Modern Data Science (SIT 742)
Assignment 2

Submitted by:
Harshdeep Singh (218318242)
Shahzeb Maqsood (218474893)
Manthan Negi (218467657)

Introduction

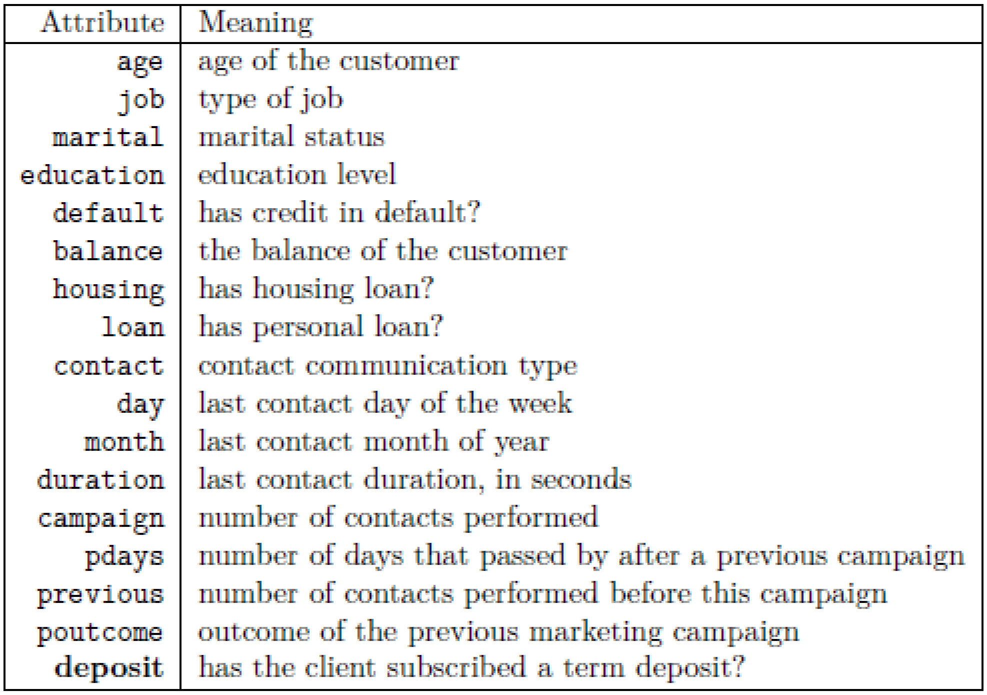
This report will present the description of the process and the result of descriptive statistics and machine learning analysis on the bank marketing dataset of a Portuguese banking institution. The results will help in making decisions by focussing on key parameters which affects the subscription of term deposit.

Executive Summary

After feature selection, data wrangling and processing both unsupervised and supervised learning algorithms are performed on the data. Decision trees algorithm shows the best accuracy of 77.7%. Analysis of the feature coefficients of the logistic regression reveals that balance and contacts performed prior to the campaign have significant effect on the objective (deposit ='Yes'). No loans and no defaults are also positive indicator that customer will subscribe. Age and marital status has no effect on likeliness. Education type and job type has some effect on the likeliness that a client will subscribe for term deposit or not.

Data Description

The data contains 16 feature columns and one label column, which is subscription to the deposit. The attributes meaning are:



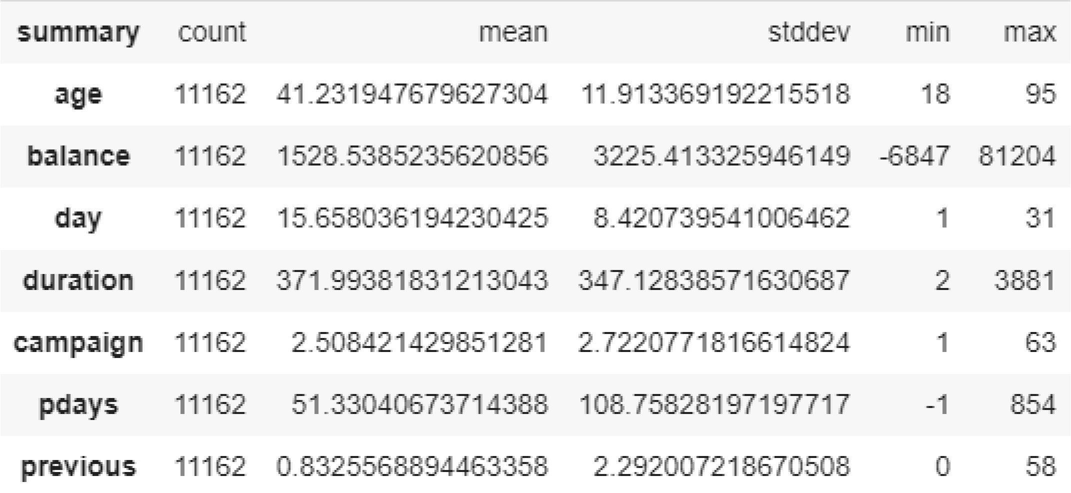
Attribute	Meaning
age	age of the customer
job	type of job
marital	marital status
education	education level
default	has credit in default?
balance	the balance of the customer
housing	has housing loan?
loan	has personal loan?
contact	contact communication type
day	last contact day of the week
month	last contact month of year
duration	last contact duration, in seconds
campaign	number of contacts performed
pdays	number of days that passed by after a previous campaign
previous	number of contacts performed before this campaign
poutcome	outcome of the previous marketing campaign
deposit	has the client subscribed a term deposit?

The features which contain numerical values are: ‘age’, ‘balance’, ‘duration’, ‘campaign’, ‘pdays’, ‘previous’.

The features which contain categorical values are: ‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’, ‘loan’, ‘contact’, ‘day’, ‘month’, ‘poutcome’.

Label column is: ‘deposit’ ,
which contains two labels: ‘YES’, ‘NO’

Numerical Features:



summary	count	mean	stddev	min	max
age	11162	41.231947679627304	11.913369192215518	18	95
balance	11162	1528.5385235620856	3225.413325946149	-6847	81204
day	11162	15.658036194230425	8.420739541006462	1	31
duration	11162	371.99381831213043	347.12838571630687	2	3881
campaign	11162	2.508421429851281	2.7220771816614824	1	63
pdays	11162	51.33040673714388	108.75828197197717	-1	854
previous	11162	0.8325568894463358	2.292007218670508	0	58

Feature	**Skewness**
age	0.86
balance	8.22
duration	2.14
campaign	5.5
pdays	2.4
previous	7.3

Age: age is positive skewed (right skewed) , therefore the data contains more customers less than 41.23 years of age.

Balance: balance is positive skewed (right skewed) , therefore the data contains more customers who has balance less than 1528.

Duration: as the duration is positively skewed, it can said that Contact duration with most of the customers is less than 372 seconds

Campaign: Most of the people have only been contacted once or twice.

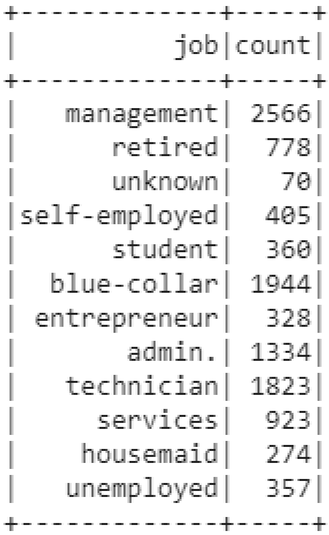
pdays: For most customers, it has been less than 51 days since the pas campaign.

Previous: mean of 0.83 and positive skewness tell us that most of the customers have not been contacted before this campaign.

Categorical Features:

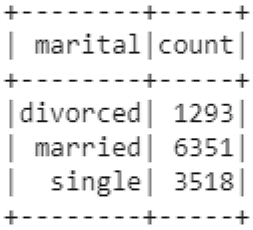
Descriptions of key categorical features:

Job: Contains 12 categories



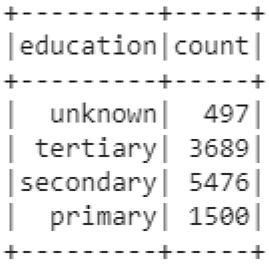
```
+-------------+-----+
|          job|count|
+-------------+-----+
|   management| 2566|
|      retired|  778|
|      unknown|   70|
|self-employed|  405|
|      student|  360|
|  blue-collar| 1944|
| entrepreneur|  328|
|       admin.| 1334|
|   technician| 1823|
|     services|  923|
|    housemaid|  274|
|   unemployed|  357|
+-------------+-----+
```

