Predictive Analysis of US Flight Delays

Brandon Boyle

bboyle@bellarmine.edu

22 Jan 2026

It seems to be that every time you go to the airport nowadays, there are always flights being delayed. This can be for a number of different reasons including weather, crew running late, quick maintenance, you name it. But it’s no fun being delayed. It messes with travel time, connection flights, future plans, and just overall convenience. But its not only the traveler that is suffering from these delays. The airlines themselves take a hit when things get pushed back. Operation cost go up, pilots have to be moved around, and sometimes reimbursements have to be given out.

All of this combined with my future career of being an airline pilot, I wanted to be able to look at some of the biggest issues in the flight industry. That’s why I decided to choose this data set and use predictive analysis to try and get to the bottom of what some of the leading factors are in flights getting delayed.

Overall the idea of this project is to be able to predict whether or not a US flight will end up getting delayed by used historical flight and scheduling data. If I am able to find trends, this would be of great use for both travelers and the airlines themselves. While the results wouldn’t guarantee accuracy, the predictive analysis would give a great baseline for what to look out for. By applying multiple machine learning techniques to the data set, I hope to be able to focus on the differences between the models and their accuracy and hopefully combine them all to find some overall patterns.

Moving onto the background, the problem that I already stated is that many flights in the US get delayed and this effects both the passengers and the airlines. The airlines do their best to combat these delays by using similar techniques. They have a bunch of data analysis tools along with lots of optimization tools that allow them to do their best to predict delays, and to do their best at getting these problems fixed. They look at both historical data and real time data.

But for my project I will be using just historical data. Taken from the Department of Transportation in conjunction with the Bureau of Transportation Statistics, this dataset looks at every major airline flight from January 2022 to July 2025. But due to the sheer size of the dataset, I might end up making the timeframe smaller. As of right now, there are over 26 million rows of data and this could possibly lead to some problems with the applications that I use. Within that, there are 9 variables. These include the date, the airline carrier (which there are 22 of in this dataset), the flight number, the airport they took off from as well as landed, scheduled departure time, actual departure time, delayed, and canceled. From these variables, I will potentially add a few more such as flight path and flight time in order to get a better understanding for some routes that might be delayed more than others. For the most part, the dataset looks pretty clean and it doesn’t look like there should be too much work to do in order to get it ready for some analysis work.

For my models, I am looking to use a random forest model, gradient boosting, and a support vector machine. I might however end up changing the SVM to another model due to the compatibility with my dataset and the tools that I have available. I will also more then likely use a linear or logistic regression but only as a baseline for my other models. The three model will give me three different perspectives of the data. Gradient boosting will be more accurate then my random forest model and the SVM will give me some good insights on the margins of my data.

My modeling will be done in Python. Within that application, some of the key tools and libraries that I will be using include pandas and numpy for data manipulation, scikit-learn for model development and evaluation, xgboost and sklearn for help with gradient boosting, and matplotlib and seaborn for data visualization.

So to wrap everything up, the goal of this project is to look at and analyze US flights in order to predict delays and some trends within the data. By looking at 3 years of historical data, I hope to get a good perspective on the industry and hopefully some key findings. Delays have a much greater effect then one might think. By effecting both the passenger and the airlines, this is pretty big deal as most airlines look to keep a high reputation and deliver their passengers safely and on time. So by using three different predictive models, I hope to compare their results and compile a bunch of finding and results.