A Framework for Airfare Price Prediction: A Machine Learning Approach

Conference Paper · July 2019

DOI: 10.1109/IRI.2019.00041

7 authors, including:

Tianyi Wang
Florida International University
11 PUBLICATIONS **198** CITATIONS
SEE PROFILE

Samira Pouyanfar
Florida International University
33 PUBLICATIONS **2,982** CITATIONS
SEE PROFILE

Haiman Tian
Florida International University
21 PUBLICATIONS **2,153** CITATIONS
SEE PROFILE

Yudong Tao
Meta
42 PUBLICATIONS **2,332** CITATIONS
SEE PROFILE

A Framework for Airfare Price Prediction: A Machine Learning Approach

Tianyi Wang*, Samira Pouyanfar*, Haiman Tian*, Yudong Tao†,
Miguel Alonso Jr.*, Steven Luis* and Shu-Ching Chen*
*School of Computing and Information Sciences
Florida International University, Miami, Florida 33199
Emails: {wtian002,spouy001,htian005,malonsoj,luiss,chens}@cs.fiu.edu
† Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124
Email: yxt128@miami.edu

***Abstract*—The price of an airline ticket is affected by a number of factors, such as flight distance, purchasing time, fuel price, etc. Each carrier has its own proprietary rules and algorithms to set the price accordingly. Recent advance in Artificial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation. This paper proposes a novel application based on two public data sources in the domain of air transportation: the Airline Origin and Destination Survey (DB1B) and the Air Carrier Statistics database (T-100). The proposed framework combines the two databases, together with macroeconomic data, and uses machine learning algorithms to model the quarterly average ticket price based on different origin and destination pairs, as known as the market segment. The framework achieves a high prediction accuracy with 0.869 adjusted R squared score on the testing dataset.**

***Keywords*-machine learning; airfare price; DB1B; T-100; prediction model;**

I. Introduction

Since the deregulation of the airline industry, airfare pricing strategy has developed into a complex structure of sophisticated rules and mathematical models that drive the pricing strategies of airfare [1] [2] [3]. Although still largely held in secret, studies have found that these rules are widely known to be affected by a variety of factors [4] [5]. Traditional variables such as distance, although still playing a significant role, are no longer the sole factor that dictate the pricing strategy. Elements related to economic, marketing and societal trends have played increasing roles in dictating the airfare prices.

Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination. Being able to predict the airfare trend at the specific market segment level is crucial for airlines to adjust strategy and resources for a specific route. However, existing studies on market segment price prediction use heuristic-based conventional statistical models, such as linear regression [6] [7], and are based on the assumption that there exists a linear relationship between the dependent and independent variables, which in many cases, may not be true.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) make it possible to infer rules and model variations on airfare price based on a large number of features, often uncovering hidden relationships amongst the features automatically. To the best of our knowledge, all existing work leveraging machine learning approaches for airfare price prediction are based on: 1) proprietary datasets that are not publicly available [8] [9] and 2) transaction records data crawled from online travel booking sites like Kayak.com [10] [11] [12]. The problem of the former lies in the difficulty of gaining access to the data, making reproducing the results and extending the work nearly impossible. The issue with the later is that the transaction records from each online booking site are a small fraction of the total ticket sales from the entire market, making the acquired data likely to be skewed, and thus, not representing the true nature of the entire market.

In this paper, we address the problem of market segment level airfare price prediction by using publicly available datasets and a novel machine learning framework to predict market segment level airfare price. More specifically, our proposed framework extracts information from two specific public datasets, the DB1B and the T-100 datasets that are collected and maintained by the Office of Airline Information within the United States Bureau of Transportation Statistics (BTS). The DB1B dataset has been utilized in various studies that assess the determinants of aircraft characteristics and frequency of flights [13], analyses for the structure and dynamics of O-D for the core of the air travel market [14], and demand-prediction [15]. The T-100 dataset includes air passenger volumes for U.S. domestic and international markets and covers large certified carriers that hold Certificates of Public Convenience and Necessity. The goal of our proposed framework is to draw a comprehensive profile of each market and uses machine learning techniques to predict the average airfare on market segment level.

The remainder of this paper is organized as follows. Section II reviews existing work that utilized either conventional

statistical or machine learning algorithms for airfare price prediction. Section III provides a detailed description of the two datasets and the proposed framework. Section IV describes the experimental setup and presents the results of applying our proposed framework, as well as a comparison with several baseline methods. In section V, we conclude the paper with a discussion of our contribution and several potential directions for future work.

