Calculating Weather Delays at Airports



In 2007 alone, airline delays cost the entire industry upwards of $8 billion in lost revenue, most of which was caused by inclement weather. For our final project, we investigate how we could create a solution for airports to mitigate financial strain that’s arises from these delays and cancellations. By narrowing our scope to six major airports nationwide, we analyzed and interpreted specific weather and airport data. Using descriptive statistics, graphic representation, and various modeling techniques, we map a strategy that would lead to a new design of flight scheduling and forecasting based off sound analytics.

Jeffrey Victor, David Mackey, Graham Latsa, Patrick Carlin

IST 718 FINAL PROJECT

**Specification**

The main goal of our project is to minimize the revenue lost due to weather cancellations and delays. This research and analysis will help us fully interpret and understand the relationship between flight cancellations and the reasoning behind them. We decided to take a sampling of six major airports nationwide to combine with each location’s corresponding weather data. Pulling daily weather and transit data from NOAA (National Oceanic and Atmospheric Administration) and the Bureau of Labor Statistics respectively our first step was accurately reading in the data and assigning significance to cancellations. Getting through the manipulation and fitting of our data, we plan to have a clear view of what variables are correlated to cancellations, and then produce recommendations based on our findings.

Data:

We were only interested in the types of weather that might affect air travel. For example, heavy snowfall might close an airport, but a high soil temperature would not. We gathered the following weather data features, as reported on a daily basis:

Weather Data

The United States federal government collects and publishes weather data through NOAA, the National Oceanic and Atmospheric Administration, which is part of the US Department of Commerce. People in the country have been collecting weather data since the 18th century, with some weather stations coming and going over the decades. By 1860, 500 weather stations supplied data to the Smithsonian Institution via telegraph, enabling daily weather reports in newspapers.

Today the data collection is automated, and immediately available in searchable databases at <https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND> .

Access

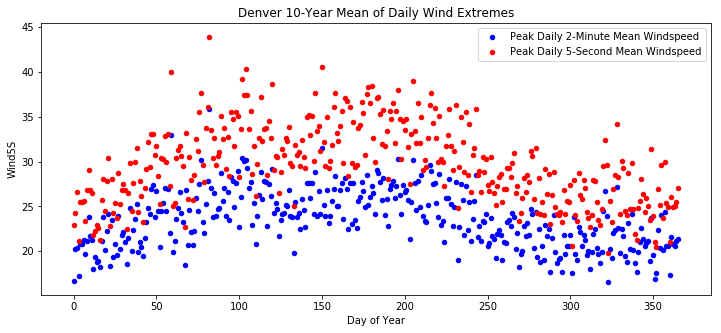
NOAA users, including the general populace, can search for weather data by city, ZIP code, or specific weather station, and across date ranges. Specific types of weather can be selected to reduce the amount of data downloaded. Also, programming interfaces are available so that weather data can be accessed over the Internet from within software.

Data Features

Each weather station reported some of the possible data, with more data fields becoming common over time. Daily high and low temperature have been basic data since the beginning, along with snow depth in those locations that commonly experience it. Some data capture requires modern technology, such as daily average temperature, daily total sunlight, and others. For these reasons and others, the data is incomplete, especially farther back in history.

Exploration

We explored the weather both for suitability in our project, completeness, sensibility, and to observe patterns that might be useful in our project. For example, this plot shows wind speed data, giving us an understanding of the periods of the year most likely to experience high wind speed that might delay flights.



Figure

Scrubbing

The data published by NOAA is fairly clean. It is delivered in CSV format, with lots of missing data. We replaced the missing categorical binary data with zeroes. (See the next section for a listing of the weather features that we used.)

Restructuring

To reduce overfitting our model, we collapsed the 16 categories listed earlier into these 5 features: fog, snow, sleet (all freezing precipitation except snow), rain, and particulates. Each of those features was a binary attribute. We also simplified the wind speed data into another binary attribute, “high wind speed,” which represented a 5-second average speed over 40 MPH, or a 2-minute average greater than 35 MPH.

