Digital Click Stream Data for Airline Seat Sale Prediction using GBT

Md Alauddin, Choo-Yee Ting

Faculty of Computing & Informatics, MMU, Cyberjaya, Malaysia

ABSTRACT

Revenue Management is important for every airline business and the seat is the main product of an airline. The purpose of the revenue management is to maximize the revenue of each airline routes based on demand. This demand, however, depends on factors such as historical demand, seasonality, seat pricing based on purchase lead days, competitors pricing and customer behaviour. Prediction of passenger demand helps to forecast revenue on future flights and thus allow the airline to generate optimal prices for the corresponding flights. Therefore, minimizing the prediction error constitute the most crucial goal of good revenue management. In this paper, A GBT based model has been proposed for airline seat sale prediction to optimize the revenue. To optimize the prediction accuracy, an analytic dataset has been developed by combining digital attributes and traditional operational and transactional attributes. This paper will also highlight an efficient data extraction and processing pipeline have been proposed to aggregate a large volume of unstructured data from various data sources. The empirical findings suggested applying GBT on transformed dataset can predict seal sales for 30 days ahead with accuracy of 93%.

Keywords : *Airline seat sale prediction, Data mining, Gradient boosting, Machine learning, Predictive analytics.*

I. INTRODUCTION

Passenger demand prediction is important for the commercial airline industry because of the increase in competition among airline companies. Predicting the demand of airline tickets (or seats) in advance and alter the ticket price with the aim to maximize the ticket sales is a challenging decision for airline companies. There are many interconnected factors like customer behavior, segmentation information, load factor, competitor price etc. Usually, intelligent dynamic pricing models are employed for revenue optimization which considers factors such as customer behavior to potentially boost up a significant percentage of airline revenue [1]. Understanding the customer behavior allows prediction algorithms to advice better pricing model [2]. However, it is quite difficult to define the pricing model since it is limited to inter-temporal price discrimination and subject to dynamic adjustment to stochastic demand. In recent years, most of the Asian airline's focus is on digital transformation [3][6]. The key objective of digital transformation is to understand the online customer acquisition, digital channel attribution, online customer segmentation, and their search patterns which are providing insight into customer behavior more than ever. There is a significant scope exists to process a huge volume of digital customer data and extract important attributes for a more accurate dynamic seat sale prediction model which would significantly benefit the revenue maximization.

The accuracy of the prediction models varies based on the dataset that has been used for the modelling and training purpose. Most of the previous research work does not report on using digital data due to the lack of proper extraction and processing pipeline of this huge volume of data. In this context, digital data can be defined as the data generated due to the user interaction with various digital platforms like website, mobile applications. Since the digital customer data provide crucial insight into customer purchasing behavior it is very important to extract useful attributes to feed the machine learning model. Furthermore, due to the various digital platform (i.e. mobile devices and website), the structure of stored data can be different. This presents a further challenge while creating an aggregated dataset by comprising different types of data sources (i.e. transactional, operational, digital) [5].

One of the main goals of this research is to identify the digital variables that can represent user behavior to improve the accuracy of demand forecasting. To accomplish this, an efficient data extraction and processing pipeline have been proposed to aggregate a large volume of unstructured data from various data sources and to create an analytics dataset for the machine learning model. The focus has been given to collect and process a large volume of raw airline online visitors click-stream or digital data. The raw data used in this research belong to a major Airline company in Southeast Asia. Finally, a gradient boosted tree (GBT)

based machine learning model has been developed by using this analytics dataset to predict airline seat sale.

II. RELATED WORK

In this section, several forecasting methods proposed for airline industry will be discussed based on current literature. Focus will be given to describe data collection, data preparation, modelling, and evaluation methods to understand research gap and relate with the contribution of this research.

A. Fuzzy prediction method

A fuzzy prediction method has been proposed in [10] for airline seat sale prediction. The authors considered day-to-day price change and linking flights also defined as round-trip flights to improve the accuracy of forecasting model. A major North American airline data has been used for that research. The dataset contains over 500,000 sales records from 22,900 flights over 20 months. There were three routes data (both direction) and arranged into two tables; general flight information and sales detail.

