BIG DATA ANALYSIS OF AIRLINE TRANSACTION DATASET

Submitted By :
Aman Chandok (2K18/IT/017)
Aman Jha (2K18/IT/018)

INTRODUCTION

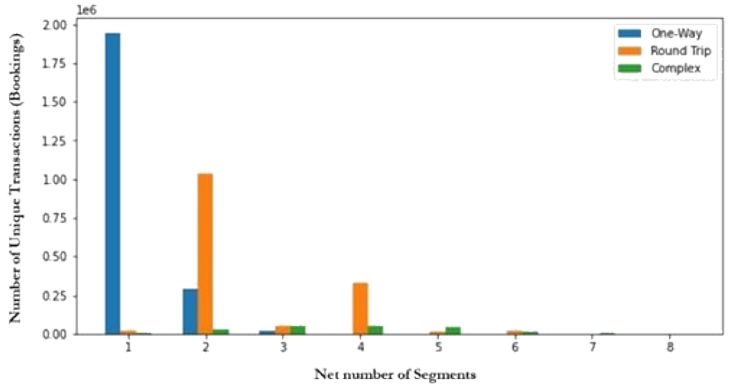
- Airline Reporting Corporation (ARC) - Provides ticket transaction settlement services between airlines and travel agencies.
- In 2019, ARC processed more than 97.4 Billion U.S dollars' worth of transactions for its customers.
- ARC's into the Travel-Verse Hackathon, hosted on HackerEarth is a global hackathon, with the objective to identify the trends that can lead to new predictions using Airline Transaction Data, which can later be incorporated into a marketable data product.
- This work aims to find ways to apply this vast data store from historical trends mapped into predictive analysis to specific recommendations for consumers and suppliers of air travel.
- We have focused our research on business applicability for three following segments::
  1. Air Travelers or Passengers
  2. Airline Companies
  3. Air Travel or Booking Agencies

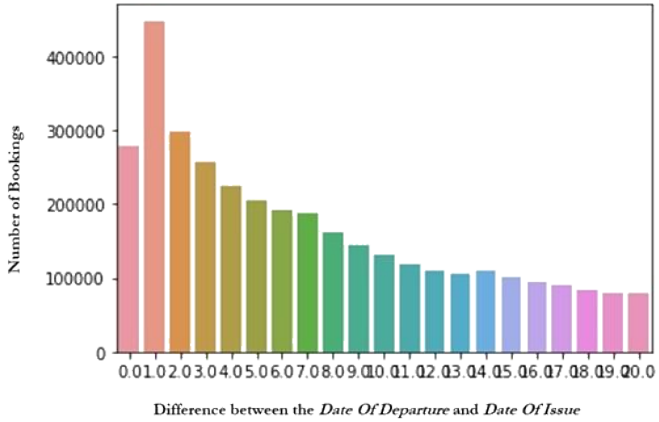
DATASET OVERVIEW

- 78 Million independent travel bookings.
- 133 Million entries
- 15 Fields - Transaction Key, Ticketing Airline, Agency, Issue Date, Country, Transaction Type, Trip Type, Origin, Destination etc.
- The dataset is big, yet possesses some challenges as listed below:
  1. No information about no. of passengers, ticket price, data of return or ticket exchanges.
  2. No information airports' country.
  3. No information about final destination of a trip as a trip is divided into various segments, which are individual entries of the dataset.

