**Predicting U.S. Airline Financial Performance Project**

This is a reject and resubmit paper. Major tasks are to collect additional data from I/B/E/S (The Institutional Brokers’ Estimate) database in WRDS and conduct additional analysis based on editor’s and reviewer’s feedback.

**WHAT YOU HAVE**

1. Current paper. So you don’t need to start from scratch.
2. Current cleaned dataset.
3. Current Python code (maybe you do not need this but just in case).
4. Links to paper, code, and data: <https://github.com/kuangwenyi/Predicting-Airline-Financial-Performance>
5. All the feedback from editors and reviewers. How to improve the paper are clearly laid out. But may need to deepen each analysis by giving it some extra thoughts.

As the econometrics analysis in the current version will be gone based on the feedback, so only look at the following ML code if needed.

[9. XGB\_DNN\_Python.ipynb](https://github.com/kuangwenyi/Predicting-Airline-Financial-Performance/blob/main/9.%20XGB_DNN_Python.ipynb)

Data used in the current analysis is *E1\_891\_Final.xlsx.* This is a very small dataset with very few selected predictors.

*E1\_Total.xlsx* is a full list of available predictors (about 60) collected from Department of Transportation (DOT). Given the current directions suggested by editor and reviewers, *E1\_Total.xlsx* might be a better choice to use to explore the ML framework (more in 2.1).

**MAJOR TASKS**

**TASK 1: Collect additional data as the Outcome variable**

This task is related to 2.1 below. Basically, collect Earnings-per-share (EPS) data for each airline in each quarter from 2004 to 2019 (more details in the two Excel files). Two data points of EPS are needed: the actual EPS and the forecasted EPS by financial analysts. These two variables are available from I/B/E/S dataset in WRDS. Maybe EPS data is not available for all airlines in each quarter. Just use whatever data is available in IBES for as many airlines as possible.

I have lost access to WRDS as our university cancelled the subscription. If your university subscribe to WRDS, most probably IBES database is included.

**TASK 2: Improve the paper**

Following is a summary of how to improve the paper. Blue texts are comments from reviewers and editor. Black texts are my own thoughts.

* 1. **Compare machine learning models VS forecasts made by financial analysts**

Comments from editor: The main result of the paper (machine learning models outperform econometric models for prediction) is not surprising and has been explained in many textbooks. Instead of using econometric models as a benchmark, it will be interesting to see a comparison of prediction power of machine learning models to forecasts made by financial analysts. Financial analysts routinely make firm level forecasts of Earnings-per-share (EPS) of firms and these forecasts are publicly available. Analysts presumably have access to all public information which they incorporate in their forecasts. It would be interesting to see if simple machine learning models are able to outperform financial analysts.

What to do:

Use actual EPS as the new dependent variable and use available predictors from DOT to compare the predictive accuracy of ML methods and the forecasts made by financial analysts. You can try both the 16 predictors in *E1\_891\_Final.xlsx* (Table 2 in the paper) and the 60+ predictors in *E1\_Total.xlsx* to see if there is a difference. The dependent variable is now Earnings-per-share (EPS) of each airline at each quarter. Both the actual EPS and forecasted EPS by financial analysts for each airline at each quarter are available from the IBES database in WRDS. The estimates made by financial analysts are already forecasts made by human. So, this is a comparison between ML accuracy and human accuracy.

Additional thoughts on top of what was suggested by the editor:

1. Compared with ML forecasts (assuming ML performs better than financial analysts), do financial analysts tend to over-forecast or under-forecast or both?

What are the potential reasons for over-forecast and/or under-forecast by financial analysts? Analyst’s years of experience? Analyst’s firms? Not sure if IBES has this info.

1. Compared with financial analyst (assuming ML does NOT perform better than financial analysts), do ML over-forecast or under-forecast or both?

What are the potential reasons for over-forecast and/or under-forecast by ML?

1. Any other observations derived from the analysis.

The undertest and overtest analysis will be challenging when using *E1\_891\_Final.xlsx* as the number of predictors is only 16. Using *E1\_Total.xlsx* should help in this analysis. The undertest and overtest ideas can be found from Section IV in Mullainathan and Obermeyer (2022), which is a good source for similar analysis. Not exactly in the same field but may be helpful to gather ideas to deepen the analysis.

If estimates made by financial analysts are available at the individual financial analyst level, and if idiosyncratic characteristics of each financial analysts are available (such as age, years of experience, firm, and etc.), extra analysis on financial analysts can also be conducted. This is Section V of Mullainathan and Obermeyer (2022).

* 1. **Improve Machine Learning Algorithms/Methods**

**Key actions are highlighted in yellow.**

1. **Use Auto-ML framework**

Comments from reviewer: I understand that the XGB provides some bit of explanation whereas the DNNs are practically 'nonparametric' from an econometric standpoint. However, in page 9, the XGBRegressor is referred to as an ensemble, which would take away from the explanation strength of the model. My question is, if anyway using ensembles, why stop there? Why not go all the way to using, say, an **Auto-ML** framework that would auto-fit and auto-tune myriad combinations of learners, including ensembles. If pure prediction is the goal, then there's little short of DNNs that can match up to Auto-ML in flexibility. A better explanation for the choice of ML methods would help.

