Daniel Mejia

DSC-680

Second Project – Flight Prices

**Business Problem**

The business problem would be to find prices according to the month in the calendar. Then, we might have an advantage choosing the best prices according to the data provided by the model. The prediction should give us an estimate mean of the prices and using the price estimated we can analyze if the price is fair or not to the current location.

**Background/History**:

According to Business Insider, the cost of flying has dip in relation to inflation. However, Airliners have taken revenues by charging luggage, meals, and other perks. The average cost of a domestic flight is about $359 dollars. The 2008 Jet fuel crisis led to airliners adding fees and cutting much of the perks from the old days like free meals. Although, fuel cost has gone down, the airline industry is not in the business of reducing revenues. Now, covid has introduce a great starvation in the industry and the perks are not close to be introduced; loses have been too great!

**Data Explanation**:

The data source will be Kaggle and for now one documents have been found further exploration for use and implementation is required. One is Flight prices prediction Dataset. The file flights.csv contains all the data to perform the analysis.

The data field were mostly in numerical type, this is optimal for AI and machine learning models. Categorical fields were found in the document and converted to integer. No cleaning was needed as the document did not have blank field or missing data. This document was cleaned before the download by an unknown entity.

The data set contains the following features for flight prediction:

**~~travelcode~~** – \*\* ***Removed*** \*\*

**~~usercode~~** - \*\* ***Removed*** \*\*

**from** – Flight departing location

**to** – Flight arriving location

**flightType** – determines the seat type: first class, economic or premium

**price** – Cost for the specific flight

**time** – The flight time to travel to destination

**distance** – Distance to travel to destination

**agency**- Name of the agency providing the service for the flight

**date** – Date the flight service was provided

**Methods**

Standard EDA will be used to explore and analyze the data to be implemented into a regression or another type of system to create a model for decision making. Visualization techniques will be chosen to improve and find patterns to choose a model which might best fit the data representation. Also, adding the correlation techniques and using internal functions such as describe will entail details of the data for analysis. Using the function info, the data types, count, and null values will be evaluated. The order of the analysis matters and will be taken into consideration to develop the machine learning model. Outliers will not be considered since this is a transaction and the price is final. These are real prices happening throughout the industry in the specific location.

**Assumptions**

The assumption made was about the pattern in the data. It was believed that the data contained a pattern, and the prices could be estimated based on the features described above. Also, correlation must have to be high for the accuracy of the data to be high. These were assumption taken while approaching the data and analyzing.

**Analysis**

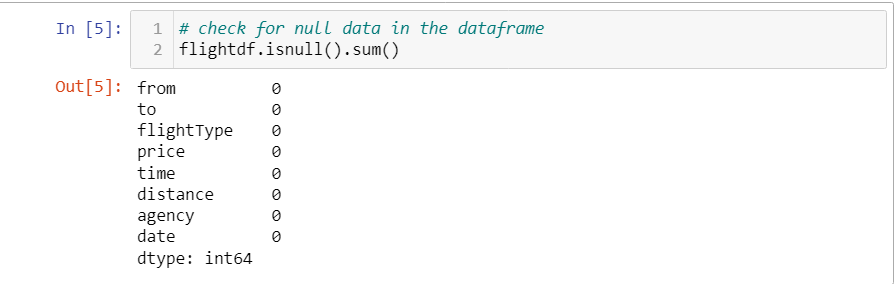
Analysis started with importing the data and what the content of the data was going to be used for the data analysis. After, determining the features travelCode and UserCode were not needed for the exploration, the columns were dropped.

*# cleaning unrequired data since all we need exist in the columns left for the analysis*

2

flightdf **=** flightdf.drop(["travelCode", "userCode"], axis **=** 1)

Then, nulls values were checked to make sure the data was clean and did not require further analysis to be used. The document contained zero nulls and did not required removal or editing in this content.



The unique values of important fields were evaluated to understand the data and what are the range of the fields. Our data is 271888 long! The image below will have key fields data:  


Using the functions described and corr, important information was found for the data. The ranges and limits were shown in the tables. Using this information, the data outliers could be found in the data. However, after a manual checkout, the data was within the range. Correlation did not seen string among any of the fields, price strongest correlation came from distance at about 64%.





After, reviewing and graphing the data for possible pattern matching and recognition. A pattern was found at first it did not look like the typical K-neighbors pattern, but a closer look and using 3-D modeling yield the pattern result typically expected in graphs. Different variables were graphed to obtain the information required to make an approximation classification system. The pictures below will graph and show different pattern depending on the variables used to model the graph.

Chart, scatter chart

Description automatically generated with medium confidenceA picture containing text

Description automatically generated Chart, scatter chart

Description automatically generated

**Limitations**

Technology limitation was a problem, I was running the ML code on a gaming computer. However, the Intel Core i7, 10th Gen was not enough to run the code without long pauses. A period of 8-12 min was required for the code to execute. Any errors would cause the whole to have to be run again and delays were not pleasant to deal. This limitation will not be seen in a more powerful system such as the cloud, multi computer server, or even quantum computers.

