**Pramod Bhandari**

**Assignment 1 Solutions**

Q1. *Domain knowledge. Based on the attribute information, which attributes, do you think, are most likely to be able to predict the outcome? Discuss two of such attributes based on your business knowledge or common sense.*

The fields that are most likely to be able to predict the outcomes are, age, job, marital, education, default, loan, previous.

I think age will be one of the important factors to decide whether someone will subscribe a term deposit or not. Old people more than 60 years of age and people younger than 25 may not be much inclined towards term deposit.

The other attribute that can predict the outcome is a default. If people already have a default, they will not have money to save, hence there will be no need for a term deposit.

*Q2. Transform the dataset using one-hot encoding to encode the discrete variables. Split the dataset into a training set (80%) and a testing set (20%). Summarize the shapes of both partitions.*

Below is the summery of the shape of both partitions:

|  |  |
| --- | --- |
| **Data** | **Shape** |
| Training (x) | (14843, 58) |
| Training (y) | (14843, 1) |
| Testing (x) | (3711, 58) |
| Testing (y) | (3711, 1) |

*Q3. Rank the discrete features using Chi-squared test (use only the training set). With a criterion of p < 0.05, which features are selected?*

Below is the ranked list of discrete features using Chi-squared test.

|  |  |  |
| --- | --- | --- |
| **feature** | **chi2** | **p** |
| poutcome=2 | 1541.35 | 0 |
| contact=1 | 407.15 | 0 |
| month=5 | 347.46 | 0 |
| month=8 | 340.46 | 0 |
| month=9 | 283 | 0 |
| default=1 | 225.29 | 0 |
| contact=0 | 203.04 | 0 |
| month=6 | 189.42 | 0 |
| job=5 | 169.7 | 0 |
| job=8 | 146.26 | 0 |
| poutcome=1 | 138.76 | 0 |
| job=1 | 133.62 | 0 |
| month=0 | 126.23 | 0 |
| month=2 | 92.89 | 0 |
| default=0 | 54.45 | 0 |
| marital=2 | 51.61 | 0 |
| education=2 | 48.58 | 0 |
| education=6 | 47.81 | 0 |
| month=3 | 30.23 | 0 |
| marital=1 | 22.53 | 0 |
| job=7 | 22.29 | 0 |
| job=0 | 20.32 | 0 |
| education=1 | 18.45 | 0 |
| poutcome=0 | 16.25 | 0 |
| day\_of\_week=1 | 13.78 | 0 |
| day\_of\_week=2 | 9.21 | 0.002 |
| job=10 | 8.69 | 0.003 |
| education=7 | 8.58 | 0.003 |
| month=7 | 6.32 | 0.012 |
| job=2 | 5 | 0.025 |
| education=0 | 4.5 | 0.034 |
| month=1 | 3.43 | 0.064 |
| education=4 | 2.69 | 0.101 |
| housing=2 | 2.36 | 0.124 |
| housing=0 | 2.35 | 0.125 |
| month=4 | 1.24 | 0.266 |
| day\_of\_week=0 | 1.1 | 0.294 |
| day\_of\_week=3 | 0.82 | 0.364 |
| education=3 | 0.81 | 0.368 |
| default=2 | 0.66 | 0.415 |
| marital=0 | 0.65 | 0.42 |
| day\_of\_week=4 | 0.55 | 0.457 |
| job=9 | 0.5 | 0.481 |
| job=6 | 0.49 | 0.482 |
| job=3 | 0.32 | 0.573 |
| education=5 | 0.29 | 0.592 |
| loan=1 | 0.26 | 0.609 |
| housing=1 | 0.26 | 0.609 |
| marital=3 | 0.17 | 0.677 |
| loan=2 | 0.16 | 0.693 |
| loan=0 | 0.07 | 0.797 |
| job=11 | 0.01 | 0.927 |
| job=4 | 0 | 0.946 |

The selected discrete features are,

job=0,job=1,job=2,job=5,job=7,job=8,job=10,marital=1,marital=2,education=0,education=1,education=2,education=6,education=7,default=0,default=1,contact=0,contact=1,month=0,month=2,month=3,month=5,month=6,month=7,month=8,month=9,day\_of\_week=1,day\_of\_week=2,poutcome=0,poutcome=1,poutcome=2

*Q4. Rank the continuous features using F test (use only the training set). With a criterion of p < 0.05, which features are selected?*

Below is the ranked list of continuous features using F test.

|  |  |  |
| --- | --- | --- |
| **feature** | **f** | **p** |
| duration | 4233.74 | 0 |
| pdays | 1938.3 | 0 |
| previous | 1096.91 | 0 |
| campaign | 131.42 | 0 |
| age | 20.5 | 0 |

The selected continuous features are,

age, duration, campaign, pdays, previous

*Q5. Train a decision tree classifier (with a fixed random state) using all of the features in the training data. What are the top 10 most important features? Are the top ranked discrete features similar to the results from Q3? What is the F1 score of this model when it is used to predict the testing set?*

The top 10 most important features are

|  |
| --- |
| **feature** |
| duration |
| age |
| pdays |
| campaign |
| month=6 |
| month=5 |
| contact=0 |
| month=4 |
| month=8 |
| month=0 |

The top ranked discrete features are different than the results from Q3.

The F1 score of this model for training set is, F1= 0.669

*Q6. Train a decision tree classifier (with a fixed random state) using the features selected from Q3 and Q4 in the training data. What is the F1 score of this model when it is used to predict the testing set? Is there a large difference between this F1 score and the one you get from Q5? If there is, why?*

The F1 score of this model when using features selected from Q3 and Q4 is, F1= 0.656

There is not a large difference between the F1 score and the one from Q5.

*Q7. Like Q5, train a decision tree classifier (with a fixed random state) and limit the max features to 58(modified to match the max columns available), record the F1 score on the testing set. Train another decision tree classifier (with the same random state) and limit the max features to 30, compare its F1 with that of the first classifier. What is the difference? Which classifier would you use? And why?*

F1 (with 58 features) = 0.6695

F1 (with 30 features) = 0.6769

The F1 score with less features is greater with a difference of 0.0073

I will use the classifier with 30 features selected because it has higher F1 score.

***Extra Credit (Additional 10%)***

*Use recursive feature elimination (RFE) to select the optimal feature subset. Evaluate this subset using another classifier and the testing data. Based on the cross validation performance, would you recommend using only this subset for prediction? (Caution: RFE can take a long time to run.)*

I wrote the code and run it for hours but it didn’t complete. So, wasn’t able to summarize the results. I have included the ipython notebook.