For this project, we were only interested in the types of weather that might affect air travel. For example, heavy snowfall might close an airport, but a high soil temperature would not. We gathered the following weather data features, as reported daily:

* Numeric data
  + High temperature
  + Low temperature
  + Snowfall
  + High wind speed, averaged over 5-second and 2-minute intervals
* Categorical data
  + Fog: 4 categories
  + Snow: 2 categories
  + Other freezing precipitation (sleet, hail, etc.): 5 categories
  + Non-freezing rain: 3 categories
  + Non-water atmospheric particles (dust, haze, smoke): 2 categories

Air Traffic Data

The Bureau of Transportation Statistics is an online open source platform where we took the following data to incorporate into our analysis. We initially had exported and were dealing with huge Air Traffic Data sets, which incorporated repetitive and unnecessary variables and data. By utilizing R for some descriptive and summary statistics we narrowed down our key variables for the Air Traffic data.

|  |  |  |
| --- | --- | --- |
| Flight Date  Reporting Airline DOT ID  Tail Number  Flight Date  Flight Number  Origin Airport  Origin City Name  Origin State  Destination | Day of week  Delay Group  Departure Block  Taxi out time  Wheels off  Wheels off  Wheels on  Taxi in time  Cancellation | Reporting Airline  Arrival Time  Distance  Distance Group  Delay code  Diverted flight Info  15-minute Delay Indicator  CRS Departure Time  Carrier Code |

Utilizing the same tools, we got rid of any NA’s as well as filled any miscellaneous data within our remaining datasets to further our predictive investigation. From here we dug deeper into significance values for regression to really get an idea of what is needed to create a solid model.

**Observation**

The first plot in this section shows the maximum amount of snow, each day of the year, across all ten years that we considered. Heavy snowfall is one of the types of weather that can cause airline flight delays.

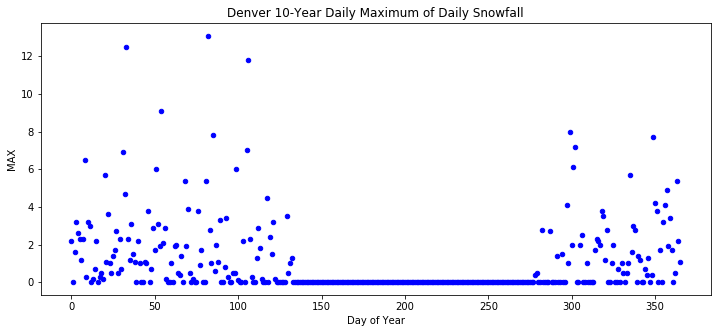


Figure 2

The next plot shows the extreme temperatures, for each day, across 10 years. Some of these, especially, extreme cold temperatures, may contribute to flight delays.

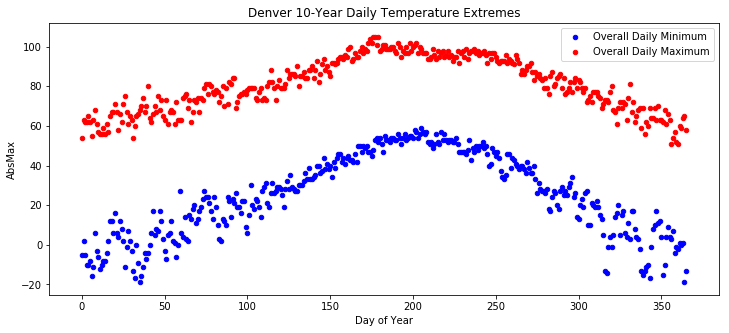


Figure 3

To follow our initial trend of view how and why the flights are being cancelled, we wanted to get a breakdown of some key variables. Figure 3 gives us a breakdown of the Boston Airport, viewing percentage of cancellations against each day of the week. Running the same against the other locations gives us an idea of location specific insights as to highest cancellation days for 2018.