DATA PROCESSING

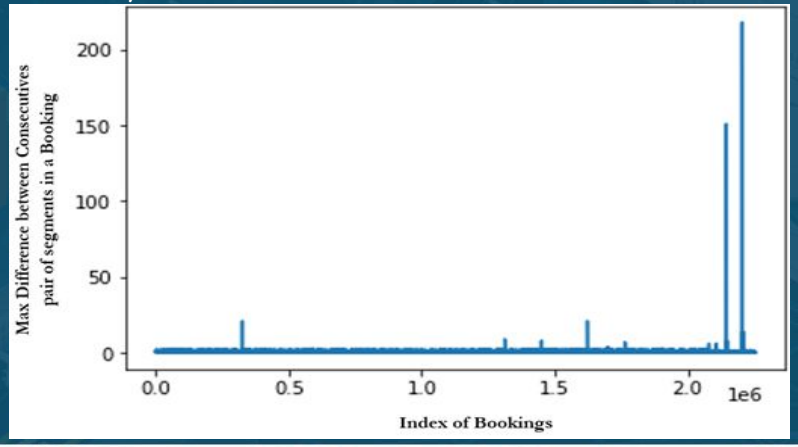
- 0.32% of dataset consists of invalid entries, which are ignored.
- Left Inner Join of this dataset is performed with Airport Codes Dataset - Associates Airport Code with Airport Country.
- Most of the data represents booking from North America.
- Exploratory Data Analysis





MERGE SEGMENT ALGORITHM

- Return the actual bookings for all *Transaction-Keys* in the dataset.
- Converts all type of bookings into 1-way bookings.
- Extended to find holiday or travel destination preferences, by considering the *Origin* of any trip as the traveler's residence and the *Destinations* being the preferred holiday/vocational/travel destination.

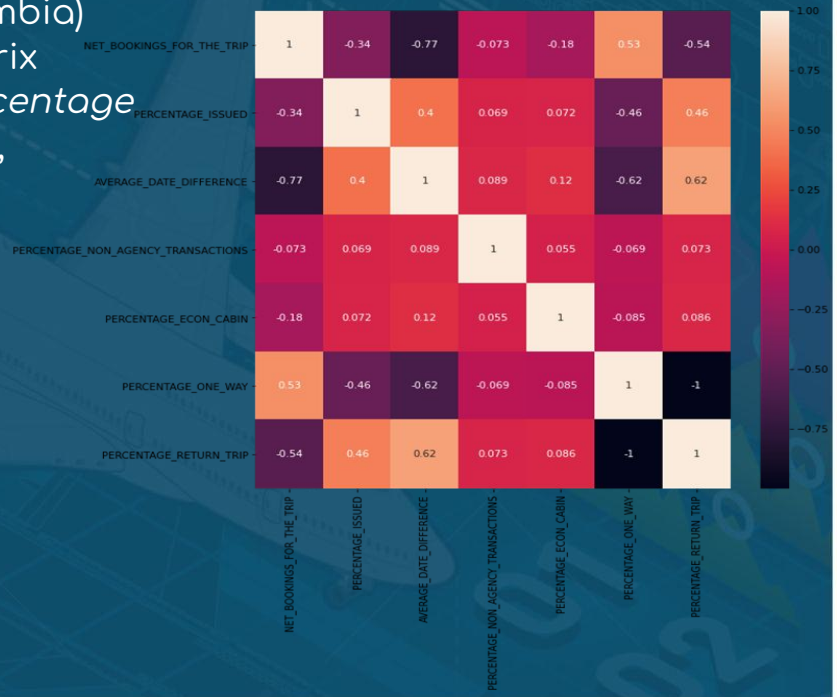



```
// Output: A Dataset containing bookings for a particular Return-Trip
or Complex-Trip type Transaction Key

Initializations:- firstRow, D_t = 1

  If df = empty:
     return {};

  initialOrigin = firstRow["Origin Airport"];
  currentDestination = firstRow["Destination Airport"];
  currentDepartureDate = firstRow["DepartureDate"];
  visitedAirport = set();
  readingIndex = 1;

  for row in df[readingIndex:]:
       dateDifference = row["Departure Date"] − currentDepartureDate;

       If (row["OriginAirport"] == currentDestination) AND
          (dateDifference > D_t AND row["Destination Airport"] not in
           visitedAirport):
             currentDestinationAirport = row["Destination Airport"];
             readingIndex += 1
       Else:
             df = {initialOrigin, currentDestination,
                   currentDepartureDate};
             Break;

  return df + MS_RT_or_XX(df[readingIndex:])
```

