**IN3062 Introduction to Artificial Intelligence**

Authors: Lewis Brooks, Ritharshan Makenthiran, Said Khan, Sourav Singh

Link to github: https://github.com/adbt118/AI\_project

*Please include the following sections. You might also include a section on data and its exploration, and/or a section for literature.*

**Introduction-**

**Motivation and description of the problem:**

The study's goal is to analyse the flight booking dataset received from the "Ease My Trip" website and run different statistical hypothesis tests to extract valuable information from it. To train the dataset and predict a continuous target variable, the 'Linear Regression' statistical procedure would be utilised. 'Easemytrip' is an internet platform for buying aeroplane tickets, and hence a platform where potential passengers may purchase tickets. A thorough examination of the data will assist in the identification of unique insights that will be extremely beneficial to travellers.

**What is our Dataset and Problem Domain?**

The problem domain we decided for our dataset has allowed us to explore and use many different artificial intelligence algorithms and analysis techniques to demonstrate and predict the labels. Initially the group decided on a mobile phone classification dataset [1], however with pre-processing of the data and omitting the anomalies, we believed the dataset was impractical for demonstrating our knowledge of using artificial intelligence models, and unrealistic to achieve a credible result. The dataset for the mobile phone prices, was also very small in size and offered no accuracy when predicting the price as it classified the prices into 3 bands, rather than as a regression problem, this didn’t provide the right amount of difficulty or depth to the data that we felt was required for the coursework assignment. The dataset had 3000 records, split 2000/1000 for a presplit training/test dataset, and compared to our chosen dataset, it is over 100 times smaller. On initial inspection we noticed some data didn’t make sense, like the screen width, or pixel resolution width.

As a result, we changed the dataset to a Flight Price Prediction dataset [2], that was obtained from the “easemytrip” internet platform, which is used to book flight tickets, and contains 300,261 records of flight bookings between February 11th to 31st March. Over the 50 days the flight bookings were collected, they were all collected via the same way using Octoparse scraping tool, and the data was collected separately for economy and business, which is useful to be able to work on the data independently and compare ticket classes.

The dataset has 10 features, including airline, flight code, source city, destination city, departure time, arrival time, stops on the way, ticket class, days left, price and duration. It is available on Kaggle.com, is being used under the license of CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. We created hypothesis questions which we used to begin our investigation and allowed us to extract relevant data which will allow us to understand correlations and potentially predict similar results. These will be demonstrated and shown throughout this report and project.

Questions we will attempt to answer:

* When is the optimum time (days) before booking before price increases drastically?
* Are there considerable differences in price when setting off during times of the day?
* Are certain airlines marketed higher, than competing airlines, if so, what features do they offer, more business class seats, afternoon departures, popular city destinations?
* What airlines have the monopoly out of the 6 cities recorded?

And overall,

* How does the 10 different features affect price?

We will need to analysis each of the features and evaluate if they can help with understanding and answering the questions asked.

We plan on setting achievable objectives which will allow us to monitor our data exploration and put to practice the artificial intelligence techniques we have learnt and apply them to our dataset and target label.

We will split the dataset into training, validating, and testing, as this will allow our artificial intelligence models to be applied appropriately, and due to the scale of our dataset, we will have more opportunity to change the dataset split, and introduce a holdout test set, which we could apply during the end of the project.

**Is our model classification or regression?**

The dataset we picked, as well as the questions we posed to ourselves, resulted in a regression model. Regression predicts a continuous result based on the input values/variables. Regression's fundamental purpose is to predict/estimate a "mapping function" based on the input and output variables. It seeks the best fit line to more precisely forecast the output, which in our instance will be the relationship between the time remaining and the price of the flights in the various periods.

**Did you have any missing, corrupt, or misleading data? If so, how did u cope it?**

Yes, there were, the data had outliers which may have resulted in incorrect learning. Therefore, we remove outliers using the z-score values. We can discard data that is outside of -2 and 2 z-score. So, we calculate z-score of prices (which is the target) and remove rows from the data that has z-score of more than 2 and less than -2.

