A Project of Applying

Different Multivariate Methods on

Bank Telemarketing Data

Submitted to

Dr. Megan Orr

Submitted By

Sharmin Hossain

Student ID: 1337949

Index

|  |  |
| --- | --- |
| Introduction | 3 |
| Data Description and Visualization | 5 |
| Research Question 1 | 9 |
| Research Question 2 | 10 |
| Research Question 3 | 13 |
| Research Question 4 | 14 |
| Conclusion | 17 |
| Reference | 18 |
| SAS Code and Output | 19 |

1.Introduction

The data I have worked on came from direct marketing campaigns of a Portuguese banking institution. Their research goal was to predict if clients will subscribe after their telemarketing.

The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The business purpose of the research was finding out the characteristics that are helping Bank to make customers successfully subscribe for deposits, which helps in increasing campaign efficiently and selecting high value customers.

Data set is taken from UCI Machine Learning repository. There is total 41188 Rows in the dataset.

**Attribute Information:**

There is total 21 columns in the main dataset.

For my project purpose, I have worked with 10 of them. Here is the brief description of those 10 variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Serial** | **Variable Name** | **Variable Type** | **Variable Description** |
| 1 | Age | Numeric |  |
| 2 | Job | Categorical | Type of Job: admin, blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown |
| 3 | Marital | Categorical | divorced, married, single, unknown |
| 4 | Education | Categorical | basic.4y, basic.6y, basic.9y, high. School, illiterate, professional. Course, university. Degree, unknown |
| 5 | Contact | Categorical | cellular, telephone |
| 6 | Duration | Numeric | Last Contact Duration |
| 7 | Campaign | Numeric | Number of contacts performed during this campaign and for this client |
| 8 | ConsumerPriceIndex | Numeric | Consumer Price Index (Monthly Indicator) |
| 9 | ConsumerConfidenceIndex | Numeric | Consumer Confidence Index (Monthly Indicator) |
| 10 | Y | Binary | Has the client subscribed a term deposit? (Yes/No) |

From this dataset, the answers I am trying to get are:

1. Are all variables equally important for this research? Or can we eliminate some of them using ‘Stepwise’ procedure.
2. What kind of classification can be used in the dataset if we consider “Job” as a grouping variable? Does error rate change if we use both resubstitution and cross validation method?
3. Using different MANOVA methods, can we decide on the equality of population mean vector of numeric variables?
4. How many factors can be used to explain the total variances of those variables? Does rotating can be used to interpret the factors properly?

2.Data Description and Visualization

For Categorical variable, from Figure 1, we can see that admin, blue-collar and technician consists of almost 65% of total frequencies of overall population.

Table

Description automatically generated

Figure 1: Descriptive Statistics of Categorical Variable

For numerical variables, as we can see from Figure 2 and 3, Age and ConsumerPriceIndex seem to be normally distributed whereas other variables are not. The variance for variables ranges from 0.33 to 67225.73 which seem to be a huge difference.

Table

Description automatically generated

Figure 2: Descriptive Statistics of Numerical Variables

Graphical user interface, diagram, application, Word

Description automatically generated

Figure 3: Graphical representation of Numerical Variables

If we make a descriptive summary of group wise numerical variables, we can see from Figure 4 & 5 that there are many outliers for duration and campaign variable. For age, there are very few numbers of outliers and consumer price index & consumer confidence index seem more stable in terms of mean, median and percentiles for every job sector.

Table

Description automatically generated

Figure 4: Descriptive Statistics of Numerical Variables using Job as a grouping variable

Graphical user interface

Description automatically generated

Figure 5: Graphical representation of Numerical Variables using Job as a grouping variable

3.Methods and Results:

3.1. Research Question 1:

Are all variables equally important for this research? Or can we eliminate some of them using some procedure?

**Method Used:** Stepwise Procedure.

**Method Description:** In many applications, there are many dependent variables available for analysis and researchers are interested in discarding redundant variables for separating the groups. Stepwise Procedure is a combination of forward selection and backward elimination. Variables are added one at a time. At each step, the variables are reexamined to see if any previously selected variable has become redundant in the presence of the recently added variables. The procedure stops when the largest partial F-statistic among the remaining variables available for entry fails to exceed the preset significance threshold.

**SAS Interpretation:** Applying Stepwise procedure in SAS, we can see that there were 6 steps in total before getting into conclusion for this dataset. And no variables were removed from the procedure. So, age, consumer confidence index, consumer price index, duration, campaign- these 5 variables will be used for the further procedures.

Table

Description automatically generated

Figure 6: Stepwise Selection Summary

3.2. Research Question 2:

What kind of classification can be used in the dataset if we consider “Job” as a grouping variable? Does error rate change if we use both resubstitution and cross validation summary?

**Method Used:** Discrim Procedure

**Method Description:** Given a set of observations that contains one or more quantitative variables and a classification variable which indexes groups of observations, the DISCRIM procedure develops a discriminant criterion to classify each observation into one of the groups. If we are not sure of the equality of population covariance matrices, we can use a Chi-square test to identify if we should use Linear discriminant function or quadratic function to classify the observations.

To judge the ability of classification procedures to predict group membership, the probability of misclassification, or error rate, is usually used. We have used two types of error rate here – one is called resubstitution and another one is cross validation method.

Using resubstitution method, the classification rule is applied to each observation vector and this observation is assigned to a group. Then, the number of misclassifications is counted. The proportion of misclassifications resulting from resubstitution is call the apparent error rate.

In the cross-validation method, all but one observation vector is used to determine the classification rule and then the omitted observation is classified into one of the groups using the classification rule based on the N-1 observations. This procedure is repeated until every observation has been classified in this manner.

**SAS Interpretation:** As the Discrim procedure shows (Figure 7) the p value of chi-square test is less than 0.0001, we can safely say that quadratic classification was applied to this dataset.

Graphical user interface, application, Word

Description automatically generated

Figure 7: Test of Homogeneity of within covariance Matrices

From Figure 8 & 9, we can see that there is not much difference between resubstitution & cross validation method in terms of overall error rate. All the variables also showed similar error rates individually.

Graphical user interface, application, table

Description automatically generated

Figure 8: Resubstitution Summary

Graphical user interface, table

Description automatically generated

Figure 9: Cross-validation Summary

3.3. Research Question 3:

Using different MANOVA methods, can we decide on the equality of population mean vector of numeric variables?

**Method Used:** GLM Procedure.

**Method Description:** I used this process to test for significant difference between at least two mean vectors. The hypothesis I want to test is

H0: µ1= µ2=………. = µk vs H1=at least two population mean vector differ

From GLM procedure, we can test this hypothesis by 4 different methods. (Wilks’ lambda, Pillai’s trace, Hotelling-Lawley Trace & Roy’s greatest root)

**SAS Interpretation:** We can see from the hypothesis test that for all methods mentioned here (Wilks’ lambda, Pillai’s trace, Hotelling-Lawley Trace & Roy’s greatest root), the P-values are significant. So, we can reject our null hypothesis and conclude that there is significant difference among the mean vectors of population variables. We can also rank the variables based on their coefficient of 1st discriminate function as:

Age>ConsumerPriceIndex>ConsumerConfidenceIndex>Campaign>Duration

Table

Description automatically generated

Figure 10: Applying MANOVA Method

3.4. Research Question 4:

How many factors can be used to explain the total variances of those variables? Does rotating can be used to interpret the factors for better?

**Method Used:** Principal Component Method in the Factor Procedure & Orthogonal Rotation using Varimax.

**Method Description:** Principal component method is a factor extraction method used to form uncorrelated linear combinations of the observed variables. The first component has maximum variance. Successive components explain progressively smaller portions of the variance and are all uncorrelated with each other. Principal components analysis is used to obtain the initial factor solution.

Varimax Method is an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors.

**SAS Interpretation:** 4 factors will be retained for this dataset.

Before rotating, we can see that duration can be explained by Factor 3.

Campaign, Age & Consumer Confidence Index is associated with Factor1, Factor2 & Factor 4 almost equally.

Consumer Price Index can be explained by both Factor 1 & Factor 3.

Table

Description automatically generated

Figure 11: Factor Procedure before Rotation

After rotating, we can see that duration is still explained by Factor 3.

Campaign & Consumer Price Index are associated with Factor 1.

Age is explained by Factor 4.

Consumer Confidence Index is associated with Factor 2.

So, except for duration, every variable association with factors are more interpretable after rotation. Also for duration, the proportion gets better after applying rotation.

Graphical user interface

Description automatically generated with medium confidence

Figure 12: Factor Procedure after Rotation

4.Conclusions:

From our multivariate analysis, we can see that all the numeric variables will be used in the process we have done so far. We can use quadratic discriminant function to classify the observations though error rates will be almost similar for resubstitution & cross-validation method. From one-way MANOVA, we can see that at least two population mean vectors differs significantly using four different methods. And four factors will be used to explain the variables. After rotation, the interpretation of factor & variable association will be more straightforward.

5.Reference:

1. UCI Machine Learning Repository: <https://data.world/uci/bank-marketing>
2. SPSS Statistics Documentation:

<https://www.ibm.com/docs/en/spss-statistics/24.0.0?topic=option-factor-analysis>

1. Performing Exploratory Data Analysis with SAS and Python:

<https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-with-sas-and-python/>

1. SAS Help Center:

<https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/casref/p1k1no304sy91on1qjvks0ggelb5.htm>

1. Data Driven Approach to Predict Success of Bank- Marketing:

<https://medium.com/mlearning-ai/data-driven-approach-to-predict-success-of-bank-marketing-31791cad8f81>

6.SAS Code & Output:

/\*Reading Text File\*/

**DATA** banknew;

INFILE 'banknew2.txt' dlm=',' firstobs=**2** lrecl=**32767**;

INPUT job$ marital$ education$ contact$ duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex y$;

/\*Categorical Variable\_Descriptive Statistics\*/

**proc** **freq** data=banknew;

tables job;

**run**;

/\*Numerical Variable \_Descriptive Statistics\*/

**proc** **means** data=banknew mean median mode std var min max;

**run**;

/\*Categorical+ Numerical Variable \_Descriptive Statistics\*/

**proc** **means** data=banknew mean median mode std var min max;

class job;

var duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;

**run**;

/\*Scatter Plot by Group\*/

**proc** **sgscatter** data=banknew;

matrix duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex /diagonal=(histogram kernel);

**run**;

/\*Stepwise Procedure\*/

**PROC** **STEPDISC** STEPWISE;

CLASS job;

**Run**;

/\*Classification Analysis'\*/

**PROC** **DISCRIM** LIST CROSSVALIDATE POOL=TEST;

CLASS job;

VAR duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;

**RUN**;

/\* MANOVA \*/

**PROC** **GLM**;

CLASS job;

MODEL duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex=job;

MANOVA H=job/PRINTE PRINTH MSTAT=EXACT;

**RUN**;

/\*Principal Component Analysis of Factor Method\*/

**proc** **factor** method=prin rotate=varimax proportion=**0.8** corr;

var duration campaign age ConsumerPriceeIndex ConsumerConfidenceIndex;

**run**;