**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 chose for our dataset enabled us to investigate and use a wide range of artificial intelligence algorithms and analytic approaches to illustrate and forecast the labels. The group first settled on a mobile phone categorization dataset [1], but after pre-processing the data and removing the anomalies, we felt the dataset was unsuitable for displaying our skills of applying artificial intelligence models and unrealistic for achieving a convincing result. The dataset for mobile phone prices was also very small in size and provided no accuracy when predicting the price because it classified the prices into three bands rather than as a regression problem, which we felt did not provide the appropriate level of difficulty or depth to the data for the coursework assignment. The dataset had 3000 records, divided 2000/1000 for a presplit training/test dataset, and was nearly 100 times smaller than our chosen dataset. On first glance, we observed several facts that didn't make sense, such as 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 analyse 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 1 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 into 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?**

For evaluation, we used root mean square error (RMSE) as a method of assessing the effectiveness of our linear regression models using no scaling, no normalisation, standard scaling and minimum/maximum scaler.

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

The criteria for selecting a model for performance evaluation, needed to accommodate linear regression and neural networks which were our key methods we used during this project, as a result, it was difficult to use a confusion matrix to evaluate the performance of our data. This suggested that we must use another method such as confidence intervals or root mean square error (RMSE) to evaluate the performance of work. We used root mean square error to assess the linear regression models, and it was successful as it returned similar results when training and testing linear regression without scaling or normalisation, and with standard and min/max scaling. We could have also used Confidence Intervals as a method of evaluation, however it wasn’t as effective as linear regression and RMSE was. Also in hindsight the data was only from February to March of 2022, which only shows a portion of the flights booked in a year so we haven’t got data to compare to across the year meaning our confidence intervals won’t be as effective.

**3. Results**

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

As displayed below, we used graphs and snippets of code to show our understanding and results from our investigation. We portray the important graphs with their respective code and display them alongside a description and breakdown of what we did and what it shows.

**Linear regression and Feed-Forward network outputs-**

Because the root mean square was low, linear regression performed well in price prediction. After creating the neural network, I compared the RMSE of the linear regression and neural network and discovered that linear regression performed better until I added some more hidden layers and epochs (I only added 2 hidden layers and 15 epochs to avoid over fitting, so that the machine learns rather than memorises the data). Because the root of the mean square was smaller than the linear regression, the feed forward model better fits the situation.

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generatedText

Description automatically generated with low confidence**

A histogram of flight duration and days left to compare. This comparison allows us to see that short duration flights are more frequently bought, and that the more days left to purchase the ticket, the more tickets sold. However, the average remains the same throughout at around 6000+ until 1-2 days prior to the flight.

**Text

Description automatically generated with medium confidence**

**Chart, line chart, histogram

Description automatically generated**

The price rises significantly with each passing day until it reaches a high 1-2 days before the flight. The price drops roughly three times on the day of the flight since many tickets are not purchased at the last minute. This graph shows that the usual price is between 20000 and that this price is purchased between 18 and 50 days before the flight.

**Text

Description automatically generatedChart, line chart, histogram

Description automatically generated**

Flight prices increases with flight time until they reach 19 or 20 hours, at which point they stabilise (or price might even drop a bit for a longer flight). Then after this trend of a drop, the price then rapidly increases. This goes to show price increases as the duration does.

Finally, the Linear Regression and Feed Forward Network was evaluated using K-Fold cross validation (5-fold to be specific). K-fold was used firstly to reduce the data as we had over 300000 values, and it helps us with problems like overfitting. We then compared the two different techniques used below with cross-validation and came to a conclusion that the feed forward network was the better out of the two.

**Text

Description automatically generated**Code used to calculate the RMSE of both linear regressions, and feed forward networks. We then show the results 5 times in the code, but only preview here 2 times, to prove its correct. We did these 5 different times for accuracy, consistency, and to make sure that feed forward networks was better than linear regression, and so we had our point proven.

**Text

Description automatically generated with medium confidence**

**Text

Description automatically generated**

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

The main problem we faced was at the beginning where we had outliers which led to incorrect learning. We also removed a column within our dataset as it had nothing to do with the price, and did not affect the price in any way.

**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.

Because the root mean square was low, linear regression predicted prices well. After creating the neural network, we compared the RMSE of the linear regression and neural network and discovered that the linear regression performed better until we added more hidden layers and epochs (I only added 2 hidden layers and 15 epochs to avoid over fitting, so that the machine learns rather than memorises the data). Because the root of the mean square was smaller than in the linear regression, the feed forward model better fits the situation.

**Lessons learned, and future work.**

Feed forward neural networks provide some predictions and classifications. The network is trained with labelled data, which trains it to spot patterns and make predictions. This is accomplished by altering the weights of the connections between neurons, which may be accomplished using several optimization techniques.

Once trained, the network may be evaluated and compared to other neural networks to see which one gives the most accurate predictions. This is accomplished by measuring the network's performance using data that the network has never seen before. The network's performance is assessed in terms of accuracy, precision, and recall. Accuracy is how effectively the network predicts the correct output; precision is how accurate the predictions are; and recall is how well the network can recognise the correct output even when the input data is noisy.

As this was completely new to us, we had to learn how to use it effectively.

If we were to do this project again, we would like to do as random forest, due to its ability to search the data set more randomly. To create predictions, random forests employ a set of decision trees. Each decision tree is trained on a subset of data, with the final prediction based on most individual trees. This makes the algorithm more resilient since it can deal with noisy data as well as data with missing values.

Random forests are a strong widely used machine learning method and are commonly utilised in a broad range of applications, including computer vision, natural language processing, and bioinformatics. They are also used in sectors including as finance and health to detect trends and make predictions in massive datasets, thus making it suitable for our dataset, as it would lead to more fair and balanced results.

We also maybe would like to work on using more AI techniques, for example Naïve Bayes, Perceptron etc.

**References**

[1] – Kaggle Mobile Price Classification Dataset <https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification/code>

[2] – Kaggle Flight Price Prediction Dataset <https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction>

[3] – Public Domain Dedication <https://creativecommons.org/publicdomain/zero/1.0/>