BIG DATA ANALYSIS AND TREND PREDICTIONS: ARC AIRLINE TRANSACTION DATASET

Aman Chandok
Department of Information Technology
Delhi Technological University
Delhi, India – 110042
amanchandok_2k18it017@dtu.ac.in

Aman Jha
Department of Information Technology
Delhi Technological University
Delhi, India – 110042
amanjha_2k18it018@dtu.ac.in

***Abstract*--- In the recent decades the use of air travel has greatly increased. According to Statista – a leading provider of market and consumer data, in the year 2020 alone the global airline industry was valued at 359.3 Billion U.S dollars and is estimated to reach 471.8 Billion dollars in 2021. In this paper, we identify trends using the newly introduced Enter the Travel-Verse dataset (2021) by Airline Reporting Corporation (ARC) representing more than Million passenger trips, to lead to new predictions that can be incorporated into a marketable data product within the Business-to-Business (B2B) and Business-to-Business-to-Customer (B2B2C) space. We have identified three predictive domains and have computed the effectiveness of Long-Short Term Memory Model (Deep Learning Technique) based on its ability to predict future trends.**

I. INTRODUCTION

The Use of commercial air travel has greatly increased since 1980s. The market capitalization of the Airline Industry it was valued at around 360 Billion U.S dollars in 2020, and is estimated to reach 472 Billion U.S dollars in 2021. There is an estimate of 5,000 airlines operating worldwide across 41,700 airports present all across the world. According to Statista, in 2019 alone the number of flights performed globally by the airline industry reached 39.8 Million, increasing continuously year-by-year for both passenger and freight. The International Civil Aviation Organization estimated around 4.4 Billion passengers carried via air transport for the year of 2019.

Airline Reporting Corporation (ARC) [3] is an organization that provides ticket transaction settlement services between airlines and travel agencies and the travel management companies that sell their services. It enables the diverse retailing strategies of its customers by providing innovative technology, flexible settlement solutions and access to the world's most comprehensive air transaction dataset. Along with the above, ARC also provides platforms, tools and insights that help the global travel community connect, grow and thrive. 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 [2], 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 within the Business-to-Business (B2B) or Business-to-Business-to-Customer (B2B2C) space.

The data acquired from ARC's transactions, representing more than 78 Million passenger trips worldwide which can provide a unique perspective on where travelers are going, when they travel and how they plan their travelling journey. Travel execution is a multi-step process that begins with the travelers being inspired their favorite destinations, looking for holiday or trip offers provided by Airline Companies and Travel Agencies, to finally booking their flights and making the trip.

This paper aims to find creative 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: Trends in Airline Booking to identify best day to issue tickets before the date of departure

2. Airline Companies: Trends in the number of passengers travelling via route, to identify the best scheduling of their resources

3. Air Travel or Booking Agencies: Trends in the number of per country, for estimating net sales and identify best strategy to market their services

Further, we have leveraged Long-Short Term Memory network models to predict future trends. We observe that the trends can be predicted with considerable accuracy and can be leveraged for two of the proposed business applications. In the end, we discuss various other methods to make use of the dataset for future work.

II. DATASET

The ARC into the Travel Verse dataset contains information about the purchase and transaction information about air travel booking.

A. OVERVIEW

The *Dataset* [2] contains information about 7,88,40,669 (78 Million) independent travel bookings. There are 13,37,22,792 (133Million) entries or rows in the dataset each containing values for 15 fields or columns, shown in *Table I.* along with their corresponding details:

S. No.	COLUMN NAME	DETAILS
1	Transaction Key	A code that identifies and allows for grouping all the segments (flight coupons) associates with a single booking or transaction
2	Ticketing Airline	The airline that issued the ticket to the travelling passenger
3	Ticketing Airline Code	A three-digit code for the ticketing airline used for accounting systems and internal revenue management at the airlines
4	Agency	Travelling agency issuing the ticket. For direct airline tickets, it's an empty string
5	Issue Date	Date the ticket was issued
6	Country	Code used to identify the country of ticker issuance
7	Transaction Type	A code that identifies the type of transaction 'I' = Issued ticket in a sale 'R' = Returned ticket for refund 'E' = Issued ticket in an exchange
8	Trip Type	Type of itinerary 'OW': One Way Travel 'RT': Return Trip 'XX': Complex
9	Segment Number	A segment in a flight coupon. It is operated by a marketing airline. Note: Collection of all segments of a ticket represents the full itinerary of the ticker purchased by the traveler.
10	Marketing Airline	The airline operating the flight between the airports on the segment or flight coupon. Note: A segment may also be a ground-travel segment (A Non-flight segment).
11	Marketing Airline Code	Code of the Marketing Airline 'V' for a Non Flight Segment
12	Flight Number	Value containing the flight number of the airline operating the flight between the airports on a particular segment of the flight coupon
13	Cabin	Premium or Economy cabin ticket purchased
14	Origin	Three-character airport code of the origin location of a flight
15	Destination	Three-character airport code of the destination location of a flight
16	Departure Date	Scheduled departure date of the flight for a particular segment (the flight between the origin and destination)

Table I: *Description of all the columns or fields present in the dataset*

Since the *Dataset* (Actual Dataset) is quite huge to process and do operations on, a *SampleDataset* [2] is also is also provided based on sampling the dataset for 39,41,207 or 3.9 Million transactions (around 2.95% of the actual dataset) with 71,88,661 or 71 Million data entries. The *SampleDataset* though not an appropriate representation of the Actual Dataset, can be used to prepare code and filter out useful data-analysis techniques.

