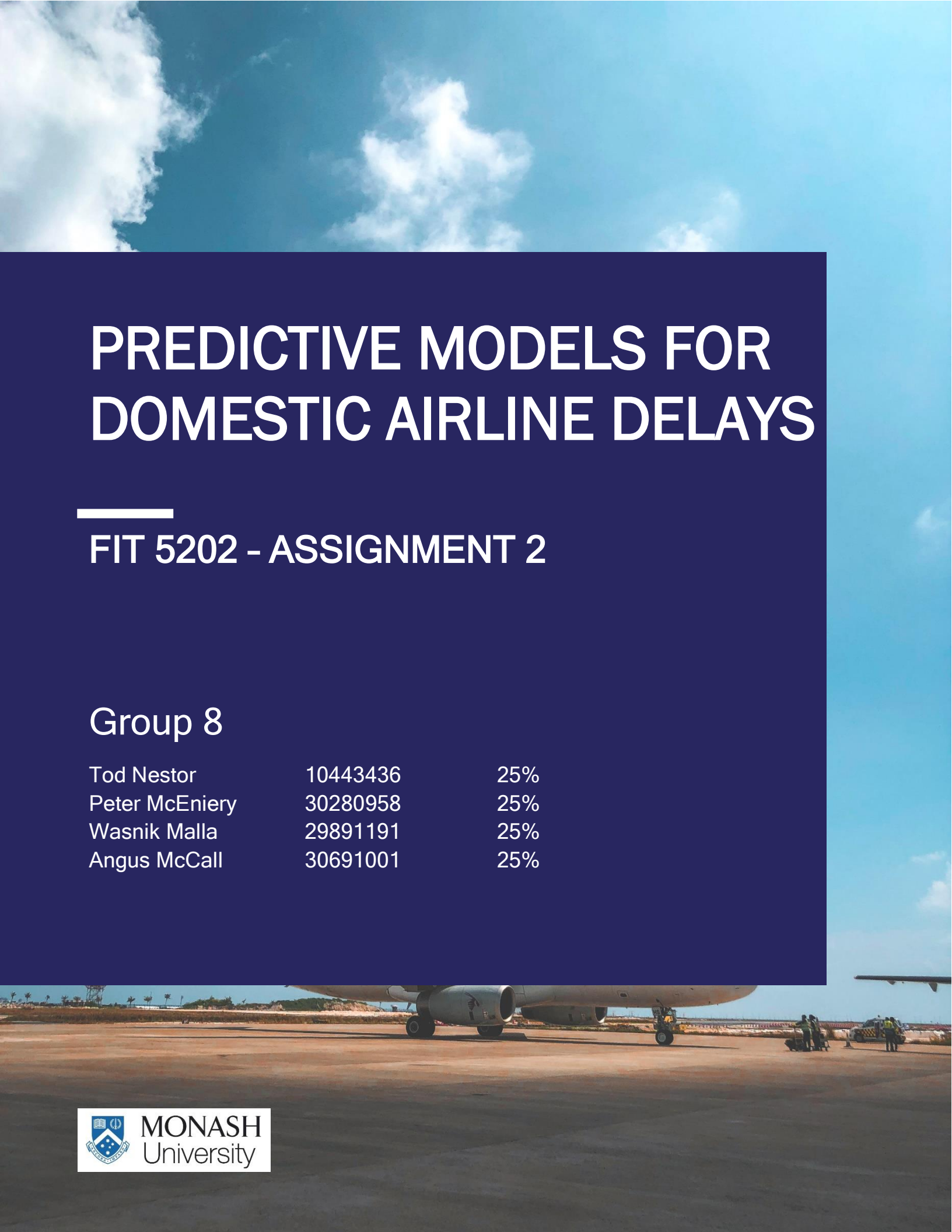

PREDICTIVE MODELS FOR DOMESTIC AIRLINE DELAYS

FIT 5202 – ASSIGNMENT 2

Group 8

Tod Nestor	10443436	25%
Peter McEniery	30280958	25%
Wasnik Malla	29891191	25%
Angus McCall	30691001	25%

MONASH University

Table of Contents

First phase

Introduction

Flying business class would be one of the most luxurious ways to travel. Being waited on hand and foot, room enough to take part in the complimentary yoga class and seats as soft as the clouds you soar above. A vivid contrast awaits passengers merely fifteen metres and a thin curtain of fabric away in economy: shoulder to shoulder seating, one blanket per row and an unspoken rule of "we're all in this together. Please… please don't recline your seat"

The above shows two completely different worlds, but the great equaliser in air travel is flight delays. Be you first class, economy or even pilot, flight delays effect all those travelling and the airlines themselves.

This project will analyse Australian flight delay data to determine correlations between airports, destinations and monthly weather cycles to create a predictive tool that enables to user to see delay possibility given their route and seasonal changes.

Business case

Flight delays can be much more than an annoyance to passengers. Delays in travel can caused missed joining flights and missed business meetings for the passenger. It also generates additional costs for the airline through passenger reimbursements, wages and fuel if mid-air delays. A study done to examine the economic impact of flight delays in the US market estimated that it cost over $30 billion per year. (Ball et al., 2010)
With a tool such as this, an airline can see any patterns in their delays and search for the reasoning behind the data. This project will look at possible weather correlations however there could be many more underlying factors causing these delays.