This feedback pretty much explained what to do: using Auto-ML and better explanation of the choice of ML methods.

1. **Interpolation versus extrapolation**

Comments from reviewer: In the regular 80-20 cross-validation procedure, the mixed effects model comes out tops but later it drops when subjected to a 'last 8 quarter' prediction test. I have to wonder why that is. Could it be that the cross-validation (CV) is essentially an interpolation exercise whereas the last 8 quarters exercise is fundamentally extrapolatory (wherein even the future values of the input variables are extrapolated?). So the prediction for the 7th (say) quarter would depend on the covariate values of quarter 7 as well as previous quarters, wouldn't it? Equations (1) - (3) do not bring out this aspect very well (seems like only contemporaneous covariates are used) whereas the text mentions AR(1) and VAR estimates. More clarity on the different nature of the CV and 8-quarter-holdout tasks would be helpful in better explaining the results.

1. **Predictions**Comments from reviewer: I would like to see prediction results from the 4 models split into 1st quarter-prediction, 2nd quarter-prediction and so on. It would help to see if the trend of results bears out as much in a one-quarter ahead prediction as when predicting 2 years out.
2. **Reporting**

Comments from reviewer: Exact numbers must be provided for how many times a procedure was run prior to its estimates being averaged. For instance, in page 15: "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."  How many is 'multiple' here?

* 1. **Improve Managerial Contributions**

Comments from reviewer 1: The study stops short of providing relevant insights or prescriptions that could be translated into meaningful managerial strategies or policy initiatives. What do we learn that is of value to firms, regulators, etc., and how can managers act upon it? For example, the focus of Ramdas, Williams, and Lipson (MSOM 2013) is to use the relationship between stock-market returns and operational performance to identify which aspects of operations to target for improvement. A very similar exercise to that of the authors, but with clear implications and insight that can be acted upon. Also, this general idea of quality being reflected in financial returns has been around a long time, e.g., Aaker and Jacobson (APA 1994), but the authors don’t cite most of the closest papers to their analysis.

Comments from reviewer 2: For managerial implications, wouldn't explanation (as opposed to pure prediction) be more useful? For instance, how is a manager to use the information that k quarters down (1 < k < 8), OPOR will be say 15% lower (or higher)? Unless managers can connect the output (OPOR) to decision inputs (Xs and Zs), how to use the predictions supplied? Some more explanation of particularly beneficial use-cases from a manager's perspective would help.

\* As the new outcome variable is EPS, so OPOR (the original outcome variable) will be replaced by EPS.

This is more like a classic sensitivity analysis in Finance: Among the 16 or 60 predictors, which predictors positively contribute to EPS and which predictors negatively contribute to EPS? For the group of positive/negative predictors, rank them based on their magnitude—similar to the “feature importance” analysis in ML. If airlines target to increase EPS by 1%, 5%, and 10% in the next eight quarters, what should airlines do with the given predictors, and etc.

* 1. **Competition Analysis**

Comments from reviewer: Inter-carrier competition appears to be completely ignored. This is now more of a general comment since I'm aware that in operations management literature unlike in (say) Marketing, competition is typically abstracted away from. While I'm not asking the authors to model inter-carrier competition, what might help (at least as a robustness test) is to incorporate past DVs and IVs of competing airlines into the covariate set of the focal airline for each observation. I would be surprised if there exists no competitive effects at all. More generally, it might capture some of the significance accruing to other time-varying covariates.

This might be an optional robustness analysis as the econometrics models are gone in the new version.

**PROJECT DELIVERY**

1. Rewrite the paper after conducting the above analysis. Take whatever is useful from the current version.

Given the econometrics analysis will be gone, now the analysis just focuses on the comparison between ML and financial analysts’ estimates, may need to come up with more ideas/figures/tables so the paper will not be too short. Maybe expand on the Auto-ML framework and etc.

1. Provide related code in any language.

**FUNDING**

Must commit to all rounds of revisions. Normally 3 rounds. 4 maximum.

1. $1200 for the initial draft.
2. $1000 for all the future rounds of revisions.

Revisions sometimes are just minor additional analysis or just explaining/defending the paper/method to reviewers. In my personal experience, 80% of the time are explaining and defending and 20% of the time are additional analysis.

How many rounds of revisions depend on the quality of the initial draft. Better initial quality, less revisions expected.

**TIMELINE**

Preferably 2 – 2.5 months from contract accepted.

**References:**

Ramdas, K., Williams, J. and Lipson, M., 2013. Can financial markets inform operational improvement efforts? Evidence from the airline industry. *Manufacturing & Service Operations Management*, *15*(3), pp.405-422.

Mullainathan, S. and Obermeyer, Z., 2022. Diagnosing physician error: A machine learning approach to low-value health care. The Quarterly Journal of Economics, 137(2), pp.679-727.