B. CHALLENGES

Though the dataset is humongous, it possesses some challenges stated below:

a) No information about the number of passengers

b) No information about the price of the ticket

c) No information about the Initial Origin and Final Destination: It's difficult to accurately decipher the actual planned trip (initial origin to final destination) due to the segment-by-segment division of a transaction.

Note, in this paper we continuously refer the actual planned trip as a ***Booking***. Thus, a booking represented by a unique *Transaction Key* may contain many sub-transactions.

d) No information about the date of return or exchange of tickets

e) No information regarding the county the airports belong to: Though the Airport Codes are provided in IATA format, there is information related to 2.7 Million distinct Airport Codes in the dataset, which makes it a difficult task to find an appropriate external dataset mapping all the IATA Codes to County Codes

The above challenges if overcame, can result in much better identification of problem statements and ease of use for the dataset, leading to discovery of more innovative ideas and solutions.

	TRANSACTION_KEY	TRIP_TYPE	SEG_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DEPARTURE_DATE
10	'T-1808639477801388902'	'RT'	1	'NNG'	'KHN'	'2018-04-08'
11	'T-1808639477801388902'	'RT'	2	'KHN'	'NNG'	'2018-04-14'
222	'T-1808639477801466628'	'RT'	1	'YWG'	'YYZ'	'2018-03-29'
223	'T-1808639477801466628'	'RT'	2	'YYZ'	'YSB'	'2018-03-29'
224	'T-1808639477801466628'	'RT'	3	'YSB'	'YYZ'	'2018-04-02'
225	'T-1808639477801466628'	'RT'	4	'YYZ'	'YWG'	'2018-04-02'
1160	'T-1808639477801808735'	'RT'	1	'YUL'	'FRA'	'2018-07-11'
1161	'T-1808639477801808735'	'RT'	2	'FRA'	'MCT'	'2018-07-12'
1162	'T-1808639477801808735'	'RT'	3	'MCT'	'CMB'	'2018-07-12'
1163	'T-1808639477801808735'	'RT'	4	'CMB'	'MCT'	'2018-07-30'
1164	'T-1808639477801808735'	'RT'	5	'MCT'	'FRA'	'2018-07-31'
1165	'T-1808639477801808735'	'RT'	6	'FRA'	'YUL'	'2018-07-31'
2045	'T-1833843851600035141'	'RT'	1	'YOW'	'CUN'	'2019-01-11'
2046	'T-1833843851600035141'	'RT'	2	'CUN'	'YYZ'	'2019-01-26'
2047	'T-1833843851600035141'	'RT'	3	'YYZ'	'YOW'	'2019-01-26'
2192	'T-1833843851600065553'	'RT'	1	'YWG'	'YYZ'	'2019-02-26'
2193	'T-1833843851600065553'	'RT'	2	'YYZ'	'DEL'	'2019-02-26'
2194	'T-1833843851600065553'	'RT'	3	'DEL'	'ZRH'	'2019-03-27'
2195	'T-1833843851600065553'	'RT'	4	'ZRH'	'YYZ'	'2019-03-27'
2196	'T-1833843851600065553'	'RT'	5	'YYZ'	'YWG'	'2019-03-27'
3580	'T-1808639478100000606'	'RT'	1	'BKK'	'HEL'	'2018-03-22'

Img I: *Representing a list of transactions with 'RT'-Return trip type. Based on the increasing segment number, the difference between departure date for two consecutive segments and the trip pattern inferred from the image, we can estimate the booking*

III. DATA PROCESSING

A. *DATA CLEANING*

Firstly, addressing the problem of erroneous values. As we know from the dataset details, that only Non-Flight Segments, or Segment with Marketing Airline Code *'V'* will contain invalid Flight-Number and Departure-date. However, we found 12,743 invalid entries in the *SampleDataset*, which amount to roughly 0.32% of the whole dataset. When, required we simply ignore these transactions. However, in case a subset of the data is acquired containing more than 5% off invalid data, we would recommend replacing the data by finding mean values for flight number and departure date based on heuristics.

Note, we use the *SampleDataset* currently, in this version of paper due to shortage of time and low computational resource. In the final version, we aim to perform required analysis over the *Dataset* (Actual Dataset).

