Predicting U.S. Airline Financial Performance Using Operational Variables

– Which is better? Econometrics Models or Machine Learning Models?

1. INTRODUCTION

Using operational predictors to predict U.S. airline financial performance is a critical issue for both airline policymakers and researchers (Alan and Lapré 2018). Heavily influenced by the finance literature, existing operations management research only focuses on predicting airline bankruptcies (Alan and Lapré 2018, Lu et al. 2015, Phillips and Sertsios 2013), which is not surprising given the high frequency of U.S. airlines filing for bankruptcies from 1980s to 2000s (Alan and Lapré 2018). However, the U.S. airline industry has become increasingly concentrated in the recent years with the Big Four (American, Delta, United, and Southwest) controlling 80% of the market share (Semuels 2023). Filing for bankruptcy has consequently become a rare event. Indeed, at the time of writing this study, only three major U.S. airlines (defined as having more than 1% domestic market share by Department of Transportation) filed for bankruptcies after 2010. American Airline filed for bankruptcy on November 29, 2011, and accordingly strengthened its position as one of the Big Four carriers. Pinnacle filed for bankruptcy on April 1, 2012, and consequently restructured itself as Endeavor Airline. While the Pandemic killed 64 airlines at a global scale (Buckley 2023), in the U.S., only ExpressJet filed for bankruptcy on August 22, 2022 (Korn 2022). Moreover, ExpressJet filed for bankruptcy purely due to losing its commercial contract with Delta in 2018 and further losing the remaining contract with United Airlines in 2020 (Yamanouchi 2022), having nothing to do with its idiosyncratic operational metrics that are frequently used to predict bankruptcy. Actually, ExpressJet is planning for a strong comeback as a charter operator in the second half of 2024 (Birns 2023).

With the market landscape and the financial landscape changed dramatically in the U.S. airline industry, coupled by the strong recovery of postpandemic travel demand (Sider 2023), concerns about bankruptcies were replaced by stronger desire to achieve better financial performance (Kletzel et al. 2023, Stalnaker et al. 2023). Consequently, how to use operational predictors to predict financial performance, rather than the probability of bankruptcy, has become more meaningful and more practical. Therefore, distinct from the extant operations management literature, the overarching goal of the current study is to predict actual airline financial performance reported as profit to the Department of Transportation (DOT), rather than the calculated probability of bankruptcy or financial distress using stock prices.

In predicting airline bankruptcies, conventional econometrics approaches have been adopted (Alan and Lapré 2018, Lu et al. 2015, Phillips and Sertsios 2013, Gudmundsson 2004; 2002), the majority of which using fixed effects models. When fixed effects models are used to predict airline bankruptcies, different operational metrics were identified, then, standard panel-data methods were used to fit the data where airline bankruptcies depend on operational metrics, along with carrier and time fixed effects to capture idiosyncratic differences of airlines and time. Despite fixed effects models are considered the gold

standard in econometrics (Schurer and Yong 2012), there are two deficiencies when using fixed effects models to predict airline financial performance. First, as fixed effects models use carriers as dummy controls, any time-invariant effects at the carrier level are also controlled out, so, these effects cannot be estimated (Bell and Jones 2015). Second, at the measurement occasion level, fixed effects models are unable to estimate the time-invariant effects (such as the differences in average yield between Delta and American) of any time-varying variable (yield itself), which is equivalent to an omitted variable bias (Bafumi and Gelman 2006, Palta and Seplaki 2003). Combined, these two deficiencies will impact the prediction accuracy when fixed effects models are used to predict airline financial performance. To deal with the deficiencies in fixed effect models, recent econometrics research (Bell and Jones 2015) proposes using mixed effects models as an alternative where all time-invariant effects and time-varying effects at any level can be separately estimated (more in Section 2.2). Accordingly, mixed effects models are expected to yield more reliable and more accurate predictions for individual units (Rubin 1980), i.e., carriers in our case. Therefore, distinct from the related airline research where only fixed models were used to predict airline bankruptcies, our study also considers mixed effects models as an alternative to predict airline financial performance to verify if indeed mixed effects models can give better prediction accuracy.

With the increasing popularity of applying machine learning models to study operations management topics in the recent years, we also examine the potential to use machine learning models to predict airline financial performance, which has not yet been explored in related airline literature. We elect to use two popular machine learning models: eXtreme Gradient Boosting (XGB) and Deep Neural Networks (DNN) due to their ability to nonparametrically estimate high-dimensional nonlinear relationships in the data (Chen and Guestrin 2016, Leshno et al. 1993). In our study, therefore, we adopt a total of four different models: fixed effects models, mixed effects models, XGB, and DNN to predict airline financial performance, the former two belonging to econometrics and the latter two coming from machine learning field. In sharp contrast to related airline literature where only fixed effects models were used, we, therefore, are able to compare the predictive performance between econometrics and machine learning using these four models. In addition, in predicting airline financial performance, only one single study (Alan and Lapré 2018) has analyzed both in-sample and out-of-sample prediction accuracy using one single econometrics model. Differently, our study compares both in-sample and out-of-sample prediction accuracy among four different models.

We collect data from DOT to conduct our analysis. Following extant airline literature (Alan and Lapré 2018, Stevens et al. 2012), we classify our operational predictors into five different operational dimensions: pure operational metrics, efficiency metrics, service quality metrics, human metrics, and market power metrics. For the outcome variable of financial performance, we use operating profit over operating revenue following extant operations management literature to better account for size differences

among carriers as well as to overcome measurement issues (Mellat-Parast et al. 2015, Tsikriktsis 2007, Dresner and Xu 1995). To avoid the impact of 9/11 and the global Pandemic as well as to study the most recent airline financial situations, we keep out data from 2004 to 2019.

We conduct our analysis at two different stages. First, we run a Monte Carlo cross-validation process using 80% of our data as in-sample and 20% as out-of-sample by randomly stratifying the data such that both in-sample and out-of-sample data contain observations from all carriers. To avoid the random stratification algorithm accidentally select easy or hard-to-fit observations, we run our analysis multiple times using different randomly stratified datasets and report the averaged results. In our Monte Carlo cross-validation process, we find that mixed effects models give the best out-of-sample prediction accuracy, followed by DNN and XGB. Fixed effects models yield the worst out-of-sample prediction accuracy. Next, we keep the last eight quarters of observations from each carrier as the to-be-predicted "future" while train our models using the rest of the data. Among the four models, the two machine learning models produce the same out-of-sample prediction accuracy, outperforming mixed effects models by 56% and also outperforming fixed effects models by 77% in terms of out-of-sample prediction accuracy. In sum, based on our analysis and given our data range, to predict future airline performance, either DNN or XGB is a better model choice compared with mixed effects models and fixed effects models.

Evaluating model performance across different non-random subsets of data is crucial to identify potential limitations or misspecifications. If a model performs poorly on certain non-random subsets, it may indicate areas where the model fails to capture or accurately represent the underlying patterns. Given that the Big Four carriers dominate over 80% of the U.S. domestic market share, we elect to use these four major carriers as our non-random subsamples to conduct all the analyses again. For the big four carriers, when predicting their "future" eight-quarter financial performance, the two machine learning models, on average, outperform mixed effects models by 93% and also outperform fixed effect models by 98% in terms of out-of-sample prediction accuracy.

As the two machine learning models outperform the two econometrics models in almost all of our analyses, we further attempt to explain the differences of the prediction accuracy between econometrics models and machine learning models. Given machine learning models are known for their ability to model complicated nonlinear high-dimensional relationships and given the existence of other nonlinear relationships in the U.S. airline industry (Steven et al. 2012), we accordingly build an mixed effects model to test if indeed the existence of non-linearity in our data can partially explain the better prediction accuracy of machine learning models. We rely on Theory of the Performance Frontier (Schmenner and Swink 1998) to select two potential variables, i.e., load factor and yield, which may demonstrate a non-liner relationship with financial performance. Our statistics results show that both load factor and yield demonstrate an inverted U-shaped relationship with airline financial performance. We can partially conclude that the non-

linear relationship in our data may help explain the superior performance of machine learning models in terms of prediction accuracy.

In summary, from a theoretical perspective, our study extends the understanding of predicting airline financial performance by predicting the reported financials, rather than the probability of bankruptcies. Our analysis is also different from the extant airline literature in that we do not solely rely on fixed effects models to predict. Rather, we adopt four different models and compare the model prediction accuracy among the four models. More importantly, we introduce two machine learning models into the related stream of airline literature. Our results indicate that the two machine learning models give superior out-of-sample prediction accuracy. We call for researchers to explore applying machine learning models to study other airline operations management topics. From a managerial perspective, given airlines are increasingly focusing on improving their financial performance in the post-Covid era (Kletzel et al. 2023, Stalnaker et al. 2023), our research can provide more meaningful insights for airline policymakers to use available operational metrics to predict their future financial performance, especially for the Big Four carriers where the out-of-sample forecast accuracy is as low as 0.0005 measured by mean squared errors. In addition, we use 15 operational variables from five different operational dimensions to predict airline financial performance. Airlines can mix and match these 15 different variables and examine how changes in each of the operational dimensions could affect their financial performance to facilitate decision making.

