

10-9-2019

## A Data Mining Approach to Building a Predictive Model of Low-Cost Carriers' Presence in the U.S. Domestic Routes

Canh Nguyen

Florida Institute of Technology, CanhNg207@gmail.com


John E. Deaton

Florida Institute of Technology, jdeaton@fit.edu

Nurettin Dinler

Florida Institute of Technology, ndinler2016@my.fit.edu

Follow this and additional works at: <https://commons.erau.edu/ijaaa>

 Part of the [Business Administration, Management, and Operations Commons](#), [Business Analytics Commons](#), [Management Sciences and Quantitative Methods Commons](#), and the [Operations and Supply Chain Management Commons](#)

### Scholarly Commons Citation

Nguyen, C., Deaton, J. E., & Dinler, N. (2019). A Data Mining Approach to Building a Predictive Model of Low-Cost Carriers' Presence in the U.S. Domestic Routes. *International Journal of Aviation, Aeronautics, and Aerospace*, 6(5). <https://doi.org/10.15394/ijaaa.2019.1354>

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in International Journal of Aviation, Aeronautics, and Aerospace by an authorized administrator of Scholarly Commons. For more information, please contact [commons@erau.edu](mailto:commons@erau.edu).

The evolution of the low-cost carriers (LCCs) model was marked by a transformation from a regional carrier (formerly known as Pacific Southwest Airlines) to a national U.S. LCC carrier, Southwest Airlines in the 1970s. From 1978 to 2013, the air transportation market witnessed a plunge by 40% in airfares, which was attributed to the impact of LCCs' pricing practice in the network (Airlines for America, 2014). Under "Southwest effect," new LCC entrants entered the market with varying degrees of success but did not experience rapid growth until the late 1990s when LCCs' flights appeared on the top 5,000 domestic routes. LCC presence continued to be bold by an increase from 1,594 in 1998 to 2,304 routes in 2003 (General Accounting Office, 2004), and to 7,915 routes in 1Q2018 as reported from the data in this study.

The cost structure is claimed to be the substantial difference between full-service carriers (FSCs) and low-cost carriers (LCCs), which is a result of several strategic pursuits. Vasigh, Fleming, and Tacker (2016) stated that LCCs are gaining advantages through: (1) lower labor cost and higher labor productivity, (2) lower ticket distribution costs, (3) no-frills service, (4) common fleet type, (5) point-to-point service, (6) use of secondary airports, and (7) higher aircraft utilization. Similarly, Belobaba, Odoni, and Barnhart (2015) explained the achieved cost advantage as the result of the productivity of both employees and aircraft. Significant higher labor productivity of LCCs lies in much more flexible rules that allow cross-utilization for all employees except those who are safety-licensed and -certified. In the meantime, the point to point flights can minimize aircraft ground times, which translates to higher aircraft utilization rates (high aircraft productivity).

By maintaining low operating cost, many LCCs were able to aggressively expand their networks to capture market share, which in turn led to bankruptcies of four of six U.S. legacy carriers between 2001 and 2005 (Belobaba et al., 2015). Ben Abda, Belobaba, and Swelbar (2012) investigated LCCs entry and growth in relation to the evolution of origin-destination air traffic and fares in the U.S. domestic market at four snapshot years, 1990, 1995, 2005, and 2008. LCCs collective share on the U.S. domestic market grew over the decades, from 10.6% in 1990 to 23.6% in 2000, and to 33.6% in 2008. The study discovered that initial strategies of LCCs in planning new services were to focus on the busiest airports that serve as large pools of local traffic as opposed to those of connecting or mixed traffic. Eventually, LCCs encountered difficulties in entering first-tier airports because of gate constraints, higher congestion likelihood, and full-service carriers' (FSC) reaction on aggressive fare matching. Secondary airports were then an ideal approach to the expansion; 18% aggregate LCC market share in 2000 soared up to 30% in 2005 in routes to second-tier airports.

Ben Abda et al. (2012) continued with the impact of LCC route penetration on the average fare and passenger volume. The average airfare decreased by 16.8% at 23 airports that experienced a substantial LCC growth in 1990-1995, while passengers who traveled on routes with LCC absence witnessed increased average fares by 1.7% in the same period. The traffic rose 28.5% at 26 airports associated with high LCC presence in 2000 and 2005; at the other airports associated with LCC absence, the reported increase was only 4.4%. The growth of LCCs manifested through its density—the coverage ratio of the number of LCCs per airport. The ratio rose steadily from 0.5 to 2.8 between 1990 and 2005, followed by a diminishing ratio due to the financial crisis, economic recession, and the saturation of the air travel market in 2007-2009. Additional difficulties in the rising price of jet fuel thereafter led to a wave of consolidation among players that consequently arrived at six major LCCs in April 2018. These six major LCCs were found to have a negative relationship to route entry and exit decisions of competing airlines (Nguyen & Nguyen, 2018). Bachwich and Wittman (2017) considered factors enabling the emergence of a new variant of the LCC model, ultra-low-cost carriers (ULCCs). The 2015 dataset indicated that the market fare was 21% lower in the presence of ULCCs and 8% lower in the presence of LCCs compared to the entire market average. Examining the trend over 2010-2015, after each one-year entry of a ULCC or LCC into any flight route, there was a 14% reduction in average fare.

Airlines are now aware of the essence of restructuring their own network in attempts of maintaining the profitability under pressure of LCC presence. Understanding the past and existing patterns of the LCCs' network structure and improving the predictability of the future presence of an LCC in the network becomes imperative for all airlines to sustain a competitive edge. Although the current literature was replete with similar studies, it is still necessary to have studies that stay current and timely to examine the presence of LCCs in the industry landscape, especially after socioeconomic volatilities and consolidations. In this study, besides examining factors reviewed in the previous paragraphs, we took advantage of data mining procedures by reconstructing the raw dataset and incorporating additional variables to the model.

### **Statement of Purpose**

The purpose of the study was to predict the presence or absence of low-cost carriers (LCCs) in the U.S. domestic network structure. Only flight routes with origin and destination airports located within the United States were included in the study, and the timeframe ranged from Quarter 1, 2016 to Quarter 1, 2018. Operational definitions of the relevant variables in this study are summarized in Table 1 and fully discussed as follows:

Table 1

*Summary and Description of Independent and Dependent Variables*

<b>Variables</b>	<b>Description</b>
<b>Dependent variable</b>	
LCC presence vs. absence	Categorical (dichotomous) variable represented whether having at least one LCC operation on a route. Dummy coding scheme with 1 as LCC presence and 0 as LCC absence.
<b>Independent variables</b>	
Average market fare	Continuous variable represented the average airfare that all airlines offered in a route.
Average connection yield	Continuous variable represented the ratio of average fare of connecting flights on average miles flown of these flights.
Destination airport	Categorical variable represented the last destination airport in a route. Unweighted effects coding scheme used for five groups: large, medium, small, non-hub, and non-primary airports.
Largest share proportion	Continuous variable represented the percentage of the largest market share for which an airline accounted over the total market in a route.
Number of carriers	Continuous variable represented the number of carriers operating in a route.
Number of connecting passengers	Continuous variable represented aggregated passengers in connecting flights in a route.
Number of total passengers	Continuous variable represented aggregated passengers carried by all airlines in a route.
Origin airport	Categorical variable represented the first departing airport in a route. Unweighted effects coding scheme used for five groups: large, medium, small, non-hub, and non-primary airports.
Route length	Continuous variable represented the geographic distance in miles between origin and destination airports.
Route type	Categorical (dichotomous) variable represented whether or not having at least one nonstop flight in a route. Dummy coding scheme with 1 as nonstop market and 0 as connection market (the reference group).

