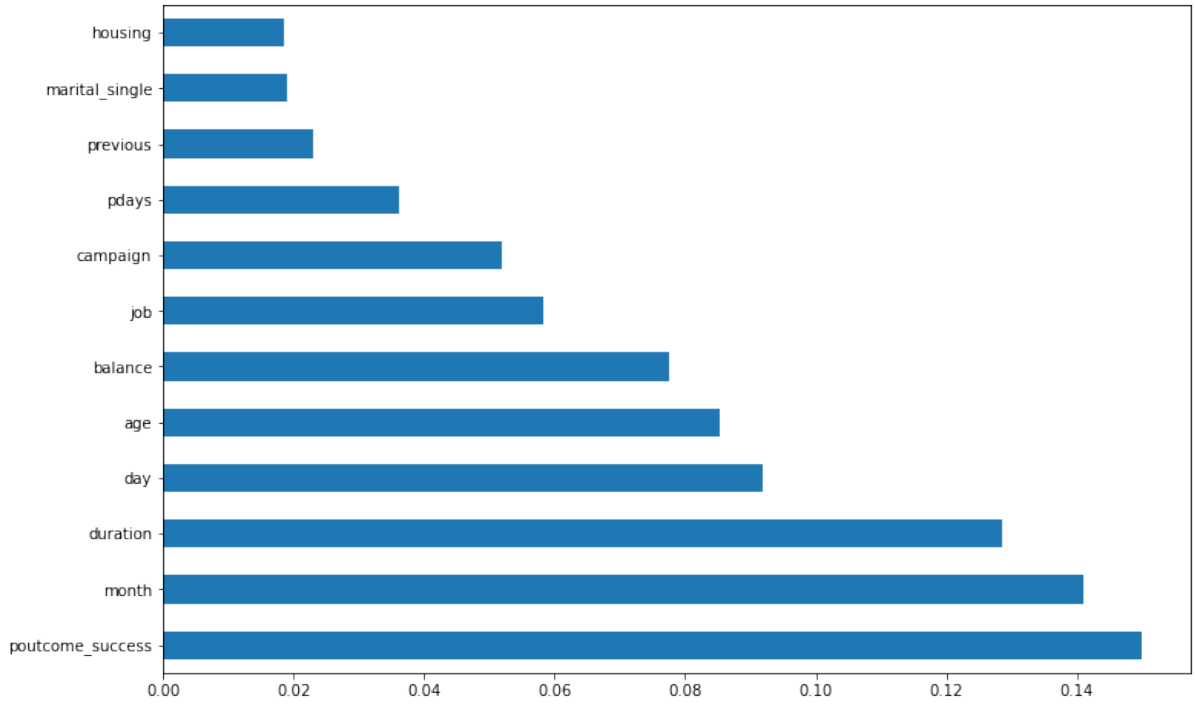


Figure 5.2 Feature Importance

For analysis, data has been split into training and testing variables which was later accorded into decision tree model for classification problem. This model makes trees by calculating entropy value of variables and higher the entropy, variable will split first and process keeps on classifying classes (see table 5.1).

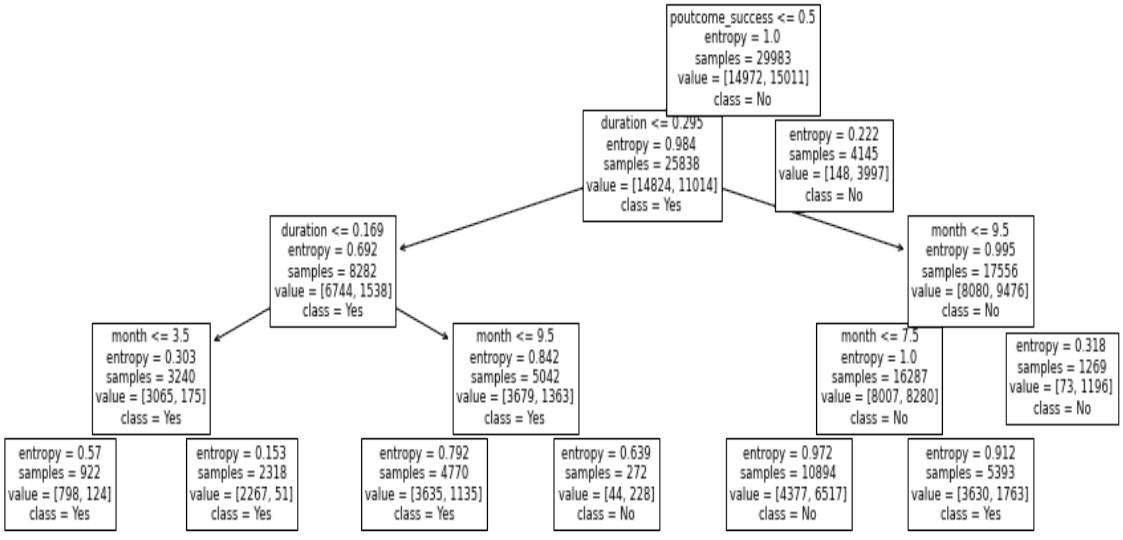


Table 5.1 Decision Tree Working

To validate the model classification report had been printed, results shows model has performed quite well (see table 5.2). All measures are quite good, there no imbalance class, f1 score for both classes for no false positive and no false negative is impressive while achieving the overall model accuracy of 74 %.

	precision	recall	f1-score	support
0	0.77	0.69	0.73	6445
1	0.72	0.79	0.75	6406
accuracy			0.74	12851
macro avg	0.74	0.74	0.74	12851
weighted avg	0.74	0.74	0.74	12851

Table 5.2 Classification Report

Another metric used was area under curve which uses true positive rate and false positive rate measures, classifier was perfectly able to distinguish between true positive and false positive data points. Area under curve measure was 0.81 which indicates model perfectly predicts (range from 0 to 1, 0 being the worst).