Algorithm 1.: *MERGE SEGMENT ALGORITHM*

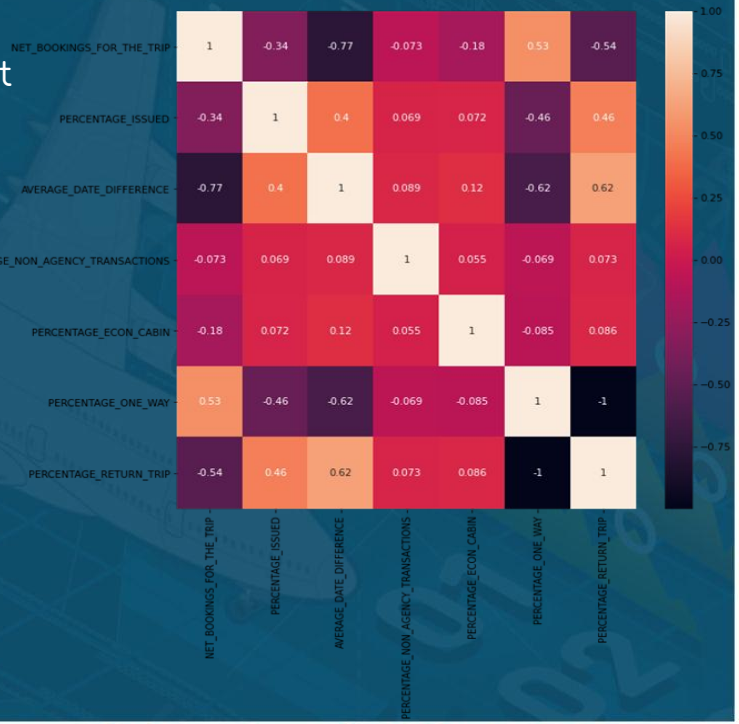
CUSTOMER TRENDS

- Most preferred trip: BEG (Serbia) to MDE (Colombia)
- It can be inferred from feature correlation matrix that *Bookings* is highly correlated with the *Percentage Agency Transactions*, *Percentage Return Types*, *Percentage One-Way Trips.*
- Most people issue tickets within a range of 10-20 days before the date of departure.



AIRLINE TRENDS

- Most frequently operating airline: *Asiana Airline Inc.* operating between *GMP* (Gimpo International Airport South Korea) as *Origin* and *CJU* (Jeju International Airport, South Korea).
- It can be seen that *Net Bookings* is correlated with the *Average Date Difference* and the *Percentage of Return Trips*.
- Significant drop in the *Net Bookings* during the end of 2018.