Marital: Contains 3 categories



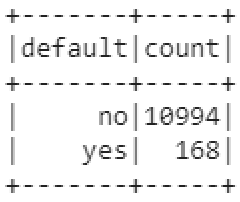
```
+--------+-----+
| marital|count|
+--------+-----+
|divorced| 1293|
| married| 6351|
|  single| 3518|
+--------+-----+
```

Education: Contains 3 categories



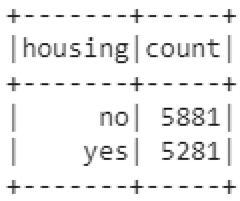
```
+---------+-----+
|education|count|
+---------+-----+
|  unknown|  497|
| tertiary| 3689|
|secondary| 5476|
|  primary| 1500|
+---------+-----+
```

Default: Contains 2 categories



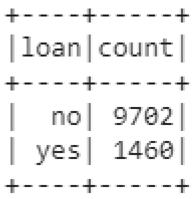
```
+-------+-----+
|default|count|
+-------+-----+
|     no|10994|
|    yes|  168|
+-------+-----+
```

Housing: Contains 2 categories



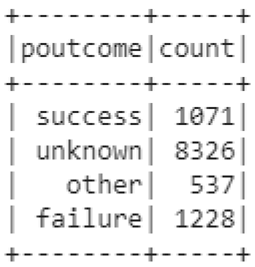
```
+-------+-----+
|housing|count|
+-------+-----+
|     no| 5881|
|    yes| 5281|
+-------+-----+
```

Loan: Contains 2 categories



```
+----+-----+
|loan|count|
+----+-----+
|  no| 9702|
| yes| 1460|
+----+-----+
```

Poutcome: contains 4 categories



```
+--------+-----+
|poutcome|count|
+--------+-----+
| success| 1071|
| unknown| 8326|
|   other|  537|
| failure| 1228|
+--------+-----+
```

Data Wrangling and Processing

Data wrangling and processing is performed to make the data suitable for a machine learning algorithm.

1. The desired features are selected from the data.
   a. New desired numerical features: ‘age’, ‘balance’, ’campaign’, ‘pdays’, ‘previous’
   b. New desired categorical features: ‘job’, ‘marital’, ‘education’, ‘default’, ‘housing’, ‘loan’, ‘poutcome’

2. Filtering is performed to remove the invalid and unknown values.

3. To feed the data into machine learning models categorical columns should be converted into format which could be understood by machine learning algorithms. For this ***OHE (One Hot Encoding)*** is performed on the data.
   The result feature vector is a sparse vector with 23 features.

4. ***Feature scaling*** is done by ***Min-max normalization*** to prevent the ***bias*** while performing the algorithm.

   The data is ready to be feed into machine learning algorithms

Machine Learning

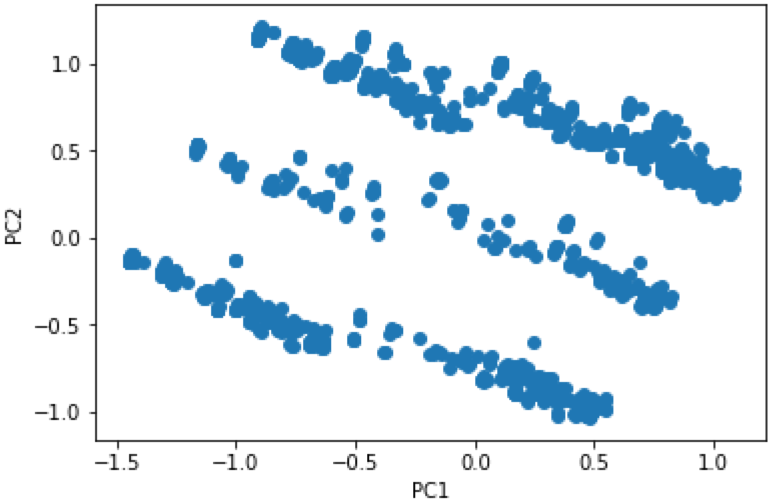
Unsupervised learning

In this we don’t know the output labels (ground truth) and we try to find the clusters in the data. K-means algorithm is used to find the clusters in the data and accuracy has been calculated.

The prediction accuracy with k-means is turned out to be 52%.

The accuracy can further be increased by reducing the dimensionality. PCA is performed to reduce the dimensionality. ***First two principal components capture feature variance of 41.2%.***

The scatter plot of first two components.