II. Related Work

Air ticket price prediction is a challenging task since the factors involved in pricing dynamically change over time and make the price fluctuate. In the last decade, researchers have incorporated machine learning algorithms and data mining strategies to better model observed prices. Among them, regression models, such as Linear Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), are frequently used in predicting accurate airfare price [10][16][17].

Early work also considered using classification models to predict the trends of the itineraries. Ren *et al.* [17] proposed using LR, Naive Bayes, Softmax regression, and SVMs to build a prediction model and classify the ticket price into five bins (60% to 80%, 80% to 100%, 100% to 120%, and etc.) to compare the relative values with the overall average price. More than nine thousand data points, including six features (e.g., the departure week begin, price quote date, the number of stops in the itinerary, etc.), were used to build the models. The authors reported the best training error rate close to 22.9% using LR model. Their SVM regression model failed to produce a satisfying result. Instead, an SVM classification model was used to classify the prices into either "higher" or "lower" than the average.

In [16], four LR models were compared to obtain the best fit model, which aims to provide an unbiased information to the passenger whether to buy the ticket or wait longer for a better price. The authors suggested using linear quantile mixed models to predict the lowest ticket prices, which are called the "real bargains". However, this work is limited to only one class of tickets, economy, and only on one direction single leg flights from San Francisco Airport to John F. Kennedy Airport. Wohlfarth *et al.* [18] integrated clustering as a preliminary stage with multiple state-of-the-art supervised learning algorithms (classification tree (CART) and RF) to assist the customers' decision making process. Their framework uses the K-Means algorithm to group flights with similar behavior in the price series. They then use CART to interpret meaningful rules, and RF to provide information about the importance of each feature. Also, the authors pointed out that one element, the number of seats left, is a key freature for ticket price prediction.

Aside from flight-specific features, many other attributes affect the competitive market. Accurately predicting the market demand, for example, can reduce a travel agency's accumulated costs, which are caused by over purchasing or lost orders. In [19], the author applied Artificial Neural Network (ANN) and Genetic Algorithms (GA) to predict air ticket sales revenue for the travel agency. The input features included international oil price, Taiwan stock market-weighted index, Taiwan's monthly unemployment rate, and so on. Specifically, the GA selects the optimum input features to improve the performance of the ANNs. The model showed good performance with a 9.11% Mean Absolute Percentage Error.

Starting from 2017, more advanced machine learning models have been considered to improve airfare price prediction [10][20]. Tziridis *et al.* [10] applied eight machine learning models, which included ANNs, RF, SVM, and LR, to predict tickets prices and compared their performance. The best regression model achieved an accuracy of 88%. In their comparison, Bagging Regression Tree is identified as the best model, which is robust and not affected by using different input feature sets. In [20], Deep Regressor Stacking was proposed to reach more accurate predictions. The proposed method is a novel multi-target approach with RF and SVM as the regressors and can be easily applied to other similar problem domains.

As airline ticket data is not well organized and ready for direct analysis, collecting and processing those data always requires a great deal of effort. For most analyses found in the literature, researchers evaluate their models' performance on different datasets by either crawling the data from the web or requesting private data from collaborative organizations. As a result, it is difficult to replicate the research and conduct comparisons of the models' performance. For U.S. airlines, the fare data is publicly available in the T100 and DB1A/1B databases. However, due to the limited association between the prices and specific flights information, these datasets are seldom used independently to generate scientific research outcomes [21]. However, researchers who are interested in analyzing the price dispersion, for example, are more likely to consider investigating the information from those datasets [22]. In Rama-Murthy's dissertation [7], the Official Airline Guide (OAG) and DB1B data are used to model the airfare prices. The author also incorporates the *Sabre AirPrice* data, which was provided by SABRE, but they only provide the information of their online users. As this online user data does not represent the whole consumer market, it can bias the results obtained from the data.

Compared to the current and recent work, our proposed framework manages to handle the price prediction task only using public data sources with minimal features. Also, not restricted by any specific market segment that usually limits the existing work, this proposed framework can be applied to predict the airfare price for any market.

Table I
SUMMARY OF DATA IN DB1B AND T-100 USED IN THE PROPOSED FRAMEWORK

Entity	Availability	Data
Ticket	DB1B	fare price, total distance, and total number of passengers
Coupon	DB1B	market segment, time of the itinerary, carrier, and seat class
Market segment	DB1B&T-100	original airport, destination airport, and segment distance
Market segment by carrier	T-100	number of passengers, and number of available seats by aircraft type

III. MATERIALS AND METHOD

A. Datasets

In order to build the airline ticket price model at the market segment level, information about both the airline traffic and passenger volume for each market segment is required. Therefore, two public datasets (DB1B and T-100) are used in our proposed framework. Data collected during 2018 are used to train and evaluate the proposed model. Table I summarizes the information of these two datasets.