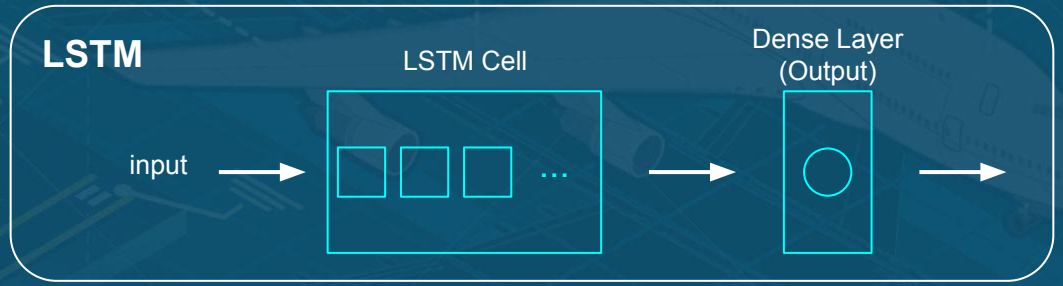


BOOKING AGENCY TRENDS

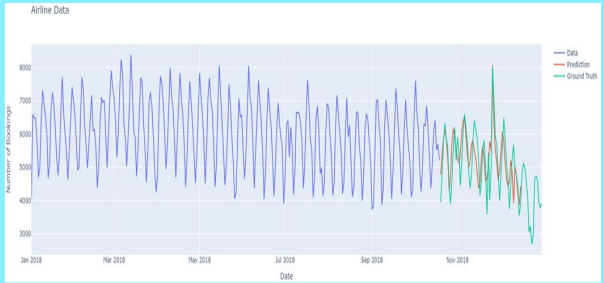
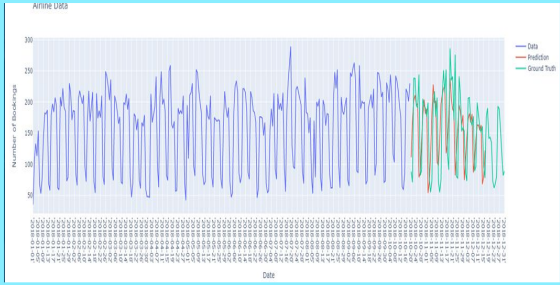
- For this purpose, we extract the of *Agency* and issuing *Country* with the highest occurrence from the *ProcessedSampleDataset*.
- Feature correlation matrix is obtained as shown.
- Result: *Agency Code '444172701161'* as Agency and *United States* as the *Country*.

PREDICTIVE MODELLING

- Implemented via LSTM - Performs well with time series forecasting problems.
- Architecture - 10 neural nodes connected with a Dense *Output* layer.
- *Look Back* parameter - 15.
- Loss function - *Mean Squared Error*.
- Optimizer - *Adaptive Moment Estimation* (Adam)
- Train-Test Split - 80:20
- Batch Size - 20
- Number of epochs - 200 (Application-I), 400(Application-II and III)
- Programming Environment - Keras framework in Python



VARIOUS APPLICATIONS

	APPLICATION - 1	APPLICATION - 2	APPLICATION - 3
AIM	Bookings on a particular day for a particular origin to destination trip	Sales for a particular flight segment	Booking Issued in a particular Country
PREDICTION			
MSE LOSS	5.994%	6.372%	1215.9%

RESULTS AND DISCUSSIONS

- *Application-I* and *Application-II* accuracy is quite high unlike *Application-III,* as Number of Bookings in a particular country via an Agency, was not correlated with other transaction related features, as compared to other application which had at least 2 highly correlating features.
- *MERGE SEGMENT ALGORITHM* is capable of providing insights regarding the Number of Bookings and the travel destinations preferred by travelers all around the world.
- 3 business predictive applications identified:
  1.Best day to book a trip (As a customer)
  2.Best way to air travel timeline (As an Airline Company)
  3.Marketing-decision related to the air travel services they provide. (As an Agency)
- Long-Short Term Memory are used to predict futuristic trends for the above application based on previous 15 days of input.
- Provides MSE loss of ~5% for two of the business applications, validating its use for trend prediction.

REFERENCES

- J. Bhattacharya, "ARC Enter the Travel Verse Dataset, HackerEarth Competition", *Kaggle – Public Datasets* (2021): https://www.kaggle.com/jayitabhattacharyya/hackerearth-arcenter-the-travelverse
- Airline Reporting Corporation (ARC), "Enter the Travel Verse Hackathon", *HackerEarth - Hackathons* (2021): https://www.hackerearth.com/de/challenges/hackathon/enter-the-travel-verse/
- *Airline Reporting Corporation (ARC) – Website*: https://www2.arccorp.com/
- Adrian Z., IATA Airport Codes – IATA Airport Codes and 3 letter code, *Kaggle – Public Datasets* (2019): https://www.kaggle.com/zinovadr/iata-airport-code
- Keras – *Website*: https://keras.io/