Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8924095, 12 pages
https://doi.org/10.1155/2020/8924095

Research Article

Air Transportation Direct Share Analysis and Forecast

Xufang Zheng,[1] **Chia-Mei Liu**,[2] **and Peng Wei**[1]

[1]*Department of Aerospace Engineering, Iowa State University, Ames, Iowa 50010, USA*
[2]*Office of Aviation Policy and Plans, Federal Aviation Administration, Washington, DC 20591, USA*

Correspondence should be addressed to Xufang Zheng; xfzheng@iastate.edu

Received 8 April 2019; Revised 14 October 2019; Accepted 4 January 2020; Published 1 February 2020

Academic Editor: Antonio Comi

Air transportation direct share is the ratio of direct passengers to total passengers on a directional origin and destination (O&D) pair. Direct share is an essential factor of passenger flow distribution and shows passengers' general preference for direct flight services on a certain O&D. A better understanding and a more accurate forecast of direct share can benefit air transportation planners, airlines, and airports in multiple ways. In most of the previous research and applications, it is commonly assumed that direct share is a fixed ratio, which contradicts the air transportation practice. In the Federal Aviation Administration (FAA) Terminal Area Forecast (TAF), the O&D direct share is forecasted as a constant based on the latest observation of direct share on the O&D. To find factors which have significant impacts on O&D direct share and to build an accurate model for O&D direct share forecasting, both parametric and nonparametric machine learning models are investigated in this research. We propose a novel category-based learning method which can provide better forecasting performance compared to employing the single modeling method for O&D direct share forecasting. Based on the comparison, the developed category-based learning model is a promising replacement for the model used for O&D direct share forecasting by the FAA TAF.

1. Introduction

In air transpiration, each trip is a connection between the origin and destination airports. The directional pair of origin and destination airports is known as the O&D pair or O&D market. For the same O&D pair, there are usually multiple itineraries provided by different carriers, which include both direct and nondirect itineraries. Passengers book different itineraries based on their own traveling preference. The passengers flying directly from the origin airport to the destination airport and the passengers taking one-connect without flight change are direct passengers. The passengers taking one-connect with flight change and the passengers taking multiple connects are nondirect passengers. Air transportation direct share (*directShare*) is the ratio of direct passengers to total passengers on a certain O&D pair.

Illustrated in Figure 1 is an example of direct and nondirect passengers' distribution on an O&D. On the O&D pair, Charlotte Douglas International Airport (CLT) ⟶ Phoenix Sky Harbor International Airport (PHX), the passengers flying directly from CLT to PHX (n_1) and the passengers taking one-connect at Hartsfield–Jackson Atlanta International Airport (ATL) without flight change (n_2) are direct passengers. The passengers taking one-connect at ATL with flight change (n_3) and the passengers taking multiple connects at Chicago O'Hare International Airport (ORD) and Denver International Airport (DEN) (n_4) are nondirect passengers. For the sake of simply and clearly describing the definition of direct share, we assume all the existing itineraries on CLT ⟶ PHX are included in Figure 1. The $directShare_{\mathrm{CLT}\longrightarrow\mathrm{PHX}}$ can be computed by equation (1), in which n_D is the number of the direct passengers and n_T is the number of the total passengers:

$$directShare_{\mathrm{CLT}\longrightarrow\mathrm{PHX}} = \frac{n_D}{n_T} = \frac{n_1 + n_2}{n_1 + n_2 + n_3 + n_4}. \quad (1)$$

O&D direct share shows the distribution of direct passengers and nondirect passengers on an O&D market, which is an essential factor of air transportation passenger distribution. O&D direct share shows passengers' general preference for direct flight services under a certain market status. A better understanding and a more accurate forecast

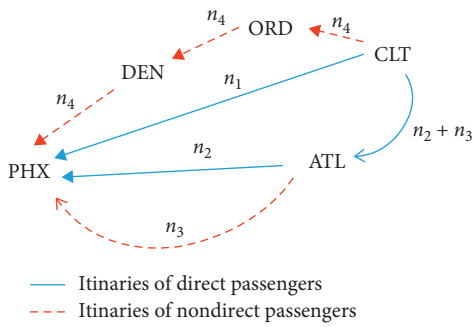

Figure 1: Illustration of the definition of O&D direct share.

of direct share can benefit air traffic planners, airlines, and airports in multiple ways.

With the increasing need for the refined forecasting of air passenger demand, air traveler itinerary demand forecasting becomes a research topic which receives much attention from both academia and industry [1, 2]. One of the major problems in this research is how to model passengers' preferences between direct and nondirect flight services. Direct share is a numeric indicator of the passengers' general preference for direct flight services on an O&D market [3]. In most previous research studies and applications, it commonly assumes the direct share as a constant percentage [4–6]. However, in air transportation practice, the assumption of constant direct share is not hold. For instance, with the increase in the low-cost carrier's market share, the direct share on Austin-Bergstrom International Airport (AUS) ⟶ Baltimore/Washington International Thurgood Marshall Airport (BWI) kept increasing from 1995 to 2005 before reaching a stable state. Another example is the direct share on Ted Stevens Anchorage International Airport (ANC) ⟶ San Francisco International Airport (SFO). Because of seasonal tourism at Alaska, there is a strong seasonality in that direct flight market. The average of quarterly $directShare_{ANC \longrightarrow SFO}$ is around 0.5 for the third quarter in each year and 0.1 for the other three quarters. The assumption of constant O&D direct share neglects the dynamic changes in the O&D direct share and the fact that other factors of the O&D market can impact the O&D direct share. This research aims to carry out a careful analysis of the characteristics of O&D direct share and find the factors which can have significant impacts on O&D direct share.

Forecast of O&D direct share is a serious concern of air traffic planners, airlines, and airports. An accurate forecast of O&D direct share is of great significance for decision-making on airport planning and investment, airline market competition, and airport labor work scheduling. The Federal Aviation Administration (FAA) Terminal Area Forecast (TAF) is the official FAA forecast for enplanements, airport operations, terminal radar approach control (TRACON) facilities operations, and based aircraft [7]. There is a wide range of applications of the FAA TAF, such as the air traffic controller workforce planning, airport long-term investment analysis [8], and airport environmental study [9]. O&D direct share forecasting is an essential component in the FAA TAF, which helps FAA making decisions on airport planning and investment. However, the model used for direct share forecasting by the FAA TAF assumes O&D direct share as a constant for forecasting. This research aims to develop a promising and reliable O&D direct share forecasting replacement for the model used by FAA TAF.

2. Literature Review

There are seldom investigations about O&D direct share in previous research and practice. To the best of our knowledge, this is the first study focusing on O&D direct share analysis and forecasting based on data mining and machine learning techniques. The air transportation itinerary demand forecasting is the most related research, which gave us inspirations on database and model development.