Motivation

A predictive tool will enable the informed traveler to apply market pressure on under performing carriers and airports. If limited to delay prone options, it will also allow the traveller to plan around possible flight delays.

Data to be studied

Analysis will be done at two data sets to determine models that predict whether or not a particular flight will be late.

The first dataset is publicly available to freely download as a CSV file from the below address.

https://data.gov.au/data/dataset/domestic-airline-on-time-performance

It contains 84777 records of 15 features as seen below.

Feature	Description
Route Departing_Port Arriving_Port	Route labelled by departure-arrival cities, e.g. (Melbourne-Sydney) Departure City, e.g. Melbourne Arrival City, e.g. Sydney
Airline	Domestic Carrier Name e.g. Qantas
Month	Month in Mmm-YY format, e.g. Jan-04
Sectors_Scheduled Sectors_Flown	Monthly Aggregate of Scheduled Flights between particular departure and arrival airports Monthly Aggregate of Actual Flights between particular departure and arrival airports
Departures_On_Time Arrivals_On_Time	Monthly aggregate of the number of on-time departures Monthly aggregate of the number of on-time arrivals
Departures_Delayed Arrivals_Delayed	Monthly aggregate of the number of delayed departures (>15 minutes) Monthly aggregate of the number of delayed arrivals (>15 minutes)
Year Month_Num	Year in YYYY format e.g. 2004 Month Number (Jan,Feb,…,Dec)->(1,2,…,12)

The second data set will be Bureau of Meteorology (BoM) data from weather stations, aggregated and averaged by calendar month. The nearest weather station to each arrival and departure airport will be determined in Phase Two. Using these additional features, the team will determine whether they can be used to augment the prediction models built in Phase Three.

For example, monthly average weather data is available for a weather station close to Brisbane International Airport here:

http://www.bom.gov.au/climate/averages/tables/cw_040842_All.shtml

and in downloadable csv format here:

http://www.bom.gov.au/clim_data/cdio/tables/text/IDCJCM0036_040842.csv

The features of the data are as follows:

Feature	Description
Max Temperature	Mean Highest Lowest Decile 1 Decile 9 Mean days over 30^{o}C, 35^{o}C and 40^{o}C
Min Temperature	Mean Lowest Highest Decile 1 Decile 9 Mean days under 2^{o}C and 0^{o}C
Ground surface temperature	Mean Lowest

	Mean days under -1^{o}C
Rainfall	Mean Highest Lowest Decile 1 Decile 5 (Median) Decile 9 Highest daily Mean days of rain Mean days of rain over 1mm, 10mm and 25mm
Other	Mean Daily Wind Max wind gust Mean daily sunshine Mean clear days Mean Cloudy days Mean daily evaporation
9am conditions	Mean temp Mean wet bulb temp Mean dew-point temp Mean relative humidity Mean cloud cover Mean wind speed
3pm conditions	Mean temp Mean wet bulb temp Mean dew-point temp Mean relative humidity Mean cloud cover Mean wind speed

Approximately 60 features will be retrieved for 45 weather stations in proximity to 45 airports.

Project schedule

Phase 1 – Due 18/05

Find a suitable dataset and business case.

At project commencement, each team member brainstormed at least two ideas for a project topic. After an initial discussion, it was agreed that Tod and Peter will develop an idea of theirs each further (this airline project by Tod and a wind power forecasting project by Peter) and present to the team. After a second discussion this airline project was agreed upon as the way forward. The primary author of this Phase 1 report was Angus.

Phase 2 – Due 25/05

Clean, reshape, and wrangle the chosen dataset.

There is a BOM station at every airport in the flight data, we'll scrape the weather data from the BOM website relative to those airports. We'll then clean the obtained BOM data, so the station names and dates are in the

same format as the airport names in our flight data. Joining them on the airport name and month/year format will allow us to have a usable and clean dataset with over 70 features.

There will then be a preliminary analysis to determine any harder to detect issues with the data such as outliers, duplicates and missing values to make sure these don't skew our analysis for phase 3.

A brief report will be submitted outlining the process taken with these.

Phase 3 – Due 01/06

Develop a machine learning pipeline, train model(s), and evaluate

Using the team members effectively each of us will attempt to build a model to evaluate delays with the given data. These models may include, binomial logistic regression, random forest or a feed forward neural network. We'll reconvene and discuss the pros and cons with our models and determine which model is most effective. The team will then see if there are any improvements that can be made together before finalising on it. Once this is done a summary report will be carried out outlining the process and results of the project.

Phase 4 – Due 08/06

A short video demonstration of the model and a walkthrough of code used for it.

References

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., ... & Britto, R. (2010, January). Total delay impact study. In NEXTOR Research Symposium, Washington DC.

Sharyaan A. (2018). Gray and white Qatar airliner during daytime photo [Online image]. Unsplash photos. https://unsplash.com/photos/IQU0Te2Wo3Q (Cover page.)