PRINCIPLES OF DATASCIENCE PROJECT REPORT: PREDICTING UK DEPARTURE FLIGHT DELAYS

Matthew Tregear

INTRODUCTION

OVERVIEW OF THE DOMAIN

The investigation of flight delays is an active area of research in data science. As stated in Sternberg (2017) ‘s review of flight delay prediction work, the current literature focuses on two problems: the prediction of new delays (including cancellations) and flight delay propagation. The former concerns the prediction of flight delays, and the factors that influence them using statistical analysis and machine learning. The latter focuses on investigating the knock-on impact of flight delays on other flights using graph theoretic modelling or operational research.

The CAA is the UK's specialist aviation regulator. As part of the CAA’s commitment to publishing information, which it believes will assist consumers make comparisons between the UK's air transport services, the CAA publishes flight punctuality data[1].

I use the CAA punctuality dataset to try and predict UK departure flight delays.

ANALYTICAL QUESTIONS AND MOTIVATION

My analysis covers two analytical questions:

- **What are the main factors that influence UK flight departure delays?**

  Understanding the key factors that influence flight delays may provide insight in several areas. Most directly, it would provide information on what factors could be optimized to reduce flight delays. However, this information may also be used more indirectly. For example, a critical factor in modelling how flight delays may propagate through an airline flight network is the original cause of the delayed flight.

  I aim to answer this question in two different ways. First, I provide an overview of how common flight delays are and how factors that may influence flight delays vary in their level of flight delays using exploratory analysis. Second, I try to isolate the impact that individual factors have on the propensity of flight delays, all other factors being equal, and the scale of these impacts using regression analysis.

- **How do flight departure delay behaviours vary between UK domestic and non-Domestic flights?**

  Flight departure delays are often modelled globally with no distinction made for behaviours of different locality. However, domestic flight delay behaviours may significantly differ from non-domestic behaviours. For example, for UK domestic flights distance may be less of a factor, as all incoming/outgoing flights are relatively

[1] https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Flight-reliability/Datasets/Punctuality-data/Punctuality-statistics-2017/

short. This may in turn mean that airlines are able to manage delays differently as they will more easily be able to divert planes to prevent delays.

I answer this question by undertaking a separate regression analysis of flight delays solely on domestic flights. In doing so, I can compare how the regression results contrast with the regression analysis of all UK departure flights.

ANALYSIS/PLAN STRATEGY

I will predict flight delays of more than 15 minutes using regression analysis on two datasets: one containing all flights and another containing only domestic flights (both supplemented with airport location data obtained from the ourairports website[2]). The all flight dataset contains around 170,000 monthly flight summaries for over 5 million UK departure flights since 2011; the domestic flight dataset contains around 15,000 monthly flight summaries for 1 million departure flights. Predicting delays in each of these datasets will involve 3 main steps:

Initial exploratory analysis. To initially explore flight delays I will look at the proportion of flight delays in my dataset, and how flight delays vary with the variables in my dataset. This will provide initial insight into which ones are better predictors. I will also briefly data visualize the proportion of flight delays as flight paths in Tableau to see if there are any spatial relationships for flight delays.

Dimension reduction using Principal Component Analysis (PCA). As both my flights datasets will likely contain a large number of attributes on flights after my categorical variables have been converted to dummies, I will need to dimensionally reduce my dataset before performing any regression to avoid the curse of dimensionality. For my regression models to perform well without dimension reduction, I would need a vast amount of data to ensure I had sufficient coverage for an adequate subgroup for each combination of attribute/predictor values. I will choose PCA over other dimension reduction techniques more targeted at categorical variables (such as Multiple Correspondence Analysis) as it is unclear how I would bin my continuous predictors in my dataset.

Logistic regression. I will train logistic regressions to investigate the main factors that influence flight delays and to see if domestic flight delay behaviour is significantly different than other flight types.

Logistic regression is better suited than linear regression for this analysis as I want to predict the probability that a flight is delayed. If I were to use a linear regression, my predictions may be biased as I may predict probabilities greater than 1 or negative probabilities for certain UK departure flights.

To validate the results of the logistic regression models, I will use a combination of out of sample testing and the data visualisation of model residuals in Tableau.

[2] http://ourairports.com/data/

ANALYTICAL PROCESS

PREPARING THE DATASET FOR ANALYSIS

Reshaping of data

The CAA punctuality database records departure and arrival flight information in the form of the table below.

Fig 1: form of original CAA punctuality dataset

Year month	reporting_airport	origin_destination_country	origin_destination	airline_name	Arrival/Departure Flight	Average_Delay_mins	No of flights in month
201104	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	Departure	7	5
201104	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	Arrival	0	6

First, to use UK arrival flight information to inform UK departure flight information on the same routes, I reshape the data from 'long' to 'wide' so that arrival flight delay information on a corresponding monthly flight route for each airline is matched to its corresponding departure flight information (as shown below).

Fig 2: Transformed departure flight monthly summaries with arrival information

…..	reporting_airport	origin_destination_country	origin_destination	airline_name	AverageDeparture_delay_mins	Average_Arrival delay_mins	No of flights in month
…..	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0	5

Second, to prepare the data for use in a logistic regression, I converted the data from 'wide' to 'long' as the monthly summary information need to be convert to 'tidy' individual flight data (as shown below) to predict individual flight delays.

Fig 3: Transformed individual flights dataset for logistic regression

….	reporting_airport	origin_destination_country	origin_destination	airline_name	AverageDeparture_delay_mins	Average_Arrival delay_mins
….	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0
…	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0
….	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0
…..	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0
….	BIRMINGHAM	AUSTRIA	INNSBRUCK	THOMSON AIRWAYS LTD	7	0

Treatment of Missing Values /redundant variables

Aside from redundant variables[3], I removed variables showing the monthly arrival/departure delay information for a route one year in the past as they contained systematic bias. They recorded a ‘0’ entry for the monthly delay one year in the past both for flights routes with no past delays AND for newly created routes.

Where departure flights had no corresponding flight in the last month for my lagged delay information variables (described below under data derivation), I used lagged flight information for the most recent previous flight summary on the same route (and flagged these cases in an additional variable).