The research of itinerary demand forecasting focuses on forecasting the passenger demand on different itineraries on a certain O&D market. The methods can be categorized into top-down and bottom-up methods. The top-down methods are based on the forecast of total passenger demand on area or O&D level. The itinerary demand is generated by multiplying a fixed share to the total demand. The fixed share is usually the current share. The bottom-up methods forecast the itinerary passenger demand directly on the market level [10, 11]. The accuracy of top-down methods depends on the assumption that the relevant airports' market share stays constant. Because of this reason, the ACRP report does not recommend this approach [11]. The most widely used bottom-up method for itinerary demand forecasting is the discrete choice model, which is based on the simulation of passengers' itinerary choice making [12–15]. The discrete choice models simulate the passengers' booking behavior based on the socioeconomic change (e.g., GDP, population, and income per capita), O&D characteristics (e.g., number of airlines, airport accessibility, and airport region), and air travel service level (e.g., travel time and distance, average airfare, and level of service). In the discrete choice models, two kinds of data are usually employed: the revealed preference (RP) data and the stated preference (SP) data. RP data are from historical air travel records, and SP data are from passenger surveys. Bias usually exists in passenger surveys because of the design of questionnaires and information collecting methods. Important information for the survey, such as personal information and experience, is difficult to be selected or quantified [16]. The RP data is a more reliable and proper source for getting information about the historical air transpiration market competition, airlines' operation, and passengers' booking behaviors.

If employing bottom-up methods (e.g., discrete choice model) for direct share forecasting, the passengers' preference model has to be developed firstly to calculate the ratio of direct passengers and total passengers. However, how to measure the modeling accuracy of the passengers' preference is of great difficulty, especially when there are lots of itineraries on that O&D. In addition to that, the accuracy of the forecasting of direct share will depend on the accuracy of the foresting of multiple variables, which makes the modeling process more complicated and unreliable. Concerning the mentioned issues, we develop forecasting model which directly forecasts O&D direct share on the O&D level, instead

of on the passenger level. To fit a model that relates the labeled numeric response to the features is a typical supervised learning problem, more precisely a typical regression problem. Machine learning models can automatically extract knowledge about the relation of response and features from data and can forecast the future of the response [17, 18]. Based on whether there is a predetermined form of the model and whether there is a fixed number of parameters, the regression models can be categorized into parametric and nonparametric regression models [19, 20].

For parametric regression models, there is a predetermined formulation of the model, and the number of parameters is predetermined as well. The most widely used parametric regression model is the linear regression [21]. Linear regression models the relation between the response and the features in a linear manner, which makes the model easily interpretable [22]. Because direct share is a continuous random variable between 0 and 1, logit and logistic transformations are necessary while applying linear regression models to guarantee the modeling boundary [23]. Beta regression is a parametric regression model specially developed for ratio and proportion modeling and forecasting. Beta regression defines a regression model for beta-distributed random variables [24], for which the logit and logistic transformations are no longer needed. Compared to the parametric regression models, there is no predetermined model formulation for the nonparametric regression models. Nonparametric models allow a more flexible regression modeling of the response that combines the features in a nonparametric manner [25]. Tree-based models are based on decision trees, which are widely employed in regression problems. The wide range of applications show the promising prediction and forecasting performance of tree-based models [26–29].

In this research, a database comprising representative features for direct share modeling and forecasting is developed first. Parametric regression models are developed to find features that impact O&D direct share significantly. Nonparametric models are investigated to forecast O&D direct share accurately. To further improve the forecasting performance, a novel category-based learning method is proposed in this paper. The remainder of this paper is organized as follows. We introduce feature engineering and data sets in Section 3. In Section 4, model development is discussed in detail with analysis. The newly proposed category-based learning method is introduced in Section 5. The modeling and forecasting performance of different models is compared and analyzed in Section 6. We draw conclusions in Section 7.

3. Feature Engineering and Data Sets

3.1. Problem Formulation. Quarterly O&D direct share is studied in this research. Denote direct share on the O&D pair $A \longrightarrow B$ at quarter $t (t \in T)$ as $directShare_{A \longrightarrow B,t}$. The model for O&D direct share can be formulated as equation (2). The direct share on O&D pair A $\longrightarrow$ B is described by the matching feature set $X_{A \longrightarrow B,t}$. The model $f(\mathbf{X})$ is a certain parametric or nonparametric regression model:

$$directShare_{A \longrightarrow B,t} = f\left(\mathbf{X}_{A \longrightarrow B,t}\right). \tag{2}$$

3.2. Features. Based on the literature review in Section 2, we include three categories of features in this study. Table 1 lists the features in each category.

The features characterize different aspects of a certain O&D market. Some features are commonly used in air transportation demand and itinerary demand analysis and forecasting. For example, *ODPaxLag* is the yearly lag of the quarter passengers on an O&D, which can show the air transpiration demand on an O&D [6]. *CarrierNumLag* is the quarterly lag of the number of carriers on an O&D market, which reflects the airlines' competition on a certain O&D. Another example is the *MidPaxLag*, which is the yearly lag of connecting passengers at the most chosen connecting airport on a certain O&D. It shows the popularity of the most chosen connect airport on an O&D market with connects. Some features are uniquely defined in this research. For instance, *RelativeFare* and *RelativeMile* are the relative average airfare and miles flown on an O&D, which are based on equations (3) and (4), respectively. A relatively small *RelativeFare* shows strong pricing competitiveness of the nondirect flight services on a certain O&D. A relatively small *RelativeMile* shows that direct flight services have a distinct advantage of flying distance and time compared to the nondirect flight services:

$$RelativeFare = \frac{mean\left(Airfare_{direct}\right)}{mean\left(Airfare_{non-direct}\right)}, \tag{3}$$

$$RelativeMiles = \frac{mean\left(Miles_{direct}\right)}{mean\left(Miles_{non-direct}\right)}. \tag{4}$$

The weighted personal income (*WeightedIncome*) is a feature that reveals the personal income level of the connect airports. For an O&D pair, if there are n one-connect itineraries, Pax_j is the passengers choosing connect airport j, and Inc_j is the personal income of the city where the connect airport j is located. The *WeightedIncome* can be computed by the following equation:

$$WeightedIncome = \sum_{j=1}^{n} \frac{Pax_j}{\sum_{j=1}^{n} Pax_j} Inc_j. \tag{5}$$

Air carriers can be categorized into legacy carriers and low-cost carriers. The operation network of legacy carriers is commonly hub-and-spoke, while the low-cost carriers prefer point-to-point operations. The level of service and airfare vary for the two categories of carriers as well. We define *LegacyShare* as the ratio of the passengers carried by legacy carriers to the passengers carried by low-cost carriers as equation (6). *LegacyShare* shows whether the O&D market is dominated by legacy carriers:

Table 1: Feature category and features.