The attributes of the dataset are flight direction, flight number, departure date, capacity, total bookings/day, number of cancelations, total bookings for regular customer and price. It has been observed that there is a significant effect of time remaining to flight on the amount of seat sold. This finding also reported in [11]. It has been concluded that the price change is an important factor on customers' decisions to buy a ticket.

B. Artificial Neural Network (ANN) based prediction

Neural network-based machine learning models are becoming popular day by day. NN based model is particularly suitable with huge volume of data. This technique has been used to forecast airline passenger demand. An ANN model has been proposed in [8] to forecast passenger and air cargo demand from Japan to Taiwan. It has been claimed that this model can overcome the shortcomings of time series analysis. It has also been reported that limitation of grey theory (which rely on historical time series without analysing the causality between the variables) can also be improved. In [7], an integrated ANN has been used to predict the number of passengers with 1993 to 2005 Iranian airline data. Another significant work has been done using ANN and Box-Jenkins for airline passenger forecasting in [9]. A major Turkey airline data has been used for this research. Daily passenger data for business class and economic class for the year 2010-2015 is taken. It has been found by the time-series analysis that the data is free of inconstant variance but contains seasonality. A Box-Jenkins method has been applied to compare the result with ANN. MAPE has been used to evaluate the performance of the model. However, there is no indication of route level data and aggregation of different data tables. Furthermore, analysis and prediction for only one route has been given which induce some confusion for the model accuracy.

C. GMM Prediction Method

Generalized method of moments (GMM) is another approach reported in the literature for airline ticket sale forecasting. A GMM based dynamic demand forecasting has been proposed in [1]. An original panel dataset of prices and seat inventories from Expedia.com has been used in this research. This dataset contains record of 228 one-way US domestic flight. It has been concluded that customers' purchasing behaviour changes as the departure date nears. It is also observed that high-valuation consumers buying earlier which is consistent with a key prediction mentioned in [12]. Furthermore, it is found that airline travellers sort themselves efficiently in equilibrium with low valuation types postponing their purchase decisions and even deciding not to buy if price closer to departure are higher than their valuation.

The effect of price dependence was also studied in [10] where a comparison was done between purchase behaviour comparing international flight and local flight. It was found that the prices for international tickets decrease with increase in time interval before departure to a certain extent. On the other hand, the prices on the local tickets behave the exact opposite. The research proposed two empirical data-driven models. Public data set was utilized for the same however the researchers pointed out some inconsistencies and incompleteness of the data.

Some research works also consider the other variables such as the passenger no-show rate. In [13], authors reported that a forecasting approach which considers passenger no-show data could yield more accurate results. The training data set which was utilized comprised of 7.7% of no-show passengers, and the probability of no-show is higher for late bookings, and passengers who book flights on the same day as the departure date.

It has been found that the work on airline ticket sale prediction or passenger forecasting so far is very limited. Furthermore, the above-discussed methods only deal with the model development. There are very little information exists on raw data collection and pre-processing method specially digital customer data. Almost all proposed forecast model so far mainly uses transactional and operational variables to predict the seat sale or passenger demand. Furthermore, there is always a high possibility that the utilized data set suffers inconsistency and incompleteness. In such a case, the

proposed model would struggle in regards to out-of-sample predictions. Consequently, it can give rise to over-fitting; a condition that can be caused by insufficient data, which leads to poor generalization of the model.

III. DIGITAL ATTRIBUTES FOR AIRLINE TICKET SALE PREDICTION MODEL

As discussed in the previous section, researchers have proposed machine learning model for airline passenger demand prediction. Mostly conventional dataset and attributes (i.e., from transactional and inventory) have been used in those models. However, it can be said that using proper feature engineering, accuracy and performance of the machine learning model can be improved further. The same thing is also true for airline passenger demand or ticket sale prediction model. Important attributes can be extracted from the processed digital data. These digital attributes can be included along with existing standard attributes to create analytics dataset for machine learning model.