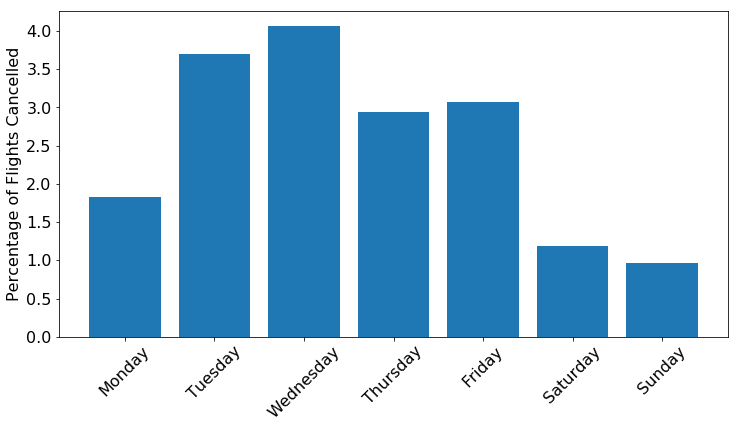
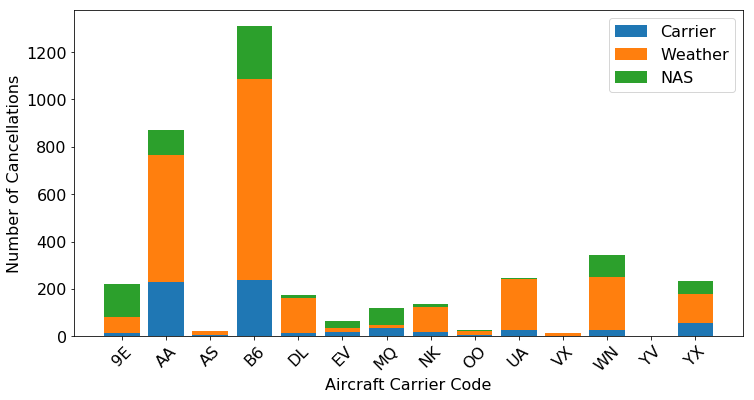


Figure 4

When looking at our Air Traffic Data, we really wanted to focus in on a couple key variables and discuss their dependencies. Figure 4 breaks down the specific types of delays within each airline carrier and compares then against flight cancellations. We see that with airline B6, with the largest number of weather cancellations you could infer that some airlines fare better in weather.



Figure

Further investigating the insights obtained from the graph above, figure 5 compares the total number of flights that each carrier takes on. The variance in aircraft carriers directly corresponds with the number of flights each Carrier takes on at the Boston Airport. When looking a little deeper we notice that the total delays in both DL and UA are attributed to weather.