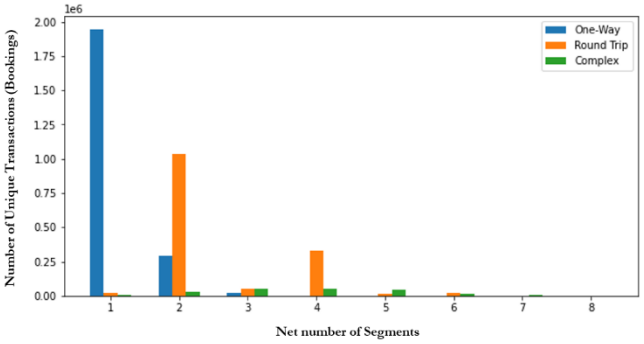
Secondly, to provide for *Challenge* (*e*). we make use of *IATA Airport Codes Dataset* [4] to associate *Origin Airport Code* to *Origin Airport Country* by performing an Left Inner Join between *SampleDataset* and the processed *IATA Airport Codes Dataset* (IATA *Dataset*) (let's call it as *ProcessedSampleDataset*). As, there are a total of, 106 out of 2,828 (~3.75%) airport codes missing in the *IATA Dataset*, we remove the entries with invalid entries from the *ProcessedSampleDataset* when appropriate.

B. *EXPLORATORY DATA ANALYSIS OVER SAMPLE-DATASET*

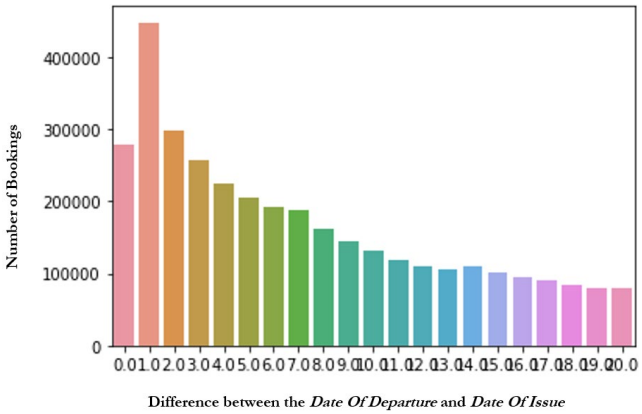
we analyze the *ProcessedSampleDataset* for various trends for which the analysis is detailed below:

a) Most Transactions have either 1 or 2 segments. However, saying that the number of segments per transaction can go upto 8 (Ref. *Img II.*). Note that, though for increasing number of segments - the number of bookings (whole trip identified by a unique *transaction-key*) are decreasing, we see a spike at 4 – this may be due to the fact there are a considerable number of *Return Trips* where in a single complete route might contain 2 segments.

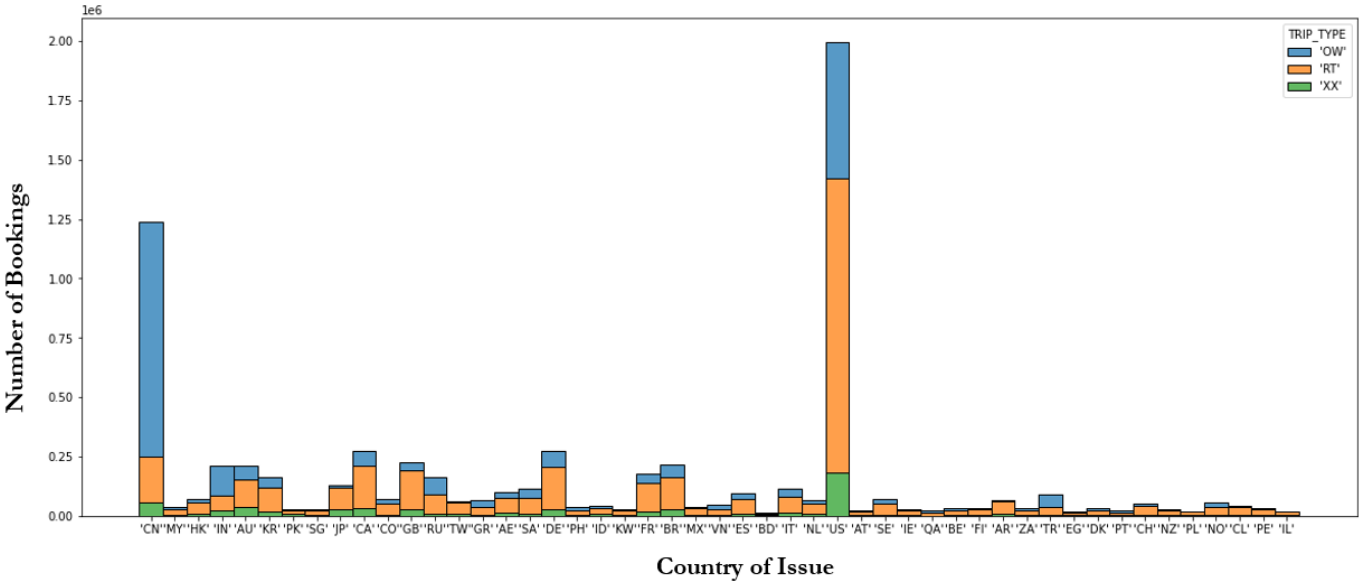
b) Most of the travelers prefer booking flights 2 days before the date of departure, with the number of bookings decreasing as the difference increases. (Ref *Img III.*)