However, tracking and processing visitors' raw events from the website or mobile app log data is complicated. The main reason is the large volume of hit level data. It has been found that one of the major Asian airlines has about 15 millions of online visitors per month, which generates roughly 3-5 billion events of unstructured or semi-structured web tracking data [5]. In this research, the online digital click stream dataset is obtained from a major Asian airline system with 50 destinations. Each route is tracked with one-way and return flights for 30 days to 120 days. The following chapters will describe the digital data collection, processing and analytics dataset.

A. Digital sources (D^G) data collection and processing

Each airline has their own digital sources such as web, mobile and tablet apps in different major platforms (i.e., iOS and Android). These digital sources are used to collect each of the raw events data of their web and mobile app visitors with details activities till e-commerce transaction completes. Airlines use following digital platform for online e-commerce transactions:

- Web/Desktop : Web/Desktop applications are accessible through browser.
- Mobile App – android : Android mobile applications are accessible through android mobile phone.
- Mobile App – iOS : iOS mobile applications are accessible through 'Apple' iPhone.
- Tablet App – android : Android tablet applications are accessible through android tablet devices.
- Table App – iOS : iOS tablet applications are accessible through iPad

Most of the digital platforms mentioned above have been used by visitors to search flights, compare prices and finally to book the tickets. Thus, data has been collected from all of these platforms to find the correct demand parameters such as visitors, flight search, transactions. The collection of digital data in real-time is a complicated process, but with the evolution of Java scripts tagging framework, it is possible to track each web page and its components based on visitor status on the internet. The passenger activities such as which page they search, how much time they spent on each webpage, how many clicks and scrolls on each page etc. Also, the e-commerce related information such as add to cart, product related information and e-commerce transaction details etc. The tracking mechanism used to collect the data from various digital platform varies slightly depending on the specific platform. Similar attributes are captured and computed using same logic from each digital source. Digital data sources are semi-structured and 6 Terabyte (13 months) digital data have been processed for this research. Table 1 shows the digital attributes captured from the raw digital data.

Table 1 Digital Platform Attributes

Variable Name	**Description**
Visitor	fullVisitorID is an ID
Visit	Session ID
Behaviour	pageview, time on site, exit page etc.
Device Info	User device information
Traffic Source	User platform before coming to website
Flight Search	The information that visitors normally enters while searching flights such as trip type, search origin, search Destination, departure date, return date and no of the passenger. These information are captured through Custom Dimensions in an Array as loop attribute

IV. AGGREGATION OF DIGITAL DATA

After cleansing, enrichment and transformation a structured and quality data set has been produced for further aggregation. The aggregation will produce desired granular dataset for machine learning model. Algorithm 1 shows the high-level process of aggregating the digital data.

Algorithm 1: generate-aggregated digital data set

Input: *Union*(D_{web}, D_{mobile}, D_{tablet})
Output: $D_{uniqueVisitorByRoute}$, $D_{uniqueFlightSearchByRoute}$, $D_{NoOfFlightSearchByRoute}$
for d in (ClickStreamRecords) **do**

1. search_timestamp ← Transform UNIX to timestamp (concat(D_{web}.visitStartTime, date))
2. visitID ← concat (sessionId, visitId)
3. visitorID ← Extract (D_{web}.fullVisitorId)
4. $D_{FlightSearch}$ ← Extract (Max(CustomDimension.index, CustomDimension.value) by iterating each items))

end for
for d in ($D_{FlightSearch}$) **do**

1. uniqueUsers ← COUNT (Distinct ($D_{FlightSearch}$.visitorID) by Routes)
2. uniqueSearch ← COUNT (Distinct ($D_{FlightSearch}$.sessionID) by Routes)
3. NoOfUsers ← COUNT ($D_{FlightSearch}$.sessionID)

end for

All digital platform (Web/Mobile/Tablet) data has been merged to make a one single data source. Since all the digital data are in the same structure, a UNION operation in BigQuery can merge multiple datasets of the same structure. This merged data table is named as 'clickStreamRecords'. Algorithm 1 takes this data as input. First step of the algorithm is to extract visitId and visitorId of the customer by hourly, daily, weekly and monthly basis and stored as $D_{FlightSearch}$. After that, data has been aggregated to get the no. of flight, no. of unique user perform flight search and no. of total search as well as group by each selected route (origin and destination), search date and departure date. Furthermore, search-lead-days have been calculated by subtracting search-date from departure-date. This will compute how many days before the departure, customer searched for the flight. Output of this algorithm has been stored as $D_{uniqueVisitorByRoute}$, $D_{uniqueFlightSearchByRoute}$, and $D_{NoOfFlightSearchByRoute}$. Aggregated final dataset sample has been shown in Table 2.

