Aviation and Airline Forecasting Projects

Riccardo Bellide, Elena Furlanis, Alessandra Meriani

July 12, 2022

Introduction

'Fore-cats' is a consultancy agency focused on Aviation and Airline projects, that assists data-driven firms in supporting and enhancing decision making process. Three of our major forecasting problems are put forward, customized on clients' needs and requests.

1 a. Problem statement

An American leader Airline wants to improve its success in terms of brand reputation. Moreover, it requires a focus on the consolidation and maintenance of its winning position in the field of business travels. Since a firm's success is critically related to their customers' experience, 'Fore-cats' decided to identify the key factors affecting the satisfaction of airline passengers and to predict it for future business clients.

2 a. Dataset

The information set that has been used for this cross-sectional study comes from the results of a *satisfaction survey*. It was conducted on behalf of the focus firm in 2015 on a representative subset of 129880 passengers of the Airline's customer base, characterized by 25 variables. During the preparatory data cleaning phase, some transformations were carried out in terms of the focus binary variable `satisfaction` (satisfied=1, not satisfied/neutral=0) or deletion of variables deemed irrelevant or unnecessary for the purpose stated (e.g. identification and partial satisfaction variables).Table 1 shows the final resulting variables. The vast majority of respondents are loyal customers (81.72%), as was to be expected from a customer survey. However, the distribution between satisfied and unsatisfied/neutral passengers is fairly balanced (43.34: 56.66%). The same applies to `gender` (50.75: 49.25%, F:M), unlike flight characteristics at the discretion of the passenger, such as `flight.distance` (38.29:26.40:35.31%) [1] and `class` (47.81: 44.98:7.21%). The latter is significantly related to `type.of.travel`, clearly unbalanced towards Business passengers (68.99: 31.01%, B:P), which tends to travel in Business class (66.30%); for Personal travel there is a significantly higher number in Economy class (82.14%).
The target variable `satisfaction` may vary, intuitively, depending on flight's and customer's characteristics. These intuitions are confirmed by a more in-depth analysis sum-

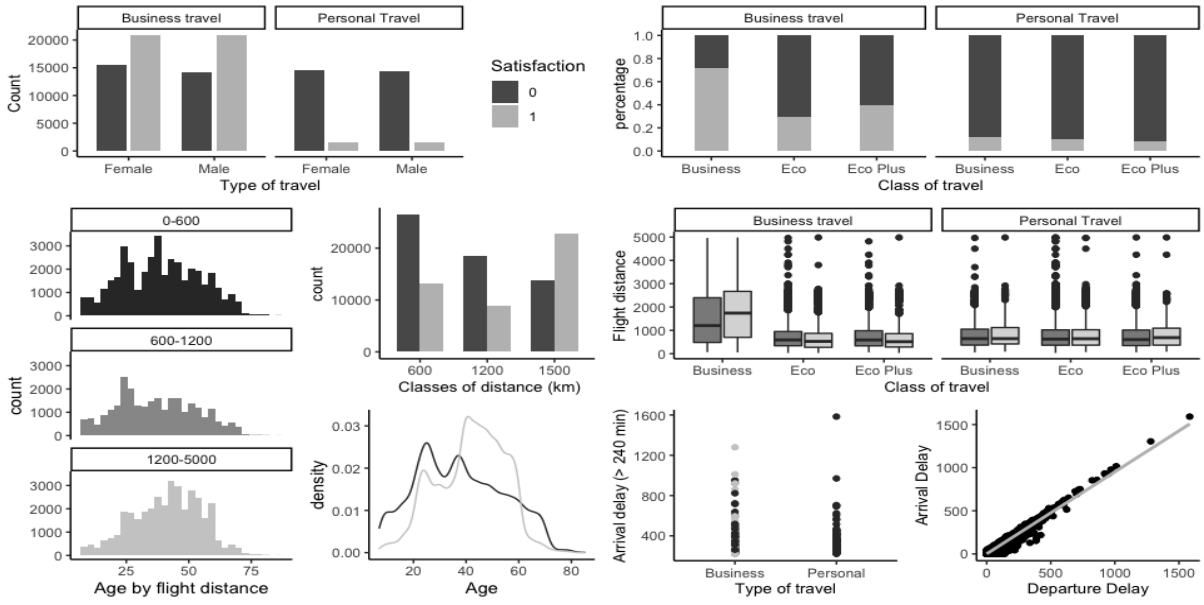
[1] For the analysis aim, the continuous variable `flight.distance` has been categorized and divided into 3 classes (0-600, 600-1200, 1200-5000).

marized by the plots in Figure 1 and corroborated by the values resulting from chi-squared test and Pearson correlation index.