*The average market fare* was defined as the averaged commercial airfare of passenger transportation service that all airlines offered on a given route. *Average connection yield* was defined as the ratio of the average airfare of connecting flights over the average miles flown of these connecting flights.

*Largest share proportion* was defined as the percentage of the largest market share for which an airline accounted in a given route based on the number

of transported passengers. For example, in the route MCO-DFW of 1Q2018 dataset, the largest share airline transported 6,480 passengers on both nonstop and connecting flights over the total of 9,846 passengers, and thus the largest share proportion was 65.81% (6,480 / 9,846).

*The number of carriers* was defined as the number of all incumbent carriers operating on a given route.

*The number of total passengers* was defined as an aggregated number of passengers transported by all airlines in a given route regardless of nonstop or connecting flights. In the meantime, *the number of connecting passengers* was defined as an aggregated number of passengers transported by all airlines only on connecting flights.

*Route length* was defined as the geographic distance in miles between origin and destination airports regardless of nonstop or connecting flights.

*Route type* was defined as the characteristic of the route market, nonstop market and connection market. It is commonly accepted in the literature that in a specific route, there is at least one nonstop flight on operations, the route is considered a nonstop market; it is considered a connecting market otherwise (Coldren, 2005; Coldren, Koppelman, Kasturirangan, & Mukherjee, 2003; Garrow, 2010). ABE-ATW in the 1Q2018 dataset was a connecting market because of no nonstop flight being operated across airlines.

*Origin and destination airports* were defined as the first and the last airports in a given itinerary. For example, in the itinerary of MCO-ATL-SEA, MCO is the origin airport while SEA is the destination airport. Federal Aviation Administration (2016a) categorized commercial service airports into primary and non-primary commercial service airports. Non-primary commercial service airports accommodate at least 2,500 and no more than 10,000 passengers. Primary commercial service airports are partitioned into subcategories based on percentage of annual enplanement, including large hub with 1% or more, medium hub with at least 0.25% but less than 1%, small hub with at least 0.05% but less than 0.25%, and non-hub with more than 10,000 but less than 0.05%.

*The presence of low-cost carriers* was defined as having at least one operation of a low-cost carrier on a given route. In the study period, a total of 36 commercial airlines reported under the name of ticketing carriers in datasets, and 7 of them corresponded to the business model of a low-cost carrier. Included in the study were Allegiant Air (G4), Frontier Airlines (F9), JetBlue (B6), Spirit Airlines (NK), Southwest Airlines (WN), Sun Country Airlines (SY), and Virgin America (VX). Virgin America was jointly reported under the name of Alaska Airlines as of 2Q2018 due to the consolidation.

## Research Questions

Research question 1: When examined from a stepwise model for logistic regression, what is the relationship between the targeted variables and the dichotomous response variable that distinguishes between the presence and absence of the U.S. LCCs in the domestic routes?

Research question 2: When examined the variable importance of the decision tree model, what are the most important factors that predict the presence or absence of the U.S. LCCs in the domestic routes?

Research question 3: In the model comparison platform, between logistic regression and decision tree, which model performs more accurately in predicting the presence or absence of the U.S. LCCs in the domestic routes?

## Methodology

### Research Design

The research methodology was ex post facto or causal-comparative, and its corresponding design was cause-type. This methodology was appropriate because the research question involved modeling the relationship of the group memberships of the U.S. LCC presence versus absence with multiple factors. The study was data-driven in nature by using a data mining approach as opposed to a theory-driven study in which theories were grounded to guide and partially answer research questions along with the support of traditional statistical analysis.

Linoff and Berry (2011) defined data mining as a business process for exploring a large amount of data to discover meaningful patterns and rules. Although statistics and data mining share numerous similar tools, they are distinguished based on the objectives and process of each discipline. In statistics-oriented studies, the objectives are well defined and driven by theories and theoretical models. The process is to make inferences to the population based on the selected sample, which is also known as inferential statistics. By contrast, in data mining-oriented studies, in many cases, the data are the entire population or a significantly large data set, and thus the inferential process is not a concern. However, the objectives of data mining studies are ill-defined and ill-directed, instead the data usually are integrated and aggregated from different sources and must be cleaned and useful variables extracted.

Two common and well-documented processes in data mining studies are SEMMA and CRISP-DM (Grayson et al., 2015; Sarma, 2013; Tufféry, 2011); the former stands for Sample, Explore, Modify, Model, and Assess, and the latter stands for Cross Industry Standard Process for Data Mining. The shared point of view was that both approaches “define a set of sequential steps that pretends to

guide the implementation of data mining applications” (Azevedo & Santos, 2008, p. 1). SEMMA schematic can be considered a general process for developing a statistical model, while CRISP-DM phases, which enumerate as business understanding, data understanding, data preparation, modeling, evaluation, and deployment, are designed to not tie to any specific tool or application and to be able to use in any industry (Chapman et al., 2000). In the current study, SEMMA was chosen as a primary and systematic approach to build the predictive model of LCC presence in the U.S. domestic network.

## **Data Preparation**

### **Target and Accessible Population**

The target population of the study was all domestic passenger flight routes that have origin and destination airports located within the United States. The accessible population of the study was 10% random census of the target population. In effect, the U.S. Department of Transportation (DOT) randomly selects 10% of all domestic recorded flights for free public access at the Bureau of Transportation Statistics (BTS) website (bts.gov). The primary database used in the study was Origin and Destination Data Bank 1B (DB1B). Quarterly datasets of 2016 and 2017 were used for developing and validating the predictive models, while the dataset of 1Q2018 was used for testing the models. All datasets were directly imported into *JMP Pro 13* software to screen and reconstruct the data before sampling and building models.

### **Data Reconstruction**

Before reconstructing, the dataset was screened for the issues of missing data. In the dataset, flights recorded under the ticketing carrier as “--” or 99 were considered missing data (i.e., no airline designator as 99 for U.S. airlines). The missing proportion was as much as 3% of all quarterly recorded flights, and thus we decided to use a list-wise deletion method for handling random missing data. Additionally, flights with bulk fares also were removed out of the datasets because bulk fares reflect airlines’ promotion programs such as flyer frequent programs (Abdelghany & Abdelghany, 2009).

Following Nguyen's and Nguyen's (2018) guideline, we reconstructed the raw datasets by sorting all information based on pairs of origin and destination airports. The purpose of reconstructing the datasets was to aggregate both nonstop and connecting flights in a specific route instead of displaying hundreds of flights in the same origin and destination airports in the raw datasets. For instance, the original 1Q2018 dataset recorded repeatedly 27 different flights (i.e., all were connecting flights) with the same ABE as origin and ATW as a destination; the reconstructed dataset now exhibited uniquely the route ABE-ATW with 27

connecting flights served by two airlines. It is noted that the route ABE-ATW characterized as a connecting market because all 27 different flights were connecting. After reconstruction, eight quarters of 2016 and 2017 were aggregated into a dataset with 498,263 routes, whereas the Quarter 1 of 2018 generated 61,024 routes.

### Sampling

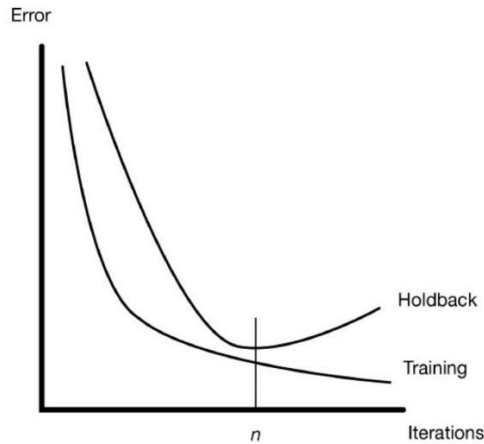
One of the drawbacks of using “big data” or analyzing data collected from the data warehouse is that too large sample sizes might lead to incorrect conclusions of significance (Paczkowski, 2018). To remedy the problem, sampling in SEMMA paradigm is necessary to this study. Tufféry (2011, p. 306) advised a critical minimum size of the training set (a) at greater than 1,000 observations, and (b) having at least 300-500 observations in each group (level) of the categorical response. To satisfy the two conditions, we randomly selected 5,000 routes from the 2016-2017 dataset for modeling and 1,250 routes from the 1Q2018 dataset for testing purposes.