Feature	Meaning
O&D feature category	
directShareQLag	Quarterly lag of *directShare*
directShareYLag	Yearly lag of *directShare*
ODPaxLag	Yearly lag of O&D quarterly total passenger
OrgPaxLag	Yearly lag of quarterly departing passengers at the origin airport
DestPaxLag	Yearly lag of quarterly arrival passengers at the destination airport
MidPaxLag	Yearly lag of connecting passengers at the most chosen connecting airport on the O&D
ConnectOrgLag	Quarterly lag of connections from the origin airport
ConnectDestLag	Quarterly lag of connections to the destination airport
ConnectMidLag	Quarterly lag of one-connect itineraries using the most chosen connect airport as the connect
SchDepLag	Quarterly lag of scheduled departures on the O&D
CarrierNumLag	Quarterly lag of number of carriers on the O&D
LegacyShareLag	Quarterly lag of the ratio of passengers carried by legacy airlines on the O&D
Air travel feature category	
AverageFareLag	Quarterly lag of average airfare on the O&D
AverageMileLag	Quarterly lag of average miles flown on the O&D
RelativeFareLag	Quarterly lag of relative airfare on the O&D
RelativeMileLag	Quarterly lag of relative miles flown on the O&D
Socioeconomic feature category	
IncomeOrgLag	Quarterly lag of personal income of origin city
IncomeDestLag	Quarterly lag of personal income of the destination city
WeightedIncomeLag	Quarterly lag of weighted personal income on the O&D

$$LegacyShare = \frac{\sum Pax_{\text{carried by legacy airlines}}}{\sum Pax_{\text{total}}}. \quad (6)$$

3.3. Data. To develop the database for direct share modeling and forecasting containing the features listed in Table 1, data mining is carried out on the Airline Origin and Destination Survey (DB1B) database, the Air Carrier Statistics (T100) database, and the Global Insight Economic database. For some features (*directShareYLag*, *directShareQLag*, and *LegacyShare*), one single database is needed. For some other features (*SchDep*, *WeightedIncome*, and *IncomeOrg*), the combination of information from the different database is needed.

The DB1B database is a 10% sample of airline tickets reported by carriers to the Bureau of Transportation Statistics (BTS) [30]. There are three tables in this quarterly database: DB1BCoupon, DB1BMarket, and DB1BTicket. DB1BMarket is the data table on the O&D market level, which contains rich air travel information including market airfare, distance flown, and passenger number. The T100 data bank contains domestic and international airline market and segment data [31]. T100 is a database built upon the data reported by air carriers as well. T100 Domestic Segment Data and T100 Domestic Market data by U.S. Carriers are explored in this research. The economic database explored in this research is the IHS Global Insight Economic database, which comprises rich economic information of a city or area [32]. The personal income information is used in this research to generate the features of *IncomeOrg*, *IncomeDest*, and *WeightedIncome*.

The developed direct share database covers 3350 O&Ds, which connecting 223 busiest airports across the U.S. [33]. To avoid the significant changes in air transportation industry by Post-9/11, the time scope of the developed quarterly database is from 2008 to 2017. Shown in Table 2 is a basic analysis of the important features, which include the minimum, average, and maximum of the features generated from the DB1B and T100 databases.

4. Model Development

For the parametric machine learning models, there is a predetermined formulation, which makes the parametric machine learning models easier to be interpreted. The parametric machine learning models can automatically identify the significance of the features' impacts on the response based on the estimation of the coefficients. With a more flexible modeling approach, the nonparametric machine learning models are powerful in predicting and forecasting, especially for problems based on real-world data. To fully exploit the interpretable and forecasting capabilities of different models, both parametric and nonparametric machine learning models are investigated carefully in this research. The feature analysis is carried out based on parametric models, while the accurate forecasting models are developed based on nonparametric models.

The entire database is randomly split into three data sets. The training set contains 60% of the observations, which is used for model fitting. 20% of the observations are employed for feature selection and parameter tuning as the validation set. The other 20% of observations are used as the testing set for forecasting performance measurement. Logit and logistic transformations are used in linear regression modeling, as shown in Figure 2. Equations (7) and (8) are the formulations of the logit and logistic transformations, respectively:

Table 2: Basic analysis of important feature.

Feature (unit)	Minimum	Maximum	Average
directShareQLag (ratio)	0	1	0.5832
directShareYLag (ratio)	0	1	0.5743
ODPaxLag (person-time)	1	21436	1582
OrgPaxLag (person-time)	216	575652	141565
DestPaxLag (person-time)	211	574131	140844
MidPaxLag (person-time)	5	610771	230743
ConnectOrgLag (count)	11	217	200
ConnectDestLag (count)	10	217	200
ConnectMidLag (count)	4	16101	7322
SchDepLag (count)	1	3300	235
CarrierNumLag (count)	1	13	2
LegacyShareLag (ratio)	0	1	0.2327
AverageFareLag (USD)	60	13033	229
AverageMileLag (miles)	116	5219	1199
RelativeFareLag (ratio)	0	74	0.9864
RelativeMileLag (ratio)	0	2.4690	0.8180

$$y' = \log\left(\frac{y}{1-y}\right), \tag{7}$$

$$y = \frac{e^{y'}}{1+e^{y'}}. \tag{8}$$

4.1. Parametric Models. For parametric models, there is a predetermined model formulation and the number of parameters is fixed. The linear regression and the beta regression are investigated carefully in this research.

4.1.1. Linear Regression. Linear regressions model the relation between the response and the features in a linear manner. Multiple linear regression (MLR) is the linear regression model with multiple features. Denote y_i as the ith response and x_{ij} as the jth feature for y_i, and the MLR model can be formulated as equation (9), in which residual sum square (RSS) is the measurement of fitting accuracy and β_j are coefficients. MLR can be easily fitted and interpreted, and one typical approach to estimate the coefficients is the least squares algorithm. The root mean square error (RMSE) is used to measure the modeling performance for different data sets, which is formulated as equation (10):

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j^2 x_{ij}\right)^2, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}. \tag{10}$$

There are 19 features in the developed MLR model. The training and testing RMSE are 0.1294 and 0.1292, respectively. Figure 3 shows the relative importance of the features based on the standardized coefficient magnitudes, based on which the *directShareQLag* is the most important features for O&D direct share modeling based on multiple linear regression. It shows that the O&D direct share is highly related to its quarterly lag.