Variable	Description	Type	Outcome
`satisfaction`	Customer satisfaction	Binary	1/ 0
`Gender`	Customer gender	Binary	Male/Female
`Customer.Type`	Customer type	Binary	Loyal/Disloyal
`Age`	Customer age	Integer	
`Type.of.travel`	Type of travel	Binary	Business/Personal
`Class`	Class of customer travel	Categorical	Business/Eco/Eco Plus
`Flight.Distance`	Flight distance in km	Integer	
`Departure.Delay`	Departure Delay in min	Integer	
`Arrival.Delay`	Arrival Delay in min	Integer	

Table 1: Description of `satisfaction` dataset.

Figure 1: A `gender`-wise distribution of `satisfaction` (*Top left*) reveals that this characteristic is irrelevant to the prediction purpose, unlike `type.of.travel` that has a clear impact. Looking at the distribution of `satisfaction` by `class` and `type.of.travel`(*Top right*), Business class passengers that travels for Business reasons have a tendency to be satisfied than passengers travelling for Personal reasons, regardless of the `class`. Also `Age`(*Bottom, second from left*) has a weight: in age range 39-60, number of satisfied passengers is higher. Its distribution varies also according to the `Flight.distance` (*Bottom left*), supported by a chi-square value that allows to reject the null hypothesis of independence. The highest class of `Flight.distance` is the one that stands out for the greatest number of satisfied customers. A more in-depth analysis revealed that this event is mainly due to Business customers travelling in Business class (*Center right*). `Departure.delay` and `arrival.delay` have a large and statistically significant positive correlation, ρ:+0.98 (*Bottom right*). For high values in min of `arrival.delay`, Personal travels record almost the totality of dissatisfied customers.



3 a. Method

The approach that has been adopted to fulfill customer requirements is strongly related to the Bernoulli distribution of the event outcome to forecast. *Logistic Regression* appears to be the most suitable model to explain the true DGP, that allows to model the probability of success (`satisfaction = 1`). Indicating with i the ith customer:

$$p_i = E(Y_i|X_i) = \frac{1}{1+\exp(X_i'\beta)}, \qquad p_i \in (0,1) \tag{1}$$

As part of a larger class of algorithms known as Generalized Linear Model (glm), logistic regression uses maximum likelihood estimation for the estimation of model parameters. The available data set has been split into train and test set (80% and 20% respectively). Since the focus the study is on people traveling for Business purposes, the test set was then purified by *Personal travel* customers. We resulted in having 103903 observations in this training set and 17980 in the *Business travel* test set.

A *full model* has been fitted on the training set and subsequently has undergone changes and improvements. The assumptions' check revealed non-linearity and multicollinearity problems: the first specification of the model (`I`) considers the interaction between predictors and the elimination of redundant information (AIC_I:99178). Taking into account the insights of the EDA and statistical tests (e.g. Analysis of Deviance, Wald's test), the model was further simplified (`II`) performing a variable selection with Lasso's penalization (AIC_{II}:99345). Another model (`III`) was fitted on a training set composed only by *Business travel* customers , with 71465 observations (AIC_{III}:78149).

`(I)`
$satisfaction_i = Customer.Type_i,\ Age_i,\ Type.of.Travel_i,\ Class_i,\ Flight.Distance_i,$
$Arrival.Delay.in.Minutes_i,\ (Age : Flight.Distance)_i,\ (Type.of.Travel : Class)_i$.