2. MODELING AIRLINE FINANCIAL PERFORMANCE

Despite a steady stream of airline literature focusing on predicting airline bankruptcy and financial distress (Alan and Lapré 2018, Lu et al. 2015, Phillips and Sertsios 2013, Gudmundsson 2004; 2002), little attention has been paid to the characteristics of airline data. We first discuss the characteristics of airline data as the data structure itself determines which modeling approach is the most appropriate.

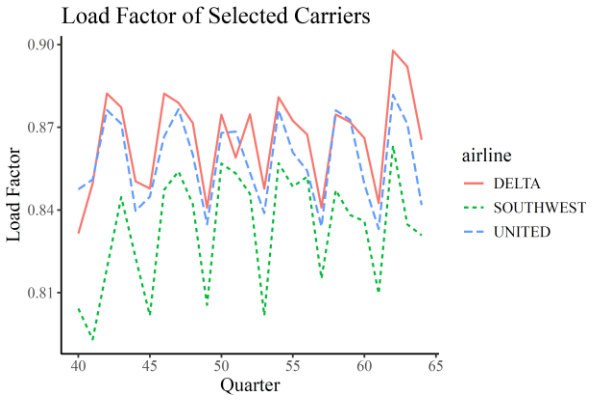
2.1 Airline Data Structure

Compared to other datasets used in longitudinal research in operations management, airline data is unique in that in the U.S. airline industry, the number of airlines is very small as a result of frequent mergers and acquisitions after the deregulation – only 15 airlines report to DOT regularly for their operational and financial metrics at the time of writing. Some of the operational metrics reported, such as load factor, may date back as early as in 1988. In other words, the data reported to DOT by U.S. airlines has comparatively fewer higher-level entities (i.e., carriers) and comparatively more lower-level measurement/reported occasions nested in each carrier. This specific structure of data is known as time-series cross-sectional (TSCS) data (Alan and Lapré 2018, Bell and Jones 2015, Beck and Katz 1995).

In TSCS data, as well as in other panel data, there are a few distinctive characteristics. First, within the same higher-level entity (airlines in our case), the repeated measures at the occasion level are marked by a high degree of dependency. For example, despite seasonal fluctuations before the Pandemic, the load

factor of Delta Airline always fluctuates between 84% – 87% quarter-over-quarter while the load factor of Southwest Airline always fluctuates between 81% – 85% quarter-over-quarter (Figure 1). In other words, the measures at each measurement occasion from each carrier are highly correlated to each other. Second, some higher-level variables are time-invariant or rarely changing. For example, Frontier and Southwest airlines are classified as low-cost carriers (LCCs) while Delta Airline is classified as a legacy carrier. The status of being a low-cost carrier or a legacy carrier does not change over time – at least during the time period reported to DOT. An example of a rarely changing variable is network sparsity and fleet heterogeneity as airlines do not frequently adjust their routes and fleets. Third, in addition to the time-invariant variables measured at the higher-level, time-varying variables measured at the occasion level for each carrier also demonstrate time-invariant and time-varying characteristics. Turning to the load factor of Delta Airline and Southwest Airline in Figure 1, the time-varying characteristics are the seasonal fluctuations while the time-invariant characteristics are their respective average load factors during the observed period – the average load factor of Delta Airline is always higher than the average load factor of Southwest Airline by about 3%.

Figure 1 Load Factor of Selected Carriers



2.2 Challenges in Modeling Airline Data Using Econometrics Models

The unique structure of airline data poses modeling challenges. Various approaches have been used in airline research to predict financial performance. We discuss the advantages and disadvantages of various approaches used in relevant literature in below.

Traditionally, standard pooled regression has been used to predict airline financial performance (Gudmundsson 2004). A "pooled" linear regression assumes the residuals are independently and identically distributed. In other words, a "pooled" linear regression assumes no difference between higher-level entities and no dependency between measures. As a result, a standard pooled regression pools all lower-level observations into a single population. However, airline data is characterized by marked dependencies over time, such as load factor in each measurement occasion for Delta Airline is related to each other. If this

dependency is not accounted for, standard errors will not be correctly estimated (Moulton 1986). The reason is that due to the dependency of lower-level measures, the effective sample size in a standard pooled regression is smaller than the number of lower-level units that a pooled regression assumes, resulting in incorrect standard errors.

To rectify the incorrect standard errors associated with a pooled regression, different approaches can be adopted. First, a random effects model (Equation 1) can be used to partition the residual variance into both higher-level μ_j for higher-level variables z_j and lower-level e_{ij} for lower-level variables x_{ij} (i represents measurement occasion, j represents carriers, z represents time-invariant or rarely changing variables, and x represents time-varying variables). The variation occurred at the higher-level variables z_j is modeled in terms of the higher-level units, thus, the standard errors are correct (Bell and Jones 2015).

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \mu_j + e_{ij} \quad \text{(Equation 1)}$$

However, a random effects model relies on two critical assumptions: $\text{Cov}(x_{ij}, \mu_j) = 0$ and $\text{Cov}(x_{ij}, e_{ij}) = 0$, the former of which often does not hold in most research settings and has been identified as one of the endogeneity issues, leading to the abandonment of random effects models and the adoption of fixed effects models instead (Bell and Jones 2015). A fixed effects model (Equation 2) controls out all higher-level variances using the higher-level entities themselves by including them in the model as dummy variables D_j (Allison 2009), eliminating the potential endogeneity issue when $\text{Cov}(x_{ij}, \mu_j) \neq 0$, and therefore, solves the incorrect standard error issue. Consequently, fixed effects models are most commonly used in related airline research to predict airline financial performance. In addition to the basic fixed effects model specification in Equation 2, other variants of fixed effects models have been applied to further address the incorrect standard error issue. For example, Alan and Lapré (2018) used Prais-Winsten regression to estimate panel-corrected standard errors.

$$y_{ij} = \sum_{j=1}^{j} \beta_{0j} D_j + \beta_1 x_{ij} + e_{ij} \quad \text{(Equation 2)}$$

Due to the above-mentioned advantage, fixed effects models have been considered the gold standard method in many disciplines (Schurer and Yong 2012). However, in recent years, fixed effect models have been criticized by the latest research in econometrics for two major pitfalls (Bell and Jones 2015). First, as fixed effects models use higher-level entities (carriers in our case) as dummy controls, higher-level differences between higher-level entities and the distinctive characteristics of higher-level entities are also controlled out. In other words, fixed effects models cannot estimate any higher-level processes because the models only deal with occasion-level processes. As a result, fixed effects models cannot measure the effects of time-invariant variables if these variables are of interest to researchers (Bell and Jones 2015). Second, as was shown in Figure 1, even time-varying variables measured at the occasion level also demonstrate both time-invariant and time-varying effects. Fixed effects models cannot specifically estimate the nuances of these two effects. In other words, β_1 in Equation 2 is a weighted average

of two different processes from the same variable and therefore is uninterpretable (Krishnakumar 2006, Raudenbush and Bryk 2002). This un-interpretability is equivalent to an omitted variable bias because the higher-level process is omitted (Bafumi and Gelman 2006, Palta and Seplaki 2003).

To solve all the issues discussed above related to standard pooled regression (incorrect standard errors), random effects models (endogeneity issue when $\mathrm{Cov}(x_{ij},\mu_j) \neq 0$), and fixed effects models (inability to model higher-level processes), Bell and Jones (2015) proposed a rectified random effects model or mixed effects model in Equation 3.

$$y_{ij} = \beta_0 + \beta_1 (x_{ij} - \overline{x}_j) + \beta_2 z_j + \beta_3 \overline{x}_j + \mu_j + e_{ij} \qquad \text{(Equation 3)}$$

A mixed effects model adds an additional term – the higher-level mean $\overline{x}_j$ (group mean centered) of each time-varying variable x_{ij} to account for the time-invariant effect of that variable. In Figure 1, $\overline{x}_j$ would become the computed average load factor for each carrier while x_{ij} is the actual load factor reported to DOT. Therefore, in Equation 3, β_1 measures the time-varying effect of x_{ij} and β_3 measures the time-invariant effect of x_{ij}. By specifying the model in this way, all the issues discussed above are solved. First, correct standard errors are automatically calculated because Equation 3 accounts for ‘‘multiple sources of clustering’’ (Raudenbush 2009, p. 473). Second, as each time-varying variable has a group mean-centered term with a mean of zero, there will be no correlation between the group mean-centered time-varying covariate x_{ij} and the higher-level variance μ_j. Hence, the potential endogeneity issue is solved and $\mathrm{Cov}(x_{ij},\mu_j) = 0$ can be guaranteed (Bell and Jones 2015). In addition, $\overline{x}_j$ is unconstrained by the time-varying effects of x_{ij}, so it is free to account for all the higher-level variances associated with x_{ij}, meaning the estimate of β_1 in Equation 3 will be identical to β_1 from Equation 2 in the fixed effects model. Third, as Equation 3 indicates, time-invariant effects of both the lower-level time-varying variables x_{ij} (β_3) and the higher-level time-invariant variables z_j (β_2) can both be correctly estimated (Bell and Jones 2015). As mixed effects models can solve all the pitfalls associated with airline data in the modeling process, more reliable predictions for individual higher-level units can also be expected (Rubin 1980), which is key to the tenet of our current study – predicting higher-level individual airline financial performance. Therefore, we elect to use mixed effects model as one of the candidates for our analysis.