MLR models generally suffer from variable redundancy. The redundant variables can introduce unnecessary model complexity and poorer forecasting performance in the MLR model. Feature selection methods, such as step-wise selection, are widely used to eliminate redundant features [34]. Forward and backward feature selection methods are the most widely used step-wise feature selection methods. For forward feature selection, starting from an empty feature set, the features are added into the feature set sequentially based on a certain model performance criterion. The validation RMSE is used as the feature selection criterion in this research to balance model prediction performance and to avoid over-fitting issue. For backward selection, starting with a full feature set, the features are eliminated from the feature set sequentially based on the validation RMSE. Shown in Figure 4 are the feature selection processes based on forward and backward feature selections for the MLR model.

Based on the selection curves shown in Figure 4, there are eight features selected for the MLR model by forward feature selection. By backward feature selection, there are nine features retained in the developed MLR model. The MLR models developed with the two feature selection methods provide equivalent fitting and forecasting performance, for which the training RMSE is 0.1264 and the testing RMSE is 0.1262.

MLR models with fewer features are shown providing better fitting and forecasting performance compared to the model with more features. Shown in Figure 5 are the feature importance plots of the two MLR models developed with forward and backward feature selections. *directShareQLag*, *directShareYLag*, *RelativeFareLag*, *DestPaxLag*, and *OrgPaxLag* are important features that have positive impacts on O&D direct share. Meanwhile, *LegacyShareLag* has a negative impact on the O&D direct share. The *directShareQLag* and *directShareYLag* are two of the most important features in the developed models, which show that the O&D direct share is highly related to the historical status of direct share on a certain O&D. The positive impact of *RelativeFareLag* shows that the passengers prefer direct flight services if the fare difference between direct and nondirect flight services is not significant, which is reasonable. The negative impact of *LegacyShareLag* indicates that, for O&D market dominated by legacy carriers, the direct share tends to be lower, which tallies with the operation characteristics of legacy carriers.

4.1.2. Beta Regression. A regression model for the beta-distributed random variable is defined in beta regression. Beta regression is a parametric machine learning model specifically applicable for modeling and forecasting ratio and proportion [24, 35]. Shown as equation (11) is the model of beta regression, in which x_{ti} is the ith feature of the tth observation and β_i is the correlated coefficient of the ith feature. μ_t is a function of y_t and $g(\cdot)$, which is known as the

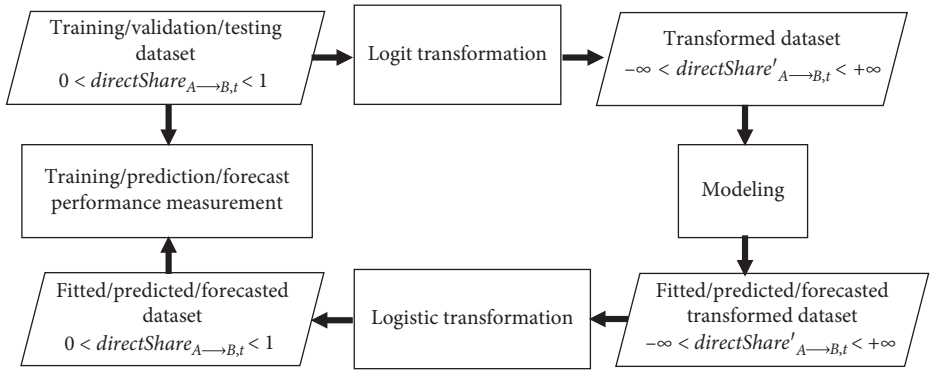

Figure 2: Direct share modeling process with logit and logistic transformations.

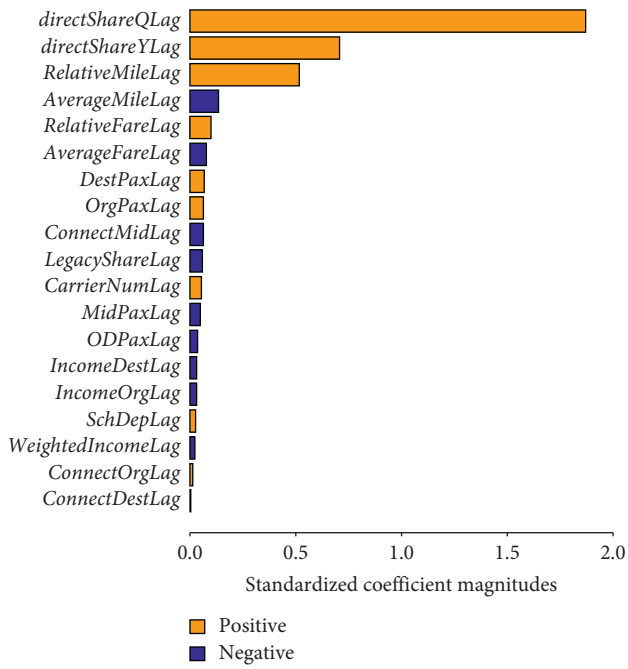

Figure 3: Variable importance of the MLR model.

link function. Shown as equation (12) is the logit link function employed in this research:

$$g(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i, \tag{11}$$

$$g(\mu) = \log\left[\frac{\mu}{1-\mu}\right]. \tag{12}$$

There is a feature selection problem for beta regression as well. The forward feature selection method is employed to search for the best feature set for the beta regression model. Shown in Figure 6(a) is the forward feature selection process for the beta regression. Seven features are selected for the developed beta regression model. The training and testing RMSE are 0.0963 and 0.0968, respectively. Compared to the MLR models, the beta regression model can provide better fitting and forecasting performance. Shown in Figure 6(b) is the variable importance of the developed beta regression model. Because the relation between the response and the features is modeled in a nonlinear manner in the beta regression, in the variable importance plot, only the level of relative variable importance can be shown. The important features in the beta regression model are very similar to the important features in the MRL models, which shows that the influential factors to O&D direct share are constant even for different modeling methods.

Based on the parametric modeling results, features have important impacts on O&D direct share including *directShareQLag*, *directShareYLag*, *RelativeFareLag*, *LegacyShareLag*, *OrgPaxLag*, and *DestPaxLag*.

(i) *directShareQLag* and *directShareYLag* are unique attributes of a certain O&D market. They indicate the air travelers' general preference for the direct flight service under a certain market scenario. The significant importance of the two features, especially the *directShareQLag*, reveals that the O&D direct share is not a factor that changes randomly. Passengers' demand for direct flight services and the supply by the air carriers determine the state of direct flight market together. Direct share is highly related to the direct flight market status, especially the recent status.

(ii) The positive impact of *RelativeFareLag* on O&D direct share is consistent with intuition. If there is no competitive pricing advantage of the nondirect flight services, direct flight services are more preferred by the air travelers.

