ETX2250/ETF5922 Data Visualisation and Analytics

Assignment 1: Analytics – *Flights*

Submission instructions

This assignment comprises 15% of the assessment in ETX2250 and ETF5922.

Your assignment submission will consist of a pdf document. The document must include both your code and graphical and other output as well as paragraph answers that provide description and discussion of the output. The presentation of graphs is important. In particular, headings and labels should be provided when necessary.

In your completed assignment, graphs should be easy to read, with appropriate use of headings, colour, formatting and text.

The pdf document must be submitted as a hard copy in class, and also must be uploaded to Moodle.

The pdf document should be titled <studentid>_*A2.pdf*

(Note: No spaces, No Names, No other characters, No extra characters)

The assignment is due before class on Monday 4th February 2019 (9.00 am), and a hard copy of the pdf document should be submitted in class, securely stapled in the top left corner and with a signed COVER SHEET on top.

If for some reason you are unable to submit the assignment personally, there will be an assignment box available on level 5 of Building H to place it there **before** the due time.

The cover sheet is available at URL:

https://www.monash.edu/__data/assets/word_doc/0004/903379/assignment-cover-sheet-fbe-1.doc you must supply the unit code and further details.

Upload your assignment to Moodle as follows:

- Go to the *Assignments* section
- Click on *Submission of Assignment 2: Analytics*
- Click on *Add submission*
- Drag and drop the file to submit it
- *Save changes*

To confirm that your upload was successful, go to the *Assignments* section, and click on the *Assignment* 2: Analytics link. The uploaded filename will be shown.

Retain your marked assignment until after the publication of the final results for this unit.

Assignment 2: Introduction

This assignment relates to the Analytics component of the unit.

The data-set is a subset of all flights departing and leaving the two major airports in Chicago, USA in the year 2008. The task is to investigate why some flights are getting delayed and build a predictive model, which flights will be delayed. This assignment will run you through the first steps of any Business Analytics projects you plan on undertaking.

Assignment 2 Part A: Explore the data (4 Points)

The data is stored in a database (as in any company you will work). The fields are:

1. **flightstatus:** 0 – normal flight 1- delayed
2. **carrier:** unique carrier code
3. **dayofweek:** 1 (Monday) - 7 (Sunday)
4. **month:** 1-12
5. **departuretime:** actual departure time (local, hhmm)
6. **destination:** destination IATA airport code
7. **origin:** origin IATA airport code
8. **weatherdelay** in minutes: Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.

a) Connect to the data (1 Point)

The data is stored in the database **summer2019** at the server with the IP **118.138.234.161.** You can access this postgres database with the username **student** and the password **4XcxqUo6AHPn.** First, create a connection with the database by reading the username and password from a csv file where you store the username and the password. (Hint: Remember Workshop02)

b) Loading data (0.5 Points)

Load the data of the table **flight_delay_chicago** in the schema **public** into R. Print out the first few rows.

c) Summary statistics (1 point)

Calculate the mean by the variable of interest flightstatus for all numeric columns with SQL, what do you notice?

d) Correlations (1 Point)

Create a correlation plot between all numeric variables by dropping all rows with missing values. Set all values in the variable **weatherdelay** to 1 if the delay is greater than 0 for this exercise. Split or analysis by **origin** airport. Display the correlation coefficients on the plots.

e) Interpretation (0.5 Point)

Analyse the first results from the correlation plots in relation to the variable **flightstatus** from d)

Assignment 2 Part B: Cluster the data (4 Points)

For this exercise, the goal is to cluster the data with **kmeans** based on **dayofweek**, **month** and **departuretime** and analyse the results.

a) **Normalize the departuretime by reducing the accuracy to an hour, e.g. 2215 becomes 22 (0.5 Points)**
b) **Write the code to iterate over different numbers of clusters. Graph your result of the appropriate metrics and justify your choice of k. (1 Point)**
c) **Use your chosen k as the dependent variable and build a decision tree classifier to explain the cluster assignment. Print out the tree as well as the decision rules. (1.5 Points)**
d) **Compare the kmeans clustering results across by cluster groups by averaging across all variables. What variables are important according to the decision tree ? (1 Point)**

Assignment 2 Part C: Build a predictive model (4 Points)

The goal of this exercise is to build a classifier that can tell you if a flight is most likely be delayed. For this, we split the data into a training set with all the data up to September, and use October, November and December as our test set. As we cannot know the delay of the weather before it happened, **recode** it as 1 if the delay is **greater than 0**. We will say in reality, we would look at the weather forecast and assign a 1 if the weather conditions are worrisome.

a) Cleaning the data and preparation (1 Point):

- Filter out all rows where the **destination** airport occurs less than 100 times.
- Filter out all rows where the **carrier** occurs less than 100 times.
- Filter out all rows with missing entries
- Recode the variable **weatherdelay** as a factor with the levels 0 and 1

b) Building (1 Point)

Split the data into a train and test set by only selecting the variables: **flightstatus, carrier, dayofweek, departuretime, destinaiton, weatherdelay.** Build a decision tree on the training test and predict on the test set (Remember the months). Print out the decision tree as well as the rules. (1 point)

c) Plot the ROC and Precision-Recall curve on the test-set. What does it tell you? (1 Point)

d) Explain the result. What are the important predictors according to the tree? How does it compare to the clustering result? Why did we exclude the month variable? (1 Point)

Overall presentation: 1 point