The United States Department of Transportation regularly updates both the DB1B and the T-100. The DB1B dataset provides quarterly-aggregated information about the airline tickets in the United States from reporting carriers and consists of 10% randomly sampled ticket data from each reporting carrier. The information in DB1B is organized in three parts, namely "Coupon", "Market", and "Ticket". A "Coupon" is an atomic unit of an airline ticket, indicating one itinerary of a passenger that is directly transported from one airport to another, while each ticket could contain multiple coupons and multiple passengers. Therefore, "Coupon" in the DB1B includes information about each leg, "Market" provides information on the market segment, such as the distance between two airports, and "Ticket" provides additional information at ticket level, such as airfare price. All the records in "Coupon" are bounded to a "Market" record and a "Ticket" record. For our proposed framework, a subset of most related data are used, including the origin airport (ORIGIN), the destination airport (DEST), time of the itinerary (QUARTER), carrier information (REPORTING_CARRIER), seat class (SEAT_CLASS) (e.g., first, business, economic, etc.), total flight distance for a ticket (DISTANCE), airfare price (ITIN_FARE), and the number of passengers in a ticket (PAX).

Different from DB1B, T-100 provides monthly domestic non-stop segment data reported by both the domestic and international carriers. It presents the number of passengers of each airline and each market segment by aircraft type.

B. Proposed Framework

Our proposed framework utilizes both the DB1B and T-100 datasets, in combination with macroeconomic data to predict the quarterly average airfare at the market segment level. Figure 1 shows an overview of the major components of the proposed framework. In the data preprocessing step, all datasets are cleaned to exclude possible erroneous

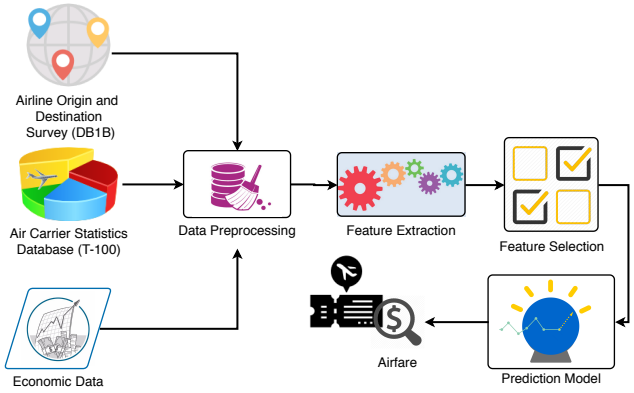

Figure 1. Proposed framework for airfare price prediction using public data sources

samples, transformed and combined based on the market segment. The feature extraction module serves to extract and generate handcrafted features that aim to characterize the market segment. The goal of the feature selection module is to optimize the performance of the prediction model by analyzing the effectiveness of the features and remove any irrelevant features. Finally, we use the selected features to build our prediction model, which generates the output value as the predicted air ticket price.

1) Data Preprocessing: In the DB1B and T-100 datasets, many attributes contain the same information. Directly merging the tables creates many duplicate fields. Also, the data reported by the airlines may include erroneous values caused by human error, currency conversion error, etc. Therefore, a properly designed data preprocessing workflow is crucial to generate accurate input data in order to build the machine learning model.

Table II shows the layout of the ticket and coupon database tables and the sample records with the same itinerary ID (ITIN_ID) in the DB1B dataset. First, the DB1B ticket and coupon tables are merged based on the ITIN_ID. The ITIN_ID is the primary key for the ticket table. In the coupon table, all of the entries that belong to the same ticket share the same ITIN_ID. Samples in the DB1B ticket table with the itinerary value (ITIN_FARE) less than $50, or distance field (DISTANCE) less than 100 miles in the Coupon table are removed because those samples in practice, are considered reporting errors. Samples with price credibility field (DOLLAR_CRED) equal to 0 are unreliable

Table II
STRUCTURE OF THE DB1B TICKET AND COUPON TABLE WITH SAMPLE RECORDS

DB1B Ticket Table							
ITIN_ID	QUARTER	ORIGIN	ITIN_FARE	DISTANCE	PAX	DOLLAR_CRED	-
2018112	1	ABE	340	1384	1	1	-
...	...	...	...	...	..	...	-
DB1B Coupon Table							
ITIN_ID	QUARTER	ORIGIN	DEST	REPORTING_CARRIER	PAX	SEAT_CLASS	DISTANCE
2018112	1	ABE	ATL	9E	1	X	692
2018112	1	ATL	ABE	9E	1	X	692
...	...	...	...	...	...	...	...