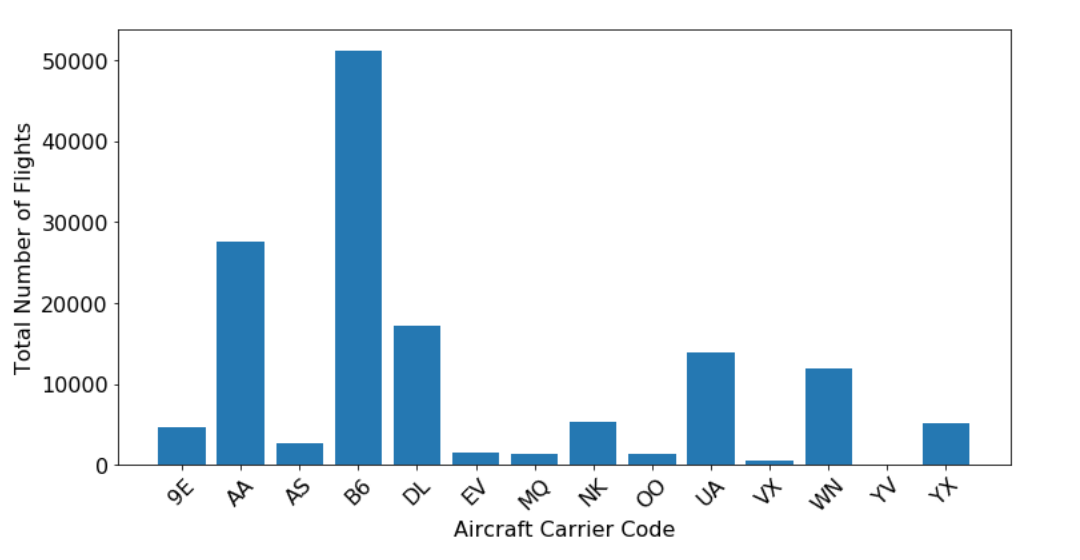


Figure 6

**Analysis**

After merging the airline and weather data, a significant amount of preprocessing had to go into prepping the data for the models used in this analysis. Preprocessing steps included but were not limited to, removing columns of missing data, converting data types, and finding columns that only included partially filled sets. After these preprocessing steps a couple of options were available on how to model this data.

It was decided that only the origination city data would be used to model how the weather would affect or cause a delay in a prospective flight. The origin was chosen over the destination data because it seemed more logical that severe weather at the time of takeoff, where a flight is taking off from, was more likely to cause a delay than weather at the destination.

Another modeling choice was how to determine the response variable of the models. The flight data already had a column (“WEATHER\_DELAY”) that represented the amount of time a flight was delayed caused by weather. The other main reason in the data that caused delays was the flight carrier meaning some sort of technical or logistic problem. Having this variable did make the option easier although, after cleaning and formatting the data, this variable was highly unbalanced as over 94% of the flight had no delays. This caused large issues when attempting to model how weather affects delays. Since the data was so unbalanced, either the model would attempt to fit and end up predicting at a very low accuracy since most of the non-zero prediction would be incorrect or the model would just predict every flight to have no delay without any actually consideration of the test data. Because of this issue, the modeling idea was changed slightly from how weather affects flight delays to searching more in the realm of what types of weather have a more serious effect on delays than other types. The data was also limited to only those flights that had non-zero delays caused by weather. This change limited the amount of data was used to fit the models although not to the point where the models would lose a significant amount of trust.

Another slight change to the modeling of this data was the transformation of time data to categorical data. The delay data ranged from 1 minute to multiple days of delay time. Since the data was so spread out and included so many unique values many of the classification models were inappropriate for the analysis. A regression test was completed as an attempt to predict the actual time of delays when a delay would exist, but the accuracy of that model was abysmal. Therefore, the delay time was transformed into a categorical variable of different ranges of delay times. These ranges are ten-minute intervals from 0 to 60 minutes, and then 1-1.5 hrs, 1.5-2 hrs, 2-3 hrs, 3-5 hrs, and >5 hrs. This allowed model techniques such as Naïve Bayes, SVM, and clustering to be implemented.

Models

The three modeling techniques that were attempted were Support Vector Machines (SVM), Naïve Bayes, and K-Nearest Neighbor (KNN) clustering. Slightly different datasets were used to fit the models that were created based on how many missing values there were in the different variables. The main variables can be found in the Appendices. The data was also split into two 60-40 training-testing sets that were randomly chosen.

KNN or K - nearest neighbor is a modeling technique similar to K-Means. Where it determines what grouping each case is in by calculating the distance of the points from each other. Where KNN is different is that K-Means calculates the mean distance of all the points in a single grouping from another grouping, but KNN calculates the distances of every point from all the other points and then groups together the points that are closest together (as in their “neighbor”).

