**Capstone Project Proposal**

**Author: Patrycja Kozik**

**Author: Eduardo Matos**

**Lecturer: James Garza**

**Subject: Strategic Thinking**

**CCT College Dublin**

Flight Ticket Price Analysis

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | **Strategic Thinking** |
| **Assessment Title:** | **Flight ticket price analysis** |
| **Lecturer Name:** | **James Garza** |
| **Student Full Name:** | **Patrycja Kozik**  **Eduardo Matos** |
| **Student Number:** | **Sba23298 - Eduardo Matos**  **Sba23087 - Patrycja Kozik** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Assessment Due Date:** | **29/10/2023** |
| **Date of Submission:** |  |

Contenido

[Introduction 1](#_Toc149586495)

[Objectives 1](#_Toc149586496)

[Problem definition 1](#_Toc149586497)

[Scope 2](#_Toc149586498)

[Methods 2](#_Toc149586499)

[Timeline 3](#_Toc149586500)

[Data Source 4](#_Toc149586501)

[Ethical Considerations 4](#_Toc149586502)

[References 4](#_Toc149586503)

## Introduction

Flying is the most practical and efficient method of travel. Currently, it's more affordable than ever to travel by plane. People are choosing this type of travel as it's easier to travel longer distances in shorter times than ferries, trains or cars.

When planning our trips, we all consider flight tickets to be a significant part of our budget. We need to plan well in advance to take advantage of lower prices and save as much as possible. There are many ways to achieve this by looking for last-minute deals, and airline promotions and by checking flight comparison websites such as Kayak.com, Skyscanner, or Google flights**.**

Airlines are currently using the “dynamic pricing” strategy which means they try to maximise profits based on customer demands. They are also using the “a la carte” model where the price shown on the website does not include the additional costs for luggage or seat selection.

Our team's aim is to investigate how controversial these ticket price increases by airlines using marketing strategies are, by comparing them with what would be normal ticket prices predicted from a database of ticket prices tracked over the last 10 years.

## Objectives

Our team aims to build a flight ticket price predictor for different airlines and destinations using machine learning techniques and in particular a regression-supervised algorithm. To achieve this, we have decided to define these three specific objectives as database modelling, exploratory data analysis and list of flight price prediction studies.

* 1. Database modelling.
  2. Exploratory Data Analysis.
  3. List of flight price prediction studies.

The final objective is to compare the price of a ticket researched three, two and one days before with what would be a predicted price based on a list of prices over the last 10 years, considering this list as an example of flight ticket prices determined without any abusive marketing strategy behind.

## Problem definition

Information is power and with it, we are able to make better decisions. In the air transport industry travellers are always at the mercy of airlines and their pricing policies and marketing techniques such as dynamic pricing strategies. Our team decided to approach this project as a study to show how much airlines increase their prices when we try to book at the last minute.

We will compare the prices of last-minute tickets with those predicted using historical data. Our project aims to reflect on the fairness of airline ticket prices between what is predicted to be a typical price for a ticket and the actual prices of those tickets when we book them as close to the time of the flight as possible.

## Scope

a. Stage 1

The scope of our project begins with learning how to implement web scraping techniques to develop a resource that allow us to build a database with flight ticket prices.

Our intention is to be able to maintain a database with historical data entries including enough no. of observations and features i.e. (Airline, Source city, Destination city, Durations, price etc.)

These data entries must be reliable enough to avoid overfitting, underfitting and overall dimensional problems when assessing our data. Our analysis, based on the chosen features, will help us to achieve an accurate predicted price.

b. Stage 2

Having achieved the first stage of our project, we define our second objective to do an exploratory data analysis of our data.

When we have our data in the format we want, we analyse the possible outliers to end with a study of the correlation of the features, which would be the first part of our exploratory data analysis, a data cleaning and a first overview. As a second part of our exploratory data analysis, we will start modelling our regression supervised algorithm, trying different algorithms and studying their accuracy scores to choose the one with better scores. Finally, we will train and test our prediction algorithm.