Plot is showing 3 clusters. However the first 2 principal components captures very less variance, hence the plot is not conclusive.

Supervised learning

Three supervised learning algorithms are performed. There accuracies are:

Algorithms	**Accuracy**
Logistic Regression	73.7%
Decision Tree	77.7%
Naïve Bayes Classifier	72.9%

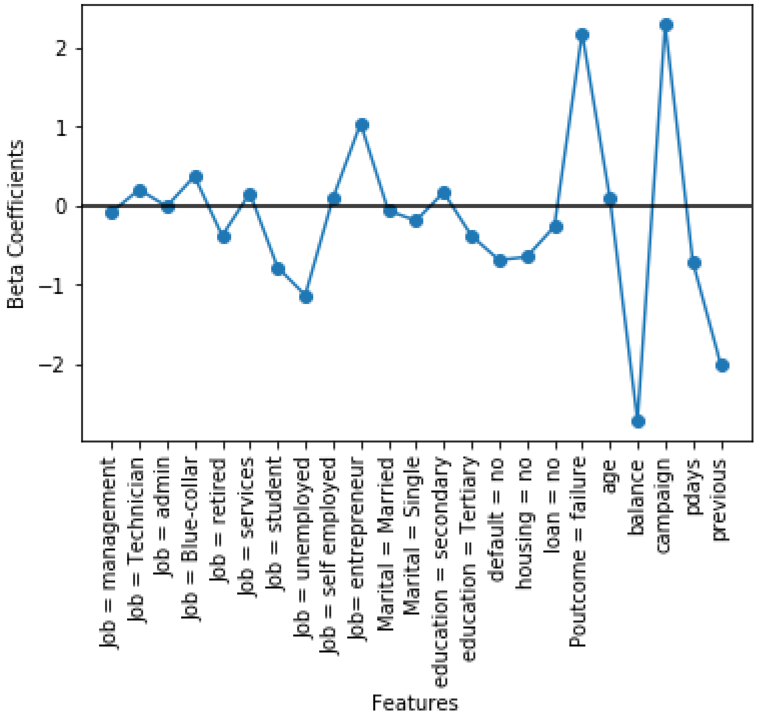
Decision tree gave the highest accuracy of 77.7%

Decision trees shows better predictions as compared to other supervised learning models because our data contains many categorical variables for which decision trees are proven to be more effective.

Logistic regression results can be improved by regularizing the logistic regression's cost function by using hyper parameters.

Important Features

Important features can be get by analyzing the feature coefficients of logistic regression.



Deposit = 'NO' has label 1

Deposit= 'YES' has label 0

This means that important features which affect the objective (deposit= 'Yes') are the ones with negative coefficients.

Results:

***Previous*: The higher the number of contacts performed earlier this campaign the higher the chance customer subscribe for term deposit.**

***Balance*: The higher the customer balance the higher the chance customer subscribe for term deposit.**

***Campaign*: The higher the number of contacts performed this campaign the higher the chance customer will not subscribe for term deposit.**

***Age*: Age has no effect on the deposit result.**

***Poutcome*: Previous campaign result highly affects deposit. If the previous campaign was success than the customer will subscribe.**

Loan: ***No personal loan prefers deposit as ‘Yes’***

Housing: No housing loan prefers deposit as ‘Yes’

Default: No default history prefers deposit as ‘Yes’

Marital: Marital status has little or no affect.

Education: Tertiary education prefers deposit as ‘Yes’

Job: Student and unemployed are likely to subscribe for term deposit.

Group Task Distribution

Harshdeep Singh (218318242)	**Shahzeb Maqsood (218474893)**	**Manthan Negi (218467657)**
Code:		
Data Distribution One Hot Encoding K-means PCA Plot Logistic Regression	Filtering Feature Scaling PCA Decision Trees	Feature Selection Naïve Bayes Classifier
Report writing and Analysis:		
Numerical Feature data description Unsupervised learning Important Features Executive Summary	Categorical Feature Data Description Supervised Learning	Introduction Data Wrangling and Processing

This assignment taught us about the functioning of spark RDDs, pyspark , pipeline RDD, spark in general. Assignment also taught us how to do data processing, especially One HOT Encoding (OHE).

We learned the practical applications of machine learning algorithms and how they can be used on real data to derive insights by finding key features and relationships and how these models can be used to predict outcomes.