After generating digital dataset, next step is to merge processed transactional and digital dataset to create a final analytics dataset with processed operational data. Using generate-aggregated transaction algorithm and digital dataset this merging process has been accomplished.

Table 2 Digital Platform Sample Data

Attributes name	Examples
fullVisitorId	1527445791
visitId	1527445791
SearchedOrigin	KUL
SearchedDestination	HKT
SearchedDepartureDate	2018-05-21
SearchReturnDate	2018-05-21
unique_search	6
NumberSearches	10

Algorithm 2: generate-aggregated transaction and digital dataset

Input: $D_{transactionRouteLevel}$, $D_{FlightSearch}$, $D_{seasonility}$
Output: $D_{transaction\_FlightSearch\_seasonilityByPurchaseLeadtime}$
for d in (D_{Routes}) **do**

1. $D_{transaction\_FlightSearch\_seasonilityByRoutes}$ ← Merge ($D_{transaction.Routes}$, $D_{FlightSearch.Routes}$, $D_{seasonility}$) on Route aggregated daily level

end for
// Transpose daily level
for d in ($D_{FlightSearch.Routes}$, $D_{transaction.Routes}$) **do**

4. $D_{transaction\_FlightSearch\_seasonilityByPurchaseLeadtime}$ ← Transposeandaggregate ($D_{transaction\_FlightSearch\_seasonilityByRoutes}$) by PurchaseLead (Weeks, Month, Quarter

end for

Algorithm 2 takes $D_{transactionRouteLevel}$, $D_{FlightSearchbyRoute}$ and $D_{seasonility}$ (which has been produced separately and the process is out of scope of this paper) as input and generate a aggregated dataset by 'PurchaseLeadTime' and store as $D_{transaction\_FlightSearch\_seasonilityByPurchaseLeadTime}$. In first step, $D_{transactionRouteLevel}$, $D_{FlightSearchbyRoute}$ and $D_{seasonility}$ have been merged and stored as $D_{transaction\_FlightSearh\_seasonilityByRoutes}$. In final step, transpose and aggregation are performed on $D_{transaction\_FlightSearh\_seasonilityByRoutes}$ to generate $D_{transaction\_FlightSearh\_seasonilityByPurchaseLeadTime}$ dataset.

A final analytics dataset has been created by merging $D_{transaction\_FlightSearh\_seasonilityByPurchaseLeadTime}$ and processed operational dataset. A sample of the final attributes has been shown in Table 3. In the next section, a GBT regression model has been described by using this final dataset to predict the seat sale for airline.

Table 3 Sample of final data for ml model

Attributes name	Examples
DepartureDate_day	1, 2
Average Base Fare: Last 31 to 45 days	92.33, 114.00
Average Base Fare: Last 181 to 210 days	50.60, 70.40
Average Base Fare: Last 241 to 270 days	0, 31.85, 50.00
Average Base Fare: Last 361 days and before	0, 50.00
Average Seat Sold: 30 days	12, 16
Average Seat Sold: 60 days	3, 5, 10
Average Seat Sold: 90 days	2, 3
Unique Search: Last 46 to 60 days	35, 45
Unique Search: Last 61 to 90 days	86, 128
Unique Search: Last 121 to 150 days	119, 41
Unique Search: Last 271 to 300 days	17, 21
Unique Search: Last 331 to 360 days	3, 2, 0
Unique Search: Last 361 days and before	15, 5, 6