2.3 Exploring Machine Learning Methods as Alternatives to Model Airline Data

We further explore the potential of using machine learning models to predict airline financial performance given the increasing popularity of applying machine learning models in operations management in the recent years. We explore two popular models: XGB (eXtreme Gradient Boosting) and DNN (Deep Neural Networks), the former belonging to the family of decision tree-based models while the latter inspired by biological neural systems consisting of interconnected layers of artificial neurons (nodes).

2.3.1 XGB – eXtreme Gradient Boosting

We use *XGBRegressor* to formulate a regression model to predict airline financial performance. Emphasizing both computational efficiency and predictive accuracy, XGBRegressor is a powerful and widely adopted machine learning algorithm that implements the gradient boosting decision tree framework (Chen and Guestrin 2016). It stands out due to its speed, scalability, and frequently superior performance in a variety of regression tasks, including operational tasks.

Following Chen & Guestrin (2016), we choose our loss function (L) to be the mean squared error (MSE): $L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. With MSE as the loss function, the first-order gradient (Gradient), i.e., the partial derivative of L with respect to $\hat{y}$ is $\frac{\partial L(y,\hat{y})}{\partial \hat{y}} = -(y - \hat{y})$. The second-order gradient (Hessian), i.e., the second partial derivative of L with respect to $\hat{y}$ is $\frac{\partial^2 L(y,\hat{y})}{\partial \hat{y}^2} = 1$. The Gradient measures the rate of change of the loss with respect to changes in $\hat{y}$ and is used to update the predictions in the direction that reduced the loss. The Hessian measures the curvature of the loss function and is used to scale the gradient updates. Being a constant, the Hessian simplifies the calculations and makes the optimization process more efficient. During tree building, these gradients are computed for each data point and then aggregated across the entire dataset to determine the best split in each decision tree. XGBRegressor seeks the split that best minimizes the loss function. The final prediction from an XGBRegressor model is the sum of the leaf scores from all the trees in the ensemble: $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ where K is the total number of trees in the ensemble.

2.3.2 Deep Neural Networks

As a nonparametric nonlinear model, DNN has been applied in certain operations management field (Qi et al. 2023, Gabel and Timoshenko 2022). However, DNN has not yet been applied to predicting airline financial performance. Given DNN is known for its ability to model complicated nonlinear high-dimensional relationships and given the nonlinear relationships found in airline literature (Steven et al. 2012), we find it necessary to consider using DNN as another alternative to conventional econometrics models to predict airline financial performance. Despite its attractive features of approximate any continuous function to arbitrary accuracy (Leshno et al. 1993), DNN is not an econometrics model per se and thus has been considered a "black box" where the numerous parameters and their nonlinear relationships (Figure 2) make it challenging for researchers to interpret the estimation results.

Figure 2 A Deep Neural Network

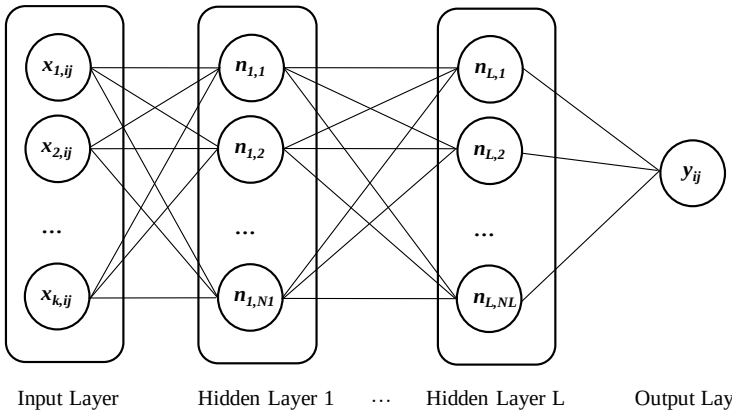


Figure 2 illustrates the architecture of a typical DNN. The input layer consists of a set of K inputs – in our case K different occasion level time-varying covariates $x_{ij} = (x_{1,ij}, x_{2,ij}, \ldots, x_{K,ij})$ specific to occasion i and carrier j. The first hidden layer in the DNN consists of N_1 neurons, the second hidden layer consisting of N_2 neurons, and the L^{th} hidden layer consisting of N_L neurons. In each hidden layer, let u represent a single neuron, each neuron u then takes a weighted sum of inputs from the previous layer. For example, the inputs feed into the neurons in the first hidden layer is $M_{1,u} = w_{1,u0} + w_{1,u1}x_{1,ij} + \ldots + w_{1,uK}x_{K,ij}$ – a linear index from the input layer. $w_{1,u0}, w_{1,u1}, \ldots, w_{1,uK}$ are parameters to be estimated. After each neuron in the first hidden layer receives the linear index $M_{1,u}$, each neuron combines the linear index $M_{1,u}$ with a nonlinear activation function to produce a scalar output $n_{1,u}$, i.e., after all neurons in the first hidden layer complete the same calculation, a vector of outputs from the N_1 neurons is $n_{1,1}, n_{1,2}, \ldots, n_{1,N1}$ as shown in Figure 2.

The neurons in the second hidden layer and the rest of the hidden layers repeat the same calculation process. For example, neurons in the second hidden layer take the outputs from the first hidden layer as the inputs and calculate $M_{2,u} = w_{2,u0} + w_{2,u1}n_{1,1} + \ldots + w_{2,uN1}n_{1,N1}$ and then pass it to a same activation function to calculate the outputs $n_{2,u}$ for $u = 1, \ldots, N_2$. Then the outputs from the second hidden layer will be used as the inputs for the next hidden layer and the process repeats itself up to the final hidden layer L.

When the network finally reaches the output layer, i.e., the outcome variable y_{ij}, the output layer simply calculates a weighted sum of all the outputs from the neurons in the final hidden layer L to estimate the outcome variable y_{ij} as in Equation 4:

$$y_{ij} = w_{L,0} + w_{L,1}n_{L,1} + w_{L,2}n_{L,2} + \ldots + w_{L,N_L}n_{L,N_L} \quad \text{(Equation 4)}$$

3. DATA

We collect airline data from Department of Transportation (DOT). To avoid the impact of DOT's report format change in October 2003, the impacts of 9/11, and the most recent impact of the global pandemic, we elect to choose our data starting point as the first quarter of 2004 and our data ending point as the last quarter of 2019. After integrating data and tracking the name changes of some airlines, our data consists of 28

carriers from 2004Q1 to 2019Q4. Some carriers span the entire 64 quarters while others have fewer quarters either due to their revenues falling below the one percent market share reporting threshold or due to merger and acquisition. A detailed summary of airlines used in our analysis is in Table 1.

Table 1 Airlines in the Dataset

No.	Airline	First quarter in the sample	Last quarter in the sample	Total quarters in the sample
1	AIRTRAN	2004 Q1	2011 Q1	40
2	ALASKA	2004 Q1	2019 Q4	64
3	ALEEGIANT	2018 Q1	2019 Q4	8
4	ALOHA	2006 Q2	2008 Q1	8
5	AMERICA WEST	2004 Q1	2005 Q4	8
6	AMERICAN	2004 Q1	2019 Q4	64
7	ATA	2004 Q1	2006 Q4	12
8	ATLANTIC SOUTHEAST	2004 Q1	2011 Q4	35
9	COMAIR	2004 Q1	2010 Q4	28
10	CONTINENTAL	2004 Q1	2011 Q4	32
11	DELTA	2004 Q1	2019 Q4	64
12	ENDEAVOR	2010 Q4	2019 Q4	28
13	ENVOY	2004 Q1	2019 Q4	56
14	EXPRESSJET	2004 Q1	2019 Q4	62
15	FRONTIER	2005 Q2	2019 Q4	59
16	HAWAIIAN	2004 Q1	2019 Q4	64
17	INDEPENDENCE	2004 Q1	2005 Q4	8
18	JETBLUE	2004 Q1	2019 Q4	64
19	MESA	2006 Q1	2019 Q4	40
20	NORTHWEST	2004 Q1	2009 Q4	24
21	PSA	2018 Q1	2019 Q4	8
22	REPLUBLIC	2018 Q1	2019 Q4	8
23	SKYWEST	2004 Q1	2019 Q4	64
24	SOUTHWEST	2004 Q1	2019 Q4	63
25	SPIRIT	2015 Q1	2019 Q4	20
26	UNITED	2004 Q1	2019 Q4	64
27	US AIRWAYS	2004 Q1	2015 Q2	41
28	VIRGIN AMERICA	2012 Q1	2017 Q4	25

Notes:

1. *RU was used from October 2003 to June 2006 by DOT to code ExpressJet. Effective July 2006, ExpressJet changed in DOT report from RU to XE. In our dataset, RU was changed to XE.*
2. *American Eagle Airlines changed to Envoy effective April 2014 in DOT report. Both Envoy and American Eagle were coded as ENVOY in our data.*
3. *Atlantic Coast Airlines changed to Independence Airline since 2004 November in DOT report. Both airlines were coded as Independence in our data.*
4. *Endeavor Air, formerly Pinnacle Airlines, was ranked for the first time in January 2013 DOT report. Both Pinnacle and Endeavor were coded as Endeavor in our data.*
5. *Atlantic Southeast (EV) was acquired by ExpressJet and changed to XE since January 2012 in DOT report.*
6. *Low-Cost Carriers: Allegiant, Frontier, JetBlue, Southwest, Spirit, and Virgin America.*

A variety of operational dimensions can impact airline financial performance. Existing airline research has adopted different operations dimensions to predict airline financial performance, such as revenue management, operational efficiency, operational complexity, and operations service failures (Alan and Lapré 2018). We review the related literature and collect relevant operational predictors from DOT. We classify our operational predictors into five dimensions: operations metrics, efficiency matrix, service quality matric, human metrics, and market power metrics (Table 2).