***Img II.**: The total number of bookings based on the number of segments per bookings for all three trip types*



***Img III:** The total number of Bookings based on the difference between the Date Of Departure and Date of Issue*



***Img IV:** The total number of bookings based on the country of Issue, grouped by Trip Types (only for the top 24)*

	TRANSACTION_KEY	TRIP_TYPE	SEG_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DEPARTURE_DATE
2283	'T-1833843851800015549'	'XX'	1	'HBA'	'SYD'	'2018-12-13'
2284	'T-1833843851800015549'	'XX'	2	'SYD'	'AUH'	'2018-12-13'
2285	'T-1833843851800015549'	'XX'	3	'AUH'	'DME'	'2018-12-15'
2286	'T-1833843851800015549'	'XX'	4	'DME'	'AUH'	'2019-01-30'
2287	'T-1833843851800015549'	'XX'	5	'AUH'	'SYD'	'2019-01-31'
2288	'T-1833843851800015549'	'XX'	6	'SYD'	'HBA'	'2019-02-01'
2561	'T-1809239577600177000'	'XX'	1	'CDG'	'HKG'	'2018-03-22'
7188479	'T-1801800000000318159'	'XX'	1	'AUS'	'DFW'	'2018-07-14'
7188480	'T-1801800000000318159'	'XX'	2	'DFW'	'FRA'	'2018-07-14'
7188481	'T-1801800000000318159'	'XX'	3	'FRA'	'MAD'	'\N'
7188482	'T-1801800000000318159'	'XX'	4	'MAD'	'DFW'	'2018-08-02'
7188483	'T-1801800000000318159'	'XX'	5	'DFW'	'AUS'	'2018-08-02'

***Img V:** Representing a list of transactions with "XX"-Complex trip type.*

c) Most of the data in the dataset represents bookings from North America (Ref. *Img IV)*. In many countries the number of return bookings is high, relating to the fact that many travelers in these countries seek vocational bookings.

C. *INFERRING NUMBER OF BOOKING – MERGE SEGMENT ALGORITHM*

To tackle the *Challenge (c)*, we device a heuristic based algorithm called *'MERGE_SEGMENT ALGORITHM'* which return the actual bookings for all *Transaction-Keys* in the dataset.

Note, the word '*Transaction Frame'* refers to a data set containing trips only for a particular *Transaction Key*.

The *MERGE SEGMENT ALGORITHM* converts:

a) One-way transactions with multiple segments: A single One-way Booking
b) Round-trip transactions: Two One-way Bookings
c) Complex-trip transactions: Multiple One-way Bookings

The Heuristics and the analysis governing the algorithm are:

a) The Booking from a *One-Way* trip can be directly looked upon by selecting the *O*rigin *Airport* from the first segment and the *Destination Airport* from the second segment, which is quite obvious.

b) For both *Return* and *Complex* trip type we base the approach on the difference between *Departure Date* of consecutive (sorted increasingly) segments (*Assumption* 1) along with keeping a check whether a previously visited *Airport* is visited again (*Assumption 2*)

Assumption 2 was validated over *One-Way Trips grouped by transactions-keys,* by checking if there were any one-way trips where a previously visited *Airport* was visited again in a later segment. No such one-way transactions were found.

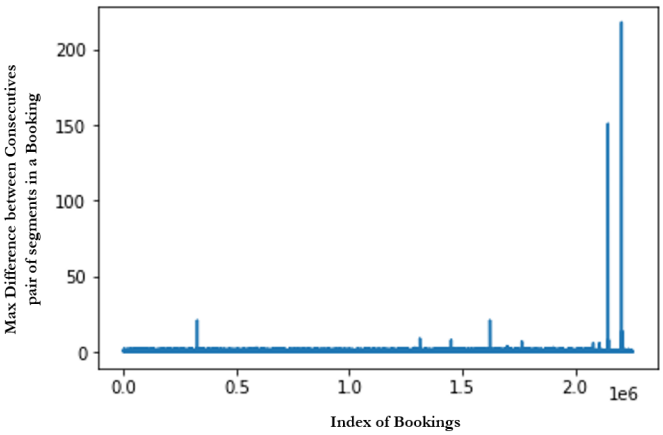
Assumption 2 is completely based on our travelling domain knowledge - that all trips from the *Initial Origin* to the *Final Destination* are tried to be placed as nearly as possible (referring to the *Departure Dates* of consecutive segments). To calculate the threshold, for the difference of *Departure Dates* between consecutive segments, we calculate *Average of Max-Difference-Between-Consecutive-Segments per Transaction-frame*, including only those *Max-Difference* that are non-zero denoted as:

$$S_m = \Sigma_{i=0}^{n} \max(a_i - a_j) \text{ for } i>j, \text{ if } > 0$$

$$D_t = S_m / m$$

Where, n : number of *One Way Transaction Frames;* a_i : a particular segment-transaction where i is the *Segment Number*; m: the number of frames for which $mam(a_i - a_j)$ and D_t is the Calculate *Date Threshold*.