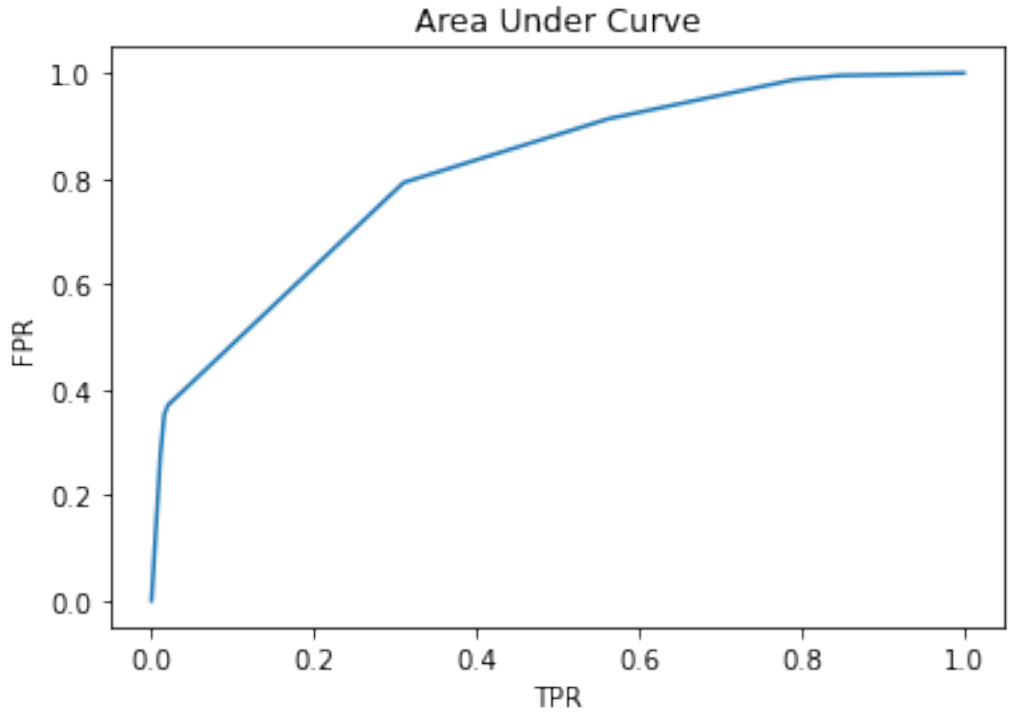


Figure 5.3 Area Under Curve

Chapter 6

Reflection of the chosen analysis methods and conclusion

The results can be concluded in many ways may be the data was not well suited for clustering and/or data has too many features. Since k-means clustering works on different distance measures but most of this dataset features were categorical, scaling the dataset does not affect either. Clustering similar features was a bad idea that would not work in this scenario, tried for various numbers of clusters using different measures but the same have proved to be useful in identifying outliers. Could have tried clustering methods.

Considering the classification analysis, results of decision tree was compared with logistic regression. Decision tree out performed logistic regression in classifying the customers who to target or call for marketing term deposit plans. Accuracy of the can model can be improved by lowering the variance which needs to be lowered by methods like bagging and boosting. Also, performance of the tree can be increased by pruning techniques, since they are not covered in our syllabus so have to follow the guidelines.

In conclusion, classification model can be implemented by improving accuracy and bank can improve their marketing campaign while targeting valuable customers which saves their resources and time. Bank or any marketing boutique can their easily control their expenses and beneficial in mitigating risks associated with marketing.

Appendix A

Software version, data collection and packages included

Python version : Python 3.10.7
Jupyter notebook version : Jupyter notebook 3.3.2

Dataset derived from : https://archive.ics.uci.edu/ml/datasets/bank+marketing

Packages used :

- Numpy
- Pandas
- Matplot library
- Seaborn
- Sklearn
- Metrics under sklearn

Bibliography

R. Greg, M N Kumar, Rahman Z Qureshi (2012) : Identifying and ranking critical success factors of customer experience in banks. URL : https://www.emerald.com/insight/content/doi/10.1108/17465661211242813/full/html

Shashmi Karanam - Curse of dimensionality : A 'Curse' to machine learning (2021) published in towards data science journal.
URL : https://towardsdatascience.com/curse-of-dimensionality-a-curse-to-machine-learning-c122ee33bfeb

Harish Reddy – K-means (2019) : URL : https://medium.com/analytics-vidhya/k-means-clustering-43d0136bf005

Scikit Learn (2022) : Decision Tree. URL : https://scikit-learn.org/stable/modules/tree.html?highlight=decision+tree

Scikit Learn (2022) : Clustering score. URL : https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html#sklearn.metrics.completeness_score

Figure 1.1 Bank dependency on big data over the years taken from google images.