Predicting potential future trends in airline profitability

Thomas Fishwick
School of Mathematics, Computer Science and Engineering
City University
London, UK
Thomas.Fishwick@city.ac.uk

***Abstract*—In this paper we look at creating various models, such as Random Forest Regression, XGB Regression and SVR, to see if we can predict the future profitability of the airline industry. We settled upon a Support Vector Regression Markov chain model and then explored potential future scenarios with it and analyse the model's performance at predicting data over the next two years. Unsurprisingly, the model was better at predicting nearer term profits (we lacked enough data to be able to test for how many years ahead its results would be meaningful). We also looked at the total number of passengers carried per year and how the change in these numbers acts as a proxy for economic growth more generally. We can also see from the passenger numbers the distinct difference between the smaller carriers and the larger airlines.**

***Index Terms*—SVR, XGB regression, Markov chains**

I. Introduction

In this paper we will be exploring potential future trends in the American airline industry, which others (Hussain 2020) have also done. However, we will be exploring trends in airlines individually rather than the average as he did. The industry has had an interesting period with the Covid 19 pandemic 2020-to date with many countries having differing restrictions disrupting international air traffic, but we do not have the data to explore this.

Fuel and labour are currently two of the biggest expenses for the US airline industry (Swelbar & Belobaba 2019). At present, you may be able predict an airline's profit from its fuel costs, but in the future, we may have electric planes or planes that run on biofuels, so models that take labour costs into account may generalise better into the future. In the further future robotics may automate more of the labour. However, in the future environmental policies may decrease customer demand or increase the costs to airlines through increased taxes.

To predict data from time-series we have various different approaches, such as ARIMA (Autoregressive Integrated Moving Average) and XGB/Random Forest Regression. Here we will be looking at SVR (Support Vector Regression) a subset of SVMs (Support Vector Machines). SVMs aim to map "the input vectors x into a high-dimensional feature space" to create a hyperplane to discriminate between classifications (Vapnik 2000, 138). Support Vector Regression aims to bring the same process as SVMs into regression problems (Vapnik 2000, 31, 208-218, 287).

By rolling one year's predictions from SVR into the next year we can use it as a means of building a Markov chain model where "each of the conditional distributions on the right-hand side is independent of all previous observations except the most recent" (Bishop 2006, 607). By altering the some of the inputs to the model we can then see how the predicted profits change in response to that.

II. Analytical questions and data

Our analytical questions are:

1. Is fuel price dependent upon the type of aeroplanes? We only have access to the general categories of aircraft that airlines have in this dataset. This will be a useful question to answer as future aircraft may fall into new categories (e.g., blended wing or spaceplanes).

2. Is fuel price dependent upon the amount used? If bigger airlines can access bulk discounts this means that the bigger you are the more profitable you should be. On the other hand, smaller airlines can only be competitive if they pay the same per unit for fuel.

3. How does profit depend upon fuel price? Here we see whether we can make a simple model to predict the profit of airlines.

4. How does profit depend upon labour costs? Here we will be building a more complicated model, using more of the features, to see if that improves our predictions over a simple model.

5. What could happen to airlines in the future? Using our models from earlier we will be trying to predict what may happen in the future with:
1. Planes that are more efficient in terms of fuel cost
2. More efficient planes with a reduced labour cost
3. Increased fuel costs due to higher taxes
We also try to predict 2018 and 2019's profits using 2017's data rolled forward in an SVR Markov chain.

6. What can we tell from passenger numbers? Here we have a look at the numbers of airline passengers and what we can see from that.

Our data is from MIT's Airline Data Program (Swelbar & Belobaba 2019) and includes data on US airlines from 1995 to 2019 including their revenue, expenses, aircraft, labour costs, profitability and details of where they operate.

III. Data (materials)

Our dataset is 400 rows of data with 160 columns (where we have selected 22 columns of interest). The data comes from MIT's Airline data project where they collected data from 1995 – 2019 from US airlines' submissions to the BTS and SEC and makes it more user friendly. The data is split by both year and airline.

Most of the operating and passenger revenue data is distributed at the lower end of the scale, but three of the airlines' revenue for the last few years is much higher than the others. The various Total Operating expenses' distribution is very close to the revenue.

The total fuel expense distribution shows all the airlines peaking and troughing together, even though some of them are a lot higher or lower than others. There is a big spike in 2008, 2012 and 2014.

The amount of fuel used varies quite a lot between the airlines, apart from the biggest three which are similar.

The total price per gallon of fuel shows that most airlines buy the fuel at roughly the same price, though a couple of huge spikes on two airlines shows that some were worse at buying cheap fuel than others.