The calculated *Date Threshold* for the *One Way Trips* belonging to *SampleDataset* is 1.03 (~1 day) (Ref. *Img V.)*:



Img. V: *The distribution of Max Difference based on the departure date of a particular booking*

Notice that in *Img. V*, though there do exist some outliers, we can simply ignore them, as they are constituting very negligible portion of the dataset (<0.1%).

An overview of the merge algorithm is depicted (Ref. *Algorithm I.*), along with *Img VI.* Corresponding to the algorithm applied over all the *Transaction Keys* in *Img I.* and *Img V.*

For future reference, we have applied the *MERGE SEGMENT ALGORITHM* over the *SampleDataset,* which provides us with a *SampleBookingDataset*.

Note that, due to the data limitations, we are not able to compare the accuracy of *MERGE_SEGMENT ALGORITHM*. We however, test it manually on around 100 unique *Transaction Frame*, randomly extracted from the *SampleDataset* based on various *Trip and Transaction Types*.To the best of our knowledge, the accuracy is 100% as compared to what a Human can infer. The merge segment algorithm can further be 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.

IV. BIG DATA ANALYSIS

Through this paper, out main motive of analyzing data is to identify trends that can lead into a marketable data product with the B2C or B2B2C space. We have identified three such segments, and the trends analysis for each of them are as followed:

A. *AIR TRAVELLERS or CUSTOMERS* (*B2B2C*) (*Application I*)

We try to analyze the trend in the number of bookings for a particular *Origin* and *Destination* pair.

Algorithm I-(a): *MS(df, tripType)*

```
// Input:-
   df: A transaction frame, sorted in order of increasing segment
      number
   tripType: The trip type of the transaction, can be OW (One-
      wa)/ RT(Return Trip) / XX(Complex Trip)
// Output: A Dataset (or Dataframe) containing bookings for a
          particular transaction Key

   If tripType = OW:
     Df = MS_OW(df);
   Else:
     Df = MS_RT_or_XX(df);

   Return Df;
```

Algorithm I-(b): *MS_OW(df)*

```
// Output: A Dataset containing bookings for a particular One-Way trip
Transaction Key

Initializations:- firstRow, lastRow

   initialOrigin = firstRow["Origin"];
   finalDestination = lastRow["Destination"];
   departureDate = firstRow["Departure Date"];

   df = {initialOrigin, finalDestination, departureDate};
   Return df;
```

Algorithm I-(c): *MS_RT_or_XX(df)*

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

Initializations:- firstRow, Dt = 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 > Dt AND row["Destination Airport"] not in
           visitedAirport):
             currentDestinationAirport = row["Destination Airport"];
             currentDepartureDate = row["Deparuture Date"]
             readingIndex += 1
       Else:
             df = {initialOrigin, currentDestination,
                   currentDepartureDate};
             Break;

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

***Algorithm I.**: MERGE SEGMENT ALGORITHM*

To perform this analysis, we extract only the *Origin* and *Destination* pairs having the highest frequency of occurrence from the *SampleBookingDataset*, which provides us with *'BEG'* Airport (Belgrade Nikola Tesla Airport, Serbia) as the *Origin* and '*MDE*' (Jose Maria Cordova International Airport, Colombia) as the *Destination*. Let's call this dataset as *Application-1 Dataset*. Next, we remove all *Bookings* having the *Transaction Type* as *Refund*.

We later process the columns of the *Application-I Dataset* to obtain a dataset, with the above features (or columns):

a) *Date*: The date of issue for the bookings

b) *Net Booking*s: The net bookings for a particular day

c) *Percentage Agency Transactions*: Total Number of *Bookings* by *Agencies* / *Net Bookings*

d) *Percentage same origin and booking counties*: Total Number of *Bookings* by *Agencies* / *Net Bookings*

e) *Percentage return trips*: Total Number of *Bookings* for *Return Trips* / *Net Bookings*

f) *Percentage one-way-trips:* Total Number of *Bookings* for *One-Way Trips* / *Net Bookings*

g) *Percentage econ cabins:* Total Number of *Bookings* for *Economy Cabin* Type / *Net Bookings*

h) *Average Data Difference*: The sum of difference between the *Date of Departure* and *Date of Issue* for all *Bookings* / *Net Bookings*

Img VI., represents the correlation between various features. We note, that *Net Bookings* is highly correlated with the *Percentage Agency Transactions*, *Percentage Return Types*, *Percentage One-Way Trips*, and a bit correlated with *Average Date Difference.* We also note that most people issue tickets within a range of 10-20 days before the date of departure (*Img VII*).

