Predicting Rideshare Probability From The Airports In Large Metro Areas Real Time

Kaiwen Zhong, Ernesto Cejas, Ganesh Rajasekharan, Azar Eyvazov

Correlation One Dataton
July 14, 2018

1 Description of end-user problem

It is 1pm in Manhattan. Janet has been driving Lyft for the past hour, and demand is slowing down in midtown. Janet decides to try her luck, and drive closer to the JFK International airport. If she is lucky, there will be scores of tired passengers getting off of their flight, eager to open up their ride sharing apps. She hopes that she would not be driving in when there are no arrivals, because that would mean a significant waste of time and toll fees.

Today, the moment Janet enters the highway towards JFK, her Lyft app immediately lights up. Pete just requested a ride and she pleasantly discovers that surge pricing is in effect. As Janet makes her way to Pete, she glances to the side of the highway. The cove where Uber and Lyft drivers usually wait for airport arrivals is empty – apparently everyone just left for passenger pickups. She is in luck!

In greater New York area, by April 2018, there are approximately 128,000 vehicles uniquely dispatched by ride sharing applications and taxi services.[**?**] While some drivers rely on intuition and experience, most drivers lack information regarding when and where to pick up passengers to optimize their earnings.

On the Friday prior to the data competition, our team interviewed Uber and Lyft drivers for their experience on airport pickups. To the drivers, the risk of a) losing money due to tolls, and b) opportunity cost of long drives are relatively high, but the reward is also significant whenever they run into scheduled arrival, delayed arrivals, or

unscheduled diverted arrivals. Some savvy drivers subscribe to port authority alerts[**?**], but the alerts tend to be delayed, which gives drivers the impression that the alerts are unreliable.

Given the high risks and rewards of airport pickups, we ask this question: How might we help hardworking ride sharing drivers like Janet, navigate the fluctuating arrival schedules in local airports, to help them decide when to drive to which airport for passenger pickup, and optimize their earnings?

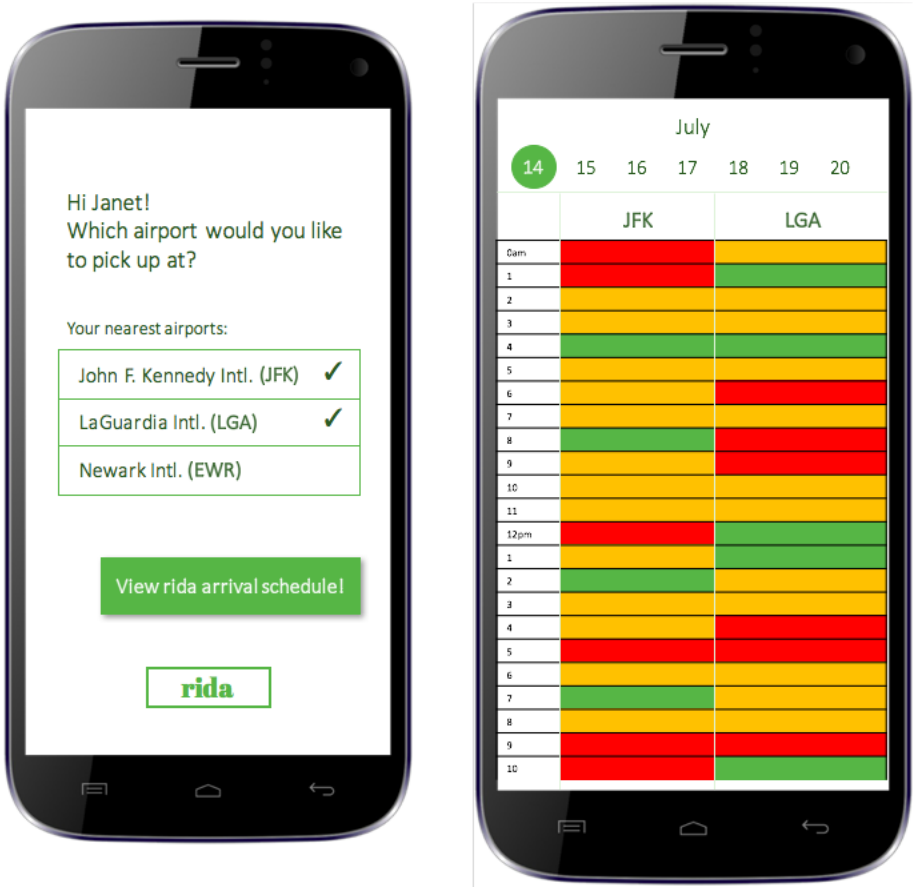

Figure 1: (Left) Setup page of the app. (Right) A typical scheduler customized to the airports picked by the user. Green windows correspond to the times of most activity, orange to the medium and red to the least expected activity.

2 What our application does and how to use it?

Our mobile application, RIDA, allows ride sharing and taxi drivers to find out what time to travel to each airport, in order to minimize opportunity cost and maximize successful pickup rate.