V. GRADIENT BOOSTED TREE (GBT) BASED PREDICTIVE MODEL

A. GBT Model for Seat Sale Prediction

In this research, several machine learning algorithms have been evaluated to predict the airline seat sale in advance. These are 1) Support vector regression (SVR) 2) Generalized Linear Model (GLM) and 3) Gradient Boosted Tree (GBT) [5]. After comparing the results from 3 algorithms, finally, GBT (or GBRT) has been chosen for this research. GBT has some advantages over the other two machine learning algorithm. It can naturally handle mixed data (categorical and numerical), predict more accurately and handle outliers through robust loss function. Since our final dataset consists of both categorical and numerical variables, thus, GBT is the preferred algorithm for this research. Figure 1 depicts the comparison result of an above-mentioned algorithm for the same analytic data set in terms of mean average percentage error (MAPE). MAPE is widely accepted for comparing error in business point of view. Thus, MAPE has been used for comparing GBT, GLM, and SVM in Fig. 1.

It has been seen that (Fig. 1), GBT based seat sale predicting model has obtained less MAPE for almost all the route. However, in the case of route CGK-DMK, PEN-SIN, LGK-SIN and SIN-LGK, GBT and SVM model showed the same MAPE. Furthermore, in the case of route CGK-PEN, SVM model shows better MAPE than GBT and GLM.

In boosting method, base models are generated sequentially. The accuracy of the prediction is improved through developing multiple models in sequence by giving emphasis on training cases which are difficult to estimate. In this process, misclassified training instances from the previous base models appear more often in the training data than the ones that are correctly classified. The aim of each additional base model is to correct the mistakes made by its previous base models.

Several hyperparameters have been tuned to obtain as close result as possible after several iteration. Fig. 2 shows the procedure of model training and hyperparameter tuning. Final model performance has been calculated based on MAPE to comply with the actual business standard.