Different from the extant airline research where the probability of bankruptcy was modeled as the outcome variable, we use profitability itself as the outcome variable. The reason is that the landscape of airline filing for bankruptcy has changed in recent years. First, the number of U.S. airlines filing for bankruptcy has dramatically decreased in recent years. Among those airlines that are required to report to DOT on a regular basis (with at least a 1% domestic market share), only three filed for bankruptcy between 2010 and 2023 (at the time of writing this manuscript): American Airline on November 29th, 2011; Pinnacle Airline on April 1, 2012, and ExpressJet on August 22, 2022. Second, although predicting the probability of bankruptcy is important, current airline literature completely ignores what happens after filing for bankruptcy. Reviewing all the bankruptcy filings by U.S. airlines after 2000, we find that the results of

filing for bankruptcy are predominantly positive instead of negative. For example, Frontier and Hawaiian airlines emerged from bankruptcy as the top performing airlines in terms of profitability; American Airline, Delta Airline, United Airline all successfully used bankruptcy to merger with other carriers and transformed themselves into the "Big Four" players (Peterson and Daily 2011); Pinnacle Airline successfully restructured itself as Endeavor Airline while ExpressJet, the most recent case, is planning for a strong comeback as a charter operator in the second half of 2024 (Birns 2023). Third, with 80% of the market share in the U.S. airline industry now controlled by the Big Four (American, Delta, United, and Southwest) (Semuels 2023), the chance of filing for bankruptcy from the Big Four has become significantly slimmer. Therefore, using operational variables to directly predict airline's profitability may provide more meaningful insights for the four major players as well as for the 20% remaining players on the market.

For the profitability measures, we further elect to use operating profit divided by operating revenue (OPOR) instead of the raw profitability numbers. First, profitability varies across years and is sometimes negative. When natural logarithms are calculated, those negative profitability values become missing data points, which is not a true reflection of airline financial status. Second, the excessive variance of profitability results from the different sizes of carriers. Ratio measures like OPOR, in this case, can better account for the size differences among carriers in comparison to other financial measures (Dresner and Xu 1995). In addition, ratio measures can also overcome the difficulty in measures associated with carriers that own aircrafts versus carriers that lease aircrafts (Mellat-Parast et al. 2015, Tsikriktsis 2007).

Table 2 summarizes all variables used in the current study, definition of variables, and data source. Financial measures, such as profitability and revenue, are reported to DOT quarterly while other measures are reported to DOT either monthly or quarterly. For those measures reported monthly, we aggregate them to a quarterly format to match OPOR. Supplemental R code details how each variable was calculated.

Table 2 Variables Used in Analysis

Variable	Formula	Data Source
OPOR	Operating profit divided by operating revenue at quarterly level	DOT Schedule P1.2
Operations Metrics		
Load Factor	Quarterly revenue passenger miles divided by available seat miles	DOT Schedule T1
Fleet Utilization	Block Aircraft Hours divided by Aircraft Days for each carrier	DOT Schedule P52
Fleet Heterogeneity	Blau index of different aircraft types within an airline's fleet in each quarter	DOT Schedule T100
Network Sparsity	Sum of squared proportions of flights originating from each airport in an airline's network in each quarter	DOT Schedule T100
Efficiency Metrics		
Fuel Efficiency	Available seat miles/gallons of fuel consumed in each quarter	DOT Schedule T1 and P12(a)
Yield	Passenger revenues divided by revenue passenger miles (RPMs) for each carrier	DOT Schedule T1 and P1.2
Average Landing Fee	Total landing fees/Total number of flights in each quarter for each carrier	DOT Schedule P6 and P52
Fuel Cost	Quarterly fuel cost for each carrier	DOT Schedule P12(a)
Service Quality Metrics		
On-time Performance	Percent of flights arriving less than 15 minutes within the scheduled arrival time	DOT Air Travel Monthly Consumer Report
Mishandled Bags	Total number of mishandled bags for each carrier in each quarter	DOT Air Travel Monthly Consumer Report
Total Delay	Total number of flights delays of each carriers in each quarter	DOT Air Travel Monthly Consumer Report
Total Consumer Complaints	Total number of consumer complaints filed to DOT for each carrier in each quarter	DOT Air Travel Monthly Consumer Report
Human Metrics		
Enplaned Passengers	Quarterly number of enplaned passengers for each carrier	DOT Air Travel Monthly Consumer Report
Full Time Employees	Quarterly number of full-time employees for each carrier	DOT Schedule P1(a)
Market Power Metrics		
Market Share	The ratio of a carrier's quarterly revenue passenger miles to the sum of revenue passenger miles of the total carriers in that quarter	DOT Schedule T1

4. METHODOLOGY

Fixed effects models are the most frequently used models in relevant airline literature to predict airline financial performance (Alan and Lapré 2018, Lu et al. 2015, Phillips and Sertsios 2013). Therefore, we first specify our model in fixed effects. Given the pitfalls of fixed effects models, recent econometrics research proposes mixed effects models that will better solve the issues associated with time-series cross-sectional data in the modeling process. Therefore, we also specify our model using mixed effects models. Lastly, both XGB and DNN are known to better capture complicated nonlinear relationships in the data whereas extant airline literature has found various nonlinear relationships in the airline data (Steven et al. 2012). Accordingly, we also use XGB and DNN as alternatives to predict airline financial performance. In sum, we compare the model performances among fixed effects, mixed effects, XGB, and DNN models to provide more meaningful and pertinent managerial insights in the post-Covid era when the focus has shifted to improve airline financial performance rather than worrying about going to bankrupt (Kletzel et al. 2023, Stalnaker et al. 2023).

Fixed Effects Models

Fixed effects models are adopted by the majority of airline research that predict airline financial performance. Hence, we first specify our models in fixed effects using Equation 2 as our benchmark model. y_{ij} is the respective OPOR at occasion i for each carrier j. x_{ij} is a vector of predictors from Table 2. D_j represents carrier dummy to control for carrier-specific effects. In addition, we include time dummy following existing airline research to control for time-specific effects (Alan and Lapré 2018).

Mixed Effects Models (Bell and Jones 2015)

Given the pitfalls of fixed effects models (Section 2.2), we next specify our model using mixed effects proposed by Bell and Jones (2015) as in Equation 3. Z_j represents time-invariant variables measured at the carrier j level as these variables do not change or rarely change over time, such as the status of being a legacy carrier. For each time-varying variable, we add a group mean-centered variable $\overline{x}_j$ to model the time-invariant effect of that time-varying variable, leaving the term $x_{ij} - \overline{x}$ only measuring the time-varying effect of that time-varying variable. As discussed in Section 2.2, mixed effects models can solve the issues of incorrect standard error as well as endogeneity where $\text{Cov}(x_{ij}, \mu_j) \neq 0$ while still being able to model the higher-level time-varying effects.

Following Bell and Jones (2015), we also calculate Intra Class Correlation (ICC) by fitting a random intercept model. This was conducted to confirm that the data is indeed longitudinal in nature as well as to ensure the variations we observe over the 16 years are not random fluctuations but indeed "meaningful individual differences" (Bliese and Ployhart 2002, p. 368). ICC results are summarized in Table 3 with the ICC being 0.61, indicating that 61% of the variation in financial performance (OPOR) is time-invariant between carriers and the remaining 39% of the variation is time-varying within carriers. The

ICC strongly indicates the nature of longitudinal data, validating the necessity to use $\overline{x}_j$ to model the time-invariant effect while the term $x_{ij} - \overline{x}$ to model the time-varying effect.

Table 3 ICC Calculation

		Parameter	OPOR
Fixed Effects			
	Intercept	β_0	0.01
			(0.13)
Random Effects			
	Level 2: Carriers	σ^2_{u0}	0.0228
	Level 1: Occasion	σ^2_{e0}	0.0147
Measures of Fit			
	Log Likelihood		569.17
	ICC		0.61

*Notes: † = p < 0.10; * = p < 0.05; ** = p < 0.01 (two-tailed).*
Z-tests are reported in parentheses for the fixed effects parameters.

Deep Neural Networks

Although DNN has been considered a "black box" where the estimation results are challenging to interpret, its ability to accurately predict outcomes (Leshno et al. 1993) fits perfectly in our research setting. Following extant DNN literature, we specify our DNN model in the following approaches.

For the nonlinear activation function that plugs into each neuron in each hidden layer, we use the exponential linear unit (ELU) activation function proposed by Clevert et al. (2015) in Equation 5. ELU is a smoothed version of the popular rectified linear unit (RELU) activation function and performs faster and yield more accurate predictions (Clevert et al. 2015).