The goal of building the model is to be able to make accurate predictions in the future for any value of predictors. It is necessary to check the models if they are overfitting the data. Overfitting occurs when the model becomes too complex such that not only the underlying model but also the random errors are explained and thus become fit with the dataset. The former might persist into future predictions, but the latter will differently deviate in the future. To detect overfitting issue in models, the cross-validation process is commonly used by data miners to determine the necessary model complexity. Figure 1 illustrated the model errors in both training and validation subsets drop down until a certain iteration  $n$ . Training model error continued to be minimized to fit the data points while the validation model error started raising, which indicates overfitting.

Holdout cross-validation method was employed in this study because of its advantage of simplification over  $k$ -fold cross-validation. In holdout cross-validation, training, validating, and testing subsets are usually generated. The function of the training subset is to fit the statistical underlying models and to estimate the models' coefficients. The function of the validation subset is to determine how much complexity is needed in the established models. More importantly, the predictive performance and model fit measures (e.g.,  $R^2$ , ROC, Lift Chart) of competing models are assessed on the validation subset to choose the best one (Klimberg & McCullough, 2016; Linoff & Berry, 2011). Sarma (2013) recommended the testing subset is the external and independent one that has no influence on the estimations and model complexity. In the current study, the 2016-2017 dataset, after a random sampling, was further partitioned into the training and validation subsets at the ratio 75:25 equivalent to 3,750:1,250 routes



in a total of 5,000 routes, while 1,250 routes in the dataset of 1Q2018 were set aside as the testing subset.



*Figure 1.* Model error curves of training and validation sets by the number of iterations (Klimberg & McCullough, 2016).

## **Descriptive Statistics**

Table 2 summarizes statistics of continuous and nominal-coded predictors relative to the dichotomous response, LCCs presence and absence, in the 2016-2017 dataset. Considering airfare-based factors, LCCs lowered the average market airfare in the domestic network by roughly \$85 from \$330 at routes having no LCCs operations to \$245 at those having more than one LCC flight. The same pattern was found in connecting routes such that the average connection yield was 0.19 dollar per mile if LCCs exist compared to 0.28 dollar per mile if any LCC was not seen.

With respect to market concentration factors, LCCs were found on routes more competitive with three or four players on average, and the largest share occupied by a carrier on LCC-present routes was 62% compared to LCC-absent routes with 85%. Demand factors showed that routes having LCC flights were markedly higher traffic with 1,187 passengers in total and 245 connecting passengers. However, the standard deviations of total traffic and connection for both LCC presence and absence were all scattered, and their ranges were large, which was an indicator for outliers that were checked in the next section. The reflection of LCC operations through route characteristics was not much different in both route length and route type.

Table 2

*Descriptive Statistics of Continuous Predictors in 2016-2017 Dataset*

Variables	LCC Presence			LCC Absence		
	Mean	SD	Range	Mean	SD	Range
Average market fare	244.92	71.02	5 – 647.07	329.94	165.67	0 – 3548.67
Average connection yield	0.19	0.10	0 – 0.84	0.28	0.20	0 – 3.85
Largest share proportion	0.62	0.21	0.27 – 1	0.85	0.19	0.30 – 1
Number of carriers	3.81	1.46	1 – 9	1.63	0.82	1 – 5
Number of connecting passengers	245.07	303.05	0 – 2,137	21.18	47.92	0 – 883
Number of total passengers	1,187	2,422.46	1 – 23,756	38.71	159.74	1 – 3,941
Route length	1,301.41	677.21	177 – 5,095	1,484.52	1,067.68	59 – 9,431
Route type <sup>a</sup>	0.60	0.49	0 – 1	0.04	0.20	0 – 1

Note. N = 5,000.

<sup>a</sup>Route type was a nominal variable coded by dummy coding scheme with the nonstop market as 1 and the connecting market as 0.

Table 3 showed statistics of five subgroups of origin and destination airports relative to LCC presence and absence. LCC flights covered the domestic network with nearly 14% (684 over 5,000 routes). Given origin airports, routes having LCC flights with departures from then large, medium, and small hub was equally prevalent and approximated at 30% each in comparison with nearly 8% of the combined group of non-hub and non-primary airports. The same pattern was observed in destination airports for both LCC presence and absence. It is noted that origin and destination airports were coded by unweighted effects coding strategy for the stepwise logistic regression model.

Table 3

*Descriptive Statistics of Airport Subgroups in 2016-2017 Dataset*

<b>Subgroups</b>	<b>LCC Presence</b>		<b>LCC Absence</b>	
	<i>N</i>	%	<i>N</i>	%
<b>Origin airport</b>	684	13.68	4,316	86.32
Large hub	209	30.56	544	12.60
Medium hub	205	29.97	492	11.40
Small hub	215	31.43	1,045	24.21
Non-hub	52	7.60	2,106	48.80
Non-primary	3	0.44	129	2.99
<b>Destination airport</b>	684	13.68	4,316	86.32
Large hub	193	28.22	541	12.54
Medium hub	197	28.80	476	11.03
Small hub	241	35.23	1,100	25.49
Non-hub	50	7.31	2,101	48.68
Non-primary	3	0.44	98	2.26

*Note.* *N* = 5,000.

### **Outliers and Multicollinearity**

Regarding the outlier issue mentioned earlier, the number of flagged cases were 559 out of 5,000 (11.18%) in the 2016-2017 dataset and 156 out of 1,250 (12.48%) in the 1Q2018 testing dataset. Random examination of these flagged cases unveiled that several flights on such routes were most likely a charter rather than commercially scheduled flights, therefore we decided to remove these flagged cases. The sample size of the training set,  $N_{\text{Training}} = 3,330$  routes,  $N_{\text{Validation}} = 1,111$  routes, and  $N_{\text{Testing}} = 1,094$  routes as shown in Table 4. Multicollinearity is an issue if two or more predictors in a model are highly correlated with one another. When severe multicollinearity issue occurs, it is difficult to determine which of the correlated predictors are most important, and it could lead to inflation in coefficients and standard errors, or even make the signs of the coefficients meaningless (Cohen, Cohen, West, & Aiken, 2003; Grayson et

al., 2015). No evidence of serious multicollinearity was found through the correlation matrix in the datasets

Table 4

*Statistics Summary of Datasets after the Preliminary Analyses*

	Removed routes	Overall	LCC presence		LCC absence	
			N	%	N	%
Training Set	420	3,330	303	9.10	3,027	90.90
Validation Set	139	1,111	89	8.01	1,022	91.99
Testing Set	156	1,094	83	7.59	1,011	92.41

## Data Analysis

### Stepwise Logistic Regression

Logistic regression is also commonly known as the linear probability model (LPM) because it is a specialized form of linear regression using to handle discrete or categorical dependent variables (Klimberg & McCullough, 2016). This was the case for the current study as the dependent variable was binary responses—the U.S. LCCs presence versus absence. Stepwise estimation was used as the primary method of selecting variables for inclusion in the logistic regression model. In the stepwise model, the variable entry order is determined based on the objective of maximizing  $R^2$  with the fewest predictors (Hair, Black, Babin, & Rolph, 2010). The model starts with selecting the best predictor that has the largest explanatory power (semi-partial correlation squared  $sr^2$ ). One at a time, an additional predictor is selected given the incremental explanatory power it can contribute to the regression model. This procedure is continued as long as their increments are statistically significant, and thus formally known as forwarding addition approach (Cohen et al., 2003). Table 5 specified the entry order of predictors for the study's forward addition stepwise model with the stopping rule of the maximum validation  $R^2$ .