`(II)`
$satisfaction_i = Customer.Type_i, Type.of.Travel_i, Class_i, Arrival.Delay.in.Minutes_i,$
$(Type.of.Travel : Class)_i.$

`(III)`
$satisfaction_i = Age_i, Customer.Type_i, Class_i, Arrival.Delay.in.Minutes_i, Flight.Distance_i,$
$(Flight.Distance : Age)_i.$

4 a. Analysis

All three models have been tested *out-of-sample* (Table 2). Performance results were compared computing several indicators[2].

Models	Accuracy	Sensitivity	1-Specificity	AUC
I	0.7288	0.7648	0.3226	0.7746
II	0.7256	0.7417	0.2972	0.7743
III	0.7288	0.7711	0.3314	0.7724

Table 2: Prediction performance evaluation, test set of business travels

[2]The *cut-off point* has been set at 0.5 to achieve the best trade-off between False Positive Rate and True Positive Rate.

With slightly lower predictive accuracy, the model with less predictors (II) is preferred over model (I), according to the KISS principle. Moreover, model (II) results in a lower false positive rate compared to the other two models. Despite the greater number of observations of the original training set, model (I) does not outperform model (III) in accuracy; this result suggests that the presence of additional information about *Personal travel* customers does not improve the prediction performance on *Business travel* customers.

1 b. Problem statement

The Atlanta Airport administration wants to optimize the schedule of near future flights. In order to find an effective solution to achieve its client's goal, 'Fore-cats' decided to predict the daily average delay of all the departing flights from Atlanta Airport.

2 b. Dataset

The data set used was collected and published by the DOT's Bureau of Transportation Statisticsis collected by the US Bureau of Labor Statistics. It contains 5819079 total flights departed from various American Airports in 2015 and a set of 31 variables regarding the characteristics of the single flight (e.g. the origin and destination airport, the departure and arrival delay in rounded minutes). The data was filtered by selecting only the flights departing from Atlanta. Cancelled and diverted flights was deleted, considered as rare exceptions ($<0.06\%$). According to the aim of the study, the original data set was deprived of useless variables and changed by modifying and creating other variables. The result was a new data set composed by 376015 observations of flights departed during the year, characterized by 4 variables (Table 5).

Variable	**Description**	**Type**	**Outcome**
`date`	Date	date	yy-mm-dd
`day_of_week`	Day of the week	Categorical	Monday/.../Sunday
`weather_delay`	Delay caused by weather	Integer	
`total_delay`	Flight arrival delay	Integer	

Table 3: Description of `flight1` dataset.

Some quantities of interest were computed, such as the average `total_delay` for day of week, the daily average `total_delay` caused by weather, the daily average `total_delay` . Basing on the administration goals and interest, the daily average `total_delay` was deprived of the daily average `total_delay` caused by weather. The reason of this choice lays on the fact that weather is nearly unpredictable and independent from the Airport's administrators will. This vector quantity, addressed from now on as `average_delay`, is the forecast object of the present study.
The vector of 365 values is then transposed into a time series format (Figure 2).

By looking at the time series behavior and going forward with the analysis it was found out that most of the spikes in the time series correspond to public holidays. Moreover, three periods of high/medium/low density of departing flights were identified.

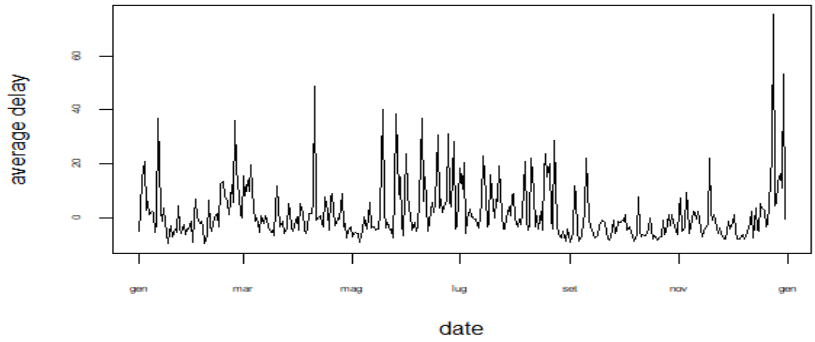


Figure 2: Time series of daily average delays.