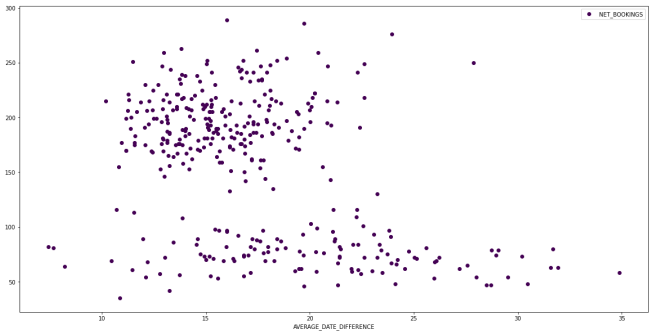
B. *AIRLINE COMPANIES (B2B)* (*Application II*)

We try to analyze the trend in the total number of sales (no. of tickets issues – no. of tickets refunded) for a particular *Flight Segment*, operated by a particular *Transaction Airline* over the year. For this purpose, we extract the pair of *Transaction Airline*, *Origin* and *Destination* with the highest occurrence from the *Sample Transaction Dataset*, which with *Asiana Airline Inc.* as the *Transaction Airline*, operating between *GMP* (Gimpo International Airport, South Korea) as *Origin* and *CJU*

	TRANSACTION_KEY	NUMBER_OF_SEGMENTS	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DEPARTURE_DATE	TRIP_TYPE
0	'T-1801800000000318159'	3.0	'AUS'	'MAD'	'2018-07-14'	'XX'
1	'T-1801800000000318159'	2.0	'MAD'	'AUS'	'2018-08-02'	'XX'
2	'T-1808639477801388902'	1.0	'NNG'	'KHN'	'2018-04-08'	'RT'
3	'T-1808639477801388902'	1.0	'KHN'	'NNG'	'2018-04-14'	'RT'
4	'T-1808639477801466628'	2.0	'YWG'	'YSB'	'2018-03-29'	'RT'
5	'T-1808639477801466628'	2.0	'YSB'	'YWG'	'2018-04-02'	'RT'
6	'T-1808639477801808735'	3.0	'YUL'	'CMB'	'2018-07-11'	'RT'
7	'T-1808639477801808735'	3.0	'CMB'	'YUL'	'2018-07-30'	'RT'
8	'T-1808639478100000606'	1.0	'BKK'	'HEL'	'2018-03-22'	'RT'
9	'T-1809239577600177000'	1.0	'CDG'	'HKG'	'2018-03-22'	'XX'
10	'T-1833843851600035141'	1.0	'YOW'	'CUN'	'2019-01-11'	'RT'
11	'T-1833843851600035141'	2.0	'CUN'	'YOW'	'2019-01-26'	'RT'
12	'T-1833843851600065553'	2.0	'YWG'	'DEL'	'2019-02-26'	'RT'
13	'T-1833843851600065553'	3.0	'DEL'	'YWG'	'2019-03-27'	'RT'
14	'T-1833843851800015549'	2.0	'HBA'	'AUH'	'2018-12-13'	'XX'
15	'T-1833843851800015549'	1.0	'AUH'	'DME'	'2018-12-15'	'XX'
16	'T-1833843851800015549'	3.0	'DME'	'HBA'	'2019-01-30'	'XX'

***Img VI.**: Booking Dataset obtained after application of MERGE SEGMENT ALGORITHM for transaction in Img I. and Img V. respectively*

***Img VII.**: Correlation between all numerical features belonging to the processed Application-I Dataset*



Img VIII.*: A scatter plot depicting the distribution of Net-Bookings based of the difference between the Date of Issue and the Date of Departure, over the Application-I Dataset*

(Jeju International Airport, South Korea) as *Destination*. Let's call this dataset as *Application-II Dataset*.

Similar to the process followed for the processing of *Application-I Dataset*, we extract a dataset with the above features from *Application-II Dataset*:

a) *Date*: *Date of Departure*

b) *Net Bookings for the Trip*: Net *Issue* type *transactions* + Net *Exchanged transactions* – Net *Refunded/Returned transactions.* An estimate for the total number of sales.