Derivation of new variables

By matching the origin destination pairs of individual flight routes to an airports database I was able to identify the coordinates of the origin and destination city[4] of each flight. Then, applying the Haversine[5] function (using the formula below) to the origin-destination coordinates, I was able to construct a distance variable.

Fig 4: Python code for the Haversine formula

```
 def haversine(lat1, lon1, lat2, lon2):
 # deltas between origin and destination coordinates
 dlat = np.radians(lat2-lat1)
dlon = np.radians(lon2-lon1)
 a = np.sin(dlat/2)**2 + np.cos(lat1) * np.cos(lat2) * np.sin(dlon/2)**2
 c = 2 * np.arcsin(np.sqrt(a))
 R =6371
 # a spherical distance between the two points, i.e. hills etc are not considered
return R * c
```

I also created a new binary dependent variable for whether an individual flight was delayed or not using the proportion of delayed departure flights in a given month and the number of departure flights recorded in each flight summary (after my data had been reshaped to individual flights data). This is the dependent variable used in my logistic regression.

[3] A full list of all removed variables is reported in my appendices.

[4] Because it was often difficult to identify the exact coordinates of each airport, I used the coordinates of the city (taken as the average coordinates of all airports in a given city).

[5] Haversine Formula (from R.W. Sinnott (1984), "Virtues of the Haversine", Sky and Telescope, vol. 68, no. 2, 1984, p. 159):

Data transformations

I converted my categorical predictors into binary variables so they can be used in a logistic regression model.

I log-transformed continuous predictors because their histogram plot were highly skewed. As shown below in the histograms and QQ plots, log transforming these continuous variables removes the skew so a variable's distribution more closely resembles a normal distribution. Not log-transforming these variables may mean that PCA and logistic regression overweight the variance in values of greater magnitudes.

Fig 5: raw distance variable

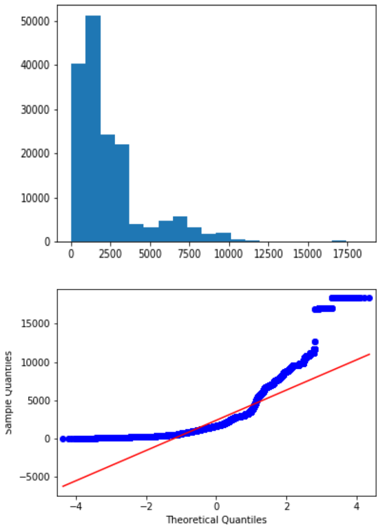
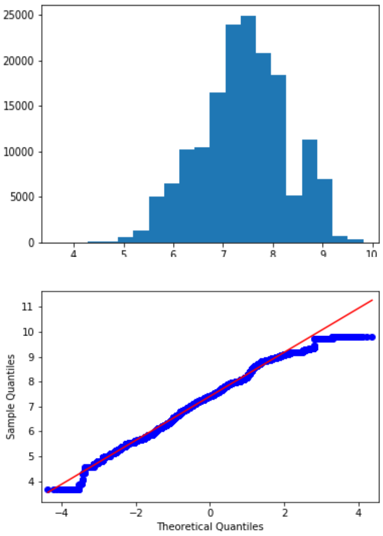


Fig 6: logged distance variable



I also normalized these continuous variables so that their values and variance are on the same scale as the other dummy categorical variables for logistic regression and PCA.

CORE ANALYSIS

Initial exploratory analysis

I undertook some initial exploratory analysis to look at which predictors vary by proportion of flight delays the most, and how this varies for domestic flights.

UK departure flights delays longer than 15 minutes tend to occur 22% of the time. This is slightly lower for UK domestic flights (19%).

Fig 7: All delayed flights

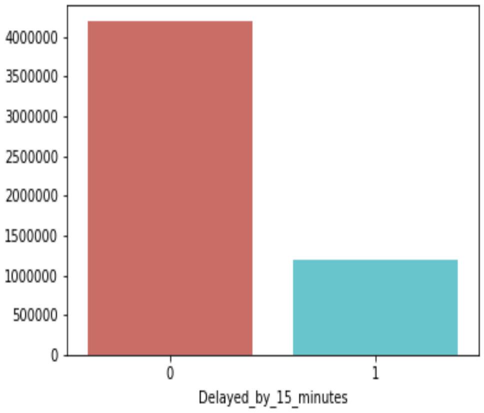
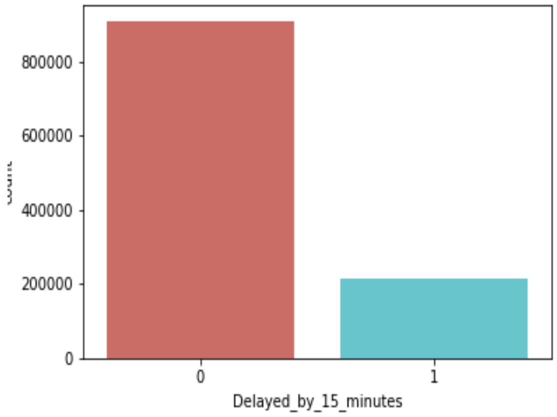


Fig 8: Delayed domestic flights



The means for the logged and normalised continuous variables[6] show that they may each affect the propensity of delays in all flights. Domestic flights data shows the same trend but, as expected, distance is less of a factor.

Fig 9: Predictor means for all flights that are delayed/not delayed

Flight delay	Distance	Arrival delay in previous month	Departure delay in previous month
0	-0.5	0.048	-0.07
1	-0.33	0.22	0.25

Fig 10: Predictor means for domestic flights that are delayed/not delayed

Flight delay	Distance	Arrival delay in previous month	Departure delay in previous month
0	-1.48	-0.035	-0.09
1	-1.47	0.13	0.14

The proportion of flight delays varies between 15% and 25% across the 10 largest airlines by number of flights. There is greater variation amongst the largest airlines for domestic flights: proportions vary between 7% and 25%.

[6] distance, arrival flight delay in the previous month, departure flight delay in the previous month.