(iii) *LegacyShareLag* shows the market share of the legacy carriers, which has a negative impact on the O&D direct share. Since the legacy carriers prefer the hub-to-spoke operation network, there is a higher possibility of flight services with connects, which can bring the O&D direct share to a lower level.

(iv) *OrgPaxLag* and *DestPaxLag* are a pair of features that show the passenger demand at the origin and destination airport. In general, airlines prefer to offer direct flight services on large hubs, which can make the carriers competitive on those O&D markets with large passenger demand.

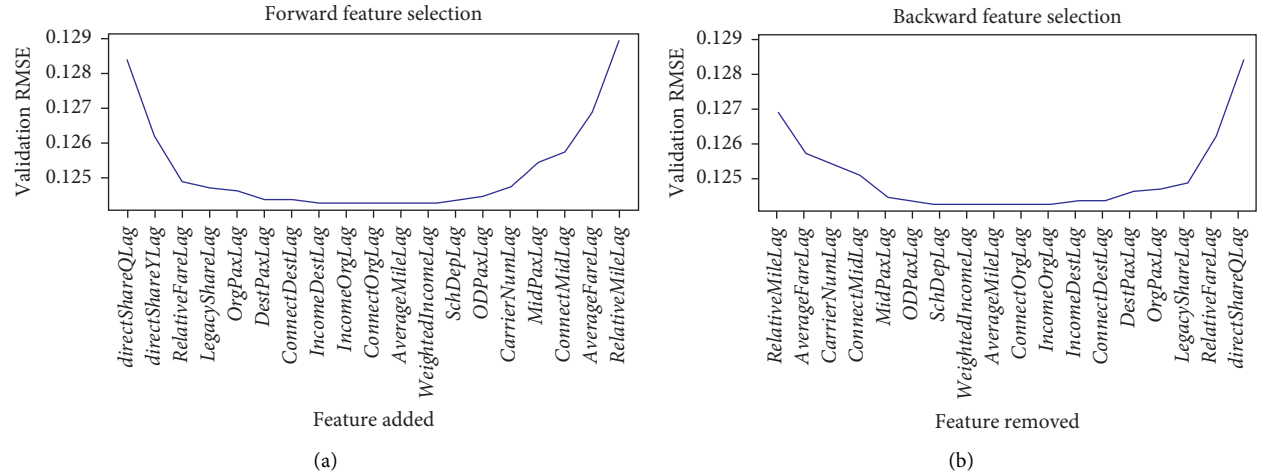

Figure 4: Forward and backward feature selection process.

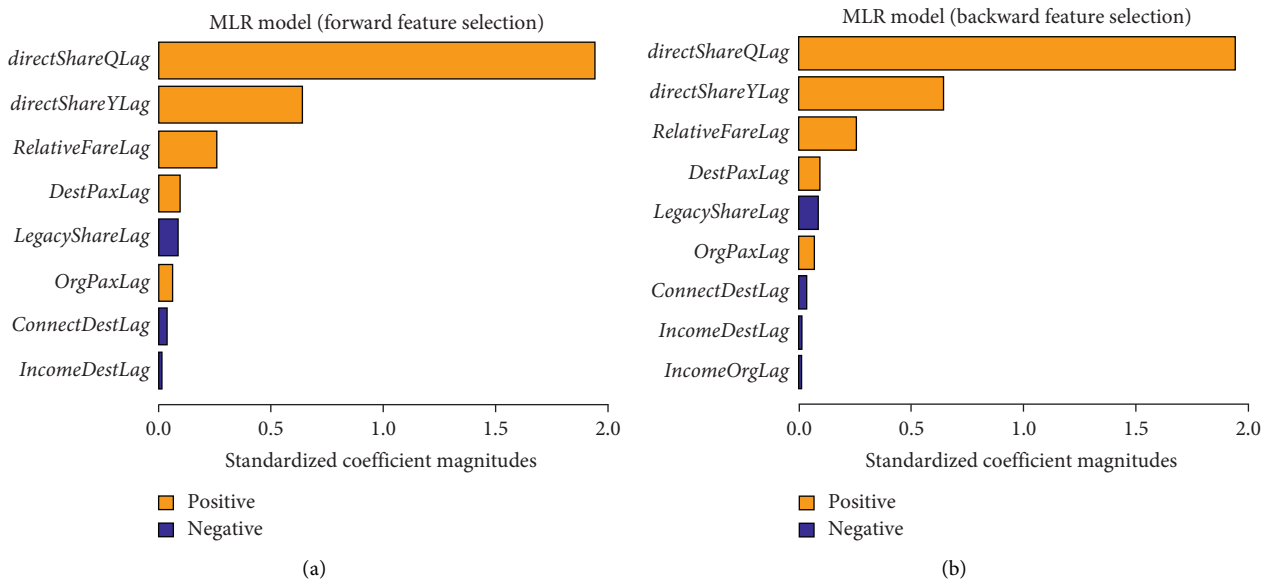

Figure 5: Variable importance of MLR models developed with forward and backward feature selections.

4.2. Nonparametric Models. There is no predetermined form of the underlying model for the nonparametric models. Comparing to the parametric models, the nonparametric models are more difficult to be interpreted, which is not appliable for feature analysis. Nonparametric models can provide better fitting and forecasting performance compared to the parametric models, which is shown in various applications. To develop an accurate model for direct share forecasting, tree-based models are investigated in this research.

Tree-based models are a category of nonparametric models based on decision trees. Decision trees involve stratifying or segmenting the feature space into a number of simple regions [36]. Tree-based models combine multiple trees to yield a single consensus prediction, which can result in significant improvement in prediction accuracy [18]. The tree-based models explored in this research are Random Forest and Gradient Boosting Machine. For the nonparametric models, the architecture of the model is determined by the hyperparameters. For example, *Ntrees* is the hyperparameter that decides how many decision trees should be grown in a tree-based model. For different nonparametric models, the hyperparameters are different. Even for the same nonparametric model, the optimal combination of hyperparameters may vary significantly for different data. The approach searching for the best combination of the hyperparameters is hyperparameter tuning. In this research, the Bayesian optimization method is introduced for hyperparameter tuning.

4.2.1. Hyperparameter Tuning. The most classic and straightforward hyperparameter tuning method is the grid searching method. All the hyperparameter combinations are tried exhaustively, and the hyperparameters are selected based on modeling performance metrics. With the increase in the number of hyperparameters and the size of the tuning

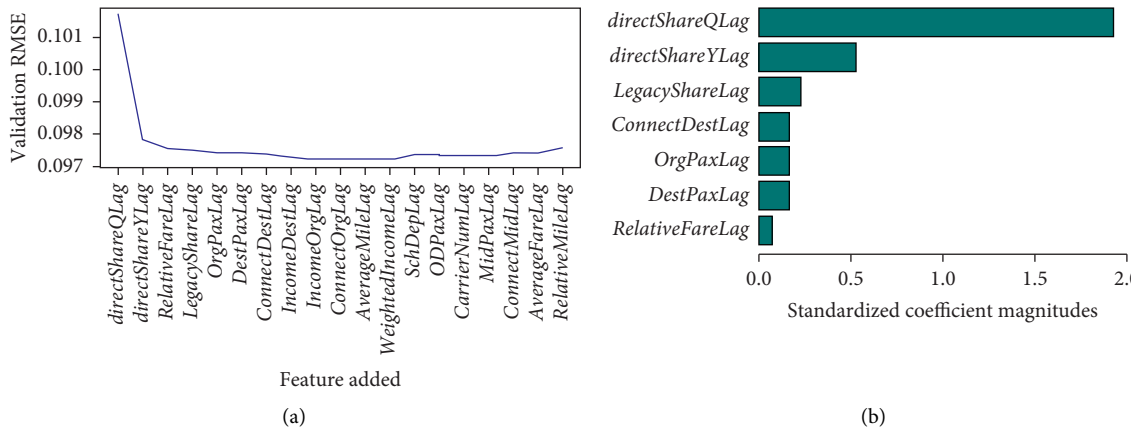

Figure 6: (a) Feature selection and (b) variable importance.

grid, this exhaustive searching method suffers from computational efficiency. An alternative method is random searching, by which the combinations of hyperparameters are selected randomly [37]. The accuracy and efficiency of random searching highly depend on the random sampling strategy. The foreknowledge about the impacts of hyperparameters on the modeling performance is necessary to build an efficient random sampling strategy, which is difficult to obtain for most of the practical problems. The same issue exists for other optimization methods, such as gradient descent searching.

The Bayesian optimization (BO) is a more reliable and practical alternative for hyperparameter tuning compared to the other methods mentioned previously. The most notable advantage of BO lies in its capability of hyperparameter optimizing for black-box functions [38]. The model performance is modeled as samples from a Gaussian process in BO, which induces tractable posterior distribution. The information obtained at the current step enables optimal choices of hyperparameters to try for the next step [39]. BO is applied for hyperparameter tuning for the tree-based models in this research. The validation RMSE is used as the tuning criterion.

4.2.2. Random Forecast. Random Forest is a tree-based model combining multiple trees to yield a single consensus prediction. When growing a decision tree in the Random Forest model, m randomly picked features from the database will be used at each split. *Mtry* is the hyperparameter indicating the number of randomly picked features. Other two hyperparameters are *MaxDepth* and *Ntrees*. *MaxDepth* determines how deep each tree can grow. *Ntrees* is the hyperparameter which indicates how many trees are grown in a Random Forest model. If there are n trees grown in a Random Forest model, the average of the n predictions will be the prediction of the Random Forest model [40, 41]. To develop the Random Forest model which can provide the best performance, the three hyperparameters are tuned together based on BO. Shown in Table 3 are the hyperparameter tuning result and the modeling performance of the developed Random Forest model.

Table 3: Random Forest hyperparameter tuning and modeling performance.

Hyperparameter tuning	Training RMSE	Testing RMSE
$MaxDepth = 20$ $Mtry = 4$ $Ntress = 400$	0.0378	0.0783

4.2.3. Gradient Boosting Machine. Gradient Boosting Machine (GBM) is another tree-based model explored in this research. Instead of growing multiple trees and taking the average of the prediction result, the GBM model takes advantage of the boosting method. Boosting is an ensemble method, which generates the predictors sequentially instead of independent. The GBM model takes information about the previously grown tree (mistakes or errors from previous predictor) to grow a new tree [36]. There are three important hyperparameters in a GBM model, which are *LearningRate*, *MaxDepth,* and *Ntrees*. *MaxDepth* and *Ntrees* play the roles in determining the architecture of a GBM model same as in the Random Forest model. Hyperparameter *LearningRate* is a value between 0 and 1, which indicates how much information should be learned from the previous tree. *LearningRate* is usually less than or equal to 0.1 [42]. The hyperparameter tuning result by BO and the modeling performance of the developed GBM model are shown in Table 4.

The nonparametric models can provide much better fitting and forecasting performance compared to the parametric models, which makes them promising models for O&D direct share forecasting.

5. Category-Based Learning

The modeling work in Section 4 focuses on developing one single direct share forecasting model which was based on the data of all the O&D pairs, which was based on the assumption that the direct share of different O&Ds is from the same population. What if the direct share from different O&D pairs are under different distributions? Will the overall forecasting performance be improved if we employ different models for different groups of data? To answer these questions and further improve forecasting performance, a

Table 4: GBM hyperparameter tuning and modeling performance.

Hyperparameter tuning	Training RMSE	Testing RMSE
$MaxDepth = 10$ $LearningRate = 0.03$ $Ntress = 600$	0.0376	0.0780

novel category-based (*C-based*) learning method is proposed in this research.

The essential idea is to split the database into different categories and develop models for each category of data individually. How to split the data efficiently into different categories is the most important problem for category-based learning. The ideal categorization can categorize the data generated from the similar underlying processes into the same category and make the difference between categories as distinct as possible. Based on previous analysis, *directShareQLag* and *directShareYLag* are the features that have significant impacts on O&D direct share. However, for different O&Ds, the influence may be different. For seasonal and nonseasonal O&Ds, the impacts of *directShareYLag* and *directShareYLag* on direct share are different. Based on this fact, the data are categorized into two categories based on the seasonality of the O&D.

The learning process of the *C-based* model is as shown in Figure 7. Based on the seasonality of the O&D pair, the training, validation, and testing data sets are split into six subsets. The training, validation, and testing subsets for each category (seasonal and nonseasonal) are employed to develop parametric and nonparametric models individually. Based on the validation RMSE, the best model is selected for each category. Instead of a single model, the result of *C-based* learning is a model set comprising the selected models for each category. The overall prediction performance, testing RMSE, is measured on the testing subsets.

Based on the seasonality of the direct share time series on each O&D, the 3350 O&Ds are split into seasonal O&Ds (834) and nonseasonal O&Ds (2516). To make fair comparison with models developed in Section 4, the training, validation, and testing data sets are kept the same. Because the training, validation, and testing data sets were generated randomly, after the categorization, the data proportion of the three data sets is still 3 : 1 : 1. The details of the models selected in the final model set are shown in Table 5.

The two model selected for the two categories are both GBM models. For the two developed GBM models, the hyperparameters are slightly different. Figure 8 shows the importance of *directShareQLag* and *driectShareYLag* in the two selected GBM models. The scaled relative influence is computed by the improvement in squared error by the selected feature [43]. Even though the architecture and forecasting performance (testing RMSE) of the two models are similar, there are significant distinctions of the two models. For the seasonal O&D category, the *driectShareYLag* plays a much more important role in the developed GBM model compared to its counterpart in the nonseasonal O&D category. In addition to it, the overall forecasting performance is improved comparing to the developed single GBM model. The improvement is not that significant because only a small proportion (24.90%) of O&Ds are categorized apart from the original data sets.

6. Modeling Performance Comparison and Forecasting

Both parametric and nonparametric models are explored in this research for O&D direct share forecasting. To further improve the overall forecasting performance, we proposed a novel *C-based* learning model in this research. One of the major objectives of this research is to develop a more accurate and reliable direct share forecasting model which can replace the model used by FAA TAF. We denote the model employed for direct share forecasting by FAA TAF as $Model_{\mathrm{TAF}}$. The O&D direct share is assumed as a constant same as the most current observation in the $Model_{\mathrm{TAF}}$. Shown in Table 6 is the forecasting comparison of different models based on the same testing data set.

Based on comparison in Table 6, even though the parametric models can provide knowledge about the important features to O&D direct share from the historical data, it failed to provide accurate forecasting of direct share which can outperform the $Model_{\mathrm{TAF}}$. The nonparametric models can provide the forecast of O&D direct share with great accuracy improvement compared to the $Model_{\mathrm{TAF}}$. The newly proposed *C-based* learning method can further improve forecasting performance. The improvement results from categorizing the data into proper categories and developing the model independently for each category.

To validate the forecasting performance of each model for application scenarios, the direct share forecasting for 2018 Q1 is generated by different models. Shown in Table 7 is the performance comparison for forecasting 2018 Q1.

Based on the comparison in Table 7, when employing different models to forecast direct share on the O&D level, the nonparametric models and *C-based* models can outperform the $Model_{\mathrm{TAF}}$ model, which shows that the *C-based* is a more reliable replacement of $Model_{\mathrm{TAF}}$.

7. Conclusions

Air transportation direct share is the ratio of direct passengers to total passengers on a directional O&D pair. It is an significant factor of air passenger distribution and indicates the passengers' general preference for direct flight services under a certain direct flight services demand and supply market status. A better understanding and a more accurate forecast of O&D direct share can benefit air transportation planners, airlines, and airports in multiple ways.

To find the factors which have significant impacts on O&D direct share, parametric regression models are investigated with feature selection methods. Based on the modeling results, O&D direct share is a predictable factor that is highly related to direct flight services market status, especially the recent status. How competitive the pricing advantage of the nondirect flight services can have impacts on O&D direct share. The O&D markets dominated by low-cost carriers tend to have relatively higher direct share. In

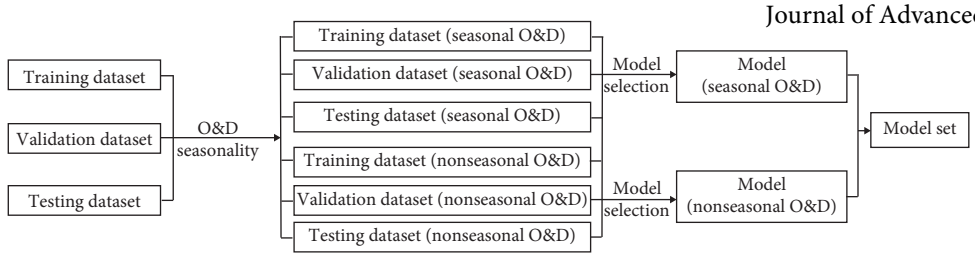

Figure 7: Category-based learning process.

Table 5: Category-based model performance.

Category	Selected model	Training RMSE	Testing RMSE
Seasonal O&D category	GBM $Maxdepth = 5$ $LearningRate = 0.07$ $Ntrees = 600$	0.0541	0.0781
Nonseasonal O&D category	GBM $Maxdepth = 10$ $Mtry = 0.06$ $Ntrees = 400$	0.0423	0.0770
		Over all	0.0772

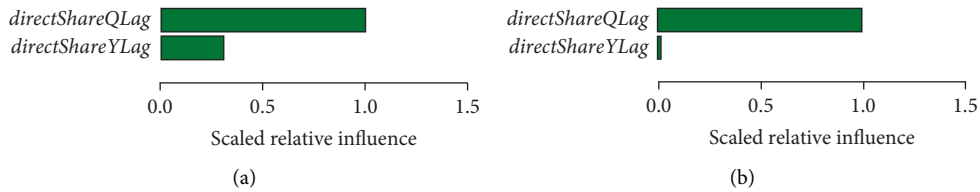

Figure 8: Importance of *directShareQLag* and *driectShareYLag* in the two selected models: (a) variable importance of *directShareQLag* and *driectShareYLag* in the GBM model for seasonal O&D category; (b) variable importance of *directShareQLag* and *driectShareYLag* in the GBM model for nonseasonal O&D category.

Table 6: Model forecasting performance comparison.

Model category	Model name	Testing RMSE
Model used by FAA TAF	$Model_{TAF}$	0.0942
Parametric models	MLR (with feature selection)	0.1262
	Beta regression	0.0968
Nonparametric models	Random Forest	0.0783
	GBM	0.0780
C-based learning	*C-based*	0.0772

Table 7: Model performance comparison for forecasting of 2018 Q1.

Model category	Model name	Testing RMSE
Model used by FAA TAF	$Model_{TAF}$	0.0988
Nonparametric models	Random forest	0.0865
	GBM	0.0865
C-based learning	*C-based*	0.0857

addition, the O&Ds connecting busy hubs tend to have a higher share of direct flight services.

To develop an accurate model for direct share forecasting, nonparametric machine learning models are explored. The Bayesian optimization method is employed for hyperparameter tuning. Both the Random Forest model and the GBM model can provide better forecasting performance compared to the model used for direct share forecasting by FAA TAF. When a single model developed, the GBM model can outperform the Random Forecast model in this research. To further improve forecasting performance, a novel category-based learning method is proposed in this research. Category-based learning method can provide better forecasting performance because of the efficient categorization and the variety in the result model set. The category-based learning method is shown as a promising replacement of the model used for O&D direct share forecasting by FAA TAF.

Data Availability

The DB1B and T100 databases used to support the findings of this study are available at Bureau of Transportation Statistics. The relevant database websites are cited in this paper as [30, 31]. The Global Insight Economic database used in this research was supplied by the Office of Aviation Policy and Plans, Federal Aviation Administration under license, and so cannot be made freely available.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Federal Aviation Administration Project Passenger Route Share Forecast. The authors thank all the colleagues at the FAA Office of Aviation Policy and Plans for their great support and help, especially David Chien, Roger Schaufele, and Li Ding. The authors also thank Thea Graham, Robert Nazareth, and James Bouse for their great help on the DB1B and T100 database data mining work.

