|  |  |
| --- | --- |
| **Student IDs:** | 218546973, 218545396 |
| **Student Name**: | Anand Bhat and Pooja Bhat |
| **Campus:** | Burwood  Waterfront  Waurn Ponds  Warrnambool  Cloud |

|  |  |  |  |
| --- | --- | --- | --- |
| **Assignment Title:** | Assessment 2: Modern Data Science | | |
| **Due Date:** | 18 May 2019 by 11.59 PM | **Assessment Item:** | Report |
| **Course Code/Name:** | S777/ Master of Data Analytics | | |
| **Unit Code/Name:** | SIT742/ Modern Data Science | **Unit Chair /  Campus Coordinator:** | AsPr Gang Li |
| **Practical Group: (if applicable)** | Not applicable |

|  |  |
| --- | --- |
| If this assignment has been completed by a group or team:   1. Each student in the group must complete and sign a separate coversheet 2. The assignment will be returned to the student in the group nominated below | |
| Assignment to be returned to: (Student name and Student ID number) |  |

**PLAGIARISM AND COLLUSION**  
Plagiarism occurs when a student passes off as the students own work, or copies without acknowledgement as to its authorship, the work of another person. Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose with the intent of obtaining an advantage in submitting an assignment or other work. Work submitted may be reproduced and/or communicated for the purpose of detecting plagiarism and collusion.

**DECLARATION**I certify that the attached work is entirely my own (or where submitted to meet the requirements of an approved group assignment, is the work of the group), except where work quoted or paraphrased is acknowledged in the text. I also certify that it has not been previously submitted for assessment in this or any other unit or course unless permissions for this has been granted by the Unit Chair of this unit. I agree that Deakin University may make and retain copies of this work for the purposes of marking and review, and may submit this work to an external plagiarism-detection service who may retain a copy for future plagiarism detection but will not release it or use it for any other purpose.

|  |  |  |  |
| --- | --- | --- | --- |
| Signed: | Anand Bhat and Pooja Bhat | **Date:** | 11.05.2019 |
| An assignment will not be accepted for assessment if the declaration appearing above has not been signed by the author. If submitting electronically, print your full name in place of a signature. | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **COMMENTS** | | | | | |
| **Mark Awarded:** |  | **Assessor’s Signature:** |  | **Date:** |  |

### Overview

### In this assignment, we are presented with a dataset about a marketing campaign of a Portuguese bank. Our goal is analyze the provided data and derive meaningful insights into what are the different attributes that influenced the subscription of term deposits by customers from this campaign in order to design strategies for future marketing campaigns.

### Data Distribution

An initial look at the data has a combination of categorical and numerical attributes. There are a total of 11162 records in the dataset. Age, balance, day, duration, campaign, pdays and previous are numeric fields. Job, marital, education, default, housing, loan, contact, month, poutcome and deposit are the categorical fields in the dataset. We will firs get the summary statistics of the dataset to see how the data is distributed.

Summary Statistics of Numerical variables:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **summary** | **age** | **balance** | **day** | **duration** | **campaign** | **pdays** | **previous** |
| **count** | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 |
| **mean** | 41.23 | 1528.54 | 15.66 | 371.99 | 2.51 | 51.33 | 0.83 |
| **stddev** | 11.91 | 3225.41 | 8.42 | 347.13 | 2.72 | 108.76 | 2.29 |
| **min** | 18 | -6847 | 1 | 2 | 1 | -1 | 0 |
| **max** | 95 | 81204 | 31 | 3881 | 63 | 854 | 58 |

Table 1: Summary statistics of numerical variables

Summary Statistics of Categorical variables:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **summary** | **job** | **marital** | **education** | **default** | **housing** | **loan** | **contact** | **month** | **poutcome** | **deposit** |
| **count** | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 | 11162 |
| **mean** | null | null | null | null | null | null | null | null | null | null |
| **stddev** | null | null | null | null | null | null | null | null | null | null |
| **min** | admin. | divorced | primary | no | no | no | cellular | apr | failure | no |
| **max** | unknown | single | unknown | yes | yes | yes | unknown | sep | unknown | yes |

Table 2: Summary statistics of numerical variables

From the summary statistics it is evident that there are some variables with “unknown” values in the categorical fields which we will need to remove in-order to have a clean data for our analysis.

We will use spark sql to select the fields such that they do not contain any invalid data as instructed in the task description and load into a data frame.

Next, we will try to explore the data to see if it is obvious if any of these features directly influence the variable of interest- deposit.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Figure 1 : Bar plots to show relationship of features with deposit

### Data Wrangling

We will need to transform this data in a manner that will be in a format that can be fed into supervised and unsupervised learning algorithms to arrive at meaningful insights.

For the dataset here, we will first need to convert the values in the categorical fields into numerical values. This has been done using OneHotEncoding. Next we have assembled the encoded features data into a single features column. The deposit field is also a categorical field which has been encoded as a label field. These features and label fields have then been normalized using MinMaxScaler. The final data frame has the attributes in the values between 0 and 1

### Performance of Supervised and Unsupervised learning

We have used the prepared dataset to run various classification methods supervised and unsupervised learning.

Un-Supervised Learning:

K-Means: We have used k-mean algorithm and Principal Component analysis methods for unsupervised learning. Since k-means is meant to cluster, rather than to compare clusters to known groups, this might not be the best way to evaluate the accuracy of model performance. For the purpose of analysis in this case, we will try to compare the predictions from k-means with the actual known deposit values. Here is a summary of the comparison of the clustering output with the actual deposit values.

|  |  |  |
| --- | --- | --- |
| **count\_matching** | **total\_count** | **percentage\_predicted\_correctly** |
| 1136 | 2181 | 52.08 |
| **count\_not\_matching** | **total\_count** | **percentage\_predicted\_incorrectly** |
| 1045 | 2181 | 47.91 |

Table 3: Summary of k-means analysis

PCA- PCA gives us a visual representation of how the data is scattered. This is a very good method of unsupervised learning and provided

Supervised:

The important features

Possible reasons for obtaining these analysis

Group activities:

Our group was made up of Anand Bhat and Pooja Bhat. We distributed our tasks between ourselves and we separately researched the various methods of supervised and un-supervised learning. We then shared our learnings with each other such that we both understand the concepts that are key to this assignment. The python code was written by both of us separately and then we picked the most concise code to create the final python notebook. We have learnt a lot of new Machine learning concepts as a part of this exercise and we hope to implement them soon for some other real-world data analysis.