In the application, drivers could set their preferred airports, and see recommended times lots during the day they would be able to pick up large amount of passengers from the airport. The application visualizes the airport arrival traffic, including and scheduled arrivals that will not be canceled, delayed arrivals, and arrivals diverted to the given airport. RIDA also provides simple explanation on why those are good times for pickup.

Reminders to pick up at an airport alerts for arrival changes will be pushed to drivers through texts or application notifications. The application refreshes constantly to ensure that the information is up to date.

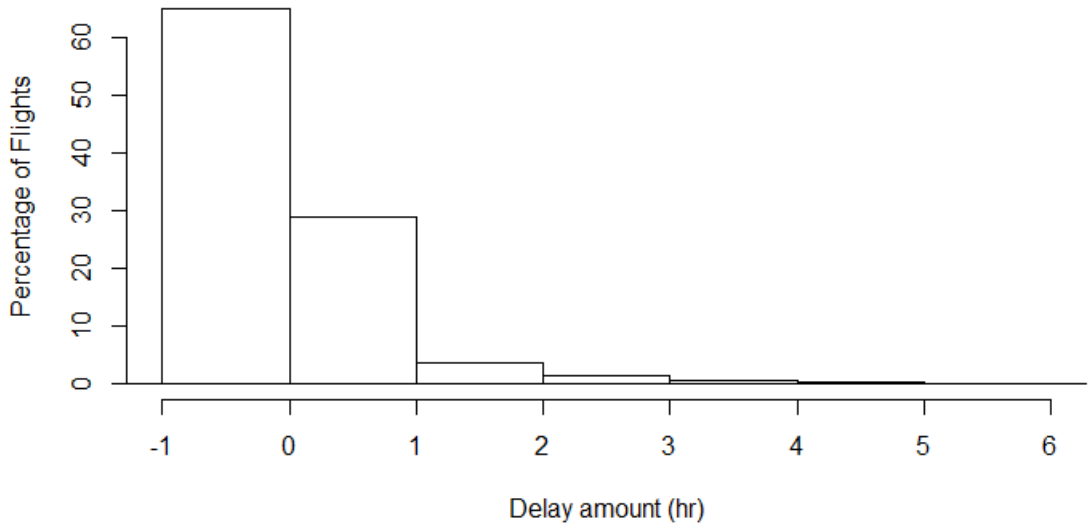

Figure 2: Percentage of the flights delayed in the United States over 2017 by the average number of hours they were delayed.

3 How we designed the application

The application developed by our group aims to estimate the real time, accurate arrival information at the three airports of the greater NYC area. Despite the information given by the airlines, the arrival times or even whether the flight will go to the intended destination is not guaranteed to follow the schedule. Figure 2 depicts the percentage of flights arrived at hourly increments from the scheduled arrival time. Weather at the origin and the destination airports, various flight information can be used to estimate the cancellation probability of the flights and predict the expected amount of delay for different flights.

We used three of the machine learning methods, support vector machines, the neural networks and the random forest for the numerical prediction of cancellation probability and the delay amount. For both of these exercises, the data was divided such that 70% was used to train the models and the 30% of the data was used to test the success of the models against the historical data. The features used in our machine learning models are grouped into two main categories. The flight specific information captures various predictors of accurate flight plan, while the airport specific information captures geographic information and the weather patterns at the origin and the destination airports.

Flight Specific Information	Airport Specific Information
Airline ID	Airport Name
Actual departure time	Elevation
Taxi time before departure	Visibility
Scheduled arrival time	Wind speed
Distance	Snow depth
	Cloud Status

The output for the model used to estimate the cancellation probability was a binary 0 or 1 output predicting whether a flight will be canceled or not. For the model built to estimate the average delay amount, the output variable has four categories: 0 (not delayed) or a number 1-3 corresponding to how many hours a flight is expected to be delayed. This output is then combined with the original flight information to calculate how many flights will arrive at a given airport and the selected time window with high confidence. The collated information will then be presented to the user

4 Technologies and tools used

In summary, we mainly used scikit-learn to run, tune, and adjust machine learning and deep learning models. Pandas library is used to clean and prepare data for the model training and testing. To optimize our model training speed, we used Google Colab, which significantly decreased time spent on training data on multiple models.

5 Next Steps

Seeing the possible business use of this application, we can improve the work by including the following expansions.

5.1 Diverted flight

Include another piece of analysis that will predict the probability that a certain flight will be diverted to a different airport. This analysis will also use supporting historical data to estimate the probability that the new airport is one of the airports in the preferred service area of the rideshare provider

5.2 Including other metro areas

This app was primarily developed to improve the rideshare productivity in the New York City area; there are number of other large metropolitan areas in the United States that can benefit from one such app. Examples for the immediate expansion are Los Angeles area, Silicon Valley-San Francisco area, Chicago metro area. Further upgrades will allow the rideshare provider to select the set of airports they prefer to operate in, and the app will customize the analysis and the presentation to suit the preference of the rideshare provider.

5.3 Commercialization

This productivity app will be first put on the Google and App store to be used by the rideshare provider and will feature the advertisements to provide for the operating and development costs. The refined and improved app will be offered with payment from the end user.