Fig 11: Delay by airline (all flights)

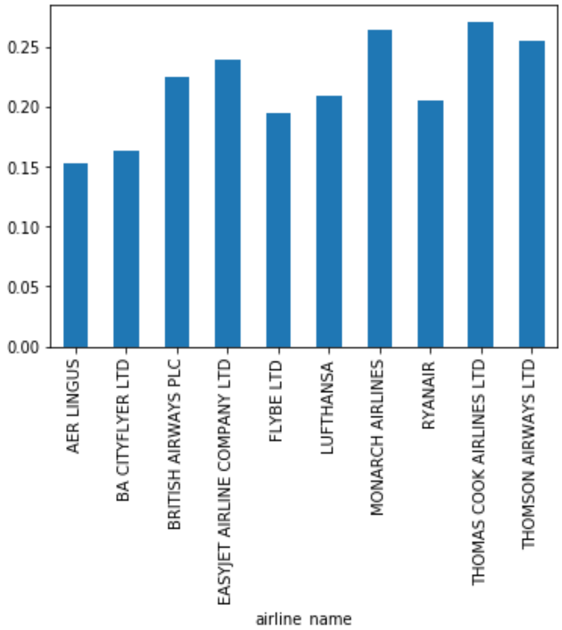
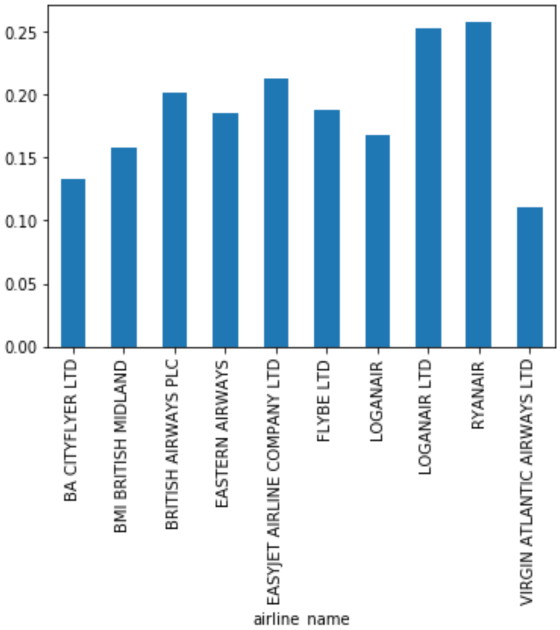
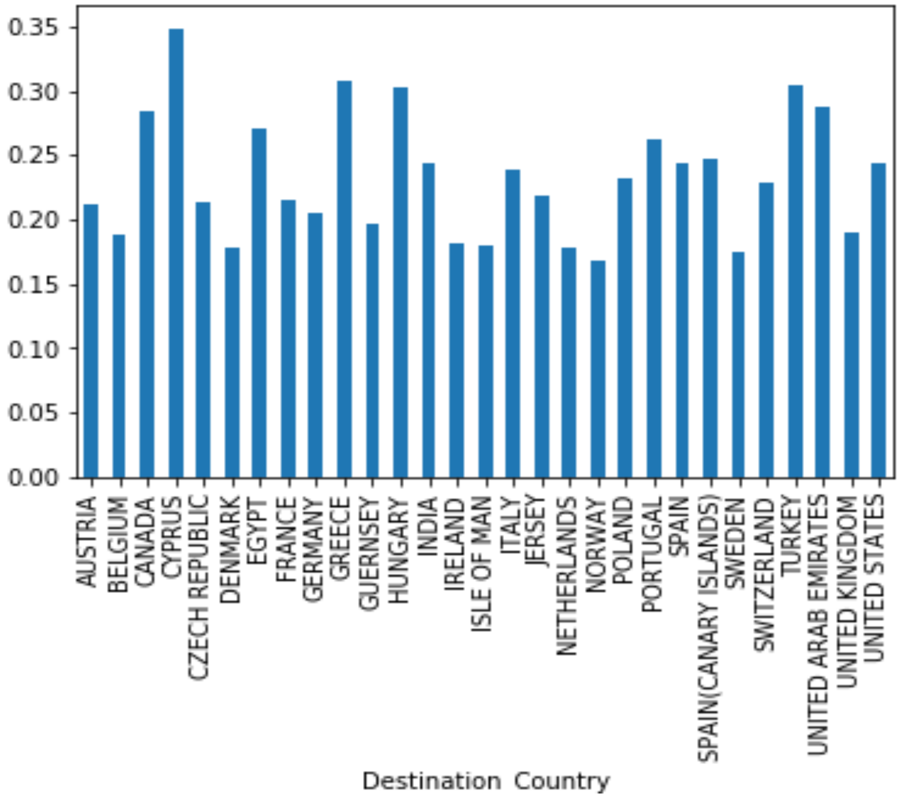


Fig 12: Delay by airline (domestic flight)



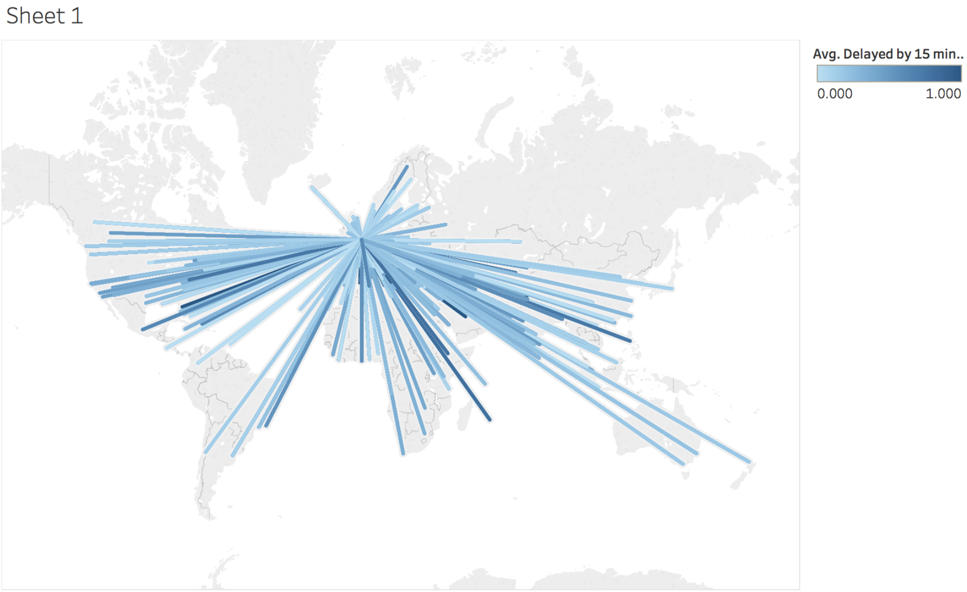
For the top 30 most popular flight destination countries the proportions of delays ranges from 13% to just below 35%. The UK doesn’t seem to be a major outlier (20% of its flights are delayed).