There is an evidence of a systematic pattern of the average `total_delay` between the week days, since it seems to be significantly lower approaching to the weekend (friday, saturday, sunday). A new dataframe `dataset2` of 365 rows (year's days) was created, consisting of three new categorical variables: the dummy variable `holiday` means whether we are in correspondence of holidays [3], the three-levels variable `period` indicates whether we are in a high/medium/low density departing flights period, the seven-levels categorical variable `weekday` refers to the corresponding day of the week.

3 b. Method

The time series of 365 observations was split into a training set and a test set, containing respectively the first 345 days and the last 20 days. By using the `auto.arima` function, which detects the models with the lowest AIC, an ARIMAX(1,0,0) was adapted, which also exploits the information of regressors in `dataset2`. The use of those categorical variables is extremely useful in explaining the strong deterministic seasonal component in time series. In order to model also the time series' stochastic seasonal component two SARIMAX models were adapted: a SARIMAX(1,0,0)(0,0,1)7 and a SARIMAX(2,0,1)(1,0,0)7 were chosen by the minimization of the AIC.

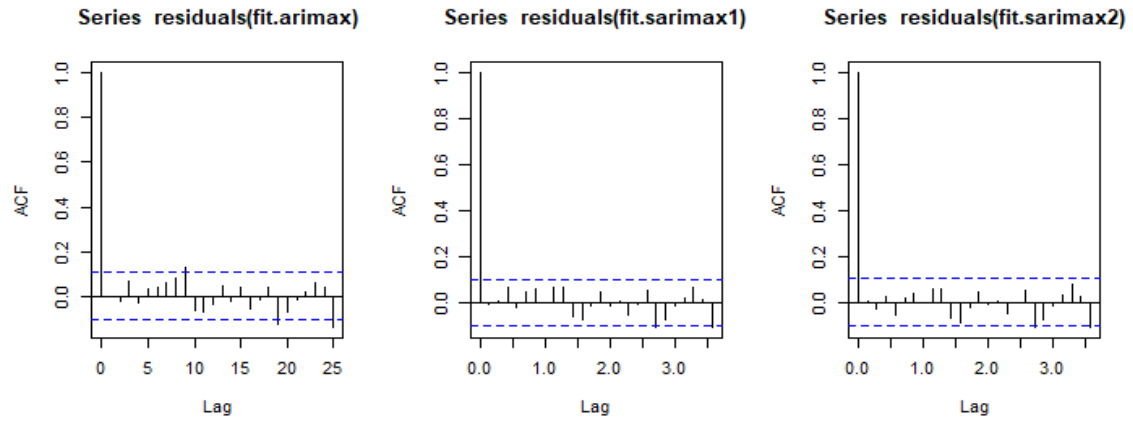


Figure 3: Models' ACF plot.

The ACFs (Figure 3) and PACFs for all three models are deemed good and the zero-

[3] By *holiday* we mean the days right before/right after holidays

mean residuals do not show any particular pattern. According to a Dickey-Fuller test executed on the residuals' series (pvalues<0.01), random values are assumed around zero. Nevertheless, the residuals' distribution is far from normal, so that the Ljung-Box test rejects the null hypothesis of residuals' uncorrelation for each of the three models. This result may be imputed to the limited number of observations available, to the numerous spikes in the time series and to the test sensitivity to the residuals' non-normal distribution.

4 b. Analysis

The best value for AIC is, as expected, the one obtained by the ARIMAX model (Table 4), the quantities for the MSE obtained on the test set result in being lower for the ARIMAX and SARIMAX1 models. Since the aim of the study is focused on prediction, greater importance must be given to the indexes evaluating the prediction performance of the model so, in this case, the MSE.

Models	**AIC**	**MSE**
ARIMAX (1,0,0)	2447.6	259.4319
SARIMAX (1,0,0)(0,0,1)7	2448.35	259.714
SARIMAX (2,0,1)(1,0,0)7	2449.6	271.9507

Table 4: MSE and AIC compared.

The forecast object was observed from a beginning time t=1 to ending time t=T=345. $\{y_t\}_{t=1}^{T=345}$. *h-step-ahead* forecasts have been performed. The short term and long term forecasts, as we expected, do not catch perfectly the time series outliers. The exploitation of the X regressors and the non-differenciated time series allow respectively the long term predictions not to assume the value of the test set mean but to adapt to the real observations, and the long term prediction interval to be less wide and almost constant over time. Basing on the MSE values and the prediction performance for each model, only ARIMAX and SARIMAX1 models' graphics are shown.