As reported in Table 5, the stepwise logistic regression model was statistically significant,  $\chi^2(12) = 1,144.82, p < .0001$ . The full model provided a predictive gain of 56.39% over the null model,  $R^2 = .5639, df = 12$ . The logistic constant in the null model that assumes the absence of information provided by the predictors was  $B_{\text{Constant}} = -2.335$ , and the corresponding odd of LCC presence

in the network was  $e^{-2.335} = 0.097$ . When applied the mathematical expression,  $e^{-2.335} / (1 + e^{-2.335}) = 0.088$ , it indicated that 8.8% of the routes had the presence of LCCs in the calendar year of 2016-2017.

The positive sign of the logit coefficient for the number of carriers,  $B_{N_{\text{Carriers}}} = 1.708$ ,  $p < .0001$ , indicated a positive relationship between the LCC presence and average market fare. The average marginal effect,  $ME = 0.137$ , revealed that for one additional competitor commencing flights in a route, there was nearly 14% more likely to have at least one LCC exited in the route. Route type had indeed a positive relationship with LCC presence,  $B_{R_{\text{Type}}} = 1.311$ ,  $p < .0001$ , and  $ME = 0.105$ . If a nonstop market, the route was 10.5% more likely to have at least one LCC operation than the one under the condition of a connecting market. Regarding airfare-related predictors, as average market fare declined by \$100 in a flight route, there was 8% more likely to have at least one LCC operation in that route,  $B_{M_{\text{Fare}}} = -0.010$ ,  $p < .0001$ , and  $ME = -8e-4$ . Meanwhile, average connection yield also had a negative relationship,  $B_{A_{\text{Yield}}} = -2.361$ ,  $p = .0065$ , and  $ME = -0.189$ , such that every decrease of 1 dollar per miles flown on connection routes, there was closely 19% more likely to have flights operated by LCCs. With respect to airport hubs, regardless of origin or destination, on routes with either departure from or arrivals to large, medium, and small hub, there was 6.5% and 5.7% more likely to have at least one LCC operation, respectively. These positive relationships were statistically significant,  $B_{\text{Origin (L\&M\&S - Nh\&Np)}} = 0.812$ ,  $p < .0001$ , and  $ME = 0.065$ ;  $B_{\text{Dest (L\&M\&S - Nh\&Np)}} = 0.716$ ,  $p < .0001$ , and  $ME = 0.057$ . Taking three types of hub (large, medium, and small) into consideration, there was 2.2% more likely to have LCC presence on routes with origin as medium hubs,  $B_{\text{Origin (M - L\&S)}} = 0.268$ ,  $p = .0145$ , and  $ME = 0.022$ . The same case happened for destination as medium hub at a slightly higher preset  $\alpha = .06$ ,  $B_{\text{Dest (M - L\&S)}} = 0.232$ ,  $p = .0547$ , and  $ME = 0.019$ . Noticing that *JMP* by default utilizes unweighted effect coding for categorical variables such that the group mean of interest was interpreted by comparing to the unweighted average mean across all groups—the grand mean (Cohen et al., 2003).

Table 5

Summary of Stepwise Logistic Regression for the Model of LCC Presence vs. Absence

	$B_i$	$\chi^2$	$p$	Average Marginal Effects <sup>b</sup>
<b>Null Model</b>				
Constant	-2.335	1,948.6	< .0001***	
<b>Stepwise Model<sup>a</sup></b>				
Constant	-5.113	34.81	<.0001***	
Number of connecting passengers	0.002	1.56	.2121	1.6e-4
Number of carriers	1.709	104.04	<.0001***	0.137
Route type	1.311	21.96	<.0001***	0.105
Average market fare	-0.010	52.20	<.0001***	-8e-4
Origin airport (L&M&S – Nh&Np)	0.812	28.44	<.0001***	0.065
Destination airport (L&M&S – Nh&Np)	0.716	24.48	<.0001***	0.057
Average connection yield	-2.361	7.41	.0065**	-0.189
Origin airport (M – L&S)	0.268	5.98	.0145*	0.022
Destination airport (M – L&S)	0.232	3.69	.0547	0.019
Largest share proportion	1.155	2.78	.0954	0.093
Origin airport (L – S)	-0.226	2.31	.1287	-0.018
Destination airport (L – S)	0.180	1.35	.2455	0.014
-Log Likelihood in Null Model			1,015.07	
-Log Likelihood in Full Model			442.66	
Difference			572.41	
$\chi^2(12)$			1,144.82***	

Note.  $N_{\text{Training}} = 3,330$ .  $N_{\text{Validation}} = 1,111$ .  $R^2_{\text{Training}} = .5639$ .  $R^2_{\text{Validation}} = .4337$ .

<sup>a</sup>The predictors of stepwise model are listed in the entry order. L = Large hub, M = Medium hub, S = Small hub, Nh = Non-hub, and Np = Non-primary airport. <sup>b</sup>JMP provided the average predicted probability of LCC

presence  $\Pr(Y = 1 | X) = 0.088$  and LCC absence  $\Pr(Y = 0 | X) = 0.912$ . Average marginal effects =  $\Pr(Y = 1 | X) \times \Pr(Y = 0 | X) \times \text{Logistic Coefficients}$ .

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$

## Decision Tree

A decision tree is a hierarchical structure of variables in which the dataset is broken up into smaller groups (child nodes) from the initial root node (parent node) based on the criterion variable (dependent variable) in logical-based rules. As illustrated in Figure 2, the percentage of groups (levels) of the categorical response in each node is represented by the gray and white shades. For example, the gray represented the percentage of LCC-present routes and the white represented the percentage of LCC-absent routes. Each node is split into either two or more than two branches, which Neville and Ville (2013) referred to binary partitions and multiple-way partitions. Common splitting criteria for each node include Chi-square, Gini, and Entropy. By default, *JMP Pro 13* use binary partition and Chi-square to build the decision tree. Chi-square statistic and its associated  $p$ -value were used to measure the dissimilarity in the proportions between the two split groups, LCC presence and absence. The lower the  $p$ -value, the bigger the difference between the groups. *JMP* adjusts the  $p$ -value to account for the number of splits by transforming to a log scale using the formula -  $\log_{10}(\text{adjusted } p\text{-value})$ , which is called the *Log Worth*; the bigger the *Log Worth* value, the better the split is (Grayson et al., 2015). Chi-square and *Log Worth* are used to rank the predictors based on their importance in explaining the categorical response.

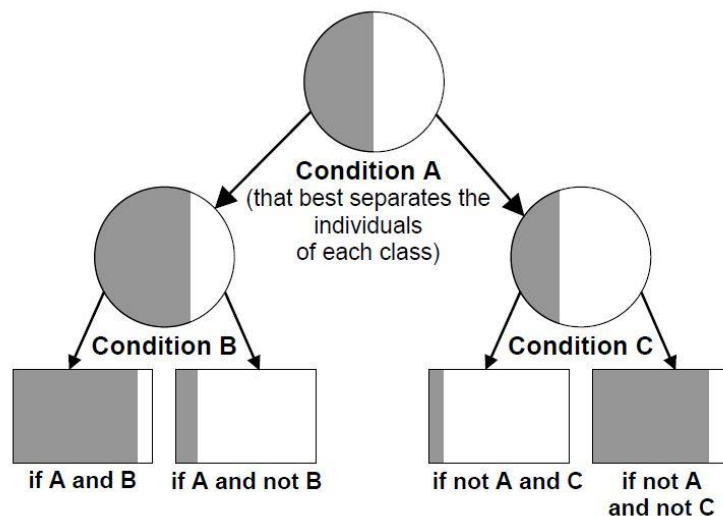


Figure 2. Decision tree (Tufféry, 2011, p. 314).