SVM or Support Vector Machines center on the idea that every dataset can be separated into at least two clusters. Now the magic of SVM is that with data that seems difficult to separate in an n-dimensional space, the data may be separable in an n+1 dimensional space. Take a two-dimensional problem as an example. If the points that are to be separated are right on top of each other on a 2D graph, then there would be no single line that could properly cluster them. But if the data was taken and moved into a 3D space using a transformation, it is likely that a hyperplane could be fit that would separate the clusters effectively. Since this transformation would be standard across all the points, the underlying data would be effectively the same but now can be clustered properly.

SVM also is a very effective and handy clustering technique as it comes with its own confidence metric. It is straightforward to calculate the distance between the hyperplane and the closest data that is correctly clustered. This distance metric is the confidence of the model. The farther away the closest data is, the more confident the model is and vice versa. Since no model is ever perfect SVM also can be programmed with a slack constraint that allows incorrectly modeled data to be acceptable to a point.

The Naïve Bayes model is based on the Bayes theorem which basically says the probability of event A depending on event B is equal to the probability of B depending on A multiplied by the probability of A divided by the probability of B. In essence the probability of a hypothesis can be broken down into the probabilities of a known similar situation and some related evidence. But the issue with this is that in any real-world situation there are many, many factors that can affect a hypothesis. This makes the full Bayes theorem extremely complicated. Here is where the naïve part comes in. The Naïve Bayes model assumes that each factor affecting the hypothesis do not have any effect on each other. Although this assumption can be completely inappropriate in some situations, the Naïve Bayes model can be quite accurate when the factors are truly independent.

After fitting these different models to the data, none of the models were at all effective at predicting the correct amount of delay time for a flight. The best model, SVM, had an overall accuracy of just under 30%, which would not be acceptable by almost any metric. One reason that these models could have such low effectiveness is that the variables that were available would not be the leading factors that would cause delays in flights. Weather factors such as lightning, storm size, and hail were not accounted for or did not meet the required amount of data to be included in the models. Another reason could be that the cause of delays was not localized to the origin of the flight. Although the destination was considered, it is quite possible that a portion of the weather delays in this data were caused by weather that occurred between the origin and destination that would have been unknown to this analysis. A final, and probably the most obvious, reason for the lack of accuracy in these models is that this kind of analysis is much more complicated than this data can support. This analysis had to limit the amount of data and the robustness of the data collected just to complete the models themselves. When a flight is delayed due to weather, countless factors are considered, such as more accurate weather models and the subsequent consequences of a delay. Not only could the delays be caused by other factors, but the opposite could also be true. Other factors that caused a flight to not be delayed that might have characteristics of a delayed flight also could very well exist in this data.

Without effective models, a slightly less complex method of analysis was completed that averaged the delay time of flights where each variable was either marked as present (True) or greater than 0. In this analysis the variables Snow Drift, Heavy Fog, and Peak Gust Time had considerably higher average delay times than the other variables. Interestingly, none of these variables were able to be included in the models although, just going off a logical approach, these variables would most likely have a considerable effect on a prospective flight.

**Recommendation**

After communicating the results of the models, the team is going to express recommendations for opportunities the team was unable to work during this phase of the project. The recommendations pertain to 3 focus areas: data sources, data storage, and implementation. The first focus area is data sources, this improvement is about the type and amount of data collected. When the team initially went out to obtain the data it was only requested to interact with flight information and weather data. The team was about to generate models from that information, however more effective models would have been created if other sources of data was identified e.g. Congested air traffic and Airline glitches data (see article in references). The team additionally went through and found that a few of the airports did not have very significant weather delays, LAX and ATL, so if the data sources were better understood at the beginning it would have saved the team the effort of going through this exercise for Airports without having a big weather issue. These were the data sources recommendations that would help other team in the future by giving them more encompassing data and having them effectively spend time on projects that have an issue