For transport related expenses the big three are much higher than others, with Alaska airlines also paying a lot in expenses (perhaps as Alaska is more remote) in comparison to its revenue.

The total number of planes shows that many airlines buy or sell planes a lot over the years. For small narrowbody planes most airlines have been shedding them over the years while Spirit has been acquiring them.

Large narrowbody planes have become more popular with most airlines over time, while widebody plane numbers have broadly remained stable over time.

IV. Analysis

A. Data preparation

1) Initial Download: To prepare the data for analysis we first had to download it, which sounds simple but involved creating a dataset just to tell the program we wrote of all the settings to use when downloading the 158 different webpages of data. These all had to be unpivoted, the columns then had to have the numbers extracted from strings and then finally the individual datasets had to be joined onto all of the other datasets to create one massive dataset (with some of the airlines' names being inconsistent and needing to be changed automatically).

2) Pre-Modelling: Before we modelled any of the data, we took a copy of the original data frame, filtered it to the target columns (such as pre modelling for Q4). Then we derived the target variable profit by calculating it from the revenue minus the expenses. We removed any null values in this field, as the final value could not be either missing or imputed as it would make the model effectively useless. Any other null values we left to be filled in with that column's mean value by imputation in the model. We also encoded all of the airline name fields as binary columns and then dropped the column with the name, this was so that the model would only see numbers and not think that one airline was inherently better than another. This was all so our modelling data would be as clean as it could be for the model. We then shifted the data to derive next year's profit so that we had a target variable.

B. Data derivation

1) Principal Component analysis (PCA): PCA is a technique to "extract the important information from the data table and to express this information as a set of new orthogonal variables" (Abdi & Williams 2010, 1). We used this technique in our question 1 to attempt to cut down the number of dimensions for the different types of plane, fuel price and year into two Principal components (which between them explained 97% of the data). These two dimensions were easier to visualise against the original data than five different dimensions.

2) Clustering: To finish off our question 1 of whether the fuel price is dependent upon the number of different types of plane we used k-means clustering. "K-means seeks an optimal partition of the data by minimising the sum-of-squared-error criterion" (Xu & Wunsch II 2008, 68). We used this technique to see whether particular categories of aircraft dominated any cluster, but the total number of aircraft seemed to be more important to the cluster which the row ended up in, rather than any individual type of aircraft (this was the same both with and without PCA first). The intention was that it would be clustered so that each type of aircraft was in its own cluster so that we could then use that further along in the analysis.

C. Construction of models

1) XGB Regression (Q3): For our third question we had to come up with a way to predict the profit of an airline based upon the price of fuel. Linear Regression isn't appropriate as it "will not take into account the temporal dimension ... because each feature is assumed to be independent of one another" (Tan et al. 2021, 1045). Random Forest manages to capture quite a lot of the variation, but isn't quite as good as the XGB Regressor. This is likely due to XGB being able to "handle sparse data well" (Chen & Guestrin 2016). Both algorithms use combinations of decision trees, Random Forest uses "bagging and random features" (Schapire 2001, 29) (bootstrap samples and then a subset of the features) and XGB uses boosting (Chen & Guestrin 2016) "which forces the learner to focus on its mistakes" (Schapire 1999).

2) Support Vector Regression Markov Chain (Q4-5): For our question five we looked at fuel price, fuel quantity, labour cost, year, airline, passenger revenue, fuel expense, number of passengers, profit and how these impacted future profit. We then rolled these up into a Markov Chain so that next year's predicted profit was then fed into the next run of the model (which is similar to this study (Dai et al. 2012) is doing except without generating data through the Monte Carlo simulation). The SVR model aims to "build a high dimension plane to fit the data to" (Vapnik 2000, 31, 208-218, 287). We tested the

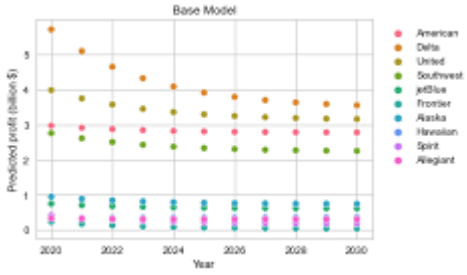

Fig. 1. SVG-HM base predictions

model by seeing how well it could predict 2018-2019's profits. Ideally, we would have given it a longer run of training and testing data, but we had very limited amounts of data to utilise. The model had an RMSE of $0.36 billion in 2018 then $0.96 billion in 2019.