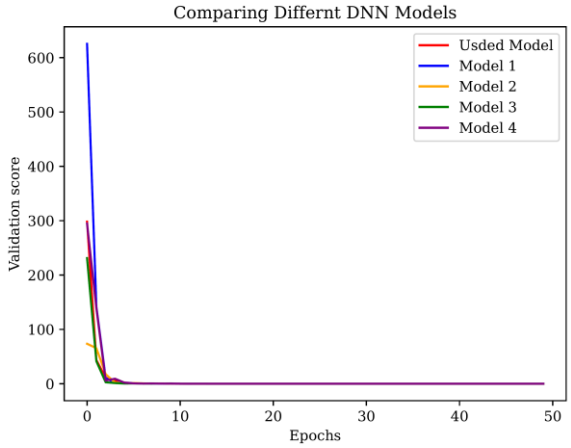
$$n_{1,u} = \begin{cases} e^{M_{1,u}} - 1, \ if M_{1,u} < 0 \\ M_{1,u}, \ if \ M_{1,u} \geq 0 \end{cases} \quad \text{(Equation 5)}$$

A DNN needs to compute different weights, i.e., the parameters, for different neurons at different layers. If the number of input variables are large, the proliferation of parameters presents a computational challenge to compute a full set of optimal weights for each neuron throughout the network. As such, LeCun et al. (2012) and Ruder (2016) proposed various methods to train DNN. We follow their suggestions to train our neural network. First, we search for weights that minimize the sum of squared errors $\sum_{j=1}^{J}\sum_{i=1}^{N}(\hat{y}_{ij} - y_{ij})$ using Adam (adaptive moment estimation) – a stochastic gradient descent proposed by Kingma and Ba (2014) and has become popular in the machine learning field ever since. Second, for the starting value of the weights, we elect to use the approach proposed by He et al. (2015), i.e., the starting values were drawn from a truncated normal distribution with mean = 0 and standard deviation = $\sqrt{2/I}$ (I is the number of inputs in the input layer). Any draws more than two standard deviations from the mean will be discarded and redrawn. He et al.'s (2015) approach helps to achieve two things in DNN model training: 1) the scale of the input variance to each neuron is constant; 2) avoid the gradients for the weights, especially the weights of the lower layers, from vanishing or exploding which will consequently slow the DNN's learning speed. Third, we use batch normalization proposed by Ioffe and Szegedy (2015) to improve model

performance. As weights are calculated for each layer and then passed on to the next layer as inputs throughout the network, a small change in the early layers may be magnified in the later layers, causing computational problems. Batch normalization helps to reduce these computational challenges by zero-centering and normalizing these weights before feeding them into each neuron as the inputs. Lastly, a common issue in DNN is overfitting due to the large number of parameters. Hinton et al. (2012) and Li et al. (2019) proposed the technique of "dropout" to mitigate overfitting by giving each neuron a certain probability that once assigned with this probably, this neuron will not be used at a given iteration. We use the probability of 0.2 in our study. As we also adopt batch normalization in our model training, we only apply dropout to the last hidden layer as was suggested by Li et al. (2019), i.e., when dropout and batch normalization are used simultaneously, dropout should be applied to the last hidden layer.

As DNN does not arrive at a single solution, we use Adam search algorithm to find the optimal weights and stop the algorithm when no further improvements in model fit can be achieved after certain iterations. In addition, we also test different model specifications to ensure the best possible DNN results. We first change the activation function to the popular rectified linear unit (RELU) in alternative Model 1; we then increase the dropout rate from 0.2 to 0.5 in Model 2; we further increase the hidden layers from 10 to 15 with a 0.5 dropout rate in Model 3; lastly, we increase the hidden layers from 10 to 11 with a 0.5 dropout rate in Model 4. We report the model comparison results in Figure 3 where no single model performs significantly better than the other. Finally, we elect to use the DNN model with 10 hidden layers and 0.2 dropout rate for the rest of our analysis. We train all our DNN models using Google's Tensor-flow package in Python and utilizes Nvidia CUDA® architectures to accelerate model training process.

Figure 3 Comparing Different DNN Models



5. MODEL COMPARISON RESULTS

We compare the fit of the four competing models using the average mean squared prediction error (AMSE) $\sum_{j=1}^{M}\sum_{i=1}^{N}(\hat{y}_{ij} - y_{ij})^2$, where y_{ij} is individual airline's financial performance. We first compare AMSEs in randomly generated samples using Monte Carlo cross-validation. Then, we compare the predictive

performance of the four models by using the last eight quarters of observations as the hold-out set to mimic the "future" to be predicted.

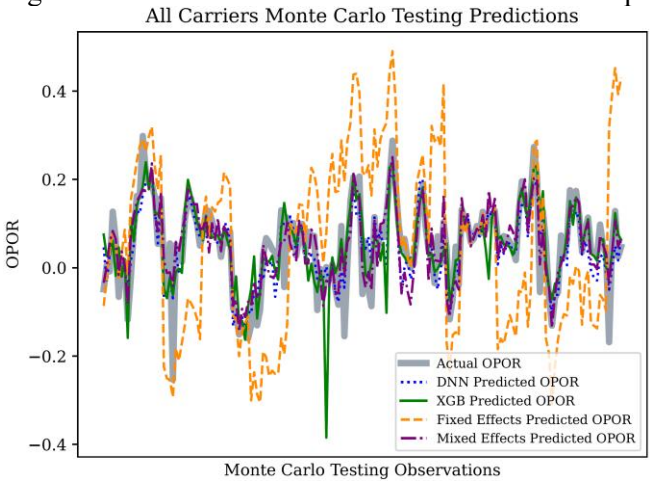
5.1 MONTE CARLO CROSS-VALIDATION RESULTS

Here we compare the out-of-sample fit of the four competing models in a Monte Carlo cross-validation process. We randomly partition our data into 80% in-sample (training) and 20% out-of-sample (testing). To make sure that our Monte Carlo process is representative of all airlines, we stratify our training and testing data to ensure that both training and testing data contain observations from all airlines. This enables us to estimate a complete set of airline fixed effects. As all models include airline and quarter fixed effects, all models are being compared on how well they predict idiosyncratic fluctuations in airline financial performance at the *i,j* level. To avoid the random stratification algorithm accidentally select easy or hard-to-fit samples, we randomly stratify training and testing datasets multiple times, test all four models on these multiple samples, and report the averaged results from all iterations. Table 4 present both in-sample and out-of-sample fit results for all four models. Figure 4 only reports the out-of-sample prediction performance of the four models.

Table 4 Monte Carlo Cross Validation – All Carriers

All Carriers	**In-sample**			**Out-of-sample**		
Estimator	N	MSE	R^2	N	MSE	R^2
Fixed Effects	692	0.0079	34.18%	173	0.0078	20.48%
Mixed Effects	692	0.0047	56.09%	173	0.0006	74.96%
DNN	692	0.0036	66.21%	173	0.0025	72.21%
XGB	692	0.000001	99.99%	173	0.0039	56.45%

Figure 4 Monte Carlo Cross Validation – Out of Sample (All Carriers)



Fixed effects models, by controlling for carriers and quarter fixed effects, give nearly identical fits (as measured by MSE) both in-sample (0.0079) and out-of-sample (0.0078). Mixed effects models, which allow for carrier fixed effects in intercepts and slopes, yield a substantial improvement in the in-sample MSE (0.0047 compared with 0.0079 in fixed effects). In addition, the out-of-sample MSE of mixed effects models is 0.0006, which is a dramatic improvement over fixed effects (0.0078), showing support for Bell

and Jones (2015) in that mixed effects models, by solving the various modeling pitfalls, can more accurately predict individual higher-level units. Turning to machine learning models, the in-sample MSE of DNN is 0.0036, a further improvement over the previous two econometrics models. The out-of-sample fit of DNN is 0.0025, slightly better than its in-sample fit of 0.0036, but significantly worse than the out-of-sample fit of mixed effects models of 0.0006. XGB generates a near perfect in-sample fit with an MSE of 0.000001. However, XGB suffers from a serious over-fitting issue: its out-of-sample MSE is only 0.0039, an astonishing 3800% reduction compared to its in-sample fit. In addition, the out-of-sample fit of XGB (0.0039) is also worse than the out-of-sample fit of both DNN (0.0025) and mixed effects models (0.0006).

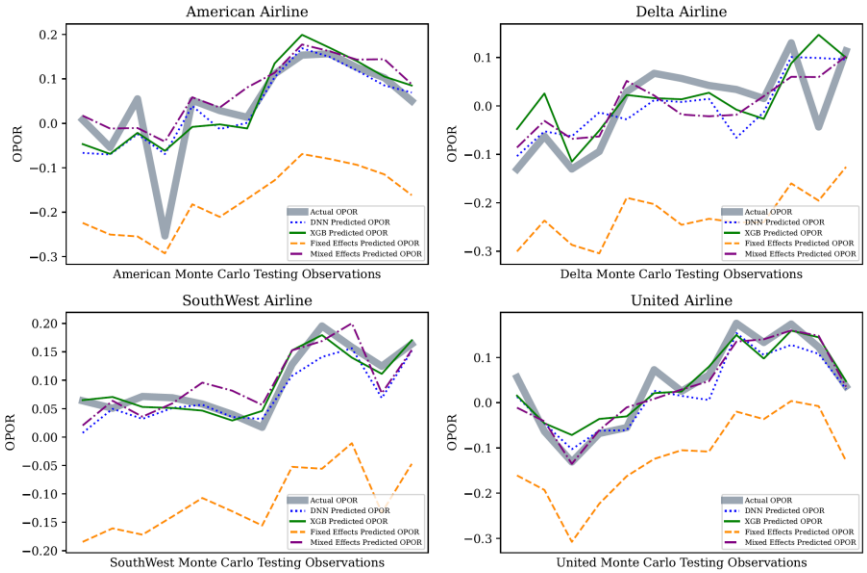
To summarize our Monte Carlo cross-validation results so far, the ranking of models in terms of their out-of-sample prediction accuracy (measured by MSE) is: 1) mixed effects models, 2) DNN, 3) XGB, and 4) fixed effects models. While fixed effects models are considered the gold standard for modelling panel data (Schurer and Yong 2012), we show fixed effects models yield the worst prediction accuracy in our analysis so far.