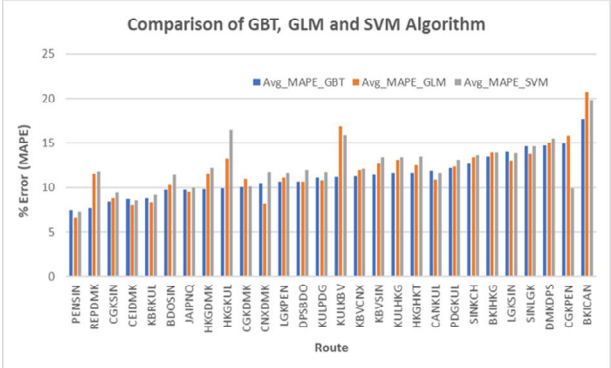


Fig. 1: Comparison of GBT, GLM and SVM model performance based on MAPE

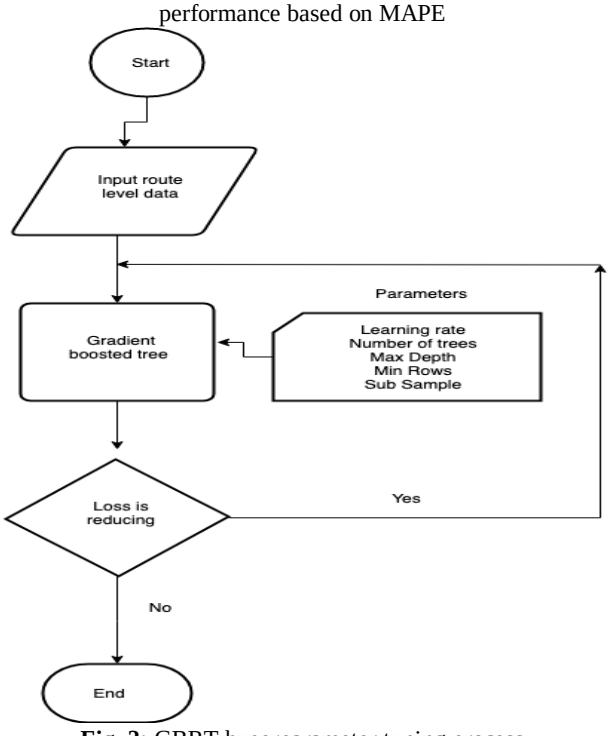


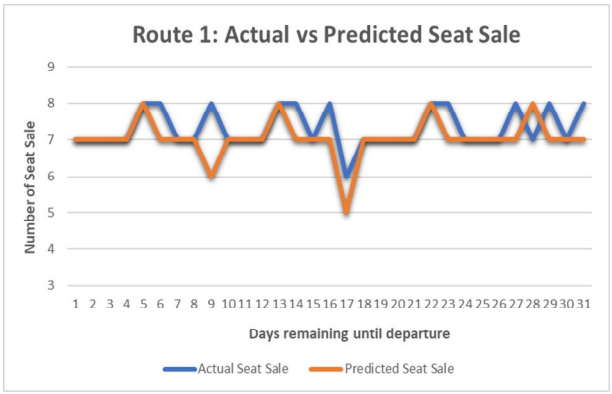
Fig. 2: GBRT hyperparameter tuning process

Table 4 Final Hyperparameter Of Gbrt Model

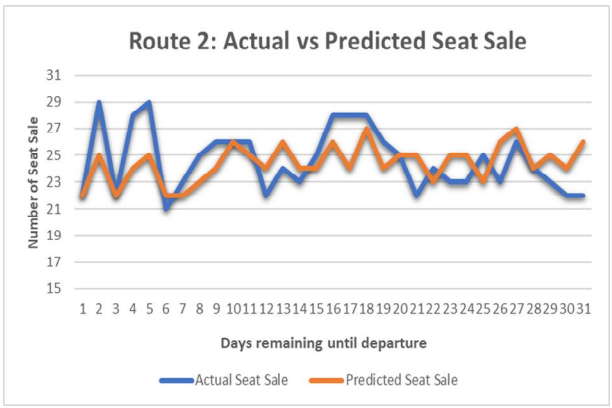
Parameters	Value
Learning rate	0.1
Number of trees	30
Max Depth	6
Min Rows	5
Sub sample	0.8

B. Result Analysis of GBRT Seat Sale Prediction Model

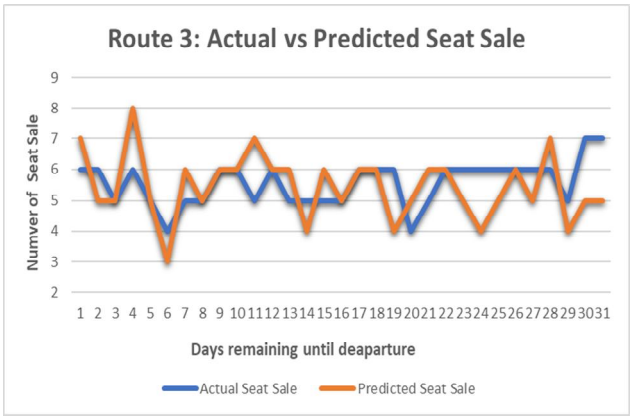
Seat sale prediction of 3 different routes for 30 and 60 days using developed GBT model have been given in figure 2 and 3. Initial GBT model with conventional attributes obtained an error rate of 11%. However, with the newly introduced digital attributes in this research, the GBT model obtained less than 10% error rate. As compared to the current running model, our predicted model exhibits better performance in terms of accuracy (1.5% increment).



(a)

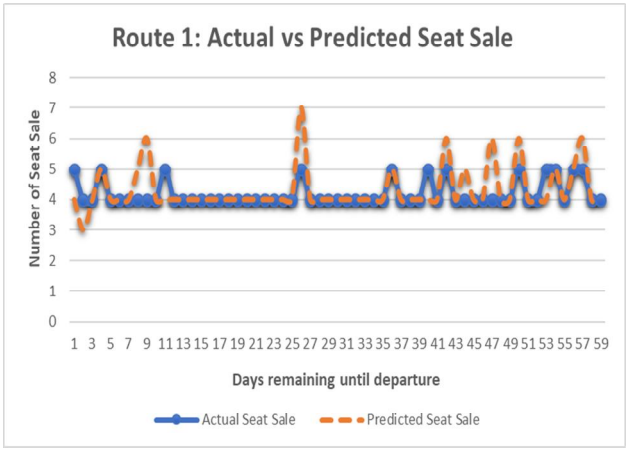


(b)

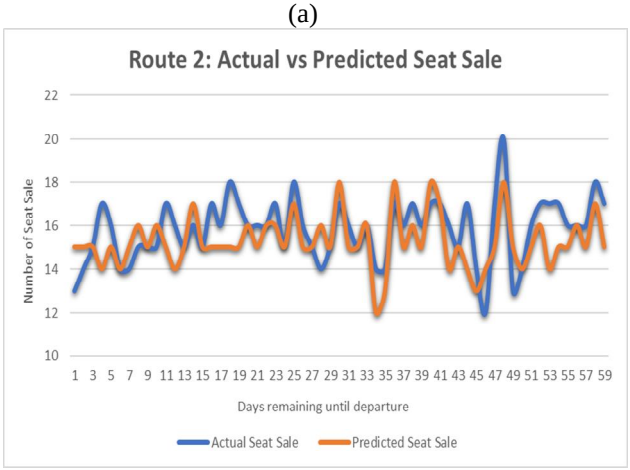


(c)

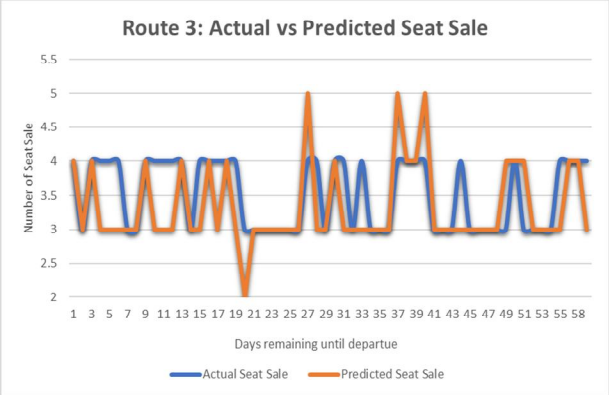
Fig. 3: Seat sale prediction for 30 days (a) Route 1 (b) Route 2 (c) Route 3



(a)



(b)



(c)

Fig. 4: Seat sale prediction for 60 days (a) Route 1 (b) Route 2 (c) Route 3

It is seen from Fig. 3 and 4 that for each route the developed model with digital attributes is able to follow the actual seat sale demand curve trend then conventional model. It is also observed that with the decrease of flight departure day (at the beginning of the curve) capturing the trend of seat demand is difficult for some routes due to price uncertainty. However, this limitation can be adjusted from the revenue management perspective, since, the price become higher during the last few days before a flight departure. It is also found that point-to-point prediction is, indeed, less suitable for time series data due to its sequential nature. Besides, the effect of one day to another is not being considered as all the observation are assumed to be independent of each other, which is rarely the case for time series data.

Figure 5 shows the accuracy comparison for 28 routes based on MAPE. First, essential and traditional attributes have been used to predict the demand using developed GBT model. After that, same GBT model has been used with essential and digital attributes for demand prediction. It can be seen that there is a significant increase in accuracy upon inclusion of the digital variables. This result validates the contribution of this research work. The current running approach (the booking curve) scores a remarkable error rate of just 8.5%, which equates to accuracy of 91.5% while the time series model we have built marginally outperforms the former, scoring us 93% in terms of prediction accuracy.

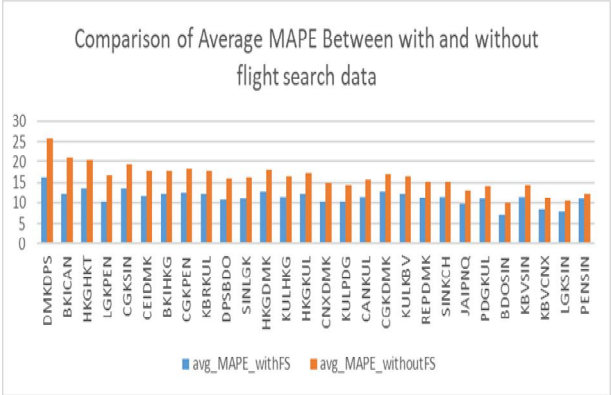


Fig. 5: Comparison of accuracy (average MAPE) for 28 routes using proposed GBT model (without digital attributes vs addition of digital attributes approach)