Fig 13 : Flight delays by destination country



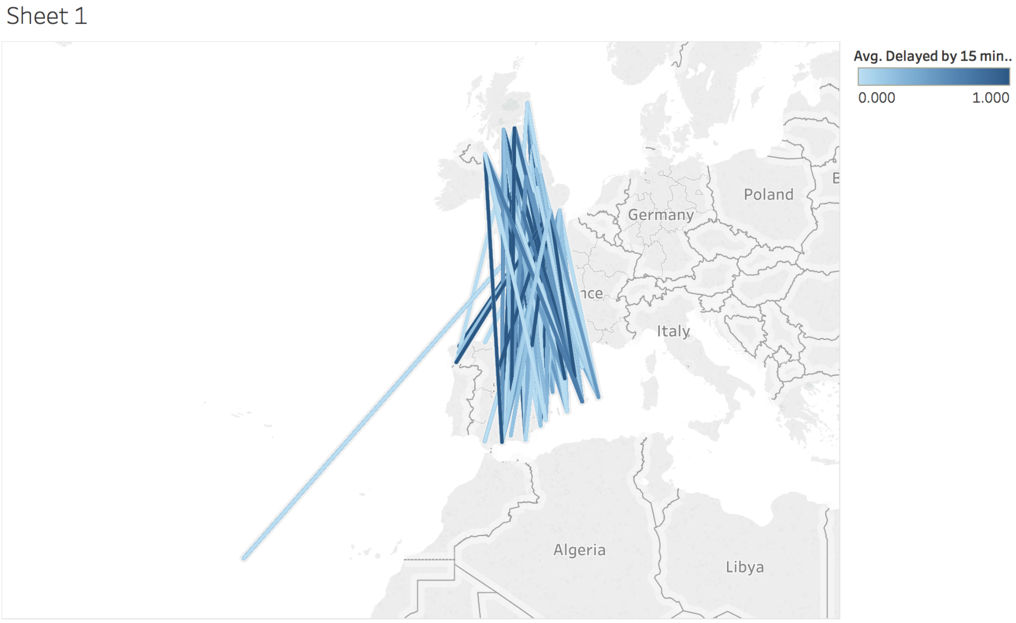
An initial data visualisation of delays in Tableau doesn’t show any clear trends how flight delays are spatially distributed.

Fig 14: Proportion of flight delays for all Heathrow flight routes



Map based on Longitude and Latitude. Color shows average of Delayed by 15 minutes. Details are shown for Index. The data is filtered on Reporting Airport, which keeps HEATHROW.

Fig 15: Proportion of delays for all flight routes to Spain



Map based on Longitude and Latitude. Color shows average of Delayed by 15 minutes. Details are shown for Index. The data is filtered on Destination Country, which keeps SPAIN.

PCA dimension reduction

I used PCA dimension (in the scikit learn package[7]) to reduce the total number of attributes from over 1,060 to 49 and 147 to 32 for my all flights and domestic flights datasets respectively. This ensured that there would be sufficient data for each subgroup of flight in my dataset for each logistic regression.

The scree plots below shows that the explained variance in both my datasets is widely spread across many principal components. Because the initial few principal components only covered limited variance in my all flights dataset - and thus limited variance for my logistic regression to exploit - I looked at the first 45 principal components that covered 80% of the explained variance. For my UK domestic flight dataset, I followed the same process and looked at the first 29 components.

Fig 16: Scree plot (all UK flights)

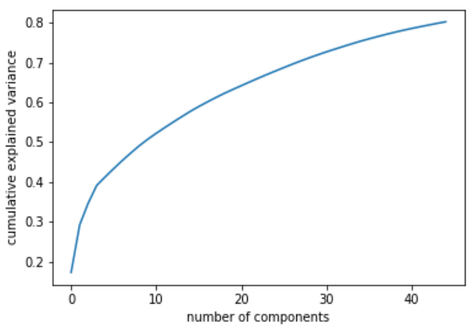
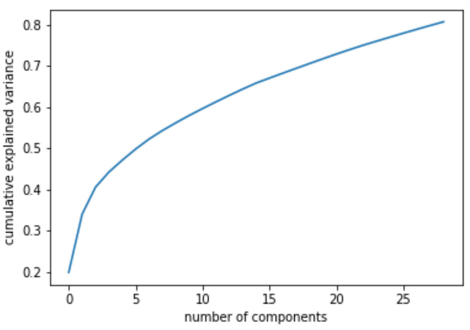


Fig 17: Scree plot (all domestic flights)



I then chose variables with loadings greater than 0.3 in a single principal component as my regression predictors. This ensured my predictors included the variables responsible for the most variance in each component.

Logistic Regression

To train my logistic regression models using the sci-kit learn package[8], I first partitioned each of my datasets into a training set (70%) and a test set (30%). I then trained each model on their respective training sets and used the trained models to predict delays in the test set.

Coefficients for my trained all flights model, converted into odds-ratios, suggest that several factors may be related to flight delays. Flying in November (as opposed to May) and flying with Wizz air or Easyjet reduces your odds of incurring a flight delay by more than 15%. Flying in August, September, October, December, June or July (as opposed to May),in 2016 (as opposed to 2014), flying from Manchester or Stansted airport, flying Ryanair and an increase in 1% of average departure delay in the previous month all increase your odds of incurring a flight delay by more than 25%. All predictors were statistically significant at the 5% level apart from the Heathrow airport and United States destination variables[9].