As a second part of our exploratory data analysis, we will start modelling our regression supervised algorithm, trying different algorithms, and studying their accuracy scores to choose the one with the best scores. Finally, we will train and test our prediction algorithm.

c. Stage 3

Finally, once we have a list of predicted flights, we compare them with the prices of the same tickets three, two and one day before the flight, with the aim of reporting and analysing the difference in price between them and whether there is a pattern to how much tickets are being increased by airlines using marketing strategies.

## Methods

1. Database’s modelling

Our database will be built on Python using Selenium library and BeautifulSoup.

1. Exploratory data analysis

Our exploratory data analysis will be built on Python using some libraries such as pandas, matplotlib.pyplot, Scikit-learn and Statsmodels and seaborn.

1. List of flight price prediction studies.

Using library visualisations, we are aiming to reflect the difference in price between a list of predicted flights, and the same tickets three, two and one day before the flight.

Our List of flight price prediction studies will be built on Python using some libraries such as seaborn.

## Timeline

* 1. Databases modelling. Week 1
     1. Web Scrapping (Eduardo)
  2. Exploratory Data Analysis. Week 2 - Week 3
     1. Converting data into numerical formats. (Patrycja)
     2. Outliers Analysis. (Patrycja)
     3. Dummy categorization. (Eduardo)
     4. Study of correlation. (Patrycja) (Eduardo)
     5. Study of the different regression supervised algorithms and their accuracy scores (Val, MAE, MSE, RMSE). (Patrycja) (Eduardo)
     6. Scatterplot visualisation of our actual and predicted data to visualise how accurate it is. (Eduardo)
  3. List of flight price prediction studies. Week 4 - Week 5 (Patrycja) (Eduardo)
     1. Linear Regression (Patrycja)
     2. Polynominal (Eduardo)
     3. Lasso (Patrycja)
     4. Ridge (Eduardo)
     5. ElasticNet (Patrycja)
     6. Random Forest (Eduardo)
     7. Github Personalisation. Week 5
     8. Readme Proposal (Patrycja)(Eduardo)
     9. Readme Databases modelling (Patrycja)(Eduardo)
     10. Readme Exploratory Data Analysis (Patrycja) (Eduardo)
     11. Readme Metric Results (Patrycja) (Eduardo)

## Data Source

Our data source is a public travel search engine website, from which data is public and therefore legally scrapable.

## Ethical Considerations

The considerations facing our project start with the fact that while it is perfectly legitimate to collect or scrape data from the internet, as it is public, companies have every right to restrict access to their information, so permission to use it may need to be sought.

A database of ticket prices scraped from a travel search engine does not prove that ticket prices are set without an unfair marketing strategy behind them, so the prices used to build our model could be another example of prices being unfairly raised by airlines.

In addition, ticket prices may be set taking into account features that are outside the scope of our database.

Finally, due to the number of features in relation to the number of records, our database could be affected by underfitting.

## References

Dave Gray 2018, Better web scraping in Python with Selenium, Beautiful Soup, and pandas, freeCodeCamp, viewed 20 Sept 2023, <<https://www.freecodecamp.org/news/better-web-scraping-in-python-with-selenium-beautiful-soup-and-pandas-d6390592e251/>>.

Dinesh Kumar 2023, A Complete understanding of LASSO Regression, Great Learning Team, viewed 20 Sept 2023, <<https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/>>.

Kevin R. Williams 2021, The Welfare Effects of Dynamic Pricing: Evidence from Airline Markets, National Bureau of Economic Research, viewed 20 Sept 2023, <<https://www.nber.org/papers/w28989>>.

Luke Bryan 2023, Flight Search Engine, Medium, viewed 20 Sept 2023, <<https://medium.com/@lukebryan2302/flight-search-engine-4e232af6c7ee>>.