**Challenges**

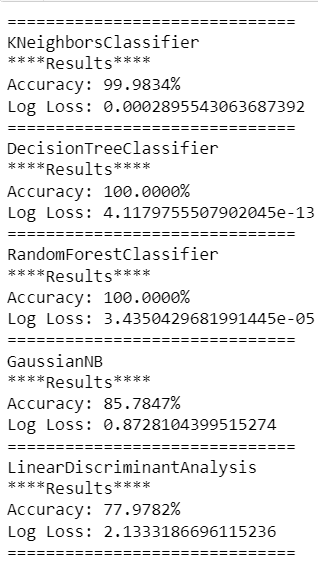
Challenges arose mostly from cleaning the data for this file. Fields required conversion to numerical values. I decided to convert fields to Int32 since this should not affect the prices guessing system, as we only required an average of the data. The loss of decimals values was not harmful to the performance of the model. In addition, the data came cleaned and this caused the challenges to be minor for this file and data.

**Future Uses/Additional Applications**

This application scope was limited to local flights of a region. But the system can be extended to international and flights from all over the world. A system containing data around the world can be advantages and create a system to pick the best flight at the best prices. Knowing the prices ahead of time could prove essential to proving customers with the best prices, beating the competition.

**Recommendations**

This system is limited by the data obtained by the data scientist. For this system to be applied in real life and useful to a business, more data must be incorporated into the system. The data was great for proving the development of similar system and more capable with some machine models yielding accuracies mostly of greater than 95%. The models and the accuracy will be shown below for understanding of the data ML accuracy.



Also, the prices system can be used to create a system of obtaining different flight for cheap and lower travel time. The data must be in the same format as it is in this file. The features used in this file should be also in the data if expansion will be done to other cities.

**Implementation Plan**

Getting the data would be the hardest part of the job once the data is obtained for the flights and prices. The data must have all the fields necessary to make the system and get the right prices, and the right flights. Then, the set up of super computers will be required to get and set data at the speeds necessary to compete and give our users the data and analysis required for them to make best trip choices and prices. Once the data is obtained and models are created, setting a website for the searches and connection of prices and flights should be as simple as calling SQL queries and displaying the correct data.

**Ethical Assessment**

The ethical implication does not exist as the model is only determining prices based on the data. The models do not provide a specific company or place to work. The determination is solely a price given approximation according to the data file provided by the data scientist.

**Conclusion**

The system was found to be highly consistent in getting the prices, graphical and numerical analysis did help to obtain information from the data. Standard deviation and mean were helpful to understand when does data seem to start getting out of the norm. Understanding the bound of the data were essential in looking at the data objectively and not with my precognition.

**Questions**:

1) How much will it cost to implement the system?

A system having multiple cloud or super computers could be costly to implement but if the system proves to be successful the revenue would be greater than the cost. Prices are hard to determine as we don’t know user use and computer uptime which cloud system charge.

2) How difficult is to implement the pipeline to the model?

The most difficult part will be to obtain the data, once the data is obtained the pipeline should be a system of databases and websites fetching data for use. Sharing information with other companies should provide extra revenue.

3) Reducing limitation on the model?

The limitation with this model would be on the data and if we have the information required to implement such a complicated system. Solution is to find the data of many flights!

4) What is the accuracy and if required how to increase it?

The accuracy for three models is above 99% and do require to be increased. However, we are not sure what would happen if models for other flights destination are added, if it would yield the same.

5) Is the business model being solve by the model?

I believe the business model is being solve by finding the pricing of destination with the accuracy required to be competitive.

6) Scalability and maintenance requirements to keep it functional?

The data will need to be checked to assure the data does not become damage and prices start to be outdated and not portraying the real world.

7) How reliable is the system?

The reliability of the system depends on the system and design chosen to hold the models and data servers fetching, cleaning and analyzing the data.

8) How will the company make profit?

The company will make profits from being able to get prices and destination desired by users at below average prices. Knowing the average price allows for this to be done easily.

9) What are the benefits of implementing the system?

It is always good to get lower prices when buying any products.

10) Are there any issue with the model and how can they be fixed?

Right now, the issue is the limitation on the amount of data and places with flight prices.

**Reference**:

*How much airfare in the US costs today compared to 10 years ago*. (2019, November 4). Business Insider. https://www.businessinsider.com/fight-prices-airfare-average-cost-usa-2019-11?international=true&r=US&IR=T#youre-also-charged-more-for-your-bags-now-according-to-a-cnn-report-from-2010-if-your-airline-charged-for-checked-baggage-it-was-between-20-and-35-depending-on-how-many-bags-you-had-3