Table III
STRUCTURE OF THE T-100 DATASET WITH A SAMPLE RECORD

SEATS	PAX	CARRIER	ORIGIN	DEST	QUARTER
150	140	9E	ABE	ATL	1
...	...	...	...	...	...

carrier reports, which are also disregarded. Since only the ticket table contains the ticket price, the price for each market segment is calculated based on the ITIN_FARE in the ticket table and the distance ratio. The distance ratio measures the proportion between the distance of each leg in the coupon table and the full length of the itinerary in the ticket table. Finally, the quarterly average fare value for each SEAT_CLASS on each specific market segment is generated.

Table III shows an example record in the T-100 database. Similar to the DB1B, the "SEATS" and "PAX" fields in T-100 are aggregated based on the origin and destination airports pair for each quarter. In the final stage, the two data sources consisting of the cleaned attributes are merged based on the market segment and on a quarterly basis.

2) Feature Extraction: Several features have been extracted from the DB1B and T-100 dataset to represent a specific aspect of the market segment. Furthermore, to exploit the relationship between the airline industry and the overall economic conditions, several macroeconomic features are also added to the feature set. Table IV describes all the features that are identified during feature extraction.

Table IV
THE LIST OF FEATURES GENERATED DURING FEATURE EXTRACTION STAGE WITH EXPLANATIONS

Feature Name	**Description**
Distance	Market distance between the origin and destination airports
Seat Class	Indicator for economy or premium seat type
Passenger Volume	Total number of passengers traveled between the origin and destination airports
Load Factor	The ratio of the total number of passenger to the total number of seat in a market
Competition Factor	The market HHI
LCC Presence	Indicator of LCC operating in the market
Crude Oil Price	Quarterly average of crude oil price
CPI	Quarterly average of Consumer Price Index
Quarter	Indicates the three month period of the year

The Load Factor (LF) is a primary metric used in the transportation industry. It represents the supply and demand relationship in a given market, which strongly influences an airline's pricing strategy. The T-100 dataset includes two features, the number of available seats and the number of actual passengers carried, that allow us to calculate the LF of a market by dividing the total passenger volume (P) by the total number of Available Seats (AS) in that market segment:

$$LF = \frac{P}{AS} \tag{1}$$

The effect of competition among airlines in a given market segment has been shown to affect the pricing strategy of the airlines [23]. In a less competitive market, the market power of a given airline is stronger, and thus, it is more likely to engage in price discrimination. On the other hand, the higher the level of competition, the weaker of the market power of an airline, and then the less likely the chance of the airline fare increases. The competition factor in the proposed model is based on the Herfindahl-Hirschman Index (HHI) [24], which measures the level of competition in a given market. It is the sum of the squared fraction of the market share of each top company:

$$HHI = \sum_{a=1}^{C} s^a, \tag{2}$$

$$s^a = \frac{v^a}{P}, \tag{3}$$

where C is the total number of companies, s^a is the market share of company a, v^a is the number of passenger carried by company a, and P is the total number of passenger in the market. We used the T-100 dataset to extract the market share of each airline in a specific market segment by calculating the ratio of the number of passengers carried by that airline to the total passenger volume of the market segment.

The emergence of Low-Cost Carrier (LCC) has revolutionized the entire operating model of the airline industry. The presence of LCC in a market has had a substantial impact on the total passenger volume and the air ticket price [25]. A "LCC Presence" field is added to indicate whether the "Carrier" field in the DB1B coupon table

contains the International Air Transport Association (IATA) code [26] related to one of the LCCs operating in the U.S. domestic markets. The six LCCs are Allegiant Air, Frontier Airlines, JetBlue, Southwest Airlines, Spirit Airlines, and Sun Country Airlines.

