TAMIDS Data Science Competition 2020

The Culprits of flight delays
- What causes flight delays?

Team: AeroNet
Members: SangHyeon Lee, Ming-Hsien Chien,
Yusu Wang, Hung-Yi Chen , Bo Wang.

Content

1. Introduction

Flight delays have been increasingly under the public scrutiny and have long been the focus of researchers in different areas. Flight delays are costly for both airlines and their passenge rs. For airlines, flight delays lead to extra operational (e.g., fuel costs and crew costs) and m aintenance costs. For passengers, flight delays can result in efficiency loss and potential opp ortunity loss. According to a report sponsored by the Federal Aviation Administration, the dat a from 2007 shows that domestic flight delays brought a $32.9 billion loss to the U.S. econo my, with $8.3 billion direct costs to airlines and $16.7 billion direct costs to passengers. The r eport also estimated a $4 billion loss of GDP due to flight delays. In this report, we want to a ddress this costly problem by generating an effective way to predict flight delays. Only after u nderstanding what are the intric drivers of flight delay can managers make better scheduling decisions to balance cost and benefits.

Flight delays can be attributed to many causes such as carriers' scheduling decisions, weath er conditions, national air system controls, security concerns, etc. Some of the causes can b e out of the carriers' control, but some others are not. However, although carriers have stron g motivations to reduce cost caused by flight delays, It is not always to the best interest of ca rriers to reduce delays, because there are also opportunity costs associated with early arriva l. For example, an early arrival can cause the aircraft to stay idle without generating revenue, or may lead flight time to be changed to an undesirable time slot. What's more, there are als o costs associated with gate and ground crew management at the destination airport. In this sense, flight delays are influenced by so many controllable and uncontrollable factors and ca n be hard to predict.

In this report, we summarized some of the most relevant causes of flight delays and utilized t hose factors to predict future flight delays. We use airline data from the US Bureau of Transp ortation Statistics. Based on our analysis, we found that flight delays are highly correlated wit h and thus can be best predicted by features including scheduled departure time, the origin a nd destination regions, carriers, and number of employees that airlines have. More specifical ly, we valued the importance of features by obtaining their coefficients in our regression mod els. We then testified their contributions to the prediction of flight delays by creating several d atasets and running a DNN model using each dataset. The managerial implications for differ ent parties is also discussed in this report.

2. Data and Variables

A total of 2 million samples out of 11 million airline data from the US Bureau of Transportatio n Statistics are used for our analysis. We get our data by randomly taking data from the shuff

led whole data set in order to reduce bias. Instead of using the continuous delay time, that is, the difference in minutes between scheduled and actual arrival time, we choose the categori cal delay time as our output, which is the Arrival Delay Intervals categorized by every 15 min utes from -15 minutes to 180 minutes, leaving us a dependent variable of delay group categ ories, only the first category indicates arrival on time. In this way, every data point, including flights arriving in advance or on time, can be utilized for analysis. Using categorical delay tim e helps us train a model with higher accuracy rate. In our data set, there are disproportionate ly more short delays than long delays, which means that long delays are underrepresented i n the total data set. To make sure that each delay category receives equal weight of conside ration, we equalized the number of data in each delay category. As a result, we use equally distributed data by the number of categories as our training data and 30% of the data as our testing data to test our model loss and accuracy.

We used individual flight instead of a flight number as the unit of analysis so that we can ana lyze attributes that may contribute to the difference in delays within the same flight numbers. For instance, flights sharing the same flight numbers but of different departure day or time co uld accordingly demonstrate different systematic delay patterns. Thus, we treated flights sep arately and used individual flight as the unit of analysis.

2.1. Data Cleaning and Normalization

Historical flight delay dataset from the Bureau of Transportation Statistics include the cancel ed flight information that is lacking any meaningful data. Also, since some routes don’t have f are information in the AirFare dataset, some flight records used those routes can't get their f are information when we merge the fare and market share information to the flight delay data set. Those kinds of deficient data would mislead our model training and data interpretation. S o, we removed some deficient flight records accordingly.

The attributes of the flight delay dataset vary depending on their types and means. Some attr ibutes are categorical values and the others are numerically related values. We can use cate gorical value by using encoding methods, but the scale of the numerically related values are highly different. These scale differences will give us biased and poor results because they wi ll be calculated together in regression or neural network models. Thus, we normalized numer ically related attributes for the model to be trained on unbiased data.

2.2. Drivers of Delay

Data visualization is the graphic representation of data. In this part, we encode numerical dat a by dots or bars, in order to make complex data more accessible, understandable and usabl e. Firstly, we list and analyze some main driving features for the flight delays, and plot the rel ationship between these features and the arrival delays from the raw data (FlightDelays.csv) of 2018Q1 to 2019Q2. Also, the relationships between arrival delays and the additional drive rs of delay are listed in the second part. If there is a great amount of dispersion on the arrival

delays on a specific feature domain, we could infer this feature to be a great indicator for pre dicting the flight delays.

2.3. Main Drivers of Delay

The arrival delay can be directly attributed to four types of delays: Carrier Delay, Weather De lay, National Air System Delay and Security Delay. Different features are related to these fou r types of delays, and impact the flight delay in turn. As the main drivers of delay, four featur es including Origin State, Destination State, Carrier and the CRS Departure time have great contributions to the degree of flight delays.

Origin State and Destination State
We include the Origin and Destination airports in our input features as potential drivers for N ational Air System Delay and Security Delay. Different Origin and Destination airports have d ifferent airport capacities, for instance, the airline congestion or the bandwidth of security che ck are different between different airports. Therefore, introducing Origin and Destination airp orts in our input features captures the variation in National Air System Delay and Security De lay.

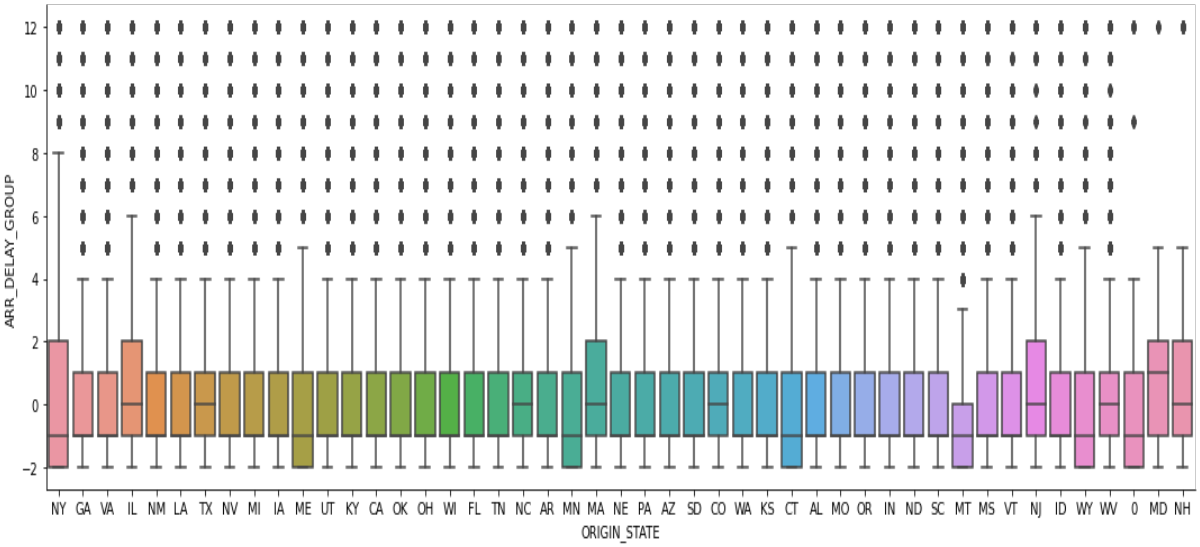

Figure 1: The delay distribution with respect to ORIGIN_STATE from FlightDelays.csv.

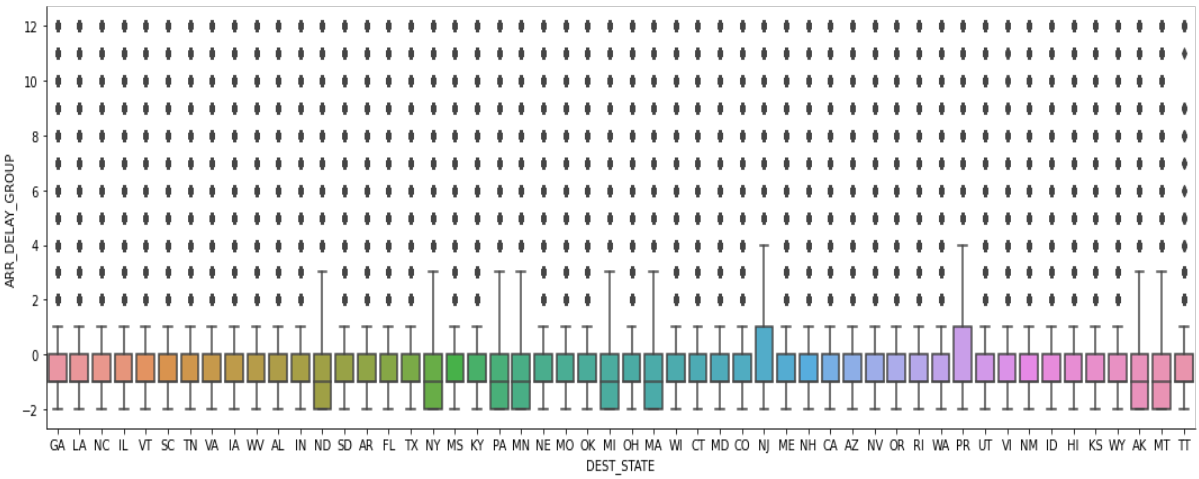

Figure 2: The delay distribution with respect to DEST_STATE in FlightDelays.csv.

Carrier

We use the Carrier information to capture the carrier specific features that could impact the a rrival schedule decision, which is directly related to Carrier Delay. We do this because the m anagement efficiency and marketing strategy differ from carrier to carrier. For example, budg et airlines have higher probabilities of arrival delay because their customers are price sensiti ve and are willing to accept delays in exchange for a better deal. Nevertheless, being punctu al is more essential for some commercial-passenger-oriented carriers, since their goodwill is built highly based on their services and customer satisfaction. Thus, Carrier information is a r easonable feature to predict Carrier Delay, which directly affects the arrival delay.

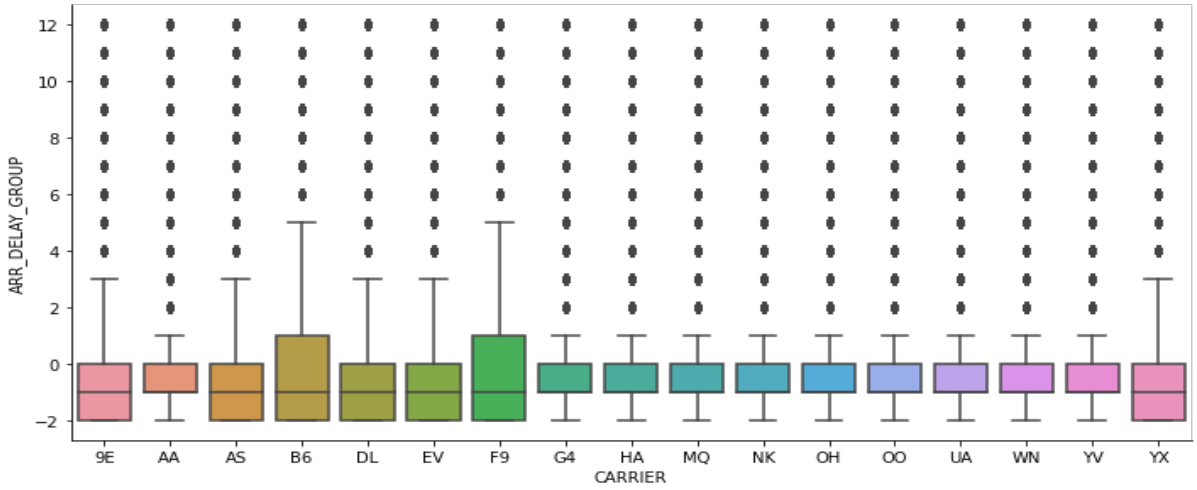

Figure 2: The delay distribution with respect to CARRIER in FlightDelays.csv.

CRS Departure time

We include the CRS Departure Time, which is a scheduled departure time for each flight, as potential drivers for National Air System Delay, Carrier Delay and Weather Delay. Periods of high traffic congestion usually occur when the demand for travel is elevated. For example, ai rlines usually schedule a large number of flights in the 1pm-9pm interval as departure time in order to meet business travel demands. Besides, the probabilities of air turbulence are differ ent in each hour of a day, and everyone knows that the turbulence is a critical factor to flight delays. Since the situations stated above can lead to National Air System Delay, Carrier Del ay or Weather Delay, it is reasonable to include CRS Departure time feature to our predictio n model.

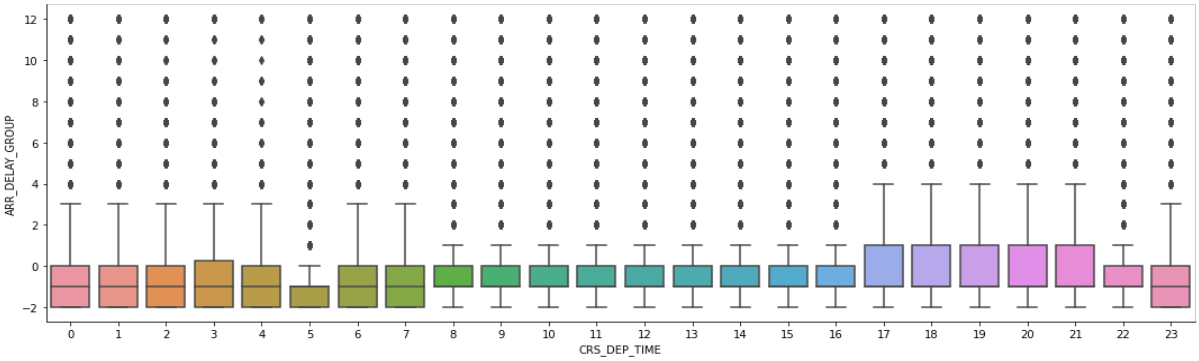

Figure 4: The delay distribution with respect to CRS_DEP_TIME in FlightDelays.csv.

Full time equivalent employees (EMPFTE)
The total number of full time equivalent employees can reflect the service quality of flights on a route. If the number of full time equivalent employees cannot afford the on/off-boarding de mands, it may affect the Carrier Delay and cause the flight delay. Thus, we include the EMP FTE as a feature to our training model.

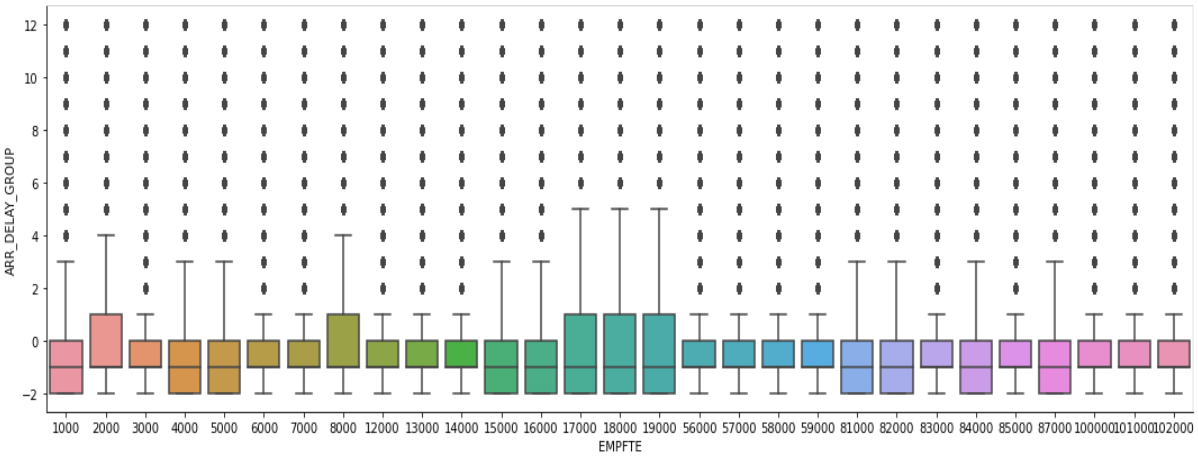

Figure 5: The delay distribution with respect to EMPFTE in FlightDelays.csv.

2.4. Additional Drivers of Delay

In terms of the additional drivers of delay, namely, drivers of the heterogeneity in airline sche duling decisions, three factors are included: operational consideration, competition, and reve nue maximization.

Hub
We added the attribute of hub (Randy Wang, Trip+ 2017) to denote the airline's operational co nsideration. It has been demonstrated that the probability of delay increases if the origin or d estination airport is a hub, because the network benefits at hubs outweigh the cost of delays (Mayer and Sinai, 2003; Deshpande and Arikan, 2012). Thus, whether the origin or destination air port is a hub is a non-negligible driver of flight delay.

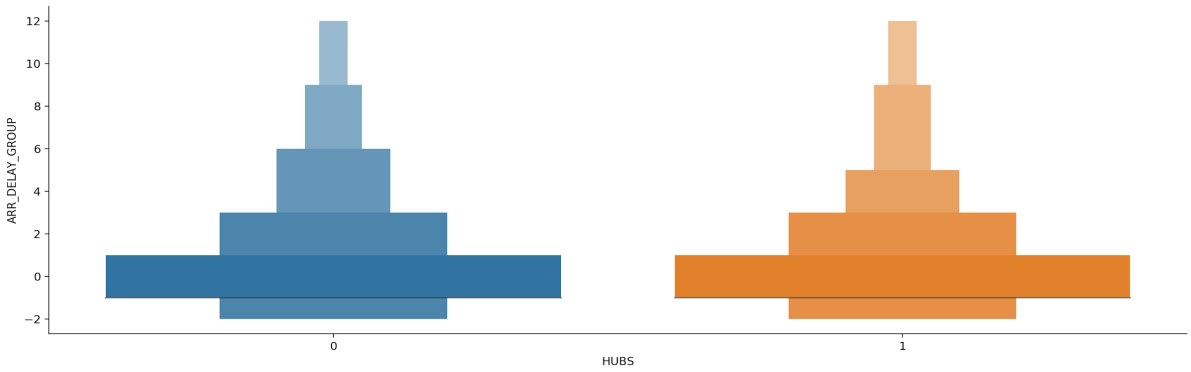

Figure 6: The delay distribution with respect to Hubs in flights (2018Q1-2019Q3).

Proportion of Market Share

We used the proportion of market covered by carriers with largest market share to capture th e competitive environment of certain routes. Competitive environment has a great impact on the airline's schedule decisions. The marginal cost of delays decreases as one carrier's mon opoly power gets stronger, because the potential revenue loss caused by disappointed pass engers switching from an airline to another airline becomes inconsiderable. As a result, the c arriers with high market share have less motivation to control delays. We chose the proportio n of market covered by the carrier with the largest market share to capture the competitive e nvironment because the largest market share covered by a single carrier reflects the market concentration. Typically speaking, the higher the market share, the higher the market concen tration. Thus, the proportion of market covered by the carrier with largest market share is a g ood indicator of market competition of a certain route.

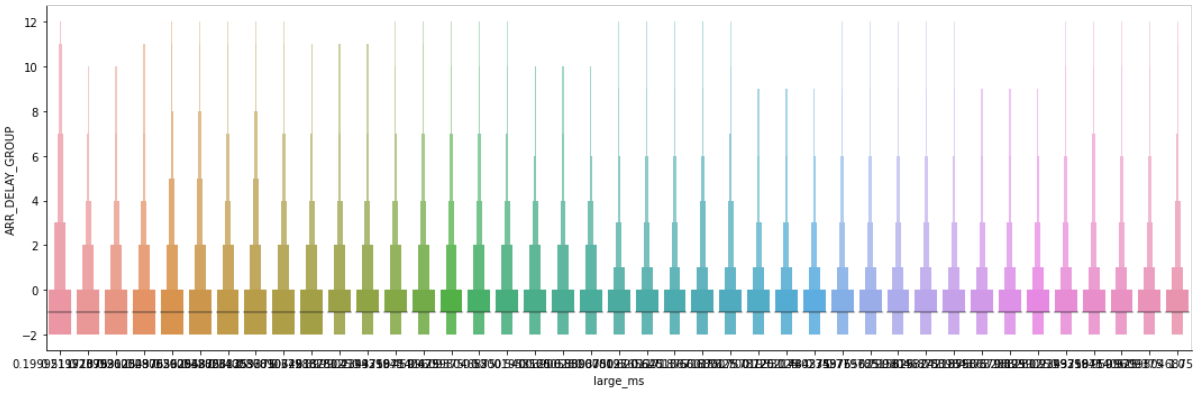

Figure 7: The delay distribution with respect to large_ms in FlightDelays.csv.

We calculated the average fare per mile to capture the airline's revenue maximization motiva tions. It is calculated by dividing the fare by the total distance between the origin and destinat ion of a certain flight. It has been shown that revenue has a significant impact on scheduled on-time arrival probabilities (Deshpande and Arikan, 2012). That is to say, revenue maximizatio n motivation is one of the most important drivers in airline scheduling decisions and is strongl y correlated to flight delays. Thus, it is reasonable to predict that the average fare per mile is a potential driver for flight delays.

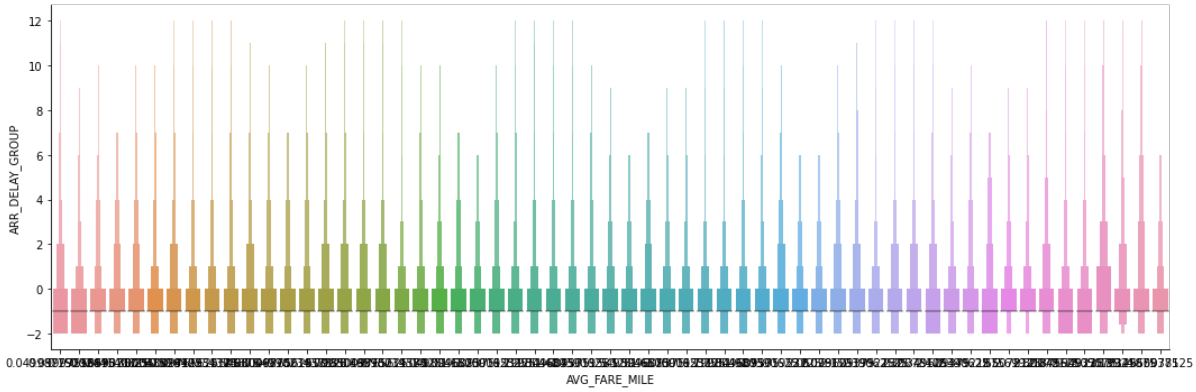

Figure 8: The delay distribution on average fare per mile vs. flights (2018Q1-2019Q3).

3. Model Specification

We used regression models to mathematically calculate which features are important for predicting the flight delays in the data and evaluated our assumption which features are important by training a neural network model with different feature combination dataset.

3.1. Regression Model

In statistics and data analysis, the regression model is a set of statistical processes for estimating the relationships between an outcome and some independent features.

Compared to the simple linear regression, adding penalty to the model helps us prevent overfitting problems and conduct calculations more efficiently. Therefore, we adopt the following two regression models with penalty .
The first model is Lasso regression. The equation is shown below.

$$\sum_{i=1}^{M}(y_i-\hat{y_i})^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

This type of regularization can reduce the feature coefficients to zero when the penalty term is sufficiently large. That is to say,some of the features are completely neglected for the evaluation of output. Thus, Lasso regression is useful not only in reducing model overfitting problems but also in performing feature selection.

The second model is the Ridge regression model. The equation is shown below.

$$\sum_{i=1}^{M}(y_i-\hat{y_i})^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

In Ridge regression, the cost function is altered by adding a penalty equivalent to the square of the magnitude of the coefficients. Compared to the simple linear regression, the penalty of the model can prevent overfitting and make the calculation more efficient. Furthermore, ridge regression puts constraints on the coefficients (w). The penalty term (insert equation later) regularizes the coefficients such that the coefficients are shrunk when they take large values. As a result, the model complexity is reduced through this process.

Finally, the coefficients of regression are used to indicate the importance of the feature. the larger the coefficient,the more crucial it is to predict flight delays.

3.2. Deep Neural Network (DNN) model

Deep Neural Network (DNN) model is used to predict flight delays and to evaluate features t hat are chosen by the Ridge regression model.

Deep Neural Network is a supervised learning algorithm used to train a function that can find a labeled output based on input feature patterns. DNN has multiple hidden layers and nodes with their own weight (W) and bias (b) values. DNN helps to find the best weight and node fo r each node by comparing the calculated output (i.e. estimated output) and labeled output. W hile the nodes transfer the weighted input, they use an activation function (f) to change the d ata into a non-linear format to keep the DNN network model. If the nodes don't use an activa tion function, the estimated output would be a linear model that is similar to a one-layer neur al network, Y=WX+b.

Figure 9 shows an example of DNN architecture with three hidden layers. X is an input vecto r and Y is a calculated output vector. In the hidden layers, each node gets input from the pre vious layer and multiplies a weight and adds a bias. Then, an activation function makes the li near equation into a nonlinear format.

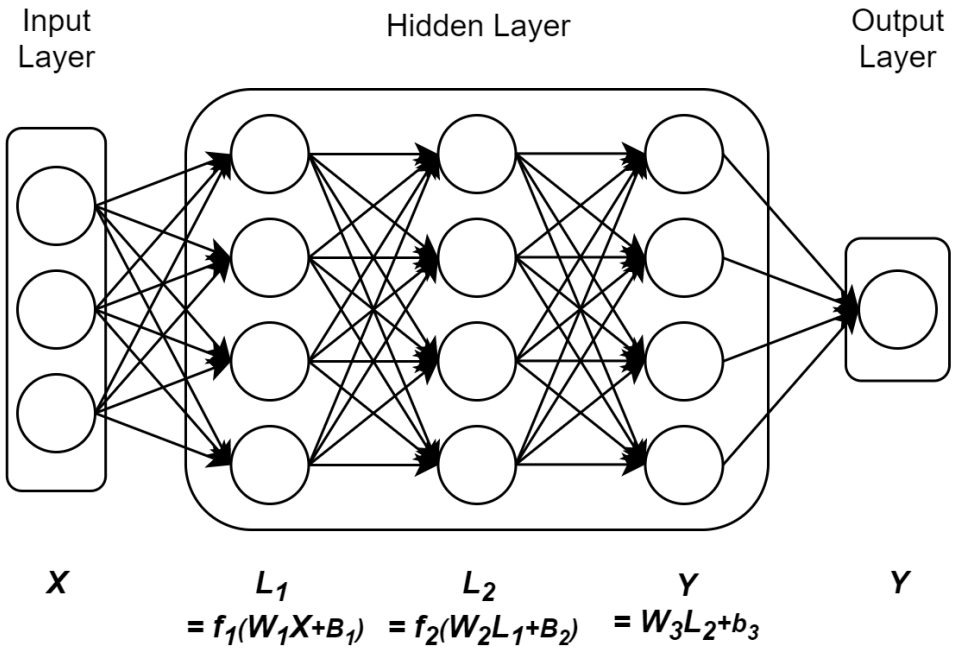

Figure 9:. The Deep Neural Network architecture

Once the model has compared the estimated output and labeled output using loss function, t he error is back-propagated to nodes in the neural network. In the back-propagation, the mo del finds the derivatives of weights and bias and then updates weights and bias to minimize l oss using an optimization algorithm.

Our cost function takes the following format,

$$cost(w) = \frac{1}{2m}\sum_{i=1}^{m}(W_x^{(i)} - W_y^{(i)})^2$$

The equation shown below is the Gradient Descent algorithm in Backpropagation.

$$W = W - \alpha\frac{\partial}{\partial W_{ij}}cost(W)$$
$$b = b - \alpha\frac{\partial}{\partial b_{ij}}cost(b)$$

To train our prediction model, We used the Multi-layer Perceptron classifier model in Scikit-le arn Library and set parameters to increase its performance.

4. Experiments

4.1. Regressions

To reduce bias, we adopt a k-fold cross-validation. It indicates that the original sampl e is randomly partitioned into k equally sized subsamples. Out of the k subsamples, one single subsample is regarded as the validation data for testing, and the remainin g (k − 1) subsamples are used as training data. The same cross-validation process is then repeated for k times, with each of the k subsamples being used exactly once as the validation data. As a result, we got k results, which were averaged to produce a s ingle estimation. The advantage of this method is that the subsample acts both as th e training data and testing data. Furthermore, we also put the original data and the d ata, which are scaled down to be in the range of 0 and 1, as the input of the models of regression to identify the critical features.To be more specific, we take the coeffici ents estimated by Ridge and Lasso to evaluate the importance of features.

4.2. Deep Neural Network (DNN)

Based on our data visualization and the regularization of the regression model, we made diff erent feature combinations of the dataset. Each dataset has different features which are com

bined by our previous data analysis. So, we made new datasets to evaluate the importance of the highlighted features by comparing the training performance.

baseline data
Raw dataset which has the following features; MONTH, DAY_OF_WEEK, CARRIER, ORIGIN_STATE, DEST_STATE, carrier_lg, CRS_DEP_TIME, HUBS, AVG_FARE_MILE, and EMPFTE. Also, CARRIER is encoded by one-hot encoder for the model to distinguish its categorical difference.

norm data
norm data is the normalized dataset from Baseline data. The normalization is only applied to continuous features, MONTH, DAY_OF_WEEK, CRS_DEP_TIME, DISTANCE, fare, large_ms, EMPFTE, and AVG_FARE_MILE.

no_crs data
no_crs data is generated from Norm data by removing CRS_DEP_TIME. We made the data to test the importance of feature CRS_DEP_TIME by comparing trained models' results.

no_carrier data
no_carrier data is generated from Norm data by removing CARRIER. We made the data to test the importance of feature CARRIER by comparing trained models' results.

no_state data
no_state data is generated from Norm data by changing state features to airport name. We made the data to test the importance of data clustering. If the input dimension is too large, the curse of dimension will make the model's performance decrease. Since we encode categorical features, the number of unique airports tremendously increased the dimension of input features which would cause the curse of dimension. However, we can simply consider the airports in a state are in a state group. So, we can compare performance differences between models trained with states and with airpots.

We used the Multi-layer Perceptron classifier of Scikit Learn library. The optimizer to reduce model's loss is Adam, a stochastic gradient-based optimizer. The hidden layer size is (20, 20, 20, 20). Also, we added the L2 penalty (regularization term) parameter to 0.01 due to the large number of input features.

5. Results

5.1. Regression Analysis to find Critical Features

By introducing the Lasso Regression and Ridge Regression analysis, we plot the relationshi p of Absolute Coefficient Values across different and Features. In Figure 9, the absolute coef ficient values are used to represent the importance of specific features. In other words, the hi gher the absolute coefficient value of a feature, the more importance of adopting this feature to predict flight delays. According to the result of two regression models, we list out the top 1 7 crucial features among 152 normalized features along with their coefficients.

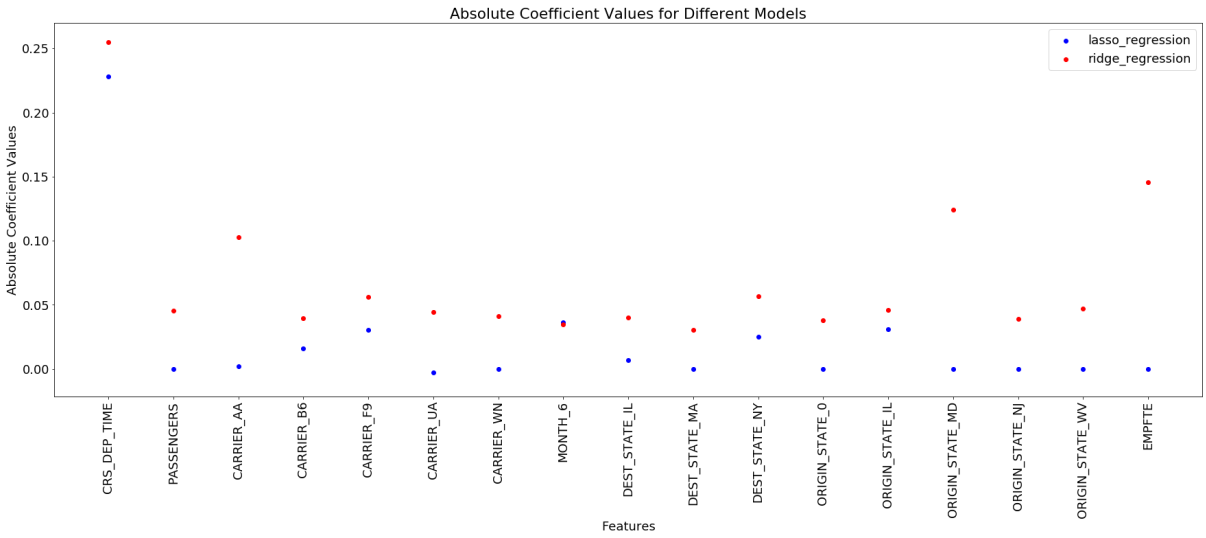

Figure 10: The Top 17 important features and their absolute coefficient values used for predi cting the flight delays. (blue: Lasso Regression; red: Ridge Regression)

The first most important feature is CRS_DEP_TIME, which implies that the later departure ti me is correlated with more flight delays. It has been demonstrated by Deshpande and Arikan (2012) that more passengers choose to take flights at night. Late flight departure time sched

ulesIt might lead to more flight delays because every airport needs to handle the higher load of passengers. In addition, the number of passengers also plays an important role in affectin g the flight delays.

Furthermore, we found that flight delays are much more pervasive in certain the Origin a nd Destination States have a significantly positive correlation with flight delays. Especially origin and destination states such aslike NY, MA and IL have several airports that ar e among some of the most bustling US airports (JFK, BOS, ORD) in 2018 and -201 9. That is to day,It implies that more flight delays can be observed when the origin or destination falls into the regions airports with having extremely heavy load transporta tion will contribute more to the flight delays.

On the other hand, EMPFTE (The total number of full time equivalent employees) is a feat ure that is most negatively correlated with flight delay. We interpret this result by suggesting t hat the more thorough services will be provided to customers when there are more employe es on a route, and probably the higher service quality incorporates enhanced efficiency in se curity check and gate management, will provide more thorough services to the customers an d reduce leading to a lower chance ofthe flight delay. We display the absolute value of EMP FTE in the above figure because it's an easier way to compare its importance with that of oth er features.

5.2. DNN Analysis to evaluate the Critical Features

We used test losss value to evaluate each model's performance by comparing the loss whic h indicates how much the model trained well and could find data patterns to predict the flight delay. The norm dataset contains important features derived from our hypotheses, and we al so made other datasets to evaluate each features' importance. It turns out that the result obt ained by trainingour model using the norm dataset is the best compared to using other datas ets.

Dataset	baseline	norm	no_crs	no_carrier	no_state
loss	0.940961	0.934431	0.934693	0.937683	0.941301
Performance rank	5	1	2	3	4

Table 1: Loss value of each model trained with different dataset and the performance rank.

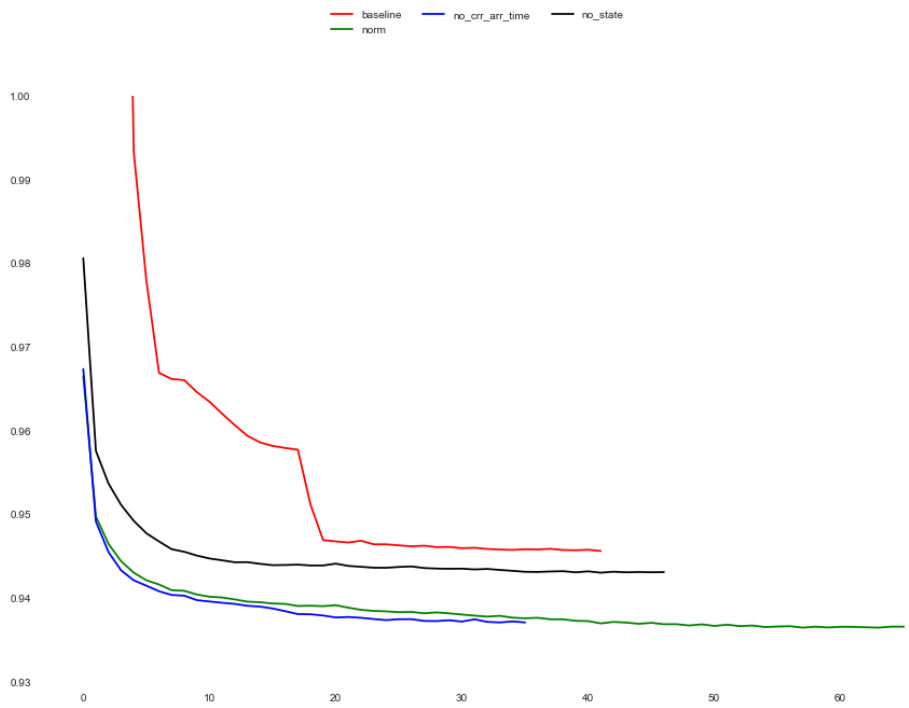

Figure 11: Training loss of models trained with different dataset. X axis is converaged epoch and y axis is the loss. A model trained with norm data (green line) converged at epoch 67 wit h minimum loss among all and another model trained with no_crr_arr_time data (blue line) c onverged at epoch 37 with second lowest loss.

Table 1 is the test results of models that trained with different dataset we made and Figure 1 1 shows the loss curve of each model. It shows that the norm data model has the minimum l oss and the others have larger loss than the norm data model.

First of all, in our analysis, we found that the flight delay frequently happened at night and we suggested that CRS_DEP_TIME is important for predicting flight delays. Since we found that the model trained with data excluding CRS departure time information performs worse than t hat including CRS information, we confirmed that the CRS_DEP_TIME feature plays an imp ortant role in flight delay.

In addition, we found a positive correlation between CARRIER and flight delays by visualizin g the relationship between CARRIER and flight delay and also by analyzing the regression c oefficient in the regression model. Here, by using the DNN model, the result also shows that data with CARRIER features performs better than the one without CARRIER features.

Finally, since the number of airports is enormous and may bring about the curse of dimensio n issue. Thus, we used state information as a parent group of the airport. In our experiment, we wanted to evaluate whether our data clustering can help our model to gain better predicti on performance or not. In our result, we confirmed that the model trained with state informati on predicted the flight delays better than the model trained with airport information.

6. Executive Summary

This report aims to understand the delay patterns of American domestic flights and predict future flight delays using several driving factors. Using the regression model and DNN model, we found and testified that flight delays are highly correlated with and thus can be best predicted by features including scheduled departure time, the origin and destination regions, and number of employees that airlines have.

Our model results have strong managerial importance for airlines to improve arrival scheduling decisions and to decrease delay-related costs for both airlines and passengers. Based on our results, in order to reduce the possibility of delays, carriers may schedule a longer block time when the flights are scheduled to depart within certain time slots of a day, or when the origin or destination airports are in certain states with a high volume of traffic. In addition, our results are also useful for airports to help improve airlines' on-time performance by spending more resources on handling heavy load in certain time or regions, including but not excluded to equipping more ground crews and staff members in the airports. We also provide helpful implications for policy makers by suggesting that certain time periods and some properties of airports may influence the flight delay probabilities, and airline system efficiency could be improved by accounting for these factors.

References

Mayer, Christopher, and Todd Sinai. "Network effects, congestion externalities, and air traffic delays: Or why not all delays are evil." *American Economic Review* 93.4 (2003): 1194-1215.

Deshpande, Vinayak, and Mazhar Arıkan. "The impact of airline flight schedules on flight delays." *Manufacturing & Service Operations Management* 14.3 (2012): 423-440.

Randy Wang, "The Complete List of Hubs and Focus Airports of Major Airlines in North America." Trip+ (2017) https://blog.tripplus.cc/en/30060/hubs-major-airlines-in-north-america