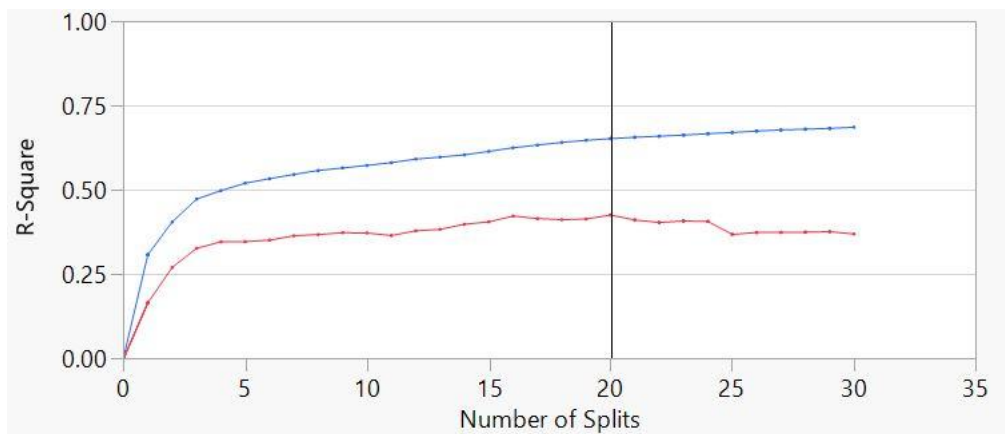


Figure 3. The training  $R^2$  (blue) and validation  $R^2$  (red) with number of splits.

Figure 3 showed the training  $R^2 = .651$  and the validation  $R^2 = .424$  with number of splits = 20. Table 6 reported the measures of how much a variable contributed to the decision tree model. One measure is the accumulated split statistic,  $\chi^2$ , and another measure is the portion of each predictor contributing to the explained variance of the dependent variable. With two times of splits and the accumulated  $\chi^2 = 673.28$ , the number of total passengers became the most important predictor in the model when accounting for 50.87% of the explained variance of LCC presence and absence. The number of carriers contributed the second largest portion after four splits and the accumulated  $\chi^2 = 361.96$ . Average connection yield, average market fare, route type, route length, and origin and destination airport added increments as least as 3% and as much as 5% to the explained variance of LCC presence and absence. Conversely, largest share proportion and number of the connecting passenger were the least important predictors in the model as they did not make any incremental contribution to the explained variance.











Figure 4 showed the full graph of the tree growth for visualization after 20 splits. Combining with the leaf reports (Figure 5) that summarize separation conditions on each node, interpretations were represented. In view of the highest probability of LCC presence, 96.79% of time it was expected to have at least one LCC operation on routes that simultaneously required (a) the number of total passenger greater than or equal to 101 passengers, (b) the number of competing carriers fewer than 4, (c) the average market fare less than \$263.97, (d) route type having the status of nonstop market, (e) destination airports being small or medium hubs, and (f) origin airports being none, small, or medium hubs. On the flipside, the highest probability of LCC absence was interpreted that 99.76% of time it was expected to have no LCC operation on routes that simultaneously required (a) the number of total passengers fewer than 101, (b) fewer 4 operating



carriers, (c) the number of total passengers fewer than 20, (d) origin airport functioning as GA or none hubs, and (e) fewer 3 operating carriers. In more simplified interpretation, if a route had fewer than 20 passengers in demand, fewer than 3 operating carriers, and arrivals from either GA or no hubs, 99.76% of time LCC operations were absent on that route.

Table 6

*Summary of Variable Importance of the Decision Tree*

<b>Term</b>	<b>Number of Splits</b>	$\chi^2$		<b>Portion</b>
Total passengers	2	673.280225		0.5087
Number of carriers	4	361.964188		0.2735
Average connection yield	3	62.5574332		0.0473
Average market fare	2	60.7443507		0.0459
Destination hub	3	45.236771		0.0342
Route type	2	42.5253527		0.0321
Route length	2	38.6484573		0.0292
Origin hub	2	38.5024921		0.0291
Largest share	0	0		0.0000
Connecting passengers	0	0		0.0000

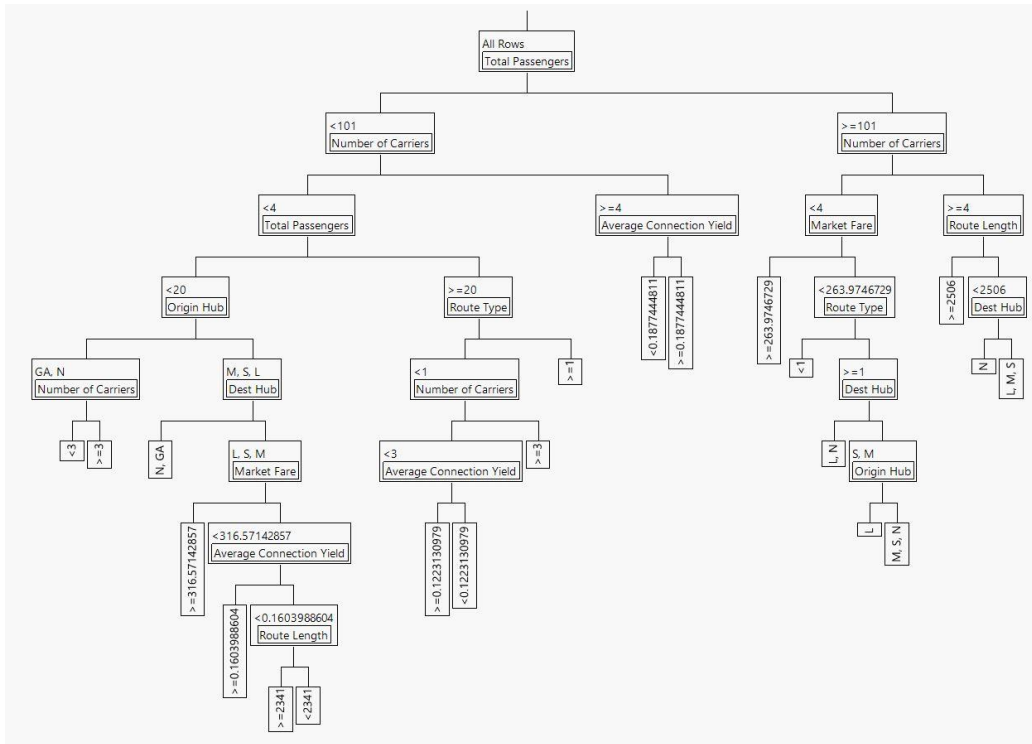


Figure 4. The decision tree view with the number of splits = 20.