[7] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[8] http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[9] only these had p-values > 0.05.

Coefficients do not significantly differ for my UK regression model, yet there are some notable differences. Having a flight on the same route on the previous month was a additional factor at the UK level that reduced odds by more than 15%. In the all flights data, flying Easyjet reduced your odds of incurring a delay by 19% whereas in the UK domestic flights data it increased your odds by 23%.

VALIDATION OF RESULTS

On first glance my regression models perform well in predicting delays: they both report a test accuracy of around 0.78. However, looking more closely, most of the accuracy is attributed to correctly classifying cases where there are no flight delays. In particular, when there is a flight delay, the models predict a very large number of false-negative cases, where the model predicts no flight delay but should predict one. So both models have very low recall of only 0.02 or less. The predictions of the models also have a precision under 0.55 for predicting delays, so much of the time when the model predict delays a delay will not actually occur.

Fig 18: All flights confusion matrix

	Flight delay	No flight delay
Predicts flight delay	8,742	7,158
Predicts no flight delay	349,26	1,253,642

Fig 19: Domestic flights confusion matrix

	Flight Delay	No flight Delay
Predicts flight Delay	126	143
Predicts no flight Delay	64,234	27,1901

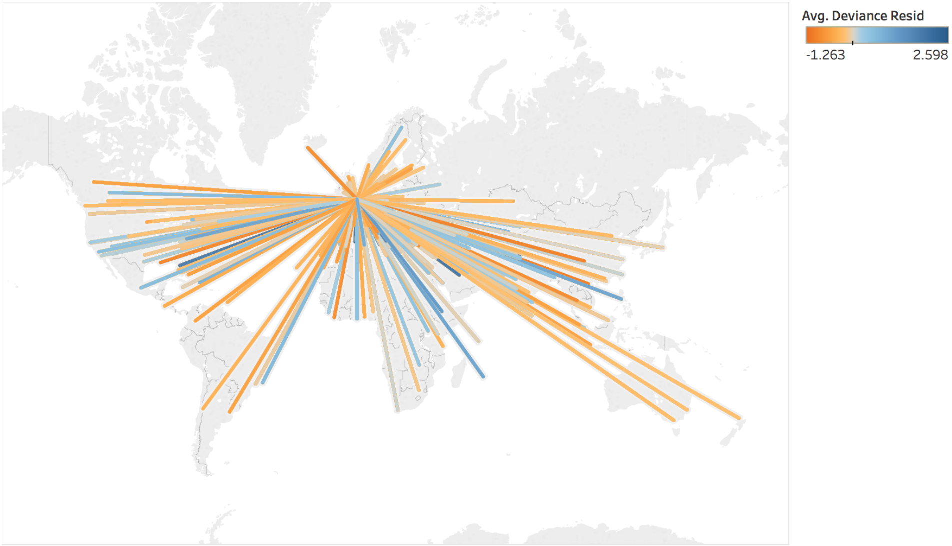
The poor model performance is confirmed by the very low pseudo r-squared of 0.0454 and 0.0314 the all flights and the domestic flights regression model respectively.

It is also reflected in the Tableau visualisations of residuals and proportions of delays below. The fit of the all flights regression model is so poor that the visualised residuals do not improve much over explaining flight delay by directly visualising the average proportion of delay on each route[10].

[10] (Darker orange residuals map very closely to light blue routes with low average delays; darker blue residuals map very closely to dark blue routes with high average delays.)

Fig 20: Tableau visualisation of residuals of all flights routes from Heathrow

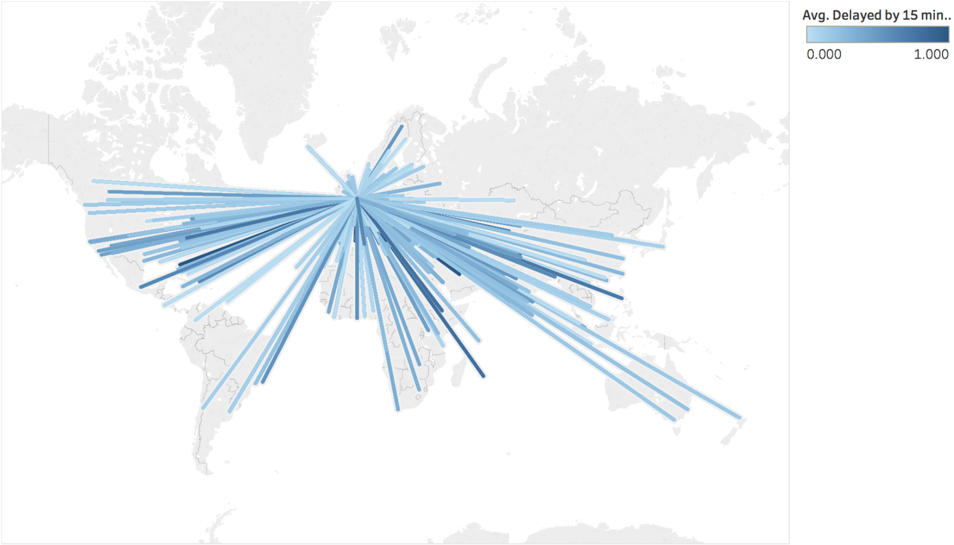
Sheet 1



Map based on Longitude and Latitude. Color shows average of Deviance Resid. Details are shown for Index. The data is filtered on Reporting Airport, which keeps HEATHROW.