c) *Percentage Issued*: Net *Issue* type *transactions* / *Net Bookings*

d) *Average Date Difference*

e) *Percentage Non-Agency Transactions*

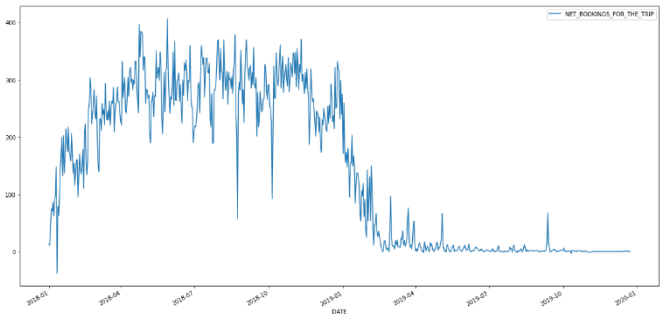
f) *Percentage Econ Cabin*

g) *Percentage One-Way*

h) *Percentage Return Type Trips*

Img X-(*a*)*.,* represents the correlation in between all the above features, while *Img XI.*, depicts the day-wise time series for the *Net Bookings*. We notice that the *Net Bookings* is correlated with the *Average Date Difference* and the *Percentage of Return Trips*. Also, there is a significant drop in the *Net Bookings* during the end of 2018. This may have caused due to various reasons, which is unrelated to the scope of our research. For further trend prediction on the *Processed Application-I Dataset*, we split the graph based on the mean value of the *Number of Bookings* (203), keeping only those entries for which, the *Number of Bookings* is greater than it. Lastly, we perform interpolation over the missing entries, for dates not present in the dataset.

C. *AIR TRAVEL BOOKING AGENCIES (B2B)* (*Application III*)



Img XI: *The time series of total sales or Net Bookings over the Application-II Dataset*

We try to analyze the trend in the total number of *Bookings* issued via a particular *Agency* for customers belonging to a given *Country*.

For this purpose, we extract the of *Agency* and issuing *Country* with the highest occurrence from the *ProcessedSampleDataset*, which provides us with *Agency Code '444172701161'* as Agency and *United States* as the *Country*. Next, we remove all *Bookings* having the *Transaction Type* as *Refund*.

Similar to the process followed for the processing of *Application-I Dataset* and *Application-II Dataset*, we extract a dataset with the above features form *Application-III Dataset*:

a) *Date*: *Date of Issue*

b) *Number of Bookings*: *Net Issue type Transaction Keys for a particular Date*

c) *Percentage Return Trip Types*

d) *Percentage One Way Trip Types*

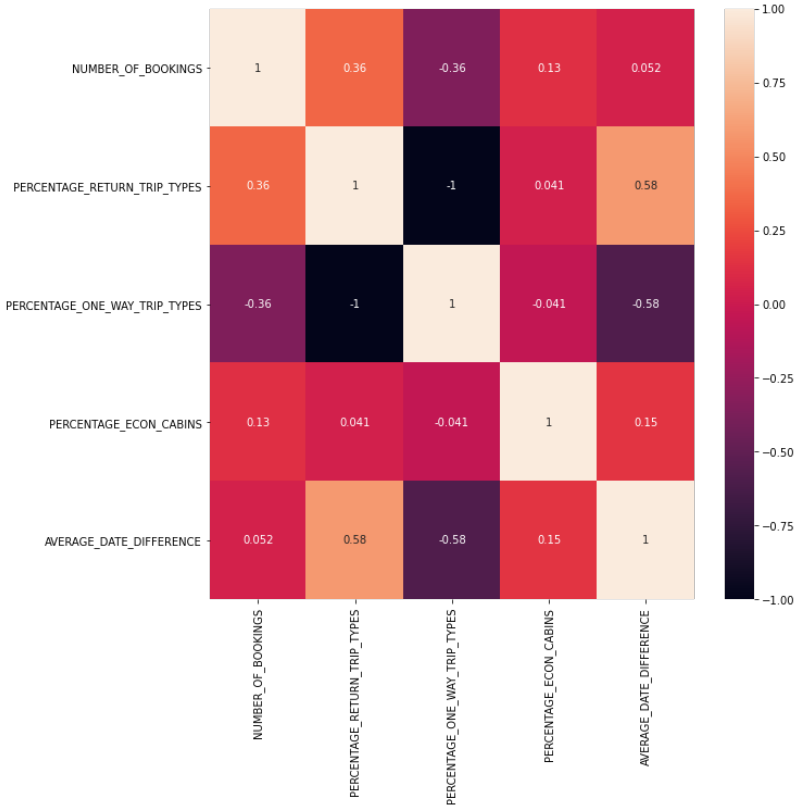
e) *Percentage Econ Cabin*

f) *Average Date Difference*

Img X-(*b*)., represents the correlation between all the above features. We note that, *Number of Bookings* is very slightly correlated with the *Percentage Return Trip Types* and *Percentage One Way Trip Types,* which are themselves completely correlated with each other, i.e., number of one-way trips decrease and the number of return trips increase. This may be due to the fact that, the tickets booked via an agency, are rarely exchanged.

V. PREDICTION MODELLING

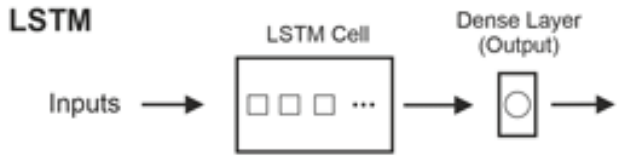
Based on the data analysis performed for the above three business segments, we now proceed to implement Long-Short Term Memory (LSTM) networks (based on Deep Learning, Recurrent Neural Networks concepts) for future trend prediction.