Leaf Label	1	0
Total Passengers ≥ 101 & Number of Carriers < 4 & Market Fare < 263.9746729 & Route Type = 1 & Dest Hub(S, M) & Origin Hub(M, S, N)	0.9679	0.0321
Total Passengers ≥ 101 & Number of Carriers ≥ 4 & Route Length < 2506 & Dest Hub(L, M, S)	0.9570	0.0430
Total Passengers < 101 & Number of Carriers ≥ 4 & Average Connection Yield = 0.1877444811	0.9152	0.0848
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(M, S, L) & Dest Hub(L, S, M) & Market Fare < 316.57142...	0.6742	0.3258
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers ≥ 20 & Route Type = 1	0.6435	0.3565
Total Passengers ≥ 101 & Number of Carriers ≥ 4 & Route Length < 2506 & Dest Hub(N)	0.4692	0.5308
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers ≥ 20 & Route Type < 1 & Number of Carriers < 3 & Average Connection Yi...	0.3901	0.6099
Total Passengers ≥ 101 & Number of Carriers < 4 & Market Fare < 263.9746729 & Route Type = 1 & Dest Hub(S, M) & Origin Hub(L)	0.3278	0.6722
Total Passengers ≥ 101 & Number of Carriers < 4 & Market Fare < 263.9746729 & Route Type = 1 & Dest Hub(L, N)	0.3069	0.6931
Total Passengers < 101 & Number of Carriers ≥ 4 & Average Connection Yield < 0.1877444811	0.3045	0.6955
Total Passengers ≥ 101 & Number of Carriers < 4 & Market Fare < 263.9746729 & Route Type < 1	0.1994	0.8006
Total Passengers ≥ 101 & Number of Carriers ≥ 4 & Route Length = 2506	0.1515	0.8485
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers = 20 & Route Type < 1 & Number of Carriers ≥ 3	0.0992	0.9008
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(M, S, L) & Dest Hub(L, S, M) & Market Fare < 316.57142...	0.0795	0.9205
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(GA, N) & Number of Carriers ≥ 3	0.0639	0.9361
Total Passengers ≥ 101 & Number of Carriers < 4 & Market Fare = 263.9746729	0.0406	0.9594
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(M, S, L) & Dest Hub(L, S, M) & Market Fare < 316.57142...	0.0285	0.9715
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers = 20 & Route Type < 1 & Number of Carriers < 3 & Average Connection Yi...	0.0158	0.9842
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(M, S, L) & Dest Hub(L, S, M) & Market Fare = 316.5714...	0.0120	0.9880
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(M, S, L) & Dest Hub(N, GA)	0.0111	0.9889
Total Passengers < 101 & Number of Carriers < 4 & Total Passengers < 20 & Origin Hub(GA, N) & Number of Carriers < 3	0.0024	0.9976

Figure 5. The leaf report of the decision tree.

### Model Comparison

Table 7 reported the measures of fit to assess two models in both training and validation datasets. For Entropy  $R^2$  and Generalized  $R^2$ , values closer to 1 indicate a better fit whereas for Mean -Log  $p$ , RMSE, Mean Absolute Deviation, and Misclassification Rate, smaller values indicate a better fit (SAS Institute Inc.,

2016). Considering the prediction models for LCC presence and absence in the validation set, binary logistic regression generated a superior prediction performance over decision tree method across fitting measures.

Table 7

Summary of Fitting Measures of Data Mining Models

Dataset	Analysis Method	N	Entropy R <sup>2</sup>	Mean		Miss Rate	AUC		
				Gen R <sup>2</sup>	-Log p				
Training	Logistic Regression	3,330	0.5639	0.6373	0.1329	0.1862	0.0728	0.0381	0.9472
	Decision Tree	3,330	0.6508	0.7175	0.1064	0.1651	0.0553	0.0327	0.9597
Validation	Logistic Regression	1,111	0.4337	0.5026	0.158	0.2016	0.0775	0.0441	0.9108
	Decision Tree	1,111	0.4241	0.4927	0.1607	0.1962	0.0691	0.0450	0.8851

Note. Entropy R<sup>2</sup> = McFadden Pseudo R<sup>2</sup>. Gen R<sup>2</sup> = Generalized (Cox-Snell) R<sup>2</sup>. RMSE = Root mean square error. Mean Abs Dev = Mean absolute deviation. Miss rate = Misclassification rate. AUC = Area under the curve.

Apart from the model fitting measures, Receiver Operating Characteristics (ROC) curve and lift chart are reported for the classification study (i.e., the dependent variable is binary). ROC curve is comprised of sensitivity in the vertical axis and 1- specificity in the horizontal axis where:

$$\text{Sensitivity (Recall)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{\text{Routes correctly classified as LCC presence}}{\text{All routes of LCC presence}}$$

$$\text{Specification} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} = \frac{\text{Routes correctly classified as LCC absence}}{\text{All routes of LCC absence}}$$

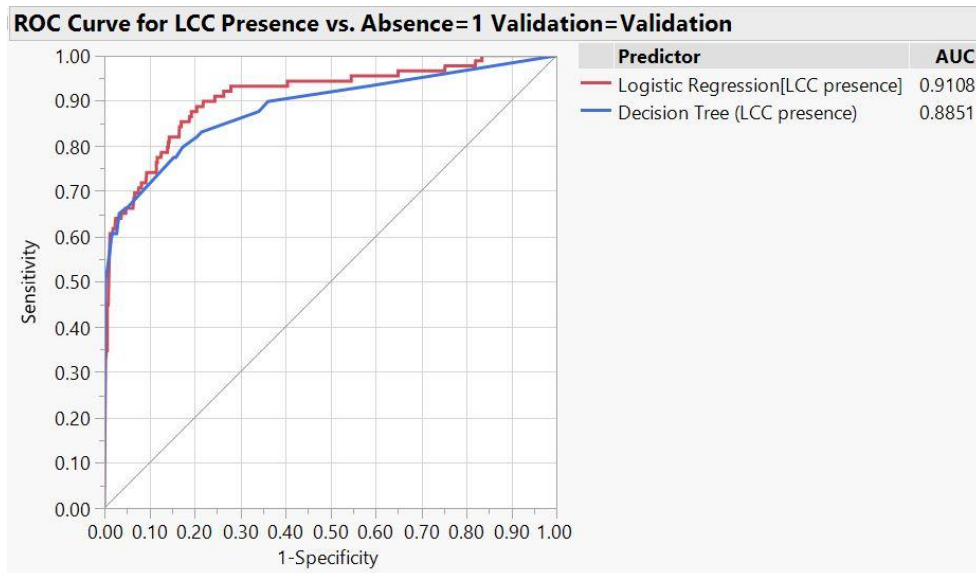


Figure 6. Receiver Operating Characteristics (ROC) curves of logistic regression (red) and decision tree (blue) on the validation set.

In the ROC graph, the vertical axis portrayed the proportion of LCC-present routes that were correctly identified, and the horizontal axis portrayed the proportion of LCC-absent routes that were misidentified as LCC-present ones. It is noticing that the coordinates (0, 1) represented a perfect classification as it always correctly identifies LCC presence routes, contradictorily the coordinate (1, 0) represented a flawed classification as it always misclassified LCC-absent routes as LCC-present routes. The dotted diagonal line represents a random guessing line, which is equivalent to flipping a fair coin to determine LCC-present and-absent routes. As such, the region beneath the dotted diagonal line is worse than random guessing while the closer to the coordinates (0, 1) the better it is. Figure 6 showed ROC graphs for the validation set, the curve of logistic regression in red was closer to the coordinate (0, 1) and thus better than that of the decision tree in blue.

The area under the curve (AUC) is another indicator for comparing ROC curves. As mentioned, the perfect classification curve passes through the coordinates (0, 1) such that AUC region equals 1. AUC for the diagonal line (random guessing line) is 0.5. Hence, a ROC with higher AUC is preferable to the one with a lower AUC. Table 7 and Figure 6 reported AUC for both models; AUC for logistic regression was 0.9108 and AUC for decision tree was 0.8851. The Chi-square test for the difference between the two AUC values. Table 8 summarized the test results showing the AUC for logistic regression was

statistically significantly higher than AUC for the decision tree,  $\chi^2(1) = 4.17$ ,  $p < .0412$ .

Table 8

*Summary of Chi-square test for AUCs of logistic regression and decision tree in the validation set*

<b>AUC Difference</b>	<b>SE</b>	<b>Lower 95%</b>	<b>Upper 95%</b>	<b>df</b>	<b><math>\chi^2</math></b>	<b>p</b>
0.0257	0.0126	0.0010	0.0504	1	4.1697	0.0412*

*Note.* AUC = Area under the curve. SE = Standard Error. AUC for logistic regression = 0.9108. AUC for decision tree = 0.8851.

\* $p < .05$

Lift curve is another plot to display the predictive ability of a classification study. It plots the lift value in the vertical axis against the portion of the observations in the horizontal axis. Each portion represents a decile (10-percentile) of the observations. The underlying idea is that each route was computed the predicted probability (posterior probability) of LCC presence and then sorted in descending order before being broken down to deciles. The lift value in the vertical axis was computed by the ratio of LCC-present routes only in that decile over the overall LCC-present routes. To interpret the lift curve in Figure 7, at the first decile (the coordinates (0.10, 6.5)), the expected number of routes having LCC presence was 6.5 over 100 routes. Similarly, at the second decile (the coordinates (0.20, 4)), for every 100 routes it was expected to have 4 routes having at least one LCC operation. Such ratios were identical in both models, logistic regression and decision tree as both lift curves virtually coincided and converged.

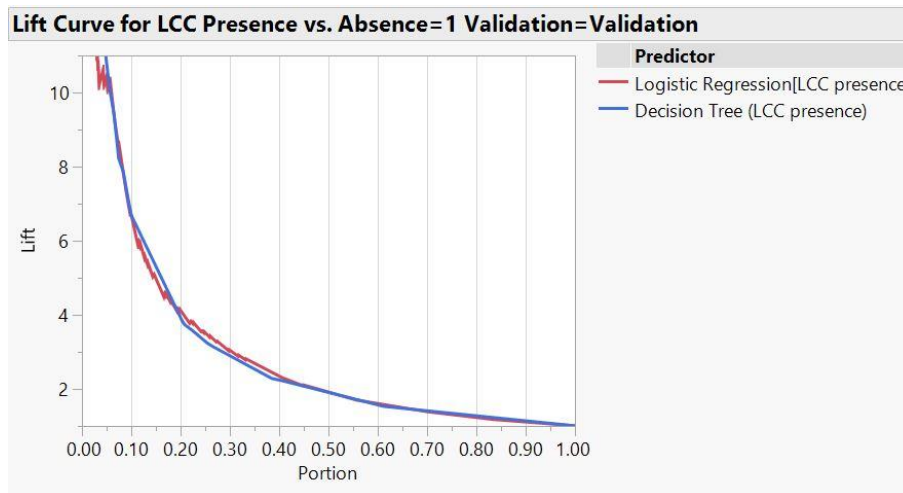


Figure 7. Lift curves of logistic regression (red) and decision tree (blue) on validation set.

## Model Deployment

The Assess phase in SEMMA scheme returned the result in favor of logistic regression over the decision tree method for modeling the LCC presence in the U.S. domestic route. For the testing purpose, the logistic regression model was thus chosen to proceed with the testing dataset 1Q2018. Eight variables significant in the Model phase entered simultaneously into a logistic regression to examine the relationship. As Table 9 reported, the whole testing model was statistically significant,  $\chi^2(8) = 274.28$ ,  $p < .0001$ . Fitting measures of the testing model were virtually identical to those of the training model. Furthermore, seven variables significant at the preset  $\alpha = 0.5$  in the training model, including number of carriers, route type, average market fare, origin airport (L&M&S – Nh&Np), destination airport (L&M&S – Nh&Np), average connection yield, and origin airport (M – L&S), were found to be significant again with the same direction sign of logistic coefficients. The variable significant at the preset  $\alpha = 0.6$  in the training model—destination airport (M – L&S)—was found nonsignificant at this stage.

Table 9

Summary of Simultaneous Logistic Regression Estimates for 1Q2018 Model Testing of LCC Presence vs. Absence

Simultaneous Model <sup>a</sup>	$B_i$	$\chi^2$	$p$
Constant	-2.21	14.30	.0002***
Number of carriers	0.903	25.78	<.0001***
Route type	1.307	8.52	0.0035**
Average market fare	-0.010	22.21	<.0001***
Origin airport (L&M&S – Nh&Np)	1.080	24.55	<.0001***
Destination airport (L&M&S – Nh&Np)	1.018	19.05	<.0001***
Average connection yield	-4.095	6.57	.0104*
Origin airport (M – L&S)	0.389	4.63	.0314*
Destination airport (M – L&S)	0.079	0.13	.7200
-Log Likelihood in Null Model		293.80	
-Log Likelihood in Full Model		156.66	
Difference		137.14	
$\chi^2(8)$		274.28***	

Note.  $N_{\text{Testing}} = 1,094$ . Entropy  $R^2 = .4668$ . Generalized  $R^2 = .5336$ . Mean  $-\text{Log } p = .1432$ . RMSE = .1990. Mean Abs Deviation = .0803. Misclassification rate = .0521.

<sup>a</sup>Seven variables significant at the preset  $\alpha = 0.5$  and one variable (Destination airport (M – L&S)) significant at the preset  $\alpha = 0.6$  in the stepwise logistic regression model were selected for the simultaneous model for testing. L = Large hub, M = Medium hub, S = Small hub, Nh = Non-hub, and Np = Non-primary airport.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$

## Discussion

With respect to research question 1, the stepwise logistic regression yielded seven significant predictors in relation to LCCs presence and absence in the U.S. domestic routes. For every additional carrier commencing flights in a route, there was nearly 14% more likely to have at least one LCC present in that route. This finding might be rooted in LCCs' cost advantage such that they tend to aggressively join head-to-head competitions to capture market share. It is noted that the respected competition might be affected by accommodating all 30 carriers

reported in the database, which in turn contained regional airlines that are feeding passengers to major airlines from spoke cities to hubs.

On the nonstop market, the route was 10.5% more likely to have at least one LCC operation than the one under the condition of a connecting market. It makes sense that LCCs strategically launch point-to-point flights to connect cities, such itineraries are thus characterized as the nonstop market. Considering airfare-related factors, for every decrease of average market fare by \$100 in a flight route, there was 8% more likely to have at least one LCC operation in that route. If just taking connection routes into account, for every decrease of 1 dollar per miles flown, there was nearly a 19% greater likelihood to have at least one flight operated by LCC. Lower airfare is a clear clue as to the presence of LCCs in a route. The relationship is especially more intensive when observing unbundling pricing practices of ultra-LCCs (ULCCs) with bare fares in the market.

On routes with either departure from or arrivals to large, medium, and small hub, there was 6.5% and 5.7% greater likelihood to have at least one LCC operation, respectively. It makes most sense when large, medium, and small hubs are designated to accommodate commercial scheduled flights with large transportation capability, while facilities in non-primary commercial service airports such as runway length and terminal capacity are primarily standardized for serving regional flights by small jets. When decomposing three types of hubs (large, medium, and small) for investigation, there was 2.2% and 1.9% greater likelihood to have LCC presence on routes with origin and destination airports as the function of medium hubs (significant at  $\alpha = .06$ ). This matches with “secondary airport” strategy of LCCs as they tend to move their operations to medium hubs for serving point-to-point flights as well as avoiding high fees, congestion in large hubs.