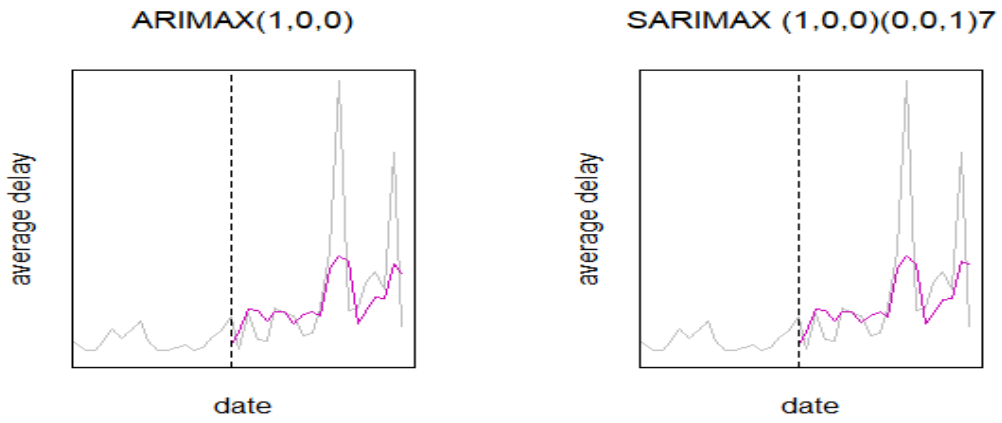


Figure 4: Short term forecasts.
plotted: last 35 days of the year

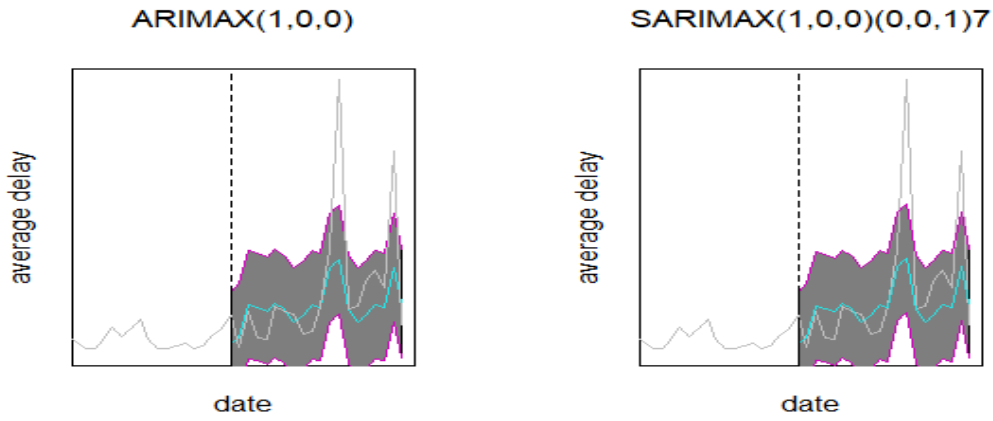


Figure 5: Long term forecasts.
plotted: last 35 days of the year

1 c. Problem statement

The Atlanta Airport administration has another long-term goal: the increase of its customer base. An additional service offered by the Airport website could boost its growth and engagement. The task of 'Fore-cats' is to create an algorithm designed to be used by business customers. Given some flight characteristics it provides an estimate of the flight delay.

2 c. Dataset

The dataset chosen to implement the cross-sectional analysis is the same used for the previous study. Again, data have been filtered by selecting only the flights departing from Atlanta Airport deprived by cancelled flights. However, different variables have been considered during the study, as the business objective changes. Due to the large number of observations, also the destination airports and the airlines were filtered. A subset of 20000 observations, starting from the original 29731 rows, was obtained.