**Did we omit any data?**

Yes, we only omitted a single column, as it was useless, and didn’t affect anything else to do with the data set, so we removed it. There is an unnamed column in the data (serial number for the rows), we can drop it as it has nothing to do with price.

**What techniques did we use to understand our dataset?**

**Scaling** is important with neural networks, which work particularly well when the data is normally distributed. The standard scaler transforms the data so that it is normally distributed and should allow the network to fit the data better. There are various scalar available we have tried some of them (such as Min-Max Scalar, Z-Score Scalar, and Standard Scalar). Figure 4 shows sample of data after scaling (using Min-Max-Scalar)

Table

Description automatically generated

Figure 4 Scaled Data using Min Max Scalar

**Splitting-**The dataset has 11 features (last one is the target: price). There are 300153 entries in the data (i.e. 300153 row). Data is usually divided into two sets (namely testing and training set). AI algorithms are applied, and models are trained on training set and for evaluation purposes testing set is used.

We have divided data in to testing and training set. Training set contains 80% of the data and testing set contains 20% of the data.

K-Fold Cross validation was also used, to do multiple training and testing to avoid biased results (i.e. getting good results by chance).

**How we encoded the input variables?**

Some of the columns of our dataset used, did include many texts format values, which we had to convert to numerical data. Some of them were the flight, source city, airline, departure time, stops, arrival time, destination city and finally class. We did this all using a label encoder class, which refined each value with a unique number and altered it within the dataset. Below is a before and after using the label encoder.

Table

Description automatically generated

Calendar

Description automatically generated with low confidence

**2. Method**

**What models did you use?**

**Linear Regression:**

Linear Regression is a supervised Machine Learning model that seeks the best fit linear line between the independent and dependent variables.

**Pros-**

* The major benefit of linear regression is the ease with which the dataset may be represented as a simple linear model. As a result, the training time for linear regression is short.

**Cons-**

* Limited to 2 class classifications

**Neural network:**

A neural network is an artificial intelligence strategy for teaching computers to analyse data in a manner inspired by the human brain.

**Pros-**

* Because neural networks are adaptable, they may be utilised to solve both regression and classification issues. Because a neural network is a mathematical model with approximation functions, any data that can be converted to numbers may be incorporated in the model.
* Neural networks are effective in modelling nonlinear data with a high number of inputs.
* Any number of inputs and layers can be used to train neural networks.

**Cons-**

* Because neural networks are black boxes, we can't tell how much each independent variable influences the dependent variables.
* Training data is extremely important for neural networks. This creates the issue of over-fitting and generalisation. The mode is highly dependent on the training data and may be modified to the data.

**Hypothesis statement:**

Tickets will be less expensive if purchased far in advance of the flight.

Tickets with fewer "extras" on airlines will be less expensive than tickets with more "extras" on airlines.

**Choice of training?**

Two techniques were used namely linear regression and neural network. Different tuning and parameters were tried (removing outliers, different scaling method, increasing number of hidden layers in ff-net, increasing the epochs, changing the activation function of ff-net).

**Evaluation Methodology?**

**What are the criteria for selecting model performance evaluation tools?**

**3. Results**

**What were your outputs & Description and presentation of the outputs?**

**Chart, histogram

Description automatically generated**

**Text

Description automatically generated with low confidence**

**Chart, histogram

Description automatically generated**

**Text

Description automatically generated with medium confidence**

**Chart, line chart, histogram

Description automatically generated**

**Text

Description automatically generated**

**Chart, line chart, histogram

Description automatically generated**

**Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

**Did you have any problems or difficulties working with the dataset?**

**4. Evaluation and Conclusion**

**Analysis and critical evaluation of results**

At first it seemed like linear regression was doing better than feed-forward network, but after some tuning (increasing hidden layers, increasing number of iterations and changing the activation function) the feed-forward network did better.

**Lessons learned, and future work.**

**References**