Fig 21: Tableau visualisation of average proportion of flight delays for all flight routes from Heathrow (repeated)

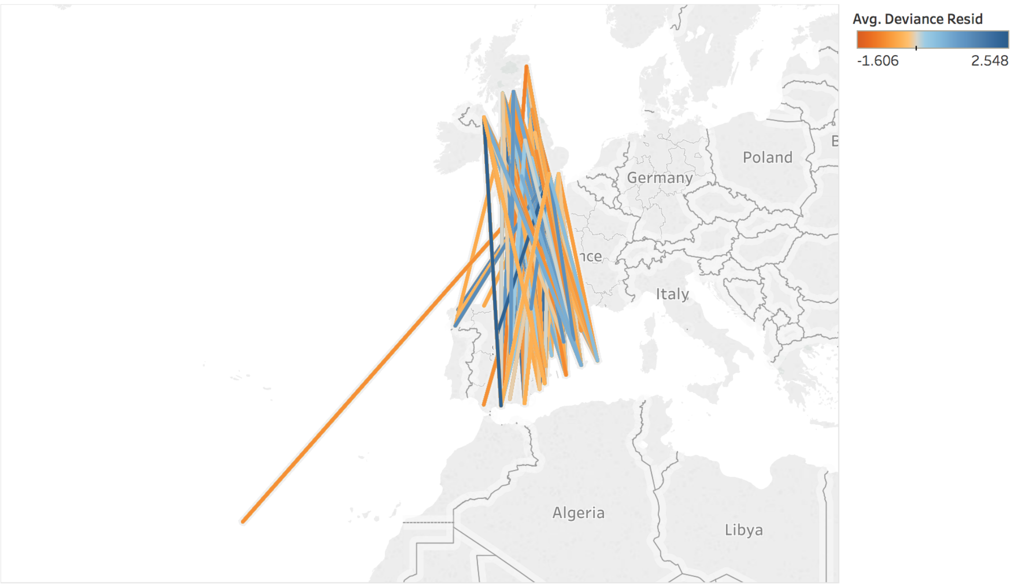
Sheet 1



Map based on Longitude and Latitude. Color shows average of Delayed by 15 minutes. Details are shown for Index. The data is filtered on Reporting Airport, which keeps HEATHROW.

Fig 22: Tableau visualisation of residuals for all flight routes to Spain

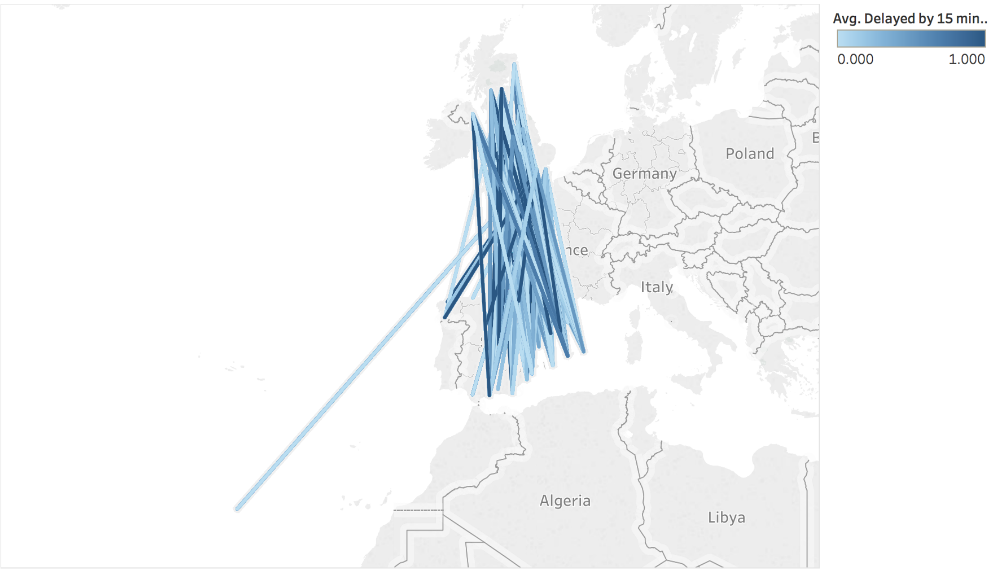
Sheet 1



Map based on Longitude and Latitude. Color shows average of Deviance Resid. Details are shown for Index. The data is filtered on Destination Country, which keeps SPAIN.

Fig 23: Tableau visualisation of average proportion of delays for all flight routes to Spain (repeated)

Sheet 1



Map based on Longitude and Latitude. Color shows average of Delayed by 15 minutes. Details are shown for Index. The data is filtered on Destination Country, which keeps SPAIN.

FINDINGS AND REFLECTIONS

MAIN FINDINGS

My analysis suggests that the main factors that are related to flight delays are flying on a route which has had a delay in the previous month, flying in a certain time of year or flying with Ryanair, Easyjet or Wizz air. They also suggest that there are few differences in behaviour between domestic and non-domestic departure flight delays (apart from when flying with Easyjet. However, given the very low recall and fit of my models, I cannot draw any conclusions on which factors have an influence on flight delays.

In particular, the poor performance of the models may mean that the odds ratio coefficients may be biased due to omitted confounding factors. For example, the lower odds of incurring a flight delay when flying from Manchester airport may mainly relate to the higher probability of adverse weather conditions at this location rather than anything intrinsic to Manchester airport.

This poor performance also means that comparing the regression results of the all flights and domestics flights models is also not informative.[11] For example, the differing impact of flying Easyjet on the odds of a delay in domestic flights and the all flights regression model may actually relate to correlation with other factors such as differing flight route capacity at the national and international level rather than anything intrinsic to Easyjet. For this reason, I cannot compare the coefficients or visualised residuals of the two models to identify differences in behaviour for domestic flights.

FURTHER ISSUES WITH THE GENERAL APPROACH

Many predictors act as useful proxies but don't directly measure potential causes of delays

Many predictors, such as the destination country, do not relate to root causes of a flight delay (or a factor of interest) but are correlated with them. They may act as useful proxies for types of root cause but they do not provide a cause –effect relationship. This is a problem because, for example, the regression results in each model can tell us that delays are more likely travelling on one airline as opposed to another - but it does not tell us why certain airlines have fewer delays.

Flight delay behaviour may vary within different flight segments

As with my hypothesis regarding differences in behaviour between domestic and non-domestic flights, there may be other segments within the dataset where flight delay behaviour varies. For example, the flight delays in smaller airports may be of a different nature due to staffing and flight maintenance differences. Or it may be the case that flights of a longer duration are intrinsically different to short duration delays – so using the same factors to predict both is incorrect. A better methodology may involve a more segmented modelling approach, such as geographically weighted regression analysis Brunsdon (1998).

[11] And for this reason I have not compared the residuals of the domestic flights dataset and the all flights datasets in data visualisations.

The methodology may lead to over-reliance on pvalues

Identifying the statistical significance of factors may lead to an over-reliance on p-values. These conventional inference statistics may provide false findings if the same analysis were to be repeated on a number of different samples. This may not be such a factor given that the sample represents large proportion of the population (the stochastic process for the distribution of delays on flights) for most analyses. However, if this analysis were to be repeated many times, the variance across different samples may mean that p-values suggest statistically significant findings that are actually attributed to sampling error rather than a true statistically significant effect. (as noted in Cumming (2013)).

THE APPLICATION OF MY ANALYSIS

Given that the regression models I have trained for my analysis perform very poorly when predicting flight delays, there is not substantial value in my results.

That said, research into understanding UK departure delay is still a useful area because there is a commercial and public need for a better understanding of flight delays solely based on publicly available information. The CAA itself needs to adequately present its findings on flight punctuality in such a way that it accurately represents the true state of the industry. Another potential application is in flight search engines, which tend to have a more restricted access to flight information than airlines. They may gain from further research into predicting delays because they will be better able to provide their customers with more detailed information on their flight choices.

My results suggest that the current level of detail available in the CAA's openly published punctuality dataset is insufficient to meet these demands.

APPENDICES

Appendix I: description of the raw dataset and full list of removals and derived variables

reporting_period	Converted into 11 month and 4 year dummy variables
reporting_airport	Converted into dummy variables for each UK origin airport
origin_destination_country	Converted into dummy variables for each destination country
origin_destination	Converted into dummy variables for each destination country
airline_name	Converted into dummy variables for each airline
arrival_departure	Removed as it is a redundant indicator of whether a summary refers to an arrival or a departure flight
scheduled_charter	Converted into a dummy variable
number_flights_matched	Consolidated into the number of flights into a given month and used to convert monthly summary into individual flights data set
actual_flights_unmatched	Consolidated into the number of flights into a given month and used to convert monthly summary into individual flights data set
early_to_15_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not
flts_16_to_30_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not

flts_31_to_60_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not
flts_61_to_180_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not
flts_181_to_360_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not
more_than_360_mins_late_percent	Consolidated into my dependent variable flight is delayed by 15 minutes or not
average_delay_mins	Used to create lagged average arrival and departure delay information in previous and then removed as functionally dependent on dependent variable
planned_flights_unmatched	Consolidated into the number of flights into a given month and used to convert monthly summary into individual flights data set
previous_year_month_flights_matched	Removed as it contains a systematic bias as described in the main text.
previous_year_month_early_to_15_mins_late_percent	Removed as it contains a systematic bias as described in the main text.
previous_year_month_average_delay	Removed as it contains a systematic bias
Distance	New variable derived from airport locations coordinates as described in the main text.
Flight on same route in previous month	New variable created to indicate whether a flight has had a flight on the same route in the previous month as described in the main text

Appendix II: All flights regression results ordered by odds ratio

	2.5% sig level	97.5% sig level	coefficient	odds-ratio 2.5% level	odd-ratio 97.5% level	odd-ratio	t-value	p-value
Intercept	-1.78	-1.73	-1.75	0.17	0.18	0.17	-160.76	0
Month11	-0.22	-0.2	-0.21	0.8	0.82	0.81	-36.62	0
Easyjet	-0.23	-0.2	-0.21	0.79	0.82	0.81	-22.05	0
Wizz air	-0.21	-0.18	-0.2	0.81	0.84	0.82	-21.35	0
2012	-0.11	-0.09	-0.1	0.9	0.91	0.9	-23.79	0
Destination Country Spain (Canary Islands)	-0.09	-0.07	-0.08	0.91	0.93	0.92	-15.36	0
Uk airport Newcastle	-0.1	-0.07	-0.08	0.91	0.94	0.92	-10.51	0
2011	-0.07	-0.05	-0.06	0.93	0.95	0.94	-14.57	0
Month5	-0.06	-0.04	-0.05	0.94	0.96	0.95	-9.02	0
airline:Flybe	-0.06	-0.04	-0.05	0.95	0.96	0.95	-10.14	0
Monarch	-0.07	-0.04	-0.05	0.94	0.96	0.95	-6.83	0
airline:British Airways	-0.05	-0.03	-0.04	0.95	0.97	0.96	-7.4	0
lagged average arrival delay	-0.03	-0.02	-0.03	0.97	0.98	0.97	-13.43	0
2013	-0.04	-0.03	-0.04	0.96	0.97	0.97	-8.43	0
Destination City Faro	-0.05	-0.01	-0.03	0.95	0.99	0.97	-3.08	0
Destination Country United Kingdom	-0.03	-0.01	-0.02	0.97	0.99	0.98	-5.44	0
Thomson airways	-0.02	-0.01	-0.02	0.98	0.99	0.98	-4.16	0
Destination Country United States	-0.03	0	-0.01	0.97	1	0.99	-1.75	0.08
Uk airport Heathrow	-0.01	0.01	0	0.99	1.01	1	-0.75	0.45
Destination Country Germany	0.01	0.02	0.02	1.01	1.02	1.02	3.81	0
Destination Country Turkey	0.02	0.04	0.03	1.02	1.04	1.03	6.19	0

Thomas Cook	0.02	0.05	0.04	1.02	1.05	1.04	4.17	0
Destination Country Spain	0.03	0.05	0.04	1.03	1.06	1.05	8.87	0
Destination Country Greece	0.05	0.07	0.06	1.05	1.07	1.06	13.13	0
Uk airport Edinburgh	0.07	0.09	0.08	1.07	1.09	1.08	14.79	0
Flight on same route in previous month	0.07	0.11	0.09	1.07	1.11	1.09	9.44	0
Distance	0.08	0.09	0.08	1.08	1.09	1.09	59.77	0
UK airport Birmingham	0.08	0.1	0.09	1.08	1.1	1.09	15.02	0
Month2	0.08	0.1	0.09	1.08	1.11	1.1	16.93	0
UK airport Luton	0.08	0.1	0.09	1.09	1.11	1.1	17.08	0
airline Jet2. Com	0.09	0.1	0.09	1.09	1.1	1.1	27.62	0
Month1	0.09	0.11	0.1	1.1	1.12	1.11	19.45	0
2015	0.11	0.13	0.12	1.12	1.14	1.13	31.72	0
Month3	0.11	0.13	0.12	1.12	1.14	1.13	22.85	0
Destination Country Italy	0.12	0.15	0.13	1.12	1.16	1.14	17.05	0
UK airport Gatwick	0.12	0.14	0.13	1.13	1.15	1.14	29.22	0
Destination Country France	0.12	0.15	0.14	1.13	1.17	1.15	17.05	0
2017	0.19	0.21	0.2	1.21	1.23	1.22	42.57	0
Charter C	0.19	0.22	0.21	1.21	1.25	1.23	26.5	0
2016	0.24	0.26	0.25	1.28	1.3	1.29	67.92	0
Uk airport Manchester	0.24	0.27	0.25	1.27	1.31	1.29	41.07	0
Uk airport Stansted	0.3	0.31	0.3	1.34	1.37	1.36	75.27	0
Month8	0.31	0.33	0.32	1.36	1.39	1.38	64.48	0
Ryanair	0.3	0.34	0.32	1.35	1.4	1.38	34.57	0
Month10	0.32	0.34	0.33	1.38	1.41	1.4	63.09	0
Month9	0.34	0.36	0.35	1.4	1.43	1.41	68.28	0
Month6	0.36	0.38	0.37	1.43	1.46	1.44	79.08	0
Month7	0.52	0.54	0.53	1.68	1.71	1.7	109	0
average departure delay in previous month	0.54	0.54	0.54	1.71	1.72	1.72	271.95	0
Month12	0.62	0.66	0.64	1.86	1.93	1.89	59.98	0

Appendix III: Domestic flights regression results ordered by odds ratio

	2.5% sig level	97.5% level	coefficient	2.5% odds ratio	97.5% odds ratio	odds ratio	t-values	p-values
intercept	-1.73	-1.62	-1.68	0.18	0.20	0.19	-60.63	0.00
Flight on same route in previous month	-0.27	-0.17	-0.22	0.76	0.84	0.80	-8.62	0.00
Month11	-0.20	-0.16	-0.18	0.81	0.86	0.84	-14.28	0.00
2012	-0.11	-0.07	-0.09	0.89	0.93	0.91	-9.24	0.00
UK airport Newcastle	-0.10	-0.05	-0.08	0.90	0.95	0.93	-6.19	0.00
2011	-0.10	-0.06	-0.08	0.91	0.94	0.93	-7.52	0.00
UK airport Glasgow	-0.03	0.00	-0.01	0.97	1.00	0.99	-1.71	0.09
Month4	-0.04	0.01	-0.01	0.96	1.01	0.99	-1.11	0.27
2013	-0.02	0.02	0.00	0.98	1.02	1.00	0.38	0.71
Distance	0.00	0.01	0.00	1.00	1.01	1.00	1.47	0.14

Destination City Aberdeen	-0.01	0.03	0.01	0.99	1.03	1.01	0.72	0.47
UK airport Edinburgh	0.00	0.03	0.02	1.00	1.03	1.02	2.29	0.02
UK airport Manchester	0.01	0.04	0.02	1.01	1.04	1.02	2.52	0.01
Destination City Glasgow	0.02	0.06	0.04	1.02	1.06	1.04	4.21	0.00
Flybe	0.03	0.06	0.05	1.03	1.06	1.05	6.59	0.00
Average arrival delay in previous month	0.04	0.06	0.05	1.04	1.06	1.05	9.60	0.00
Month1	0.04	0.09	0.06	1.04	1.09	1.06	5.07	0.00
Destination City Edinburgh	0.05	0.08	0.06	1.05	1.08	1.06	7.71	0.00
Month2	0.04	0.09	0.06	1.04	1.09	1.06	5.11	0.00
Month3	0.08	0.13	0.10	1.08	1.13	1.11	8.58	0.00
2015	0.11	0.15	0.13	1.12	1.16	1.14	14.66	0.00
British airways	0.17	0.20	0.19	1.19	1.23	1.21	25.29	0.00
2017	0.17	0.22	0.20	1.19	1.24	1.22	17.99	0.00
Easyjet	0.20	0.23	0.21	1.22	1.25	1.23	26.46	0.00
2016	0.21	0.25	0.23	1.24	1.28	1.26	26.22	0.00
Month8	0.27	0.32	0.30	1.31	1.38	1.34	24.85	0.00
Month12	0.25	0.36	0.30	1.28	1.43	1.35	10.85	0.00
Month6	0.35	0.39	0.37	1.41	1.48	1.44	33.50	0.00
Month10	0.36	0.41	0.39	1.44	1.51	1.47	31.65	0.00
Month9	0.38	0.43	0.41	1.47	1.54	1.50	34.36	0.00
average_delay_minsDep_lagged	0.40	0.42	0.41	1.50	1.53	1.51	81.81	0.00
Month7	0.50	0.54	0.52	1.64	1.72	1.68	44.34	0.00

REFERENCES

Brunsdon,C. Fortheringham,S and Charlton, S,1998 . Geographical Weighted Regression Modeling Spatial Non-stationarity. *Journal of the Royal Statistical Society.Series D* (The Statistician) Vol. 47, No. 3 (1998), pp. 431-443.

Cumming, G. 2013. The new statistics why and how. Psychological science.

Sinnott,R.W. .1984. Virtues of the Haversine.*Sky and Telescope*, vol. 68, no. 2, p. 159

Sternberg, A, Soares, J.,Carvalho D., Ogasawara, E.. 2017 .A Review on Flight Delay Prediction. *CEFET/RJ Rio de Janeiro Brazi*l.