



Recruiting Process: Machine Learning Engineer

Step 3: Challenge

Dear candidate,

in this challenge we will test your ability to handle data, answer questions to specific data sets and present your results. Bear in mind that the questions are vague and unstructured – as is the nature of our work. There is no single way to tackle a problem. If you have any questions while taking this challenge, don't hesitate to ask us.

You can use whatever programming language you like for completing this challenge, although we prefer Python or R. You should take about eight hours of work to finish this task, but you can distribute your time over the coming week. Once you completed the challenge, please send us all your code and slides (as pdf). We will set up an appointment with you where you will present your results. This presentation should be addressed to someone who has no statistical or mathematical background, should be held in English and take 15 minutes max.

We are looking forward to your presentation, as well as meeting you in person, and wish you the best of luck!

The idalab team

NYC Flights

The first data set records over 330 000 flights that departed from New York in 2013 for 105 different destinations in the USA. There are also three metadata sets that contain the full names of the airlines, detailed information about the destinations and the hourly weather data for New York.

Questions:

- 1) Are there any anomalies in the data? If so, how did you handle them?
- 2) How does the distance of a flight influence its delay? What might be the reason for this relationship?
- 3) Which New York airport is the best and why (according to this data only)?