Variable	**Description**	**Type**	**Outcome**
`total_delay`	Flight arrival delay	Integer	
`destination_airport`	Flight destination airport	Categorical	Orlando/.../Chicago
`airline`	Flight airline	Categorical	Spirit/.../Delta
`day_of_week`	Day of the week	Categorical	Monday/.../Sunday
`month`	Month of the year	Categorical	January/.../December
`holiday`	Holiday proximity	Binary	Yes/No
`Departure_hours`	Flight departure hour	Integer	
`weather`	Weather during the flight	Binary	Good/Bad

Table 5: Description of `flight2` dataset.

The forecast object `total_delay` corresponds to the flight arrival delay increased by 70 minutes, to make the delay always positive: flights in advance were registered with negative delay. The distribution of `total_delay` presents many value close to 70, that correspond to the scheduled arrival (null delay). Taking the log, a distribution as much normal as possible was obtained, even though it is still significantly skewed. The exploratory analysis highlighted especially the strong association of flight delay and the holidays' imminence, and the delay trend on `departure_hour` and `day_of_week`. As expected, early morning flights do not collect a large quantity of delay thanks to lack of traffic.

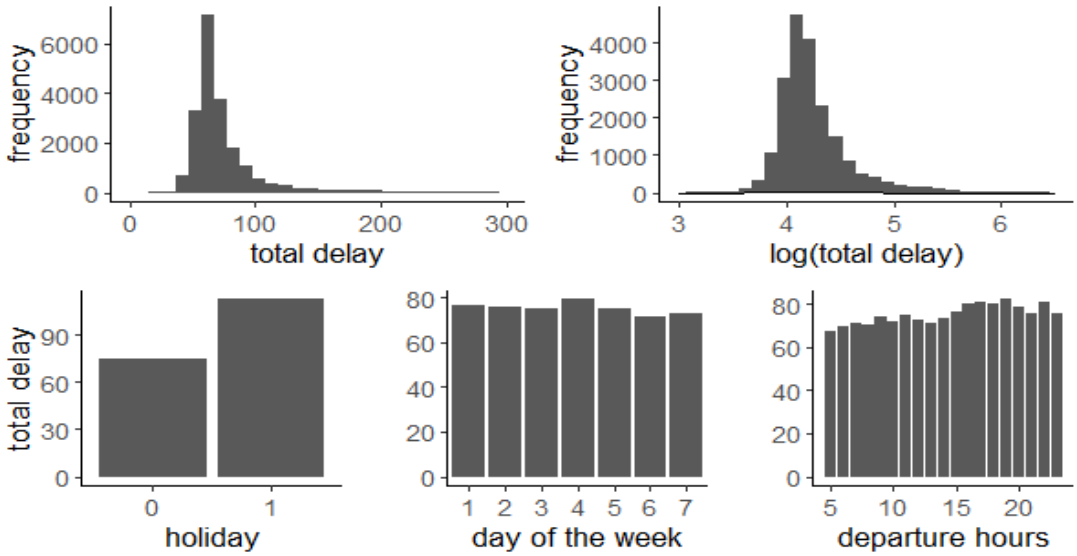


Figure 6: `total_delay` and log(`total_delay`) histograms (*Upper side*)
Average of the `total_delay` variable by `holiday`, `day_of_week`,`departure_hours` (*Bottom side*).

3 c. Method

Taking into account Airport's customers preferences and needs, 'Fore-cats' decided to provide as algorithm output the maximum delay of the customer's plane, instead of the average delay. If the choice was to predict the latter, even if the predictive model was unbiased, it would be correct only on average. This could be a problem for the output interpretation by the future passenger. On the contrary, showing the maximum delay, the website user will be able to observe the worst scenario according to the characteristics required.

Hence 'Fore-cats' decided to adapt two models for the prediction of the 90th conditional quantile.
The first one is a *Gaussian Linear Regression* model estimated by OLS. Than the 90th quantile was computed:

$$\hat{q}_{r_i}^{LR} = x_i^T \hat{\beta} + \hat{\sigma}\phi^{-1}(90) \tag{2}$$

where $\phi^{-1}(90)$ is the 90th quantile of the standard normal distribution, y_i is the dependent variable and x_i a vector of covariates.