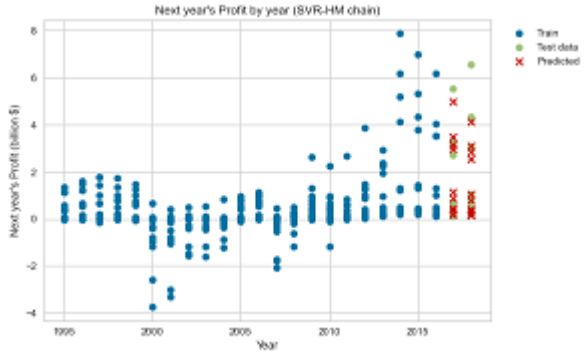

Fig. 2. SVG-HM chain test

D. Validation of results

1) XGB Regression (Q3): For the validation of the results for our XGB regression model for question 3, we calculated its R squared value (0.722), RMSE (0.542), plotted its histogram of residuals and plotted the predicted values against the actual values. We repeated this for the Random Forest Regressor (R squared 0.574, RMSE 0.671). The R squared term is a measure with "1 meaning that the data perfectly fit the model and 0 meaning that the model ignored the data" (Hamilton et al. 2015). The Root Mean Squared Error (RMSE) takes the root of the sum of the squared differences between the true and predicted values and so penalises higher value errors more than the Mean Absolute Error (MAE). The histogram of residuals shows the distribution of the residuals which should be normally distributed around the mean. Plotting the true values against the predicted values allows you to see where mistakes are being made (ideally all would be exactly on the line).

2) Support Vector Regression (Q4): To validate the results of our SVR model for Q4, we held back the data for 2018 (profit in 2019) and trained the model on the rest of the data (ideally, we would have tested a longer output, but had limited data to train on). To see how good this model really was, we trained an XGB Regressor model on the same data and tested it the same way. The XGB Regressor wasn't quite as good as the SVR model (it had an R squared of 0.879 compared with SVR's 0.907 and an RMSE of 0.706 versus 0.619 (in billion dollars)). The XGB regressor also did not seem to take into account things that it should have taken into account (it only looked at profit).

V. Findings, reflections and further work

For our second question we can also say that the amount used doesn't have an effect on price, but the airline's operating area does. From this article (America 2018) we can see the locations of America's oil pipelines and consequent fuel price zones.

For our third question we can see that the amount of fuel used is very important for the profitability of the next year, with the fuel price not being that important. As the airlines have a "very active pricing policy for tickets" (Abdella et al. 2021) they would normally be able to pass on changes to the cost of fuel to their passengers.

In our answer to question 4 we see that the XGB Regressor simply draws a straight line between the last year's profit and the next year's profit. These are highly correlated but there is not much a company can proactively do to become more profitable next year if they were not already under that model. For the SVR model we can see that it takes into account more features, so a company can see that if they reduce the labour cost then next year's predicted profit will rise and if they can cut fuel consumption it will rise.

Fig. 3. Higher price fuel effect on predicted profit

Question 5's SVG Markov chain model's main problem is that any error in one prediction is promptly fed straight through into the next year's predictions. As we can see in the test section on it, the RMSE for 2018 is $0.36 billion and $0.96 billion for 2019. This appears to be true of all forecasting techniques though, as the further into the future you go the more likely you are to be caught out by unpredictable events. One such unpredictable event was the COVID 19 pandemic, which appeared in early 2020 and is still ongoing now. This has caused governments around the world to impose various travel restrictions upon each other, causing widespread disruption to airlines and a reduction in air travel generally. For the lower profit companies, the model appears to be fairly accurate (though whether this is true proportionately is another

matter). Without any actual data for 2020 onwards, the model predicts a decline in profits.

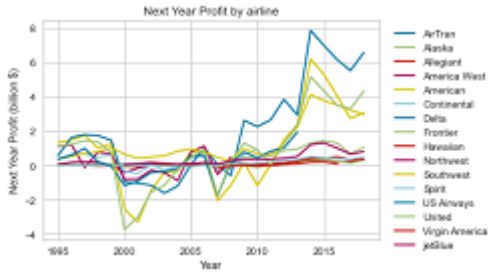

Fig. 4. Airline's next year profits

Another thing that our various models cannot easily detect is the predatory nature of the airline business, as we have 16 airlines in our dataset and 6 of them get taken over. Potentially, you could find out if airlines have announced merger plans and take that into account in the forecasting process, but this is fraught with difficulties and secrecy.