The next area of recommendation is about data storage. The team had to go to government website to get the data and their search querying features robust and it was time intensive to obtain the flight data, which ended up being nearly three quarters of a GigaByte of data to sort through. For the team it was frustrating to have to obtain the data in a very manual process. Therefore, the team would recommend getting/creating an internal database, recommending a distributed database. This would allow consultants, like our team, as well as your internal organization to have a better picture of what is going on with airline delays and cancelations. It would allow for easier live streaming of the data to get more accurate data, as well as create algorithms that will automatically run, allowing a for more opportunity to create data centric decisions. The last recommendation the team has to offer is pertains to implementation. This section does reference the last two ideas; however, this is a more integrated perspective to your mission. The team believes data awareness for you organization is pivotal, even though weather delays/cancelations happen far less than the team expected these are costly decisions, both with customer reimbursement and loss of customers. To make and integrated data organization the team recommends having data be easier to extract and play with for any and all employees. Additionally, providing training to the employees and make it part of the culture. All of these things will create a common work language as well as a driver that can analytical show progress to the team.

The models the team is providing intend to be near-team airline management tools. When there is a forecast of bad weather, an airline would be able to plug in the weather conditions and the model would estimate the potential of a weather delay. This tool is not intended as long-range planning tool. The team did not create a time series forecast for a year, with given weather and environmental data. That tool would be a helpful enterprise forecasting and financial tool, which the team highly recommends, and would support the model that this team is providing.

While attempting to create models, we found that some variables were identified as stronger factors for predicting a more severe weather delay than others. These factors were Snow Drift, Heavy Fog, and Peak Gust Time. The models themselves could most likely become more accurate with both the inclusion of these variables. However, that would require a more comprehensive collection of data by the weather associations since most of this data was missing. A recommendation from this part of the analysis would be to use the main factors as a rudimentary baseline to predict more severe delays in flights but without more historical data this can only be trusted so much. A final recommendation, after this reports and seeing a very low cancelation/delay rate, it would be best if you do use any forecasting models that would not be normally distributed models but more tail heave models like Gaussian, since the tail situations is where these cancelations/delays will happen.

**Retrospective**

At the end of this project we see a lot of things we didn't see at the beginning of this project. The first one is that we obtained too much data at the beginning of the project. The team went out and obtained a lot of data (that was also because we were scared of being unable to get to it due to the potential government shutdown). The issue there is that the team didn't know how big of an issue this was, we didn't realize the trouble it would cause in our modeling phase. Also we just got 2 types of data, we didn't look for airline data from the carriers, we didn't look up research studies that would aided in this process. The team believe it would have been better for us to do a little more up front research to see if we this was the project we were expecting as well as if we wanted to deal with the difficulties of modeling. Another difficulty the team faced was changing the focus on the model part of the way through. The team talked a lot about how to do a model in a very specific way as a time series forecast. However, after looking into this it was a lot harder than expected so towards the end we shifted our focus on the type of model that was being created, so there was a little wasted time there. But all in all, the team worked together well, collaborated well, helped each other well, and shared information well. This was a good team experience and even though the data didn't come out the way the team expected at the beginning of the project. The team did have a good system in place helping everyone out.

**References**.

Data Sources:

* <https://www.transtats.bts.gov/Fields.asp>
  + Airline data for domestic flights. Database contains delay times, reason codes, carrier, flight duration, etc..
* <https://www.ncdc.noaa.gov/cdo-web/datasets>
  + Weather data based upon weather data stations. Database contains temperature, precipitation, air quality, wind speeds

Articles:

* <https://usatoday30.usatoday.com/travel/flights/2007-12-20-flight-deldys_N.htm>
  + This article describes airline delays as a self-inflicted issue due to glitches in the system.
* <https://www.transportation.gov/sites/dot.gov/files/docs/mission/office-policy/aviation-policy/328666/us-international-air-passenger-and-freight-statistics-june-2018.pdf>
  + This is the current DOT report about airline trends (the team understands this is for international flights and the team doesn't have international data) the reports shows an increasing trend.
* <https://news.berkeley.edu/2010/10/18/flight_delays/>
  + Discusses the Cost of flight delays