Macroeconomic data, such as crude oil price and Consumer Price Index (CPI), can also be utilized to uncover the hidden trend in airline fares. Fuel costs can take up to 50% of the total operating cost of an airline [27]. Hence, the level of crude oil price plays an essential rule of formulating the airline's pricing strategy. It is a common practice for airlines to pass the cost of aviation fuel to the customer by adjusting the fare to compensate for the fluctuation of crude oil price. In this paper, we used the West Texas Intermediate (WTI) crude oil price data and calculated its quarterly average value. Furthermore, the CPI measures the weighted average prices of various types of consumer goods and services, which include the prices in the transportation industry [28]. Therefore, we exploit its potential to measure the current level of air travel cost. The monthly CPI data is acquired from the Organization for Economic Co-operation and Development. Similar to the crude oil price, we calculate the quarterly averaged value. Figure 2 depicts the quarterly value trend of crude oil price, CPI, and airfare from 2006 to 2017. It demonstrates a clear relationship between the three types of data.

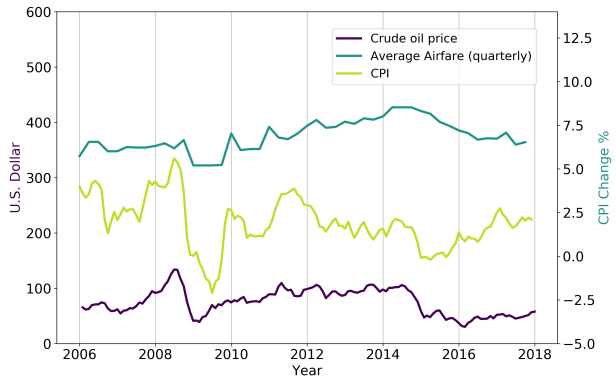

Figure 2. A comparison between the crude oil price, CPI and the quarterly averaged airfare from 2006 to 2017

3) Feature Selection: A feature selection technique is applied to improve the model performance by investigating the degree of impact of each feature on the prediction result. We utilize RF to construct an automated feature selection module. RF is a tree-based ensemble learning algorithm that builds multiple decision tree classifier during the training phase and outputs the predicted results based on either the majority vote (classification) or the average (regression) of the predictions of all decision trees. After training the RF model with the entire feature set, it ranks all the features by their importance. A feature's importance is measured by the average decrease in impurity. It is the total decrease in the node's impurity caused by the corresponding feature, weighted by the chance that the decision path includes this node. There are several ways of choosing the impurity metric, and because our target is a continuous value, Sum of Squared Errors (SSE) is chosen as the impurity metric. The SSE for node o can be calculated as:

$$SSE_o = \sum_{j=1}^{S} \epsilon_j^2, \tag{4}$$

where S is the number of splits from the node, and ϵ is the error between the true value and predicted value. The Node Importance (NI) for node o can be calculated as (assuming the parent node splits into two child nodes):

$$NI_o = w_o SSE_o - w_l SSE_l - w_r SSE_r, \tag{5}$$

where w_o, w_l and w_r are the weighted number of samples pass through node o and it's left and right child node. Then, the Feature Importance (FI) for feature x can be calculated as

$$FI_x = \frac{\sum_{b, b \in nodes\ split\ on\ feature\ x} NI_b}{\sum_{k, k \in\ all\ nodes} NI_k}. \tag{6}$$

Generally, a feature gains more importance when it has a greater effect of reducing the prediction error.

In the next step, the feature selection module applies Recursive Elimination (RE) to select the best set of features for the prediction model. More specifically, for each iteration, the feature with the lowest feature importance is eliminated, and the model will be retrained on the updated input. This process terminates when the act of removing more feature does not improve the model's performance.

4) Machine learning model: When developing the ML model, we chose RF as the learner for the airfare price prediction task. Based on our empirical study, the RF model demonstrates the best performance on the data as compared to several ML techniques including LR, SVM, and neural networks. Comparison results are explained in Section IV.

IV. Experiments and Analysis

A. Experimental Setup

For our experiments, we collected 16,577,497 and 41,360,566 samples from the 2018 DB1B ticket table and coupon table, respectively. The T-100 dataset contains 329,426 samples. We tested several well-known machine learning models as baselines to compare with the RF model. In particular, LR, SVM, Multilayer Perceptrons (MLPs), and XGBoost Tree were used for the evaluation. For the SVM model, the radial basis function kernel is used, the tolerance for stopping criterion is set to 0.001, the penalty parameter for the error term is set to 0.1. For the MLPs, three hidden layers are used with 30 neurons per layer. The Rectified Linear Unit (ReLU) [29] is used as the activation function