Img X-(a) *(Top)**: Correlation between all numerical features belonging to the Processed Application-II Dataset*

Img X-(b) *(Bottom)**: Correlation between all numerical features belonging to the Processed Application-III Dataset*

We choose LSTM networks as they are known quite well to perform for time series forecasting problems, as it learns the function that maps a sequence of past observations as input to an output observation. Our network consists of an LSTM layer with 10 neural nodes along with Rectilinear Unit (ReLU) activation, connected finally with a Dense *Output* unit (Ref. *Img XI*). The total Number of Previous Input Data to predict the current Output is referred as the *Look* Back parameter. The *Look Back* parameter is chosen to be 15, while the loss function is calculated by *Mean Squared Error*. The model is fit using the efficient *Adaptive Moment Estimation* (Adam) *Optimizer*.



Img XI.*: Image depicting LSTM model Architecture. Note: LSTM has 10 units with a look back of 15*

For training the model on the datasets for each application, we have selected appropriate input features based on the correlation matrix (Ref. *Table II.*), and based on the previous input vectors for the past 15 days the model predicts the output. For all purposes, we have split the dataset into 80:20 ratio for obtaining the Train and Test set. The training batch size per epoch is taken as 20, while number of epochs is taken as 200 for *Application-III* and 400 for both *Application-I* and *Application II.* The programming of the architecture was done via *Keras* [5] framework in *Python*.

Application	Input Features	Output
Application I	Net Bookings for the Trip, Percentage Non-Agency Transactions, Percentage One Way, Percentage Return Type Trips, and Average Date Difference	Total Bookings on a particular day for a particular origin to destination trip
Application II	Net Bookings, Average Date Difference, and Percentage One-Way Trips	Total Number of Sales for a particular flight segment
Application III	Number of Bookings	Total number of Booking Issued in a particular Country

Table II: *The Input and Output features for all Applications*

VI. RESULTS AND ASSESMENT

Table III. shows the accuracy of the LSTM network for all three applications.

Application	MSE Loss
Application I	5.994 %
Application II	6.372 %
Application III	1215.9 %

Table III: *Accuracy achieved by LSTM models for the various Application Datasets*

We note that, the accuracy for *Application-I* and *Application-II*, both are quite high (very low MSE loss)) while that for *Application-III* is not considerable for real-time use in business applications. We suspect the reason being that, for *Application-III* the Output, i.e, the Total 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, which made them easier for trend modelling along with the fact that the network was trained for 200 epochs. *Img XII.* Shows depicts the predictions made by the model

VII. CONCLUSIONS

In this paper we identify trends in the newly introduce ARC Airline Transaction Dataset. Our core objective was to identify business areas where predictive modelling can be applied based on day-to-day transactional data, which will further lead to creation of innovative marketable data product within B2B and B2B2C segments.

First, we introduce our *MERGE SEGMENT ALGORITHM* which based on the domain heuristics, is capable of providing insights regarding the Number of Bookings and the travel destinations preferred by travelers all around the world.

Secondly, we identify three business application which are mentioned below along with their predictive applications:

1. Airline Customer: Best day to book a trip
2. Airline Company: Best way to air travel timeline
3. Agencies: Data driven Marketing-decision making related to the air travel services they provide

Lastly, we leverage Long-Short Term Memory networks to predict futuristic trends for the above application based on previous 15 days of input. We found that LSTM network provide a MSE loss of ~5% for two of the business applications, validating its use for trend prediction.

ACKNOWLEDGEMENTS

The authors would like to thank Assistant Prof. Priyanka Meel of Department of Information Technology – Delhi Technological University for providing us valuable insights into the fundamentals of Data Mining, its principles and applications and providing us valuable feedback on and related to our work.

Img XII (a:Top), (b:Middle); (c:Bottom):– *Depicting time series prediction for various proposed Applications. The Curve in Green depicts the Ground-Truth whereas, the curve in Orange depicts the predictions made by LSTM model*

REFERENCES

[1]. J. Bhattacharya, "ARC Enter the Travel Verse Dataset, HackerEarth Competition", *Kaggle – Public Datasets* (2021): https://www.kaggle.com/jayitabhattacharyya/hackerearth-arcenter-the-travelverse

[2]. Airline Reporting Corporation (ARC), "Enter the Travel Verse Hackathon", *HackerEarth - Hackathons* (2021): https://www.hackerearth.com/de/challenges/hackathon/enter-the-travel-verse/

[3]. *Airline Reporting Corporation (ARC) – Website*: https://www2.arccorp.com/

[4]. Adrian Z., IATA Airport Codes – IATA Airport Codes and 3 letter code, *Kaggle – Public Datasets* (2019): https://www.kaggle.com/zinovadr/iata-airport-code

[5]. Keras – *Website*: https://keras.io/