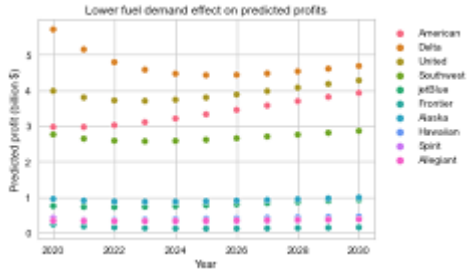

Fig. 5. Lower fuel demand effect on predicted profit

As a proof of concept, it appears that the SVR Markov chain model could be quite useful in predicting future profits. With more data (either by adding more years or going down to month level) you could get much more accurate results. In the real world the labour cost, passenger numbers, etc. would be subject to inflation/deflation over time. To keep this model simple, we used the last year's numbers in the predictions. Another refinement would be to have a different model per company (for those companies with less data available (new starters) they could supplement with comparable companies' data).

For question 6 can see from the graph of the total passenger numbers the impact of worldwide events, with the 2000 peak of the dot-com bubble and its bursting in 2001 (and the 9-11 attacks), then the 2008 market crash and slow recovery, followed by another market boom. The gradient in the rises is rather similar, possibly due to capacity constraints. Although we can hope for a continuous market rise for eternity, history teaches us that for every market boom there is a corresponding bust (Holland 2004).

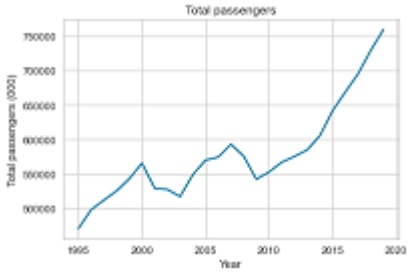

Fig. 6. Total Passengers

References

Abdella, J. A., Zaki, N., Shuaib, K. & Khan, F. (2021), 'Airline ticket price and demand prediction: A survey', *Journal of King Saud University - Computer and Information Sciences* **33**(4), 375–391.
URL: *https://www.sciencedirect.com/science/article/pii/S131915781830884X*

Abdi, H. & Williams, L. J. (2010), 'Principal component analysis', *WIREs Computational Statistics* **2**(4), 433–459. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101*

America, A. F. (2018), 'Jet Fuel: From Well to Wing'.
URL: *https://www.airlines.org/media/jet-fuel-from-well-to-wing/*

Bishop, C. (2006), Pattern Recognition and Machine Learning, *in* 'Pattern Recognition and Machine Learning', Springer Science+Business Media LLC, Chapter 13.
URL: *https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf*

Chen, T. & Guestrin, C. (2016), 'XGBoost: A Scalable Tree Boosting System', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794. arXiv: 1603.02754.
URL: *http://arxiv.org/abs/1603.02754*

Dai, H., Zhang, H., Wang, W. & Xue, G. (2012), 'Structural Reliability Assessment by Local Approximation of Limit State Functions Using Adaptive Markov Chain Simulation and Support Vector Regression', *Computer-Aided Civil & Infrastructure Engineering* **27**(9), 676–686. Publisher: Wiley-Blackwell.
URL: *https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=79650124&authtype=shib&site=ehost-live&scope=site&custid=s1089299*

Hamilton, D. F., Ghert, M. & Simpson, A. H. R. W. (2015), 'Interpreting regression models in clinical outcome studies', *Bone & Joint Research* **4**(9), 152–153.
URL: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678365/*

Holland, T. (2004), *Rubicon: The Triumph and Tragedy of the Roman Republic*, Abacus.

Hussain, A. (2020), 'MIT Airline Data Analysis'.
URL: *https://kaggle.com/xan3011/mit-airline-data-analysis*

Schapire, R. (2001), 'Random Forests'.
URL: *https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf*

Schapire, R. E. (1999), 'A Brief Introduction to Boosting', *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999* p. 6.
URL: *http://rob.schapire.net/papers/Schapire99c.pdf*

Swelbar, W. & Belobaba, P. (2019), 'Airline Data Project'.
URL: *http://web.mit.edu/airlinedata/www/AboutUs.html*

Tan, C. W., Bergmeir, C., Petitjean, F. & Webb, G. I. (2021), 'Time series extrinsic regression', *Data Mining and Knowledge Discovery* **35**(3), 1032–1060.
URL: *https://doi.org/10.1007/s10618-021-00745-9*

Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, 2 edn, Springer-Verlag.
URL: *https://statisticalsupportandresearch.files.wordpress.com/2017/05/vladimir-vapnik-the-nature-of-statistical-learning-springer-2010.pdf*

Xu, R. & Wunsch II, D. (2008), Partitional Clustering, *in* 'Clustering', John Wiley & Sons, Ltd, 68, pp. 63–110. Section: 4 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470382776.ch4.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470382776.ch4*

VI. Word counts

Section	Word Count
Abstract	138
Introduction	303
Analytical questions and data	288
Data (Materials)	289
Analysis	998
Findings, reflections and further work	602