The proposed GBT model with digital attributes has obtained an accuracy of 93% for 30 days ahead seat sale prediction while suffer from 2% drop in accuracy (91%) for 60 days ahead prediction. Compared to point-to-point forecast, which only managed to score 37% for 30 days ahead prediction, the current model was able to perform direct route flight prediction with good enough accuracy. One argument that can be made on Time Series Model is that, it has around 70% certainty of correctly predicting overselling. 'Correctness' is determined by examining total capacity and actual seat sold. Actual seat sold has to be indeed more than or equal to total capacity or cannot go > 2% below total capacity for the prediction to be considered 'correct'.

VI. CONCLUSION

A GBRT based seat sale prediction model for airline industry has been presented in this research. New digital attributes from user click stream data have been processed and combined with traditional transactional variables to create a dataset for the model. Digital data sources for data collection has been defined and dataset preparation procedure has been described. It has been found that due to inclusion of digital attributes, the proposed model was able to obtain 93% accuracy in case of 30 days ahead seat sale prediction which is the highest to the best of author knowledge. The accuracy is dropped to 91% while predicting more than 30 days prior seat sale. However, improving this model accuracy by combining other new digital attributes and new machine learning model could be a future research goal.

REFERENCES

[1] D. Escobari, "*Estimating dynamic demand for airlines*," in Economics Letters, vol. 124, 2014, pp. 26–29.

[2] S. Puller, L. Tayloar, "*Price discrimination by day-of-week of purchase: Evidence from the U.S. airline industry*," in Journal of Economic Behavior and Organization, vol.84, 2012, pp. 801-812.

[3] K. R. Williams, "*Dynamic Airline Pricing and Seat Availability*," unpublished
[4] R.P. McAfee, V. Velde, "*Dynamic Pricing in the Airline Industry*," unpublished
[5] J. A. Abdella, N. Zaki, K. Shuaib, F. Khan, "*Airline ticket price and demand prediction: A survey*", in Journal of Kind Saud University – Computer and Information Sciences, Available: https://www.sciencedirect.com/science/article/pii/S131915781830884X.
[6] W. L. Ong, A K.G. Tan, "*A note on the determinants of airline choice: The case of Air Asia and Malaysia Airlines*," in Journal of Air Transport Management, vol.16, 2010, pp. 209 – 212.
[7] S. M. T. Fatemi Ghomi, K. Forghani, "*Airline passenger forecasting using neural networks and Box-Jenkins*," in 12th International Conference on Industrial Engineering, 2016.
[8] S. Y. Kuo and S. Chen, "*Air passenger and air cargo demand forecasting: Applying artificial neural networks to evaluating input variables*," in 12th World Conference on Transport Research, 2010, pp. 11-15.
[9] M. Zandieh, A. Azadeh, B. Hadadi and M. Saberi, "*Application of Artifical Neural Networks for Airline Number of Passenger Estimation in Time Series State*," in Journal of Applied Sciences, vol. 9, pp. 1001 – 1013.
[10] M. Varedi, "*Forecasting seat sales in passenger airlines: introducing the round-trip model*", Available: https://www.semanticscholar.org/paper/Forecasting-seat-sales-in-passenger-airlines%3A-the-Varedi/18fef81419225cedaa1d71f380f1b040b7edbe80.
[11] K. Oded, E. Muller and N. J. Vilcassim, "*easyJet pricing strategy: Should low-fare airlines offer last-minute deals*," in Quantitative Marketing and Economics 6, vol. 3, 2008, pp. 279-297.
[12] Deneckere, R. Peck and J. "*Dynamic competition with random demand and costless search: a theory of price posting*," in Econometrica, vol. 80, 2012, pp. 1185-1247.
[13] L. A. Garrow and F. S. Koppelman, "*Predicting air travelers' no-show and standby behavior using passenger and directional itinerary information*," in Journal of Air Transport Management, vol. 10, 2004, pp. 401-411.