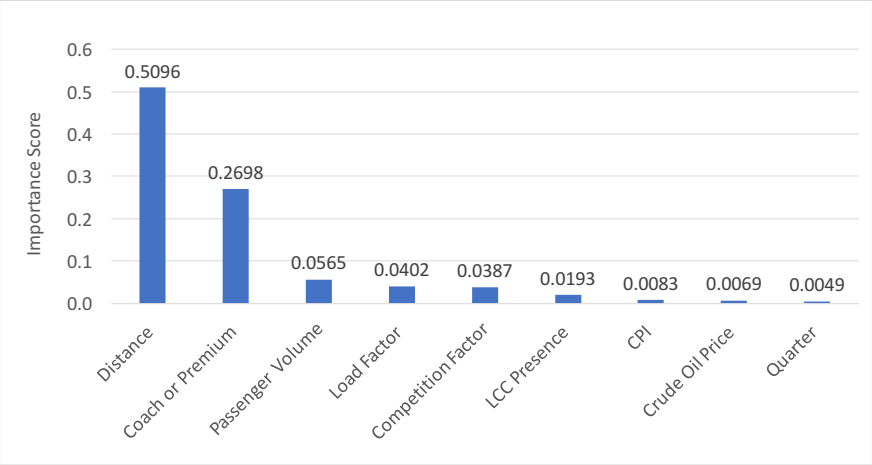

Figure 3. Importance score value for each feature generated by RF

and Adam is the optimization function [30]. The learning rate is set to 0.0001 with momentum enabled set to 0.9. For the XGBoost model, the number of estimators is set to 100 with a learning rate as 0.1, and max depth equals to 5. For the RF model, the number of estimators is also set to 100 with the minimum number of samples to split set to 2. To evaluate the proposed price prediction model, two popular metrics for regression analysis are used: the Root Mean Square Error (RMSE) and the Adjusted R Squared.

RMSE calculates the differences between the observed values, y, and predicted values, $\hat{y}$. This difference for each data point is also called the residual. Thus, RMSE is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{7}$$

where N is the total sample size. The lower the RMSE value is, the higher performance the regression model has.

The Adjusted R Squared, (R^2_{adj}), is usually used to explain how well the independent variables fit a curve or line. Adjusted R^2 also adjusts for the number of variables in a model. The higher the Adjusted R Squared is, the better the result of regression is. It is calculated as follows:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(N - 1)}{N - p - 1}\right] \tag{8}$$

where p is the number of predictors and R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{9}$$

here $\bar{y}$ is the mean value of y.

Table V
PERFORMANCE COMPARISON FOR DIFFERENT REGRESSION MODELS WITH AND WITHOUT FEATURE SELECTION

Method	without feature selection		with feature selection	
	RMSE	$\mathbf{R^2_{adj}}$	RMSE	$\mathbf{R^2_{adj}}$
LR	111.000	0.612	110.284	0.618
SVM	112.963	0.587	108.358	0.626
MLP	88.447	0.754	85.832	0.766
XGBoost	83.481	0.778	80.447	0.797
RF	**66.584**	**0.858**	**62.753**	**0.869**

B. Experimental Results

In order to demonstrate the importance of each feature for airfare price prediction, we extracted the importance scores generated by the feature selection module. Figure 3 depicts the importance value for each feature. As shown in the figure, "Distance" and "Seat Class" (Economy or business) are the most important factors for airfare price estimation followed by "Passenger Volume", "Load factor", and "Competition Factor". Although the "CPI" and "Crude Oil Price" do not have very high importance scores, they can still help the model predict a more accurate estimation of the airfare price. However, based on our experiments, "Quarter" does not help the regression model. Including the variable "Quarter" does not reduce the error during the training phase. Thus, it is automatically removed by the RF feature selection module. The goal is to identify the features that improve the model's performance and adding irrelevant features deteriorates the model's performance, as the model learns an irrelevant pattern.

The results comparing various regression models with feature selection and without feature selection are shown in Table V. As can be seen from this table, LR and SVM have the lowest performance compared to other ML methods with respect to the RMSE and R^2_{adj} metrics. The performance of all of the models improves after applying feature selection,

Table VI
PERFORMANCE COMPARISON FOR DIFFERENT REGRESSION MODELS WITHOUT LOAD FACTOR, COMPETITION FACTOR, CPI, AND CRUDE OIL PRICE FEATURES

Method	RMSE	$\mathbf{R^2_{adj}}$
LR	112.039	0.599
SVM	109.914	0.615
MLP	94.569	0.715
XGBoost	90.419	0.739
RF	**70.575**	**0.804**

which illustrates the importance of this module. XGBoost performs better than MLP, SVM, and LR, but does not outperform RF for airfare price prediction. Therefore, we utilize RF in the proposed framework, which achieves the highest performance compared to other baselines for this dataset. Specifically, it reaches 62.753 and 0.869 RMSE and R^2_{adj}, respectively. In other words, it improves the R^2_{adj} by 40% compared to the LR model, which is extensively used in the previous studies for airfare price prediction.