The second is a *Quantile Regression* model:

$$y_i = x_i^T \beta + \epsilon_i \tag{3}$$

with $Q_{90}(\epsilon_i) = 0$. It has been estimated by minimizing the check loss function.

To complete the analysis the prediction performances of the linear regression model were compared with those of others regression models for average flight delays.

4 c. Analysis

The available sample was split in training set and test set (70% and 30% of the total observations respectively). The specification of the regression models uses the logarithm of the total delay as a dependent variable with the shown regressors.

$$\log(total\_delay)_i = \\ destination\_airport_i, airline_i, day\_of\_week_i, month_i, holiday\_i, Departure\_hours_i, weather_i.$$

Interaction terms are not included in the model as they did not improve the accuracy. In both models statistically significant relationships were found between the dependent variable and all the regressors taken into account.
. The `destination_airport` is statistically significant in determining the flights delays. Moreover, flights with weathers issues, flights departing during holidays or in summer register greater delays. The same pattern seems to exist for flights managed by Spirit Airlines. A passenger departing during the weekend would expect a lower delay. The `departure_ hour` presents a linear relationship with the flights delay (for this variable polynomial trends were also included in the models without success).
Using *in-sample* estimates of linear and quantile regressions, *out-of-sample* predictions were obtained for the 90th quantile. To evaluate the prediction accuracy of the models, the following indicator variable is defined:

$$I_i = 1(y_i < q_{\tau,i})$$

If $q_{\tau,i}$ is the 90th conditional quantile of y_i given x_i , then I_i follow a Bernoulli distribution with τ as a parameter. To test the goodness of predictions a likelihood ratio test was performed on the null hypothesis that $E(I)_i$ is equal to τ (Table 6).

Model	**Percentage**	**P-value**
Linear regression	0.0874	0.001
Quantile regression	0.1001	0.9658

Table 6: Percentage of out of sample violation and p-value for the test $H_0 : E(I_i) = \tau$

Predictions obtained from the quantile regression outperform those obtained with the gaussian linear regression model.
As expected, the linear regression doesn't fit well available data since the dependent variable has a skewed not normal distribution. This intuition is also confirmed by the heavy tails of the QQ-Plot (Figure 7).
Different type of models for the average delay predictions were also adapted. Several linear model specifications and other models (e.g. Poisson regression) were fit but their prediction performance didn't differ significantly from the gaussian linear regression. Moreover,

considering as main business objective the prediction of the maximum delay, relying on quantile regression seems a good choice since it doesn't assume any parametric distributions.

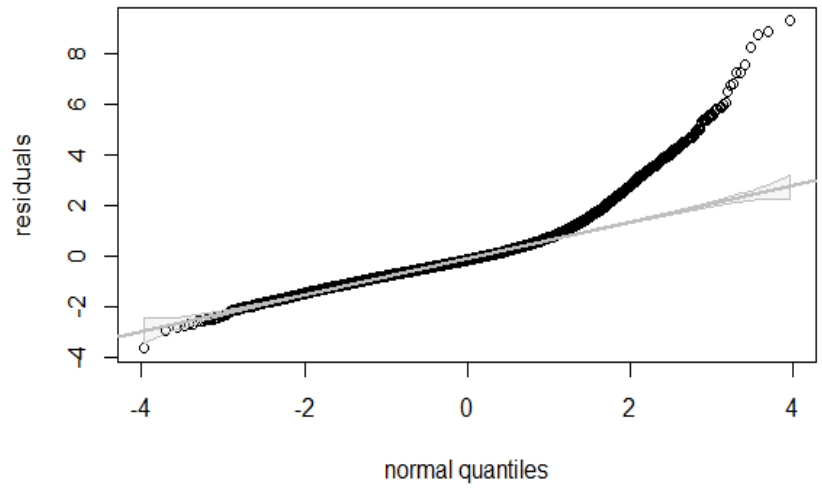


Figure 7: Normal QQ-plot of the residuals from the linear regression model

References

https://par.nsf.gov/servlets/purl/10203524
http://www2.uaem.mx/r-mirror/web/packages/robustbase/robustbase.pdf
https://tech.instacart.com/how-instacart-delivers-on-time-using-quantile-regression-2383e2e03edb
https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf