Predicting Flight Delays

Group 5

Michael Carvalheiro, William Friedeman,

Garrett Szczucki, Aidan Surowiec

Table of Contents

1. Abstract:

1.1. *Overview:*

The advancements in air travel have produced a heightened interconnectivity in our society, yet even with the massive efficiency benefits that have been produced it is still not uncommon to run into a number of flight issues. One of the most prevalent issues within this service is encountering a flight delay. Every one of us has experienced a flight delay at some point, and the consequences can range from minor to severe depending on the circumstances. Regardless of the degree, avoiding any flight delay event would be ideal.

There are a number of features and characteristics that could allow us to potentially identify the likelihood of a delay for any given flight. The intention of this project is to determine whether or not an accurate flight delay prediction model can be developed with aims to help end users mitigate this risk of delay as much as possible.

1.2. Data Set Description:

The uncleaned data set has 28 columns and 7,213,446 rows. Some of the predictors in this data set are origin, destination airport, date of flight, airline carrier, planned departure time, distance, and a few others. The primary data set that we are working with will be the flight data from 2018. Due to the size limitation of Github uploads, we took a random sample of data from the original file with 7 million rows. The maximum reduced file that we could upload was 190,000 random rows. This data was obtained from Kaggle at this link Airline Delay and Cancellation Data.

1.3. Predictions:

- For the end user/travel booking sites - predict how likely a flight departure is to be delayed based on airline, season, origin, time of day & more
- For end user/airlines - predict how likely a flight arrival is to be delayed based on airline, season, origin, time of day & more

- For end user/travel booking sites - predict how long a flight is likely to be delayed based on predictors
- For end user/airlines - predict how likely a flight will have a weather delay

1.4. Inferences:

- Does time of day impact the chance a flight will be delayed?
- Does the time of year/date impact the likelihood of a delay?
- How do airlines compare in terms of how often delays occur?
- Are origins/destinations a major factor when calculating a delay likelihood?
- If an origin flight is in an area known for bad weather, will there be more delays?
- Is there an optimal time of day that flights will be least likely to delay?
- Which destinations are highly likely to have a delay?

1.5. Conclusion:

The results of this project are promising, multiple successful models were produced for each prediction goal listed above. The classification model with the highest AUC pertains to our prediction goal of predicting the likelihood of a flight experiencing a weather delay. These models however still have much room for improvement and could be adjusted further. In terms of inferences, we were only able to answer five out of our seven above. To answer the last two points, an Apriori model would need to be introduced. The future flight analysis goals include the addition of an association element, and to have the capacity to run a larger data set so that sample size is not an obstacle. Taking the models a step further with deep learning or neural networks would also be an intriguing next step.

2. Data Collection/Cleaning/Exploration:

2.1. Expanded Dataset Description:

The data in this file was originally collected by the Bureau of Transportation Statistics, of the 28 variables included there are five object types, three integer types, and twenty float types. Eighteen of the float type variables represent either time of day in military format, or minutes. There are multiple variables in this data that could be used as targets, the ones we are focusing on will be; ARR_DELAY: Flight Arrival Delay in Minutes, DEP_DELAY: Flight Departure Delay in Minutes, and WEATHER_DELAY: Flight Weather Delay in Minutes. There are four other delay variables, and there is also data that indicates if a flight was canceled as well as the corresponding cancellation code. These variables are out of the scope of our intended investigation, they will not be included in any of the analysis provided. Certain feature variables were also deemed to be out of the scope of what we want to accomplish. For example, variables such as WHEELS_OFF, and DEP_TIME (actual departure time) both are values that are recorded after the plane has been boarded which practically would not help the end user proactively avoid flights with delay risk.

The desired insights that we plan to uncover stem from variables such as airline, distance, airport origin & destination, expected arrival & departure time, and date. Even with the consideration of boosting our accuracies by using additional variables like wheels off, we concluded that practicality is more important than overfit accuracy. As a result of the ideology, 16 of the original 28 variables were dropped.

2.2. Results of Data Exploration:

Once the data was loaded in, the first piece of notable information gathered was that there was an abundance of null values present in our random sample. For the weather delay variable, there were only 35,602 non-null values. The arrival and departure delay variables also had nulls present, but each had over 186,000 non-null observations. To handle this null issue, nulls in all three delay variables were first filled with 0. Our thought

process was that if the value was null then there was no delay, filling with 0 would allow us to keep the observation. Any other nulls were dropped from the data frame, to illustrate why - filling with 0 doesn't apply in the circumstance of air time, that data point would then become invalid.

It is also noteworthy that some of the variable names are difficult to immediately interpret, specifically CRS_DEP_TIME, CRS_ARR_TIME, and OP_CARRIER. These variables simply mean scheduled (or expected) departure and arrival time of day, and as mentioned earlier OP_CARRIER signifies airlines. Additionally, data within the OP_CARRIER variable cannot be fully interpreted without being decoded. The values within this variable are two character carrier codes, and without an external source there are some code values that are not common knowledge (YX, F9, MQ, etc.). The same type of issue is present in the ORIGIN and DEST columns - some airport codes may be familiar, like ATL or ORD. However, we saw that there are 351 and 354 unique origin and destination values respectively - meaning that the majority of these airport codes will need to be cross referenced in order to extract any useful insights.

The delay variables are each in the format of minutes delayed (or ahead of schedule). In order to predict likelihood of delay, a binary target needs to be introduced. For our three targets (arrival delay, departure delay, weather delay) a binary variable was created that returns a value of 1 if the delay is larger than 0 minutes, and naturally 0 otherwise. If a delay is ahead of schedule and departed or arrived early then the original value in minutes will be negative.

2.3. Data Manipulation:

Following the initial exploration and cleaning of the data set, the next step to understanding and interpreting the information on a deeper level was to manipulate the data through aggregation and binning. The first added columns were extracted from date. Since we cannot run a date value through our prediction models we determined that pulling the month and day of month from the date would be beneficial additions that will allow us to gain insights from the date. There were then three variable bins

performed, the first was creating the flight season variable which was binned based on the newly introduced flight month variable. The resulting bins were winter, spring, summer, and fall. The second and third binning tasks were to feed in the expected arrival and departure time, and to bin in the categories of morning, afternoon, evening, and night.

Now that all the necessary variables were introduced, it was time to develop some aggregations. The primary purpose of this project is to gain insights about factors influencing flight delays, the aggregations were focused on the distribution of delays. The general format of each aggregation was to group by a variable, gather the sum of the binary target value for each delay type, and obtain the count of each group. A percentage was then calculated by taking the sum of the delay target, dividing it by the count of the group and then multiplying by 100. The count signifies the total number of flights observed in each group, and the sum represents the total number of delays observed in each group. This aggregation was executed on the following variables; airline, season, origin, and time of day.

An interesting point that surfaced from examining Figure 5 in the Appendix section was that the majority of the arrival delays were in the range of 0 to 50 minutes, regardless of the time of day. Even though a delay of this duration may seem small, it may still make a difference, especially for travelers with connecting flights. To add to the examination of delays by time of day, we found that the expected arrival or departure in the morning resulted in the lowest delay percentage for all three delay categories.

Origin was another aspect that we wanted to analyze. Based on our aggregations, we saw that of all origin airports, six of the top seven in total as well as percentage of flights with weather delays were in southern states. This negates our first thought that colder states would lead to more weather delays due to snow. Reference Figure 2 to examine how weather delays compare by flight origin.

3. Data Exploration Insights:

One of the most surprising takeaways from the aggregations was that the percentage of delays by season were much closer than anticipated. Initially the consensus was that winter would have the most delays, but from a count perspective winter seems to generally be on the lower end for this sample. To get a better understanding of how departure delays are distributed by season see Figure 3. Arrival and departure delays are all in the 30-40 percent range for every season, this leads us to believe that season may not be as significant of a factor as we believed. It was also startling that certain airlines were experiencing delays in close to half of all their observed flights. The sample may skew these numbers, but the fact that Frontier Airlines, and JetBlue Airways were recording arrival delay percentages of 45.17%, and 41.96% respectively indicates that airline could potentially be the most prominent factor in predicting a flight delay. Perhaps this is a contributing factor to why these airlines are priced relatively lower than others. Refer to Figure 4 to see a visualization of how the percentage of flights that have a departure delay compare by airline. It is also extremely interesting as to why the morning records a significantly lower percentage for all delay types. The underlying reason was at first thought to be weather, but only 0.77% of all flights in the morning experienced a weather delay.

4. Methodology:

To achieve the goals of this project we created four separate prediction sections that use the same variables but have unique targets. There are three classification objectives that require a binary target and prediction, and we have one regression objective that aims to predict the duration of an arrival delay in minutes. The processes between the two forms of prediction did not vary by much, but certain pipeline steps were necessary to test and implement in only one form. For example, the StandardScaler was only implemented in the regression prediction component because variables such as distance and time of day are of completely different scales. Reference Figures 6 and 7 to view the general workflow block diagram, the variations between the processes can

be visualized clearly here. Once an optimal model was identified, the feature importance/weight was extracted to provide further insights. An activity specific to the random forest regressor and gradient-boosted tree regressor was the additional step of indexing the features column; this action was not listed in the block diagram. Each prediction section tested a regression model and a random forest model. The weather delay classification prediction and the arrival delay duration regression prediction both tested a gradient-boosted tree model as well.

5. Model Prediction/Model Inference:

5.1. Predicting How Long an Arrival Delay Will Be Using Regression

This model is predicting the Arrival Delay variable using the following predictors: Airline, Origin Airport, Destination Airport, Season, Distance, Day of Month, Expected Departure Time of Day, Expected Arrival Time of Day, and Month.The goal for these models was to get a picture of which predictors influence the Arrival Delay variable the most.

The initial model used was a Linear Regression model and separately, we created a scaled Linear Regression model using a Standard Scaler to scale all predictors in the model. We then performed a grid search to both Linear Regression models to hyper tune them. The supplementary models used were a Random Forest Regression model and a Gradient-Boosted Tree Regression model, where we implemented another grid search to each model for hyper tuning purposes.

Each of the four regression models were scored with Mean Squared Error (MSE). The best scoring model was the random forest regression after grid searching, with a MSE score of 2217.6132. The parameters for the random forest regression grid were the number of trees equal to 70 and a max depth equal to 5.

5.2. Predict if Flight Will Have an Arrival Delay - Will:

These models predict the likelihood of an arrival delay. The goal is to see which predictors are the most significant in determining this likelihood. The predictors that we

used in these models were: Airline, Origin Airport, Destination Airport, Season, Distance, Day of Month, Expected Departure Time of Day, Expected Arrival Time of Day, and Month.

The initial model we used was a Logistic Regression model. Since this model did not score very well at first, we performed a grid search to hyper-tune it. We then created a Random Forest model, which scored better than our hyper-tuned Logistic Regression model.

For data transformations, we created a boolean variable called IsArrivalDelay, which we renamed as “label”, that was used as the target variable.

We scored the models using AUC scores. As mentioned previously, the Random Forest model had a higher AUC score than the Logistic Regression model, therefore, we can conclude that the Random Forest model was a better model to predict the likelihood of an Arrival Delay. Moreover, we also created ROC graphs for both models, which supported the claim that the Random Forest model was better than the Logistic Regression model.

5.3. Predict if Flight Will Have a Departure Delay - Garrett:

These models predict the likelihood of a departure delay. The goal is to see which predictors are most influential when determining this likelihood. The predictors that we used in these models were: Airline, Origin Airport, Destination Airport, Season, Distance, Day of Month, Expected Departure Time of Day, Expected Arrival Time of Day, and Month.

The initial model we used to predict the likelihood of a departure delay was a Logistic Regression model. We then decided to implement a grid search to hyper-tune the model. We also created a Random Forest model, which, once again, ended up scoring better than our hyper-tuned Logistic Regression model.

For data transformations, we created a boolean variable called IsDepartDelay via binning the DepartDelay variable, and this was used as our target variable.

We scored both the Logistic Regression model and the Random Forest model using area under the curve scores. As mentioned above, the Random Forest model was better than the Logistic Regression model because it had a higher AUC value. Once again, we also created an ROC graph that verified the claim that the Random Forest was the better model.

5.4. Predict if Flight Will Have a Weather Delay - Aidan:

This model is predicting whether any given flight will experience a weather delay. The original weather delay variable observed the length of a weather delay in minutes. To simply predict whether a flight will have a weather delay or not, a binary value was captured to indicate if a weather delay was larger than 0 minutes.

The inference goal that can be achieved through this model relates to origin. The inference goal was to determine if an origin with bad weather will lead to a higher chance of delay. We can assess this by examining whether origin is a significant feature for the models created. If origin is highly significant in a model that performs well then we can conclude that origin influences the chance of delay and assume that origins with notoriously bad weather consequently will increase the likelihood of a weather delay occurrence.

The major data transformation that was required specifically for this model was a sampling adjustment on the target value (binary for if there was a weather delay). The ratio of observations without a weather delay compared to the observations with one was 80:1. This issue caught our attention because there were no resulting false positives from the logistic regression model, which means that we were unable to compute the precision of the model. Then it was clear that there were also no true positives recorded, so our model was only predicting 0 (no weather delay) for every observation. Since the number of weather delay observations was relatively very low,

predicting 0 every time still resulted in an AUC score of close to 0.6. In order to get the model to predict positive cases the data needed to be balanced out. The two approaches are to either oversample the minority classification (weather delay: 1), or to undersample the majority classification (no weather delay: 0). We undersampled the majority data by first creating one data frame that only had observations of 0 and another that filtered to only have observations of 1. Then the majority data frame was sampled at the percentage of 1/80, which is the original ratio of minority to majority classifications. From there, the newly undersampled majority data frame was fully joined with the original minority data frame. This new smaller data set was then used in the three prediction models.

The label variable was altered specifically for this model as well. The target of this model is the binary variable IsWeatherDelay, so this column was renamed “label” so that the models could read and understand what the target was.

Each model was scored using AUC, the primary goal was to find a model that resulted in the highest AUC. To progress towards that objective, a grid search was implemented for each of the three models.

6. Conclusion:

6.1. Predicting How Long an Arrival Delay Will Be Using Regression

In Table 1 in the Appendix section, you can see the MSE of each of the 4 regression models. As mentioned previously, the best scoring model was the Random Forest Regression after grid searching, with an MSE of 2217.6132. Following the Random Forest Regression, we have the Unscaled Linear Regression, the Scaled Linear Regression, and the Gradient-Boosted Tree Regression models. It was interesting to see that the Unscaled Linear Regression model scored slightly better than the Scaled Linear Regression. Also, as you can see, the MSE scores for each model weren’t that good. A way to improve these scores would have been to include the highly correlated variable Departure Delay. This wouldn’t make sense to include in the models if you were trying

to predict if a flight would arrive late prior to it departing from the origin airport; however, if you were trying to track a flight when it has already departed from the origin apart (tracking apps), including the Departure Delay would make more sense intuitively and would decrease the MSE scores on each model significantly.

In terms of feature weights and importances, both the Unscaled Linear Regression and the Scaled Linear Regression models had the same order feature weight; the same goes for the Random Forest Regression and the Gradient-Boosted Tree Regression models, the order of feature importances was the same on both these models. I will compare the Unscaled Linear Regression feature weights (Table 2 in the Appendix section) with the Random Forest Regression feature importances (Table 3 in the Appendix section) since these two models are the best two models in terms of MSE scores. In the Unscaled Linear Regression model, the features with the most weight were Season, Airline, and Flight Month; while in the Random Forest Regression model, the features with the most importance were Destination, Expected Departure Time of Day, and Flight Month. It was interesting to see how some of the variables held equal value in both models, like Flight Month being the third most important feature in both models, while other variables were important in one model, but not that important in the other model. For example, Day of Month was the fourth most important feature in the Unscaled Linear Regression model, however, it was the seventh most important feature in the Random Forest Regression model. We can say confidently that Flight Month is a very important predictor of Arrival Delay because, as mentioned above, it has the third highest weight in the Linear Regression model and it holds the third most importance in the Random Forest Regression model.

6.2. Predict if Flight Will Have an Arrival Delay - Will

Looking at the feature importances of the Random Forest model in Table 5, the features that hold the most importance when predicting the likelihood of a flight having an arrival delay are Airline and Flight Month. None of the other predictors had an importance greater than 0.1. We weren't surprised that Flight Month had such a high importance in

this model because of the regression models that we ran first. In those regression models, Flight Month was also a very important feature, so it makes sense for that to carry over to this model since we were predicting the same target variable, the only difference being that we are predicting the probability that an arrival delay occurs, not the actual duration of an arrival delay. It's interesting how the month of a flight matters so much more in this case than the day of the flight.

6.3. Predict if Flight Will Have a Departure Delay - Garrett

We wanted to investigate the feature importance of the Random Forest model since it scored better than the Logistic Regression model. The most important features that predict the likelihood of whether or not a flight will have a departure delay are Origin, Flight Month, and Airline. When thinking about what causes a flight to depart late, these three features being extremely important is not at all surprising. Most times when a flight is delayed from departing, it's usually an issue with the airport it is departing from, not the destination it is flying to. Next, oftentimes airlines have issues at the gate, whether it be with the plane or the crew, causing them to delay the flight in order to get everything situated. In addition, the month affects more than just potential weather issues, it affects the volume of people in airports and the volume of flights departing. For example, in months like July and December, when many people go on vacation or go home for the holidays, it's not unreasonable to think that flights may be delayed simply because there are more flights that are available to be delayed.

6.4. Predict if Flight Will Have a Weather Delay - Aidan

The results of this prediction task are shockingly good. The highest AUC score of all models was achieved in this section through a random forest model. Multiple grid searches were required but the work was worth the 0.7536 AUC score. Refer to Table 8 in the appendix to see how the other models compared. The most interesting aspect of the outcome here lies within Table 9 - the feature importances. Since the GBT and random forest models scored much better than the logistic regression it makes sense to rely more on these importance scores. Destination and expected departure time were

the top two most important features in both models. When you think about it, this adds up - however before obtaining the results, we believed that season, month, and origin would hold the most weight. The destination could have rough weather and would require the pilot to circle the airport while things clear up. Time of departure also can make intuitive sense if bad weather follows a time trend. Overall this aspect of prediction was easily one of the more successfully completed objectives. A lot of interesting information can be drawn from here.

6.5. Inference Results

Inference 1: Does time of day impact the chance a flight will be delayed?

To obtain the answer to this question we looked at the rank of the feature importance for both the expected arrival and departure time on each of the best performing models. The breakdown can be seen in Table 10 (Inference 1), but we can conclude that departure time has a big impact on prediction with an average rank of 3.75, but we cannot do the same for arrival. With an average rank of 5.75, this is considered a below average (> 4.5) feature.

Inference 2: Does the time of year/date impact the likelihood of a delay?

The features month, and season were ranked in terms of their feature importance in order to determine this answer (see Table 10 - Inference 2). Surprisingly, the season feature is generally not important at all with an average rank of 6.75; however, month is an extremely important feature for all prediction models with an average rank of 2.5. This proves time of year is impactful, but should be more granular than binned seasons.

Inference 3: How do airlines compare in terms of how often delays occur?

This was answered through feature importance ranking as well because an association model was not implemented. The average rank for the airline feature is 3.75 (see Table 11 - Inference 3), which validates that airlines do impact the occurrence of delays.

Inference 4: Are origins/destinations a major factor when calculating a delay likelihood?

This was one of the major questions on our end, we imagined that these features would indeed be very important to prediction. That is why we ensured to keep the variable in even though there were 354 unique values. The hypothesis was confirmed, the average rank for origin reached 3.5, and destination proved to be even more important with an average importance rank of 3.25. Refer to Table 11 - Inference 4 to see where these location variables rank for our best performing models.

Inference 5: If an origin flight is in an area known for bad weather, will there be more delays?

We were not able to extract specific values for origin to test this theory, so as a substitute we continued the trend of feature importance rank. In Table 12 only the weather delay classification models are being compared so we can attempt to bridge the gap between origin and weather. The origin feature earned an average importance rank of 3.33. We cannot conclude that origins with bad weather equal higher delay frequency, but we can say that origin is fairly important for predicting weather delays.

Remaining Inferences:

The last two inferences are; 1) Is there an optimal time of day that flights will be least likely to delay?, 2) Which destinations are highly likely to have a delay?. This is where we were unable to meet our initial objectives. In order to reach a confident conclusion for both of these questions we would most likely need to add an apriori model that checks for association between variable values. In a deeper analysis of flight delays, the first thing that we would add is an apriori model. In a practical sense, an association model would allow us to create a flight flagging system for values that are highly associated with a delay.

7. Appendix:

Table 1: MSE of Regression Models (Michael)

Model	**MSE**
Random Forest Regression	2217.6132
Unscaled Linear Regression	2223.3123
Scaled Linear Regression	2223.4633
Gradient-Boosted Tree Regression	2501.9447

Table 2: Unscaled Linear Regression Feature Weights (Michael)

Feature	**Weight**
Season	1.7402
Airline	0.2929
Flight Month	0.2044
Day of the Month	0.0464
Destination	0.0170
Expected Departure Time	0.0062
Origin	0.0041
Expected Arrival Time	0.0034
Distance	0.0018

Table 3: Random Forest Regression Feature Importances (Michael)

Feature	**Importances**
Destination	0.3273
Expected Departure Time	0.2094
Flight Month	0.1477
Season	0.0986
Airline	0.068
Origin	0.065
Day of the Month	0.0393
Distance	0.0286
Expected Arrival Time	0.016

Table 4: AUC Scores (Will): Arrival Delay Classification

Model	**Test AUC**
Random Forest	0.5911
Logistic Regression	0.5733

Table 5: Random Forest Feature Importances (Will): Arrival Delay Classification

Feature	**Importance**
Flight Month	0.370477
Airline	0.315228
Expected Arrival Time	0.091305
Origin	0.077972
Expected Departure Time	0.065174
Destination	0.060678
Season	0.008514
Distance	0.007461
Day of the Month	0.003191

Table 6: AUC Scores (Garrett): Departure Delay Classification

Model	**Test AUC**
Random Forest	0.6545
Logistic Regression	0.6282

Table 7: Random Forest Feature Importances (Garrett): Departure Delay Classification

Feature	Importance
Origin	0.350919
Flight Month	0.313462
Airline	0.243818
Expected Arrival Time	0.034619
Destination	0.033775
Expected Departure Time	0.016292
Distance	0.005003
Season	0.001682
Day of the Month	0.000431

Table 8: AUC Scores (Aidan): Weather Delay Classification

Model	Test AUC
Random Forest Classification	0.7536
Gradient-Boosted Trees Classification	0.7027
Logistic Regression	0.5946

Table 9: Feature Importances (Aidan): Weather Delay Classification

Feature	Random Forest Classification Importance	GBT Classification Importance	Logistic Regression Importance
Destination	0.399602	0.439768	0.000619
Expected Departure Time	0.339241	0.491692	0.000958
Origin	0.062935	0.024036	0.005122
Month	0.060711	0.025774	0.000281
Airline	0.042912	0.005473	0.000185
Distance	0.034996	0.000000	0.000134
Expected Arrival Time	0.025727	0.007342	0.046113
Season	0.021856	0.004241	0.009275
Day of the Month	0.012020	0.001673	0.000666

Table 10: Inference 1- Time of Day Impact Chance of Flight Delay, Inference 2 - Time of Year / Date Impact Chance of Flight Delay

Best Model	Inference 1: Time of Day Impact		Inference 2: Time of Year Impact	
	Expected Arrival Feature Rank	Expected Departure Feature Rank	Month Feature Rank	Season Feature Rank
Duration Regression: Random Forest	9	2	3	4
Arrival Delay Classification: Random Forest	3	5	1	7

Departure Delay Classification: Random Forest	4	6	2	8
Weather Delay Classification: Random Forest	7	2	4	8
Average Rank	5.75	3.75	2.5	6.75

Table 11: Inference 3 - Airline Impact on Delay Occurrence, Inference 4 - Origin & Destination Impact on Predicting Delays

Best Model	Inference 3: Airline Impact	Inference 4: Origin & Destination Impact	
	Airline Feature Rank	Origin Feature Rank	Destination Feature Rank
Duration Regression: Random Forest	9	3	1
Arrival Delay Classification: Random Forest	3	1	6
Departure Delay Classification: Random Forest	4	2	5
Weather Delay Classification: Random Forest	7	4	1
Average Rank	3.75	3.5	3.25

Table 12: Inference 5 - Origin Airport Influence Likelihood of Weather Delay

Weather Delay Classification Models	Origin Feature Rank
Logistic Regression	3
Random Forest	3
Gradient-Boosted Trees	4
Average Rank	3.33

Figure 1: Correlation Matrix

	Origin	Dest	ExpectedDepartTime	DepartDelay	ExpectedArrivalTime	ArrivalDelay	AirTime	Distance	DayOfMonth	FlightMonth	Airline	Season	ExpectDepartTOD	ExpectArrivalTOD
Origin	1.000000	-0.165484	-0.125196	-0.013488	-0.112306	-0.001442	-0.188498	-0.197046	0.002275	0.003218	0.163656	-0.003982	-0.096262	-0.100527
Dest	-0.165484	1.000000	0.133337	-0.006180	0.094049	0.005278	-0.211027	-0.197394	0.002568	0.000922	0.160754	-0.005227	0.100143	0.043951
ExpectedDepartTime	-0.125196	0.133337	1.000000	0.103228	0.687924	0.093404	-0.017657	-0.011236	-0.003690	-0.003457	-0.001190	-0.003106	0.896887	0.642350
DepartDelay	-0.013488	-0.006180	0.103228	1.000000	0.087716	0.956544	0.006631	0.007127	-0.006266	-0.000869	0.004670	-0.031829	0.096610	0.078908
ExpectedArrivalTime	-0.112306	0.094049	0.687924	0.087716	1.000000	0.082182	0.018203	0.015353	-0.005164	0.001826	-0.000876	0.015627	0.568802	0.495624
ArrivalDelay	-0.001442	0.005278	0.093404	0.956544	0.082182	1.000000	-0.004288	-0.021323	-0.007137	0.009965	0.027253	-0.040815	0.085791	0.071661
AirTime	-0.188498	-0.211027	-0.017657	0.006631	0.018203	-0.004288	1.000000	0.984836	0.000992	-0.002207	-0.038960	-0.004457	0.001107	0.058914
Distance	-0.197046	-0.197394	-0.011236	0.007127	0.015353	-0.021323	0.984836	1.000000	0.000945	0.001840	-0.044566	-0.013391	0.012657	0.065059
DayOfMonth	0.002275	0.002568	-0.003690	-0.006266	-0.005164	-0.007137	0.000992	0.000945	1.000000	0.004605	0.000286	-0.016043	-0.000566	0.000898
FlightMonth	0.003218	0.000922	-0.003457	-0.000869	0.001826	0.009965	-0.002207	0.001840	0.004605	1.000000	-0.011996	0.029587	-0.003438	-0.003247
Airline	0.163656	0.160754	-0.001190	0.004670	-0.000876	0.027253	-0.038960	-0.044566	0.000286	-0.011996	1.000000	-0.001223	0.005311	-0.012295
Season	-0.003982	-0.005227	-0.003106	-0.031829	0.015627	-0.040815	-0.004457	-0.013391	-0.016043	0.029587	-0.001223	1.000000	-0.010420	-0.008105
ExpectDepartTOD	-0.096262	0.100143	0.896887	0.096610	0.568802	0.085791	0.001107	0.012657	-0.000566	-0.003438	0.005311	-0.010420	1.000000	0.670727
ExpectArrivalTOD	-0.100527	0.043951	0.642350	0.078908	0.495624	0.071661	0.058914	0.065059	0.000898	-0.003247	-0.012295	-0.008105	0.670727	1.000000

Figure 2: Weather Delay Percentage by Origin Airport

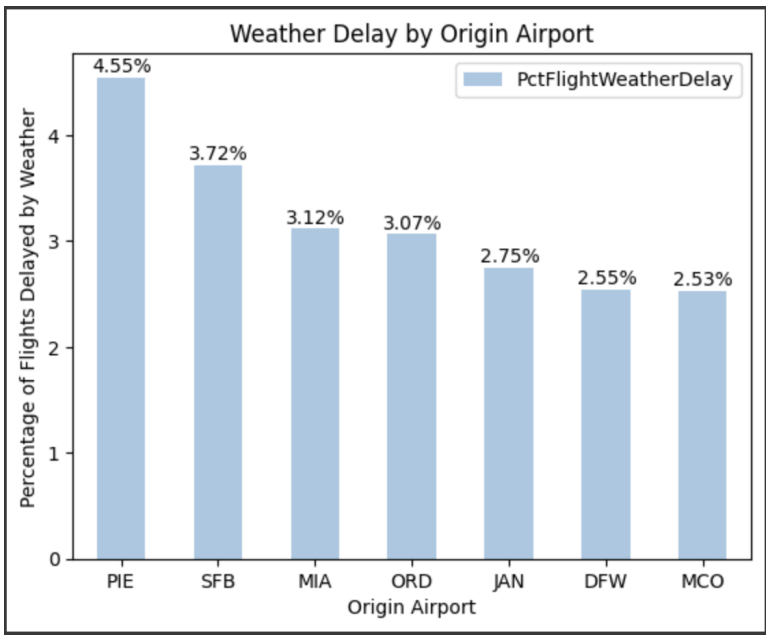


Figure 3: Arrival Delay Count by Season

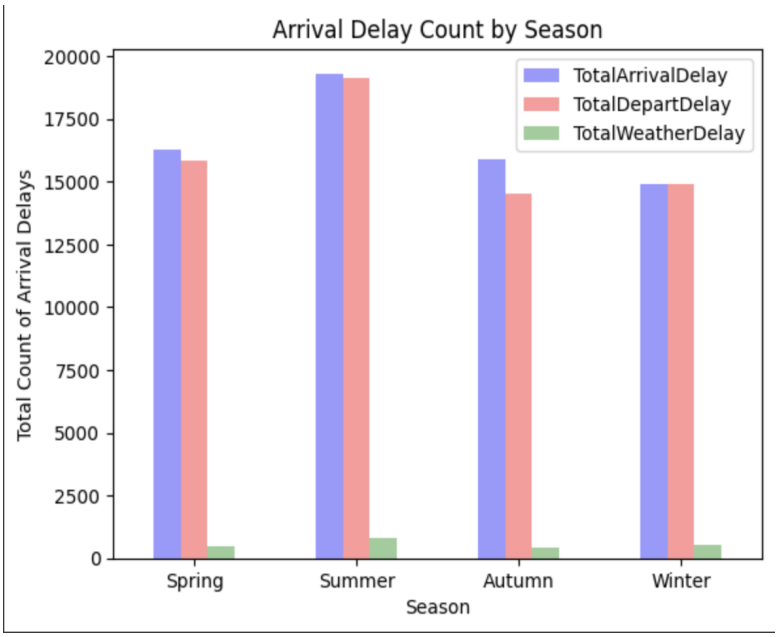


Figure 4: Diverted Bar Chart of Departure Delay Percentage by Airline

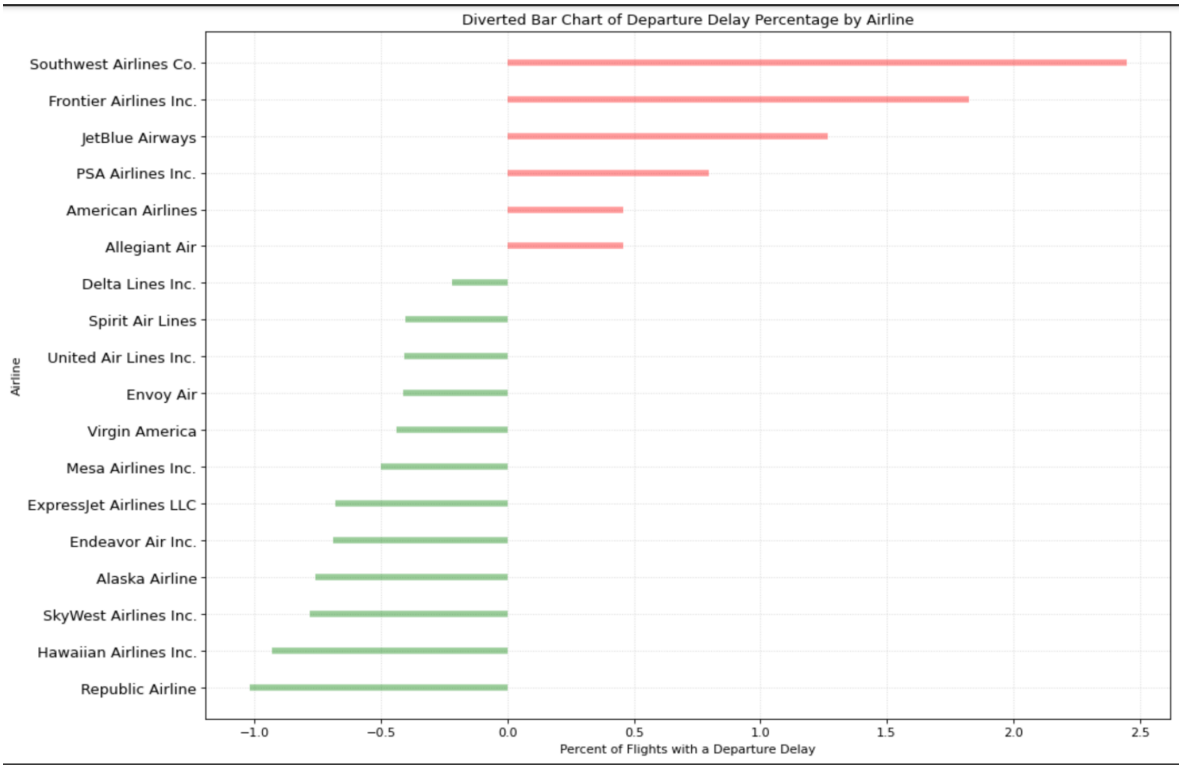


Figure 5: Density Plot of Arrival Delay (in Minutes) by Time of Day

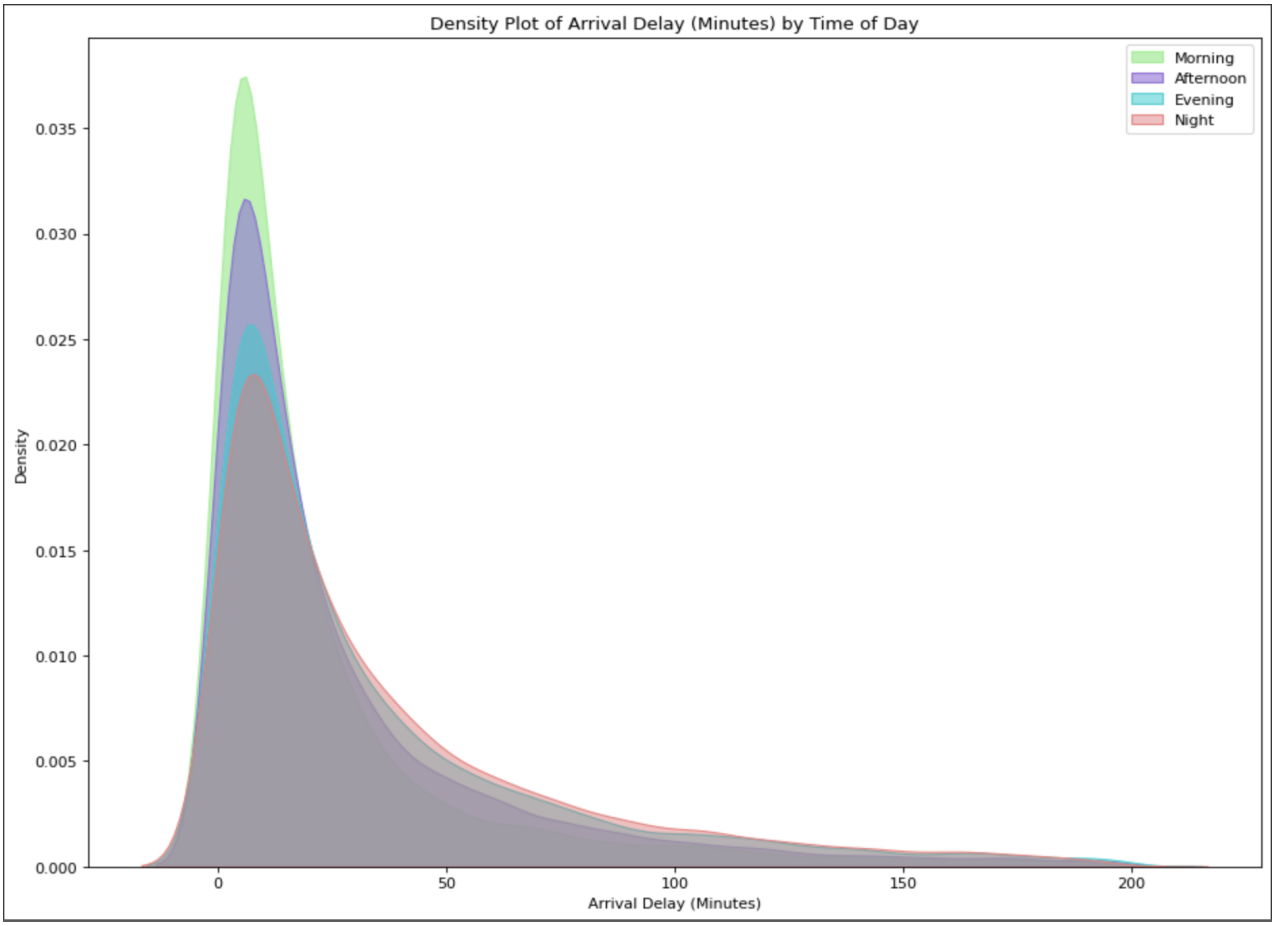


Figure 6: Classification Model Workflow

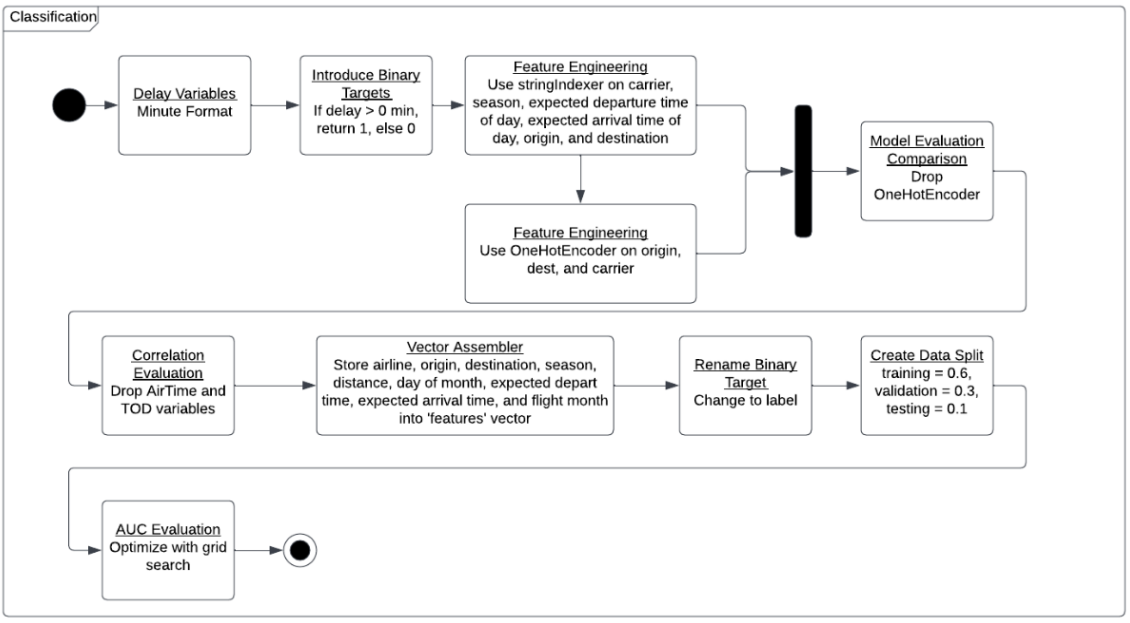
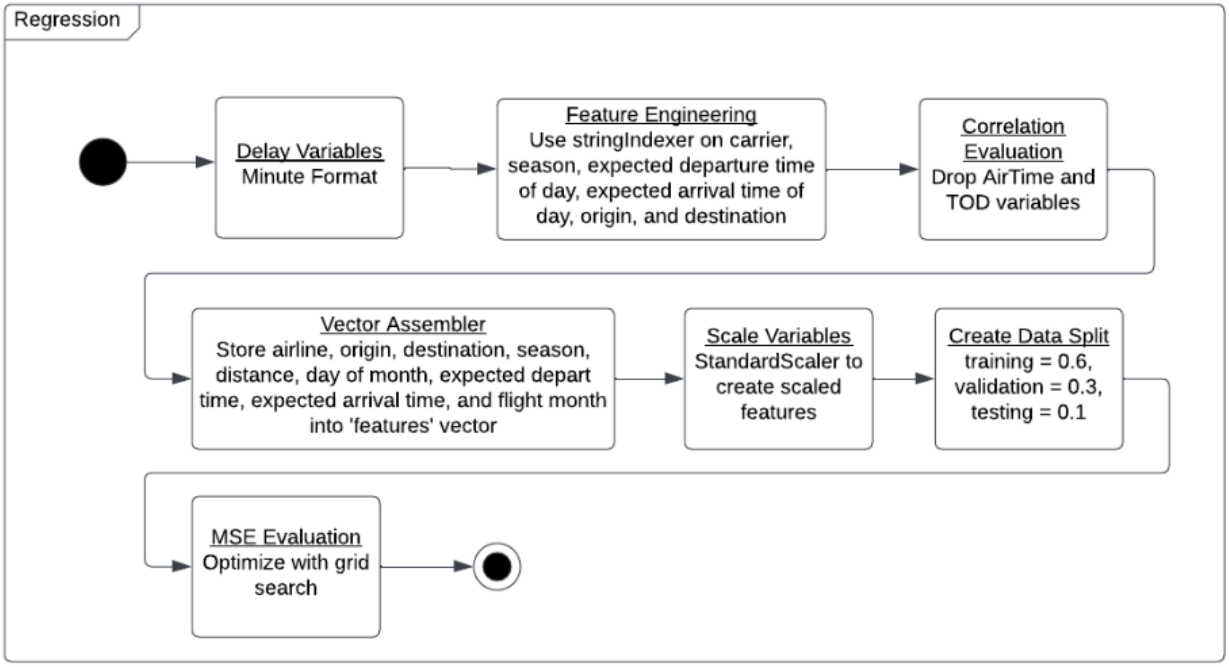


Figure 7: Regression Model Workflow



Additional Resources - Variable Definitions:

[Bureau of Transportation Statistics: Variable Definitions]