To show the importance of features specifically employed for our regression model, another experiment was conducted. In this experiment, we only used common features with very high importance scores such as "Distance", "Seat Class", and "Passenger Volume". The results are presented in Table VI. Again, we find that LR and SVM have lower performance compared to other models, and RF reaches the highest performance with respect to both RMSE and R^2_{adj}. However, the performance (R^2_{adj}) dropped by almost 7% for the RF model when the less important factors are removed. Similarly, the performance for other models dropped as well. Although the less important factors may not contribute significantly to the performance, these results show that to achieve the best performing model, one should include the "Load factor", "Competition Factor", "CPI", and "Crude Oil Price" as features. Consequently, the proposed framework utilizes all of these features to achieve the highest airfare price prediction performance.

V. CONCLUSION AND FUTURE WORK

In this study, a machine learning framework was developed to predict the quarterly average airfare price on the market segment level. We combined the U.S. domestic airline tickets sales data and non-stop segment data from two public datasets (DB1B and T-100). Several features were extracted from the datasets and combined together with macroeconomic data, to model the air travel market segments. With the help of the feature selection techniques, our proposed model is able to predict the quarterly average airfare price with an adjusted R squared score of 0.869.

To the best of our knowledge, most of previous studies on airfare price prediction using the DB1B dataset have focused on conventional statistical approaches, which have their own limitations of problem estimations and assumptions. Also, to our knowledge, no other studies have integrated the information from DB1B, T-100, and macroeconomic data to model the air travel market segment. Thus, our study demonstrates the effectiveness of machine learning algorithms and techniques, as well as compares the performance of various machine learning classifiers and finds the best one for the airfare price prediction task by leveraging the information from the DB1B and T-100 datasets.

However, there are several limitations with the techniques caused by the nature of the DB1B and T-100 datasets that are worth noting. For example, none of the datasets have detailed information related to the ticket sales such as the time and day of the week for departure and arrival. Also, because of the frequency of the dataset release, the fare prediction can only be calculated on a quarterly base.

In the future, our framework can be extended to include air ticket transaction information, which can provide more detail about a specific itinerary, such as time and date of departure and arrival, seat location, covered ancillary products, etc. By combining such data with the existing market segment and macroeconomic features in the current framework, it is possible to build a more powerful and comprehensive airfare price prediction model on the daily or even hourly level. Furthermore, airfare price in a market segment can be affected by a sudden influx of large volume of passengers caused by some special events. Thus, events information will also be collected from various sources, which include social platforms and news agencies, as to complement our prediction model. Additionally, we will investigate other advanced ML models, such as Deep Learning models, while working to improve the existing models by tuning their hyper-parameters to reach the best architecture for airfare price prediction.

ACKNOWLEDGMENT

The authors would like to thank Tim Reiz (Chief Technology Officer), David Welborn (Business intelligence Architect), and Diana Porro (Data Scientist) from Farelogix Inc. for providing input and support.

REFERENCES

[1] J. Stavins, "Price discrimination in the airline market: The effect of market concentration," *Review of Economics and Statistics*, vol. 83, no. 1, pp. 200–202, 2001.

[2] B. Mantin and B. Koo, "Dynamic price dispersion in airline markets," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no. 6, pp. 1020–1029, 2009.

[3] P. Malighetti, S. Paleari, and R. Redondi, "Has ryanair's pricing strategy changed over time? an empirical analysis of its 2006–2007 flights," *Tourism Management*, vol. 31, no. 1, pp. 36–44, 2010.

[4] T. H. Oum, A. Zhang, and Y. Zhang, "Inter-firm rivalry and firm-specific price elasticities in deregulated airline markets," *Journal of Transport Economics and Policy*, vol. 7, no. 2, pp. 171–192, 1993.

[5] B. Burger and M. Fuchs, "Dynamic pricing – A future airline business model," *Journal of Revenue and Pricing Management*, vol. 4, no. 1, pp. 39–53, 2005.

[6] T. M. Vowles, "Airfare pricing determinants in hub-to-hub markets," *Journal of Transport Geography*, vol. 14, no. 1, pp. 15–22, 2006.

[7] K. Rama-Murthy, "Modeling of united states airline fares–using the official airline guide (OAG) and airline origin and destination survey (DB1B)," Ph.D. dissertation, Virginia Tech, 2006.

[8] B. Derudder and F. Witlox, "An appraisal of the use of airline data in assessing the world city network: a research note on data," *Urban Studies*, vol. 42, no. 13, pp. 2371–2388, 2005.

[9] A. Mottini and R. Acuna-Agost, "Deep choice model using pointer networks for airline itinerary prediction," in *the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 1575–1583.

[10] K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," in *the 25th IEEE European signal processing conference*, 2017, pp. 1036–1039.

[11] Y. Chen, J. Cao, S. Feng, and Y. Tan, "An ensemble learning based approach for building airfare forecast service," in *the IEEE international conference on big data*, 2015, pp. 964–969.

[12] T. Liu, J. Cao, Y. Tan, and Q. Xiao, "ACER: An adaptive context-aware ensemble regression model for airfare price prediction," in *the international conference on progress in informatics and computing*, 2017, pp. 312–317.

[13] V. Pai, "On the factors that affect airline flight frequency and aircraft size," *Journal of Air Transport Management*, vol. 16, no. 4, pp. 169–177, 2010.

[14] M. S. Ryerson and H. Kim, "Integrating airline operational practices into passenger airline hub definition," *Journal of Transport Geography*, vol. 31, pp. 84–93, 2013.

[15] H. Baik, A. A. Trani, N. Hinze, H. Swingle, S. Ashiabor, and A. Seshadri, "Forecasting model for air taxi, commercial airline, and automobile demand in the united states," *Transportation Research Record*, vol. 2052, no. 1, pp. 9–20, 2008.

[16] T. Janssen, T. Dijkstra, S. Abbas, and A. C. van Riel, "A linear quantile mixed regression model for prediction of airline ticket prices," *Radboud University*, 2014.

[17] R. Ren, Y. Yang, and S. Yuan, "Prediction of airline ticket price," *University of Stanford*, 2014.

[18] T. Wohlfarth, S. Clémençon, F. Roueff, and X. Casellato, "A data-mining approach to travel price forecasting," in *the 10th international conference on machine learning and applications and workshops*, vol. 1, 2011, pp. 84–89.

[19] H.-C. Huang, "A hybrid neural network prediction model of air ticket sales," *Telkomnika Indonesian Journal of Electrical Engineering*, vol. 11, no. 11, pp. 6413–6419, 2013.

[20] E. J. Santana, S. M. Mastelini, and S. Barbon Jr, "Deep regressor stacking for air ticket prices prediction," in *the XIII Brazilian symposium on information systems: information systems for participatory digital governance*. Brazilian Computer Society (SBC), 2017, pp. 25–31.

[21] S. Mumbower and L. A. Garrow, "Data set – Online pricing data for multiple us carriers," *Manufacturing & Service Operations Management*, vol. 16, no. 2, pp. 198–203, 2014.

[22] M. Dai, Q. Liu, and K. Serfes, "Is the effect of competition on price dispersion nonmonotonic? evidence from the us airline industry," *Review of Economics and Statistics*, vol. 96, no. 1, pp. 161–170, 2014.

[23] K. S. Gerardi and A. H. Shapiro, "Does competition reduce price dispersion? new evidence from the airline industry," *Journal of Political Economy*, vol. 117, no. 1, pp. 1–37, 2009.

[24] S. A. Rhoades, "The herfindahl-hirschman index," *Federal Reserve Bulletin*, vol. 79, p. 188, 1993.

[25] G. Francis, A. Fidato, and I. Humphreys, "Airport–airline interaction: the impact of low-cost carriers on two european airports," *Journal of Air Transport Management*, vol. 9, no. 4, pp. 267–273, 2003.

[26] International Civil Aviation Organization, "List of low-cost-carriers (LCCs)," cited July 2018. [Online]. Available: https://www.icao.int/sustainability/Documents/LCC-List.pdf

[27] C. Koopmans and R. Lieshout, "Airline cost changes: To what extent are they passed through to the passenger?" *Journal of Air Transport Management*, vol. 53, pp. 1–11, 2016.

[28] S. Lee, K. Seo, and A. Sharma, "Corporate social responsibility and firm performance in the airline industry: The moderating role of oil prices," *Tourism Management*, vol. 38, pp. 20–30, 2013.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *the 27th international conference on machine learning*, 2010, pp. 807–814.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *the 3rd international conference on learning representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980