**Appendices**.

Table of modeling variables

|  |  |
| --- | --- |
| **Variable** | **Name** |
| TMAX | Temperature Max |
| WSF5 | Daily peak 5 second wind speed |
| PRCP | Precipitation |
| FZFOG | Freezing Fog |
| SMOKE | Smoke |
| TMIN | Temperature Minimum |
| TAVG | Temperature Average |
| AWND | Wind Direction |
| SNOW | Snow |
| WSF2 | Daily peak 2 Minute Wind Speed |
| SNWD | Snow Depth |

Naïve Bayes Results Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Prediction | | | | | |
|  |  | 0-10 MIN | 10-20 MIN | 20-30 MIN | 3-5 HR | 30-40 MIN | >5 HR |
| Truth | 0-10 MIN | 27 | 36 | 41 | 6 | 14 | 59 |
| 1-1.5 HR | 4 | 4 | 11 | 3 | 7 | 7 |
| 1.5-2 HR | 1 | 4 | 11 | 4 | 0 | 4 |
| 10-20 MIN | 24 | 49 | 65 | 3 | 19 | 38 |
| 2-3 HR | 3 | 3 | 11 | 3 | 1 | 7 |
| 20-30 MIN | 12 | 32 | 50 | 2 | 16 | 12 |
| 3-5 HR | 0 | 3 | 4 | 2 | 0 | 1 |
| 30-40 MIN | 12 | 17 | 32 | 1 | 6 | 12 |
| 40-50 MIN | 3 | 10 | 14 | 3 | 3 | 4 |
| 50-60 MIN | 2 | 7 | 10 | 5 | 2 | 4 |
| >5 HR | 0 | 0 | 6 | 0 | 1 | 5 |

KNN Clustering Results Table

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Prediction | | | | | | |
|  |  | 0-10 MIN | 1-1.5 HR | 1.5-2 HR | 10-20 MIN | 20-30 MIN | 50-60 MIN | >5 HR |
| Truth | 0-10 MIN | 71 | 0 | 0 | 100 | 12 | 0 | 0 |
| 1-1.5 HR | 8 | 0 | 0 | 23 | 3 | 1 | 1 |
| 1.5-2 HR | 7 | 0 | 0 | 14 | 3 | 0 | 0 |
| 10-20 MIN | 49 | 1 | 1 | 128 | 14 | 0 | 5 |
| 2-3 HR | 8 | 0 | 0 | 15 | 2 | 2 | 1 |
| 20-30 MIN | 19 | 1 | 0 | 98 | 6 | 0 | 0 |
| 3-5 HR | 7 | 0 | 0 | 3 | 0 | 0 | 0 |
| 30-40 MIN | 14 | 0 | 0 | 54 | 9 | 0 | 3 |
| 40-50 MIN | 5 | 0 | 0 | 26 | 5 | 0 | 1 |
| 50-60 MIN | 8 | 0 | 0 | 16 | 4 | 1 | 1 |
| >5 HR | 1 | 0 | 0 | 7 | 2 | 0 | 2 |

SVM Results Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Prediction | | | |
|  |  | 0-10 MIN | 10-20 MIN | 20-30 MIN | >5 HR |
| Truth | 0-10 MIN | 92 | 88 | 3 | 0 |
| 1-1.5 HR | 9 | 22 | 4 | 1 |
| 1.5-2 HR | 12 | 9 | 3 | 0 |
| 10-20 MIN | 62 | 124 | 10 | 2 |
| 2-3 HR | 13 | 12 | 3 | 0 |
| 20-30 MIN | 28 | 88 | 8 | 0 |
| 3-5 HR | 4 | 5 | 0 | 1 |
| 30-40 MIN | 16 | 55 | 7 | 2 |
| 40-50 MIN | 9 | 25 | 3 | 0 |
| 50-60 MIN | 12 | 17 | 1 | 0 |
| >5 HR | 4 | 4 | 2 | 2 |