An important approach to assess model performance is to consider model fit in non-random subsamples. If a model fits poorly in certain subsamples, it may reveal dimensions in which this model misses to capture or mis-specifies. As the Big Four carriers account for more than 80% of the U.S. domestic market share, we elect to use these big four carriers as our non-random subsamples to test the model performance of the four competing models again. The results are summarized in Table 5. Figure 5 plots the out-of-sample prediction performance for the Big Four. We see that XGB still produces the best in-sample fit with a MSE of 0.000001 for all the four carriers. However, the out-of-sample fit of XGB is similar to mixed effects models and DNN, again indicating the overfitting issue with XGB. Similar to our all-carriers sample in Table 4, fixed effects models still give the worst prediction accuracy for the Big Four carriers. Unlike all-carriers sample in Table 4 where mixed effects model is the clear winner, for the big four carriers, mixed effects, DNN, and XGB are almost equal or similar in terms of their prediction accuracy.

Table 5 Monte Carlo Cross Validation – Big Four Carriers

Big Four Carriers	**In-sample MSE**				**Out-of-sample MSE**			
Estimator	American	Delta	SouthWest	United	American	Delta	SouthWest	United
Fixed Effects	0.0029	0.0127	0.0016	0.0026	0.0491	0.0555	0.0443	0.0273
Mixed Effects	0.0019	0.0033	0.0014	0.0028	0.0046	0.0029	0.0011	0.0011
DNN	0.0022	0.0021	0.0011	0.0019	0.0027	0.0019	0.0013	0.0021
XGB	0.000001	0.000001	0.000001	0.000001	0.0024	0.0019	0.0011	0.0011

Figure 5 Monte Carlo Cross Validation – Out of Sample (Big Four Carriers)



5.2 PREDICTING AIRLINE FINANCIAL PERFORMANCE

In this section, we compare the out-of-sample fits of the four competing models using observations of last eight quarters of each carriers as a holdout sample. We use the holdout sample to mimic the future quarters to be predicted. We train our models using data excluding the last eight quarters and then use the trained model to forecast airline financial performance for the last eight quarters. To predict financial performance for the holdout sample for fixed effects and mixed effects models, we need to first forecast future values of time effects that capture idiosyncratic carrier changes to the future.

For the fixed effects model, we forecast the future values of quarter effects in the intercepts in Equation 2. We obtain the forecasts of time effects using AR(1) models by including lagged quarter effects along with time trends. To prevent exponential growth in levels of our dependent variable when forecasting, we elect to use two time variables *log(t)* and $log(t)^{1/2}$ to capture time trends following econometrics literature (Wooldridge 2010). To forecast time effects for mixed effects model, we use VAR(1) system using a vector of lagged time effects along with the two time variables of *log(t)* and $log(t)^{1/2}$ for Equation 3. DNN generates its own forecast of future time effects by allowing the quarterly inputs evolve over time, which is a strong advantage to use DNN models. However, as was mentioned by the literature, this is also a "black box", a weakness of DNN models as it is hard to interpret what the time trends actually captures.

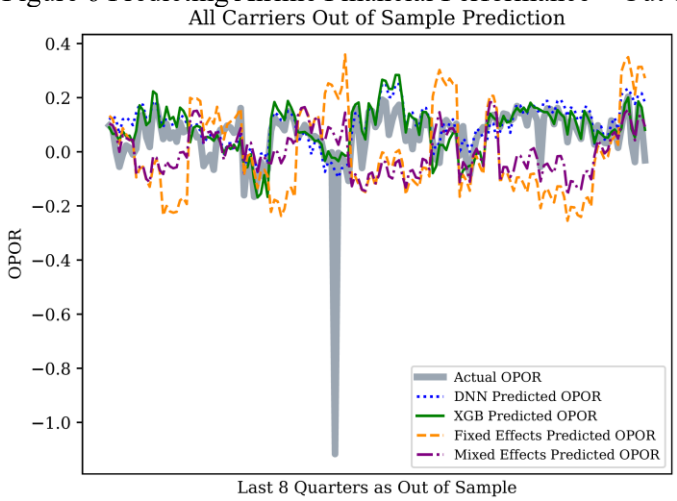
We first report the results of the out-of-sample fits for all carriers in Table 6 and Figure 6. Strikingly different from Monte Carlo cross-validation process where mixed effect models yield the best out-of-sample prediction accuracy, the two machine learning models now have the best (tied) out-of-sample prediction accuracy. Mixed effects models rank the second and fixed effects models again yield the worst prediction accuracy. A consistent finding is that XGB still has a serious over-fitting issue, with a near-

perfect in-sample MSE of 0.000001 while its out-of-sample MSE is the same as DNN at 0.0111. In sum, the two machine learning models outperform mixed effects models by 56% and outperform fixed effects models by 77% in terms of out-of-sample prediction accuracy.

Table 6 Predicting Airline Financial Performance – All Carriers

All Carriers	**In Sample**			**Out of Sample**		
	N	MSE	R^2	N	MSE	R^2
Fixed Effects	705	0.0096	35.86%	160	0.0494	2.62%
Mixed Effects	705	0.0032	67.19%	160	0.0252	10.28%
DNN	705	0.0022	76.65%	160	0.0111	19.88%
XGB	705	0.000001	99.98%	160	0.0111	20.49%

Figure 6 Predicting Airline Financial Performance – Out of Sample (All Carriers)

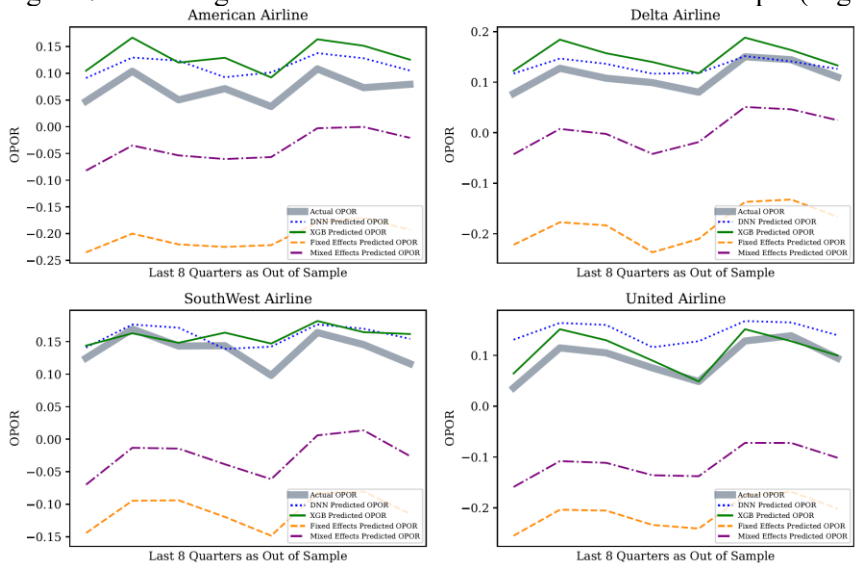


Next, we report the out-of-sample fits for non-random subsamples of the Big Four carriers in Table 7 and Figure 7. Turning to the out-of-sample MSE, we see that the findings are consistent with the findings from all carriers sample in Table 6. In terms of MSE, the two machine learning models still produce the best forecasting accuracy: DNN has the best prediction accuracy for American Airline (0.0021), Delta Airline (0.0006), and Southwest Airline (0.0007, tied with XGB) while XGB has the best out-of-sample prediction accuracy for United Airline (0.0005). Mixed effects models rank the third for all four carriers while fixed effects models again yield the worst prediction accuracy. In sum, for the Big Four carriers, the two machine learning models, on average, outperform mixed effects models by 93% and outperform fixed effect models by 98% in terms of out-of-sample prediction accuracy.

Table 7 Predicting Airline Financial Performance – Big Four Carriers

Big Four Carriers	**In Sample (MSE)**				**Out of Sample (MSE)**			
Estimator	American	Delta	SouthWest	United	American	Delta	SouthWest	United
Fixed Effects	0.0103	0.0233	0.0027	0.0081	0.0769	0.0875	0.0616	0.0919
Mixed Effects	0.0026	0.0047	0.0015	0.0026	0.0126	0.0122	0.0271	0.0420
DNN	0.0022	0.0015	0.0008	0.0014	0.0021	0.0006	0.0007	0.0033
XGB	0.000001	0.0000008	0.000001	0.0000008	0.0037	0.0016	0.0007	0.0005

Figure 7 Predicting Airline Financial Performance – Out of Sample (Big Four Carriers)



Combining the results from all-carriers analysis and the Big-Four analysis, our conclusion is that to predict airline financial performance, machine learning models, either DNN or XGB, is a better choice as both of these two models outperform classic econometrics models by a large extent when measured by the out-of-sample MSE.

6. WHY DO MACHINE LEARNING MODELS PERFORM BETTER?

We see from the previous section that the two machine learning models are superior to econometrics models in terms of prediction accuracy in both analyses. Given the extant findings of the nonlinear relationships in airline literature (Steven et al. 2012) and given machine learning models are known for better capturing the complicated nonlinear high-dimensional relationships, we now test if indeed there exist any nonlinear relationships in our data, hoping to provide a possible explanation for the differences in prediction accuracy between machine learning models and econometrics models. In other words, we test if the superior prediction accuracy of machine learning models may indeed be partially explained by the nonlinear relationships in our data.

Note that we have 15 operational predictors (Table 2). Testing nonlinear relationships between each single predictor and airline financial performance may be a too random practice. Accordingly, we leverage current operations management theory and literature to help us narrow down the potential candidates. Schmenner and Swink (1998) developed the Theory of Performance Frontier, defined as "the maximum performance that can be achieved by a manufacturing unit given a set of operating choices" (p. 108). A performance frontier is made up of an operating frontier (i.e., frontiers formed by choices in plant operations) and an asset frontier (i.e., frontiers formed by choices in plant design and investment). The concept of Performance Frontier was accordingly applied in airline literature, such as the design load factor can be viewed as the asset frontier and the actual/effective load factor as the operating frontier (Lapré and Scudder

2004). Schmenner and Swink (1998) proposed that if a firm operates close to its asset frontier, the firm will be likely to operate under the law of trade-offs where "no single plant can provide superior performance in all dimensions simultaneously" (p. 110). But if a firm operates away from its asset frontier, the firm will be likely to operate under the law of cumulative capabilities and it may simultaneously improve its performance in several dimensions.

Applying this concept to load factor, carriers with higher operating load factors can be viewed as operating closer to their asset frontiers (the design capacity) while carriers with low operating load factors can be viewed as operating away from their asset frontiers. Based on the Performance Frontier Theory, if a carrier operates at a high load factor, it will operate under the law of trade-offs, i.e., it is challenging to increase both its load factor and financial performance simultaneously. In addition, this carrier has to spend "more and more resources … in order to achieve each additional increment of benefit" (Schmenner and Swink 1998, p. 110) to further improve its financial performance while increasing its load factor. At some point, the resources spent may eventually outweigh the synergistic benefits garnered from these resources. From that moment onwards, increasing load factor will only dampen its financial performance. However, if a carrier operates away from its asset frontier at a lower load factor, it will operate under the law of cumulative capabilities. In this case, this carrier will be able to improve both dimensions simultaneously, i.e., increasing load factor will also increase its financial performance. Therefore, the effect of increasing load factor on carrier's financial performance should demonstrate an inverted U-shaped relationship. A similar inverted U-shaped relationship can also be argued between yield and airline financial performance. Therefore, we test the nonlinear relationships for load factor and yield in the current section just as two examples to illustrate the potential causes of the superior performance of machine learning models.

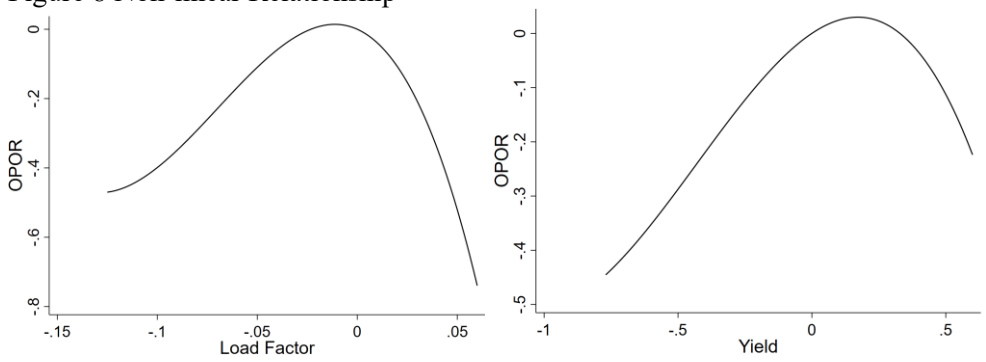
We specify our econometrics models using Equation 3 by adding a squared term and a cubic term for both load factor and yield in our model. Table 8 reports the statistical results while Figure 8 plots the relationships between load factor, yield and OPOR. In Table 8, all the coefficients for the three terms of load factor and yield are statistically significant, indicating an inverted U-shaped relationship. Therefore, we can partially draw the conclusion that the superior performance of the machine learning models may be indeed due to the high-dimensional nonlinear relationships in our data. Note that we did not add the squared terms and cubic terms to the two econometrics models to re-conduct all the tests as our research purpose is to compare the prediction accuracy among the four models using the raw operational metrics as they are.

Table 8 Testing Non-linear Relationships

Variables	Parameter	Load Factor Model		Yield Model	
Fixed Effect					
Intercept	β_0	-43.13***	(-3.28)	-29.27**	(-2.29)
Load Factor	β_{1_1}	**-2.59*** **	(-2.82)		
Load Factor^2	β_{2_1}	**-126.27*** **	(-2.65)		
Load Factor^3	β_{3_1}	**-601.98*** **	(-2.66)		
Yield	β_{1_2}			**0.33***	(1.88)
Yield^2	β_{2_2}			**-0.80***	(-1.79)
Yield^3	β_{3_2}			**-0.62*** **	(-2.59)
Fleet Utilization	β_4	0.03	(0.18)	-0.48**	(-2.07)
Fuel Efficiency	β_5	0.66***	(3.98)	1.05**	(2.14)
On-time Performance	β_6	-2.66**	(-2.43)	-2.71*	(-1.76)
Miss-handled Bags	β_7	0.05	(1.01)	-0.03	(-0.58)
Total Delay	β_8	-0.89***	(-2.74)	-0.44**	(-1.83)
No. of Complaints	β_9	0.05*	(1.90)	0.06	(1.13)
Fleet Heterogeneity	β_{10}	-0.19**	(-2.14)	0.20*	(1.76)
Avg Landing Fee	β_{11}	0.50***	(3.03)	0.35**	(2.11)
Network Sparsity	β_{12}	-0.07	(-1.42)	-0.02	(-0.53)
Market Share	β_{13}	-2.29***	(-3.13)	-1.41**	(-2.27)
Fuel Cost	β_{14}	0.27**	(2.16)	0.74**	(2.33)
No. Enplaned Passenger	β_{15}	2.09***	(2.90)	0.98**	(2.17)
Full Time Employee	β_{16}	0.46**	(2.02)	-0.05	(-0.77)
(Year and Quarter Fixed Effects Included)					
Random Effect					
Level 2: Carriers	σ^2_{u0}	0.0028		0.0025	
Level 1: Occasion	σ^2_{e0}	0.0041		0.0042	
Measures of Fit					
Log Likelihood		1136.78		11122.46	
Total R^2		58.30%		58.27%	

*Notes: * = p < 0.10; ** = p < 0.05; *** = p < 0.01 (two-tailed).*
Z-tests are reported in parentheses for the fixed effects parameters.

Figure 8 Non-linear Relationship



7. CONTRIBUTION AND CONCLUSION

Our study is distinctive from the extant airline literature that used operational metrics to predict airline financial performance. First, we do not use operational metrics to predict the probability of bankruptcy. Instead, we use operational metrics to predict the actual financials reported to DOT given the most recent financial landscape in the U.S. airline industry – only three major airlines (with more than 1% domestic market share) filed for bankruptcy between 2010 and 2023. Our study, therefore, is complimentary to the extant airline literature by extending this stream of literature from predicting probability of bankruptcy to predicting actual financial performance. We contribute to knowledge accumulation in this stream of literature. Second, the related airline literature only adopted fixed effects models to predict airline financial performance without testing the performance of other competing models. In contrast, our research compares

the predictive accuracy of four competing models: two econometrics models and two machine learning models. To this end, we expanded the scope of the current research stream from using one single model to comparing different models. Third, in contrast to extant airline literature where only econometrics models were used to predict airline financial performance, we introduced two machine learning models to forecast airline financial performance. Our results show that machine learning models demonstrate superior prediction accuracy when predicting future financial performance in both all sample analysis and in nonrandom subsample analysis. In all sample analysis, the two machine learning models outperform mixed effects models by 56% and outperform fixed effects models by 77% in terms of out-of-sample prediction accuracy. Our results reinforce the potential of adopting machine learning models to better solve business problems in the airline industry. In addition, despite almost all related airline research adopted fixed effects models to predict bankruptcy, our results show that fixed effects models yield the worst prediction accuracy. We advise researchers to use fixed effects models with caution in similar research settings. Fourth, among the related literature we surveyed, only one research compared in-sample and out-of-sample prediction accuracy (Alan and Lapré 2018) while all other related research (Lu et al. 2015, Phillips and Sertsios 2013, Ciliberto and Schenone 2012, Gudmundsson 2004; 2002) only modeled in-sample prediction accuracy. Our results show that modeling in-sample performance only may not present a holistic picture and may not provide meaningful guidance for practitioners. For example, turning to American Airline in Table 7, the in-sample MSE for fixed effects models is 0.0103 while the out-of-sample MSE for fixed effects models is 0.0769 – meaning the out-of-sample performance is seven times worse than its in-sample performance; the in-sample MSE for XGB is a near-perfect 0.000001 while the out-of-sample MSE for XGB is 0.0037 – a staggering 3700 times worse out-of-sample performance compared with its in-sample performance. Hence, modeling in-sample performance only without benchmarking out-of-sample forecasting is not only misleading but also statistically flawed. We call for researchers to conduct both in-sample and out-of-sample comparisons in similar operations management applications.

Our research also provides meaningful and important guidelines for airline practitioners, especially in the post-Covid era when airlines are focusing on improving their bottom-line financial performance (Kletzel et al. 2023, Stalnaker et al. 2023). We see two fundamental applications of our research in the airline industry. First, the excellent out-of-sample prediction accuracy can provide airline practitioners with an effective tool to predict individual airline's future financial performance. For example, for the Big Four carriers, the best out-of-sample forecast accuracy in terms of MSE is: American Airline 0.0021, Delta Airline 0.0006, Southwest 0.0007, and United Airline 0.0005 (Table 7 and Figure 7). These forecasts are incredibly close to the actual airline financial performance and can be used as a reliable tool by these Big Four carriers to predict their future financials using their operational metrics. Second, we include 15 different operational predictors covering five different categories in airline operations. Airlines have a wide

variety of choices to mix and match how changes in these 15 variables will impact their financials. For example, an individual airline may plan to purchase new fleets and restructure its whole network. Accordingly, its market share, number of enplaned passengers, load factor, yield, and other factors are also expected to change. Given the changes in these factors, what will be the impact to its financials? The two machine learning models can then be readily applied to reliably forecast future financials given these changes in operational metrics. By introducing XGB and DNN models into airline operations applications, we show that machine learning models, compared with econometric models, can better forecast and simulate such business cases to a larger extent.

Financial distress or bankruptcy studies are popular in the finance field. Influenced by finance studies, the airline research we survey also studied financial distress and bankruptcy (Alan and Lapré 2018, Lu et al. 2015, Phillips and Sertsios 2013, Gudmundsson 2004; 2002). However, from the time DOT started reporting airline financial performance in 1988 up to 2023, there are only 21 bankruptcy filings in these 36 years. Moreover, in the most recent years from 2010 to 2023, there are only 3 bankruptcy filings in DOT data. The number of bankruptcy filings are extremely small compared with other industries (Alan and Lapré 2018). Therefore, we elect to use four competing models to predict actual airline financial performance instead of bankruptcies, aiming to provide more meaningful managerial insights and tools for decision makers to forecast financial performance. In doing so, we overcome the pitfalls of previous airline bankruptcy studies by 1) comparing model performance between econometrics models and machine learning models whereas only econometrics models were used in previous airline bankruptcy studies; 2) showing that fixed effects models, used by previous airline bankruptcy research, produce the worst out-of-sample forecast accuracy; 3) introducing DNN and XGB into airline bankruptcy research and demonstrating the superior out-of-sample forecast accuracy of these two machine learning models. We call for future research to continue investigating the potential of applying machine learning models in both airline industry and in other industries to solve operations cases. We also call for researchers to adopt fixed effects models with caution as our result indicate that fixed effects models yield the worst in-sample fit and out-of-sample fit across all analyses.

REFERENCES:

Alan Y, Lapré MA (2018) Investigating operational predictors of future financial distress in the US airline industry. *Production Oper. Management* 27(4):734-755.

Allison PD (2009) Fixed Effects Regression Models. London: Sage.

Bafumi J, Gelman A (2006) Fitting Multilevel Models when Predictors and Group Effects Correlate. Paper presented at the *Annual Meeting of the Midwest Political Science Association*. Chicago, IL.

Beck N, Katz JN (1995) What to do (and not to do) with timeseries cross-section data. *Am. Polit. Sci. Rev*. 89(3): 634–647.

Beck N, Katz JN (2011) Modeling dynamics in time-series cross-section political economy data. *Ann. Rev. Polit. Sci*. 14:331–352.

Bell A, Jones K (2015) Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Polit. Sci. Res. Methods* 3(1):133-153.

Birns H (2023) US's ExpressJet plans restart as charter carrier using B777s. https://www.ch-aviation.com/portal/news/129952-uss-expressjet-plans-restart-as-charter-carrier-using-b777s.

Bliese PD, Ployhart RE (2002) Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organ. Res. Methods* 5(4):362-387.

Buckley J (2023) How the pandemic killed off 64 airlines. https://www.cnn.com/travel/article/pandemic-airline-bankruptcies/index.html

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*:785-794.

Clevert DA, Unterthiner T, Hochreiter S (2015) Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Dresner M, Xu K (1995) Customer service, customer satisfaction, and corporate performance. *J. Bus. Logistics* 16(1):23-40.

Gabel S, Timoshenko A (2022) Product choice with large assortments: A scalable deep-learning model. *Management Sci*. 68(3):1808-1827.

Gudmundsson SV (2002) Airline distress prediction using nonfinancial indicators. *J. Air Transport.* 7(2): 3–24.

Gudmundsson SV (2004) Management emphasis and performance in the airline industry: An exploratory multilevel analysis. *Transport. Res. Part E Logistics Transport. Rev*. 40(6):443–463.

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*:1026-1034.

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*:448-456.

Kingma DP, Ba, J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kletzel J, Stillman A, Wysong R (2023) How can airlines return to profitability? By following these 5 steps. https://www.pwc.com/us/en/industries/consumer-markets/library/how-can-airlines-return-to-profitability.html.

Korn J (2022) ExpressJet Airlines, formerly under United, files for bankruptcy. https://www.cnn.com/2022/08/23/tech/expressjet-bankruptcy/.

Krishnakumar J (2006) Time Invariant Variables and Panel Data Models: A Generalised Frisch-Waugh Theorem and its Implications. In *Panel Data Econometrics: Theoretical Contributions and Empirical Applications, edited by Badi H. Baltagi*:119–32. Amsterdam: Elsevier.

Lapré MA, Scudder GD (2004) Performance improvement paths in the US airline industry: Linking trade-offs to asset frontiers. *Production Oper. Management* 13(2):123-134.

LeCun YL, Bottou GO, Muller K (2012). Efficient backprop. In *G. Montavon, G. Orr and K. Muller (Eds.), Neural Networks: Tricks of the Trade, Vol. 7700 of Lecture Notes in Computer Science.* Springer, Berlin:9–48.

Leshno M, Lin VY, Pinkus A, Schocken S (1993) Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks* 6(6):861-867.

Li X, Chen S, Hu X, Yang J (2019) Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*:2682-2690.

Lu C, Yang AS, Huang J (2015) Bankruptcy predictions for US air carrier operations: A study of financial data. *J. Econ. Finance* 39(3):574–589.

Mellat-Parast M, Golmohammadi D, McFadden KL, Miller JW (2015) Linking business strategy to service failures and financial performance: Empirical evidence from the US domestic airline industry. *J. Oper. Management* 38:14-24.

Moulton BR (1986) Random Group Effects and the Precision of Regression Estimates. *J. Econom*. 32(3):385–97.

Palta M, Chris S (2003) Causes, Problems and Benefits of Different between and within Effects in the Analysis of Clustered Data. *Health Serv. Outcomes Res. Methodol*. 3:177–193.

Peterson K, Daily M (2011) American Airlines files for bankruptcy. https://www.reuters.com/article/idUSTRE7AS0T7/.

Phillips G, Sertsios G (2013) How do firm financial conditions affect product quality and pricing? *Management Sci.* 59(8):1764–1782.

Qi M, Shi Y, Qi Y, Ma C, Yuan R, Wu D, Shen ZJ (2023) A practical end-to-end inventory management model with deep learning. *Management Sci*. *69*(2):759-773.

Raudenbush SW (2009) Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-varying Treatments in School Settings. *Educ. Fin. Policy* 4(4):468–91.

Raudenbush SW, Bryk A (2002) Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed. London: Sage.

Rubin DB1 (1980) Using Empirical Bayes Techniques in the Law-school Validity Studies. *J. Amer. Statist. Assoc.* 75(372):801–16.

Ruder S (2016) An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Schmenner RW, Swink ML (1998) On theory in operations management. *J. Oper. Management* 17(1):97-113.

Schurer S, Yong J (2012) Personality, well-being and the marginal utility of income: What can we learn from random coefficient models? Research paper.

Semuels A (2023) Blame The Airlines for the American Inequality. https://unitedafa.org/news/2023/1/16/airlines-are-terrible/

Sider A (2023) U.S. Airlines Expect Strong Demand as Travelers Find Postpandemic Routines. https://www.wsj.com/articles/u-s-airlines-expect-strong-demand-as-travelers-find-postpandemic-routines-3c64057f.

Stalnaker T, Usman K, Alport G, Buchanan A, Taylor A (2023) Airline Economic Analysis 2022-2023. https://www.oliverwyman.com/our-expertise/insights/2023/may/airline-economic-analysis-2022-2023.html.

Steven AB, Dong Y, Dresner M (2012) Linkages between customer service, customer satisfaction and performance in the airline industry: Investigation of non-linearities and moderating effects. *Transport. Res. Part E Logistics Transport. Rev.* 48(4):743-754.

Tsikriktsis N (2007) The effect of operational performance and focus on profitability: A longitudinal study of the US airline industry. *Manufacturing Service Oper. Management* 9(4):506-517.

Wooldridge JM (2010) Econometric analysis of cross section and panel data. MIT press.

Yamanouchi K (2022) ExpressJet files for Chapter 11 bankruptcy after failed Reno operation. https://www.ajc.com/news/atlanta-airport-blog/.