References

[1] M. G. Lijesen, P. Nijkamp, and P. Rietveld, "Measuring competition in civil aviation," *Journal of Air Transport Management*, vol. 8, no. 3, pp. 189–197, 2002.

[2] OECD, *Airport Demand Forecasting for Long-Term Planning, ITF Round Tables*, OECD, Paris, France, 2016.

[3] J. Dumas and F. Soumis, "Passenger flow model for airline networks," *Transportation Science*, vol. 42, no. 2, pp. 197–207, 2008.

[4] G. M. Coldren, F. S. Koppelman, K. Kasturirangan, and A. Mukherjee, "Modeling aggregate air-travel itinerary shares: logit model development at a major us airline," *Journal of Air Transport Management*, vol. 9, no. 6, pp. 361–369, 2003.

[5] V. Warburg, C. Bhat, and T. Adler, "Modeling demographic and unobserved heterogeneity in air passengers' sensitivity to service attributes in itinerary choice," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1951, no. 1, pp. 7–16, 2006.

[6] L. A. Garrow, *Discrete Choice Modelling and Air Travel Demand: Theory and Applications*, Routledge, Abingdon, UK, 2010.

[7] Office of Aviation Policy and Plans, Terminal Area Forecast Summary, Fiscal Year 2017–2045, 2017.

[8] R. Sims, *A review and application of aviation forecasting for airport planners*, Ph.D. thesis, University of North Dakota, Grand Forks, N. Dak., USA, 2016.

[9] D. J. Monteiro, S. Prem, M. Kirby, and D. N. Mavris, "React: a rapid environmental impact on airport community tradeoff environment," in *Proceedings of the 2018 AIAA Aerospace Sciences Meeting*, p. 0263, Kissimmee, FL, USA, January 2018.

[10] M. C. Gelhausen, P. Berster, and D. Wilken, "A new direct demand model of long-term forecasting air passengers and air transport movements at German airports," *Journal of Air Transport Management*, vol. 71, pp. 140–152, 2018.

[11] D. Thompson, S. Perkins, and W. de Wit, *Summary of Discussions*, in *Airport Demand Forecasting for Long-Term Planning*, OECD Publishing, Paris, 2016.

[12] U. Freund-Feinstein and S. Bekhor, "An airline itinerary choice model that includes the option to delay the decision," *Transportation Research Part A: Policy and Practice*, vol. 96, pp. 64–78, 2017.

[13] J. Viken, S. Dollyhigh, J. Smith et al., "NAS demand predictions, transportation systems analysis model (tsam) compared with other forecasts," in *Proceedings of the 6th AIAA Aviation Technology, Integration and Operations Conference (ATIO)*, p. 7761, Wichita, KS, USA, September 2006.

[14] M. C. Gelhausen, "Modelling the effects of capacity constraints on air travellers' airport choice," *Journal of Air Transport Management*, vol. 17, no. 2, pp. 116–119, 2011.

[15] C.-W. Yang, J.-L. Lu, and C.-Y. Hsu, "Modeling joint airport and route choice behavior for international and metropolitan airports," *Journal of Air Transport Management*, vol. 39, pp. 89–95, 2014.

[16] S. de Luca, "Modelling airport choice behaviour for direct flights, connecting flights and different travel plans," *Journal of Transport Geography*, vol. 22, pp. 148–163, 2012.

[17] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, Cambridge, MA, USA, 2009.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, Berlin, Germany, 2013.

[19] W. Hardle and E. Mammen, "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, vol. 21, no. 4, pp. 1926–1947, 1993.

[20] B. McCune, "Non-parametric habitat models with automatic interactions," *Journal of Vegetation Science*, vol. 17, no. 6, pp. 819–830, 2006.

[21] G. A. Seber and A. J. Lee, *Linear Regression Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 2012.

[22] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 59, no. 1, pp. 3–54, 1997.

[23] C. F. Baum, "Stata tip 63: modeling proportions," *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 8, no. 2, pp. 299–303, 2008.

[24] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[25] J. J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman and Hall, London, UK, 2016.

[26] V. Svetnik, A. Liaw, C. Tong, and T. Wang, "Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules," in *Proceedings of the International Workshop on Multiple Classifier Systems*, pp. 334–343, Springer, Cagliari, Italy, June 2004.

[27] L.-Y. Chang and W.-C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," *Journal of Safety Research*, vol. 36, no. 4, pp. 365–375, 2005.

[28] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, 2011.

[29] Z. Quan, G. Gan, and E. A. Valdez, *Tree-based models for the efficient valuation of large variable annuity portfolios*, 2018.

[30] Bureau of Transportation Statistics, "Data profile: airline origin and destination survey (db1b)," 2018, https://www.transtats.bts.gov/DatabaseInfo.aspDB_ID=125&DB_Name=Airline%20Origin%20and%20Destination%20Survey%20%28DB1B%29.

[31] Bureau of Transportation Statistics, "Database name: air carrier statistics (form 41 traffic)- u.S. carriers," 2018, https://www.transtats.bts.gov/Tables.aspDB_ID=110&DB_Name=Air%20Carrier%20Statistics%20%28Form%2041%20Traffic%29-%20%20U.S.%20Carriers&DB_Short_Name=Air%20Carriers.

[32] IHS, "Ihs global insight," 2018, https://globalsso.ihs.com/KeystoneSTS/SSOLogin/Login.aspx?theme=IGI&ReturnUrl=https%3a%_2f%2fglobalsso.ihs.com%2fKeystoneSTS%2fKSFed%2fDefault.aspx%3ftheme%3dIGI.

[33] Federal Aviation Administration, "Airport categories airports," Federal Aviation Administration, Washington, DC,

USA, 2018, https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/categories/.

[34] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[35] A. K. Gupta and S. Nadarajah, *Handbook of Beta Distribution and its Applications*, CRC Press, Boca Raton, FL, USA, 2004.

[36] J. Friedman, T. Hastie, and R. Tibshirani, "The elements of statistical learning," in *Springer Series In Statistics*, Vol. 1, Springer, New York, NY, USA, 2001.

[37] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[38] D. J. Lizotte, *Practical Bayesian Optimization*, University of Alberta, Edmonton, Canada, 2008.

[39] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, pp. 2951–2959, Princeton University, Princeton, NJ, USA, 2012.

[40] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[41] J. Albert, E. Aliu, H. Anderhub et al., "Implementation of the random forest method for the imaging atmospheric cherenkov telescope magic," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 588, no. 3, pp. 424–432, 2008.

[42] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[43] H2O.ai, Docs.algorithms.distributed random forest (drf) (2019), http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html.