With respect to research question 2, the decision tree model disclosed eight predictors contributing to the predictive model of LCC presence and absence in the U.S. domestic routes. More specifically, the number of total passengers was the most important predictor in the model when accounting for 50.87% of the explained variance of LCC presence and absence. Followed by this was the number of carriers adding an incremental portion of 27.35% to the model. Airfare factors, route characteristics factors, and airport factors added increments from 3% to 5%.

With respect to research question 3, both logistic regression and the decision tree consistently showed the significant relationships of the number of carriers, two airfare-related factors, route type, and two airport factors with LCC presence and absence in the U.S. domestic routes. On the validation set, model comparison tests unveiled a superior performance of logistic regression over the



decision tree in predicting the presence of LCCs in the network. The higher predictability of logistic regression was reflected in fitting measures, ROC curves, AUC comparison test, and Lift chart. Model testing was then deployed, showing stability and consistency of the logistic regression method.

### **Conclusions**

In the aviation industry, predictive modeling has proven to be important and widely used in supporting decision-making. This study represented two data mining methods, logistic regression and decision tree to predict the presence of LCCs in the U.S. domestic network. Data in the period of 2016-2017 and 1Q2018 from DB1B database revealed that market concentration was an important predictor positively related to LCC presence. This finding was somewhat contradictory to the conventional wisdom that a firm is more likely to do business in the monopolistic market with fewer competitors to leverage the bargaining power of suppliers. The study's finding did not support Nguyen's and Nguyen's (2018, p. 112) finding saying that "on routes with at least one operations of a LCC, airlines were 6% less likely to make an entry decision." Findings on both average market fare and average connection yield indicated the negative relationship with LCC presence. The findings were parallel with those in prior studies reporting that the market fare was lower in routes having the presence of LCCs than the average of the entire network (Bachwich & Wittman, 2017; Ben Abda et al., 2012). LCCs were more likely to appear on nonstop market by serving the nonstop flights connecting cities as opposed to flying to hubs. This finding concretely supported the "point-to-point" strategy aligning with the LCC business model (Belobaba et al., 2015; Vasigh et al., 2016). First finding on airport factor made the most sense when LCCs operations were more likely to be present in primary commercial airports (large, medium, and small hubs). Second finding on airport factor implied a shift of LCCs' focus to medium hubs rather than maintaining their operations in large or small hubs. This finding was consistent with the secondary airport strategy of LCC business model (Ben Abda et al., 2012; Vasigh et al., 2016). The study's findings have implications to activities in network planning of airlines and airports relative to understanding characteristics of the LCC operations. Enhancing the prediction on the presence of LCCs in a route could help airlines avoid head-to-head competition on airfare with LCCs. Airport personnel in an air service development department may gain insights about reallocating LCCs operations away from or to their airports.

The study had a limitation pertaining to the data integrity that we had no control over; that is, how the data were recorded and stored in the DB1B database. The delimitation of the study reflected on the data collection period of 2016-2017 and 1Q2018, and thus similar studies conducted on different periods might not generate the same results. Other delimitations referred to our choices for number

of routes for sampling, handling data missing, removing outliers, coding techniques, and thus replicative studies using different techniques for data analysis might not get the same results.

Future studies should limit the dataset to major airlines truly involved in the competition. In certain routes, the competition level was somewhat distorted by counting operating carriers or regional carriers that are feeding passengers to hubs under the ticketing name of major airlines. This study failed to find the significant relationship of LCC presence with demand factors, which may be inconsistent with previous research. Future research before reconstructing the dataset should remove flight records in a route that have fewer than 90 passengers per quarter (Berry, 1990) or fewer than 260 passenger per quarter (Aguirregabiria & Ho, 2012), as such traffic would be reflected more accurately in routes. Because of the sampling delimitation future research should expand the sampled population to include the full data set.

## References

- Abdelghany, A., & Abdelghany, K. (2009). *Modeling applications in the airline industry*. Abingdon, U.K.: Routledge.
- Aguirregabiria, V., & Ho, C.-Y. (2012). A dynamic oligopoly game of the US airline industry: Estimation and policy experiments. *Journal of Econometrics*, 168(1), 156–173.
- Airlines for America. (2014). Domestic round-trip fares and fees. Retrieved February 20, 2019, from <http://airlines.org/dataset/annual-round-trip-fares-and-fees-domestic/>
- Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *IADS-DM*.
- Bachwich, A. R., & Wittman, M. D. (2017). The emergence and effects of the ultra-low cost carrier (ULCC) business model in the U.S. airline industry. *Journal of Air Transport Management*, 62, 155–164.  
<https://doi.org/10.1016/j.jairtraman.2017.03.012>
- Belobaba, P., Odoni, A. R., & Barnhart, C. (2015). *The global airline industry* (2nd ed.). In *Aerospace Series; Aerospace Series*. (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc. WorldCat.org.
- Ben Abda, M., Belobaba, P. P., & Swelbar, W. S. (2012). Impacts of LCC growth on domestic traffic and fares at largest US airports. *Journal of Air Transport Management*, 18(1), 21–25.  
<https://doi.org/10.1016/j.jairtraman.2011.07.001>
- Berry, S. T. (1990). Airport presence as product differentiation. *The American Economic Review*, 80(2), 394–399.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. 76, 3.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Coldren, G. M. (2005). *Modeling the competitive dynamic among air-travel itineraries with generalized extreme value models*. Northwestern University.
- Coldren, G. M., Koppelman, F. S., Kasturirangan, K., & Mukherjee, A. (2003). Modeling aggregate air-travel itinerary shares: Logit model development

at a major US airline. *Journal of Air Transport Management*, 9(6), 361–369.

- Federal Aviation Administration. (2016). Airport Categories. Retrieved November 21, 2017, from [https://www.faa.gov/airports/planning\\_capacity/passenger\\_allcargo\\_stats/categories/](https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/categories/)
- Garrow, L. A. (2010). *Discrete Choice Modelling and Air Travel Demand*. Farnham, England: Ashgate.
- Grayson, J., Gardner, S., & Stephens, M. (2015). *Building better models with JMP Pro* (1st ed.). Cary, NC: SAS Institute.
- Hair, J. F., Black, W. C., Babin, B. J., & Rolph, A. E. (2010). *Multivariate Data Analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education.
- Klimberg, R., & McCullough, B. D. (2016). *Fundamentals of predictive analytics with JMP* (2nd ed.). Cary, NC: SAS Institute.
- Linoff, G. S., & Berry, M. J. (2011). *Data mining techniques: For marketing, sales, and customer relationship management* (3rd ed.). New York: Wiley.
- Neville, P., & Ville, B. de. (2013). *Decision trees for analytics using SAS Enterprise Miner* (1st ed.). SAS Institute.
- Nguyen, C., & Nguyen, C. (2018). Understanding determinants of making airline route entry and exit decisions: An application of logit models. *The Collegiate Aviation Review International*, 36(2).
- Paczkowski, W. R. (2018). *Market Data Analysis Using JMP* (1st ed.). Cary, NC: SAS Institute.
- Sarma, P., Kattamuri S. (2013). *Predictive modeling with SAS® Enterprise Miner™ practical solutions for business applications*. (2nd ed.). Cary, N.C: SAS Institute Inc.
- SAS Institute Inc. (2016). *JMP® 13 Predictive and specialized modeling*. Cary, NC: SAS Institute Inc.
- Tufféry, S. (2011). *Data Mining and Statistics for Decision Making* (1st ed.). <https://doi.org/10.1002/9780470979174>
- Vasigh, B., Fleming, K., & Tacker, T. (2016). *Introduction to air transport economics: from theory to applications* (2nd ed.). Farnham, England: Ashgate.

