Florida Atlantic University
COLLEGE OF BUSINESS

ISM 6136 – Datamining/Predictive Analytics

Class Assignment 1
5 points

Task: Apply Principal Component Analysis (PCA) to reduce dimensions and predict the sales price for Toyota cars listed in the 'new data'.

Dataset: Download the dataset Class Assignment 1 – Toyota.xls from Canvas and perform the following steps.

Procedure to be followed:

Follow the following data mining steps:

1. Understand the problem and purpose of data mining task
2. Obtain the dataset for analysis
3. Explore, clean and preprocess data
    - *Perform 'Missing Data Handling'.*
    - *Create dummies for any categorical data (use 'Transform Categorical Data') otherwise PCA will not work.*
    - *Explore the data with Scatterplot Matrix.* Provide screen shot of the 'scatter plot matrix'. You can do two matrices or more (otherwise Excel might run forever) – first one - outcome by first 5 predictors and then second one – outcome and remaining set of predictors.
    - Reduction of data dimension – **Use PCA technique** to reduce the dimensions. Select top 5 dimensions/PCA components (PC1, PC2, PC3, PC4, PC5) and the Smallest number that show strong variance *(When you do PCA - Smallest# that give 95% variance ...you might get upto 34 or 36 predictors ....that will be too much to go through so you can **select the top 15 predictors** in that case).*

Highlight those variables and their weights on the PCA Components table (PCA Output Tab)

4. Determine the appropriate data mining task
5. Partition the data – *(using only the PCA components and Price (output))*
6. Choose the appropriate data mining techniques/algorithm – *under 'Predict' tab.*
7. Build the model by interpreting results of algorithm *(Predict > linear regression - (using only 5 dimensions as input and price as output))*
8. Try 3 models for each (Top 5) and (Top 15) – that is total 6 models – select the best one – highlight/name the best model output Excel tab
9. Deploy the best model - *(Use Score tab)*
10. Upload your Excel file to Canvas
11. On a separate Word document – explain the following with screen shots:

    a. How did you clean, explore the data? Were there any categorical variables in the dataset?
    Explain the scatter plot matrix results in terms of correlation amongst predictors and with the outcome. Provide screen shot of the matrix.

    b. Which top 5 variables did you select using PCA technique and which Smallest # gave you max. variance, how much variance did the Top 15 actually cover?

    c. How did you select the best model - explain the 'full selection criteria' you followed for comparison between the two models? (Refer to the Lecture 2 slides for 'full model selection criteria' for Linear Regression)
    Prepare a Table showing – all the models, PCA, Partitioning, Training Validation RMSE, R squared.

13. Submit only 'one' Excel file with all 6 models – highlight the 'best' model tab and the Word document for the explanation of #12.