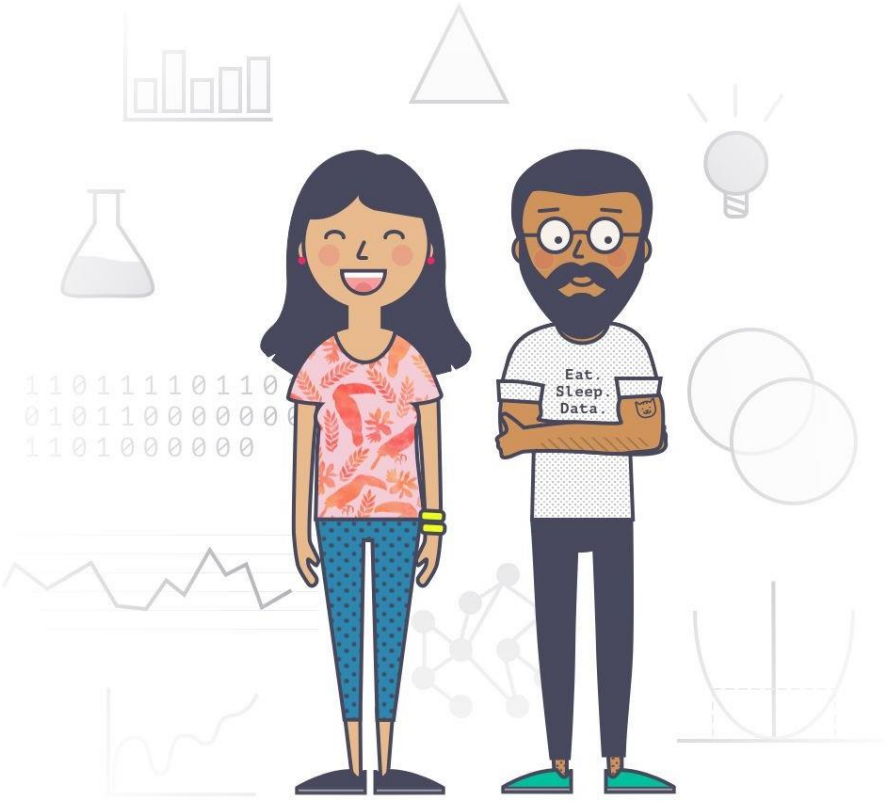
Data science assignment



2017

Dear data science enthusiast,

We're happy you finally reached the document. We trust you, we respect you, and we want you to join our family! Please, do not spread this assignment, our dataset nor your solution.
We expect you to work hard and productively spend around 3-4 hour solving the problems and 10 minutes relaxing and enjoying the situation while giving us feedback about the assignment. When you finish, send us the results in a presentable format along with all code. (e.g. in the format of a jupyter notebook). Don't forget to comment and plot graphs - you have to convince us! Also, one of your task will be to build a classifier.

Here your XXXXXX story
begins! Good luck!

You're given our real dataset containing the information about trips completed for 30 days in some of our city. We randomly picked up all the rides taken by 9000 joyful customers and now we are giving them to you. Also, we've hidden 3000 customer rides as a testing dataset.

The dataset:

customer_id	Unique customer identifier. If a customer took more than one trip within 30 days, there will be more than one entry with the customer's id
driver_id	Unique driver identificator. If a driver made more than one trip within 30 days, there MAY be more than one entry with the driver's id. But be careful, we filtered the data by the customers, not drivers, so a drive may have a ride that's not shown in the dataset
creation_date	Date and time when the customer booked the ride
booking_source	The application type via which the customer booked the trip. It can be Android/IOS App, web/mobile web, etc.
car_type	Type of the car used in the trip. There are different prices and service provided by the different car type. It can be economical, luxury, minivan, etc.
estimated_distance	Estimated distance between pick-up and drop-off location according to our algorithms. Can be empty, if the customer didn't put the drop-off location in the app. Measured in kilometers
distance_travelled	Real trip distance calculated after the trip finished. Measured in kilometers
distance_travelled_while_moving	Distance driven when the car was running fast enough (eg. not stuck in a traffic)
estimated_duration	The number of minutes that we predict the trip will take. Can be empty, if the customer didn't put the drop-off location in the app
duration_time	The number of minutes that the trip really took
wait_time_initial	The number of minutes between the driver arrives to the pick-up location and customer gets into the car
wait_time_in_journey	The number of minutes during the trip when the car's speed was extremely slow (eg stuck in a traffic)
estimated_price	Price that our algorithms predict. Can be empty, if the customer didn't put the drop-off location in the app

price	Real trip price calculated after the trip completed
is_cancelled	Shows if the trip was cancelled
rating	1-5 stars the customer rated the trip. 0 is when there is no data, 1 is a minimal rating 5 is a maximal one
was_rated	Shows if the customer rated the trip

1. Help us to understand our data!

We want you to get a bit deeper in our business process. To be frank we need your help! We have some ideas:

1. We think, the higher a captain's average speed the higher his/her rating. We think customers like speedy rides!
2. Our customers can request a trip price estimation by putting the destination address but it's not required. If a customer requests the price estimation, estimated_price column isn't empty and contains the price our algorithms predict for the trip. We suppose, there is a dependence: customers are more likely to request a trip price if they are going far.

Can you check it out for us and explain the dependencies? If you have a couple of ideas how to check and prove - go ahead, we would like to listen all of them!

2. Hypotheses and insights bringing

Cool to warm up, isn't it? Your understanding of the business process and creativity are rapidly growing and urgently needed at XXXXXX!
Inspired by the previous examples, can you come up with a couple of interesting hypotheses and check them out?
Explain what you've found to your team leader and product manager, convince them that you are right.
For example, you can cluster the trips and then give a meaningful descriptions to the groups you found.

3. Rating prediction

In order to give the best possible service, we want to understand our customers better. We really like feedbacks and to improve the interaction with our customers we need a model to predict if they will rate a trip.
You are given training (the one you already know well) and testing datasets.
Using this datasets, build a model to predict if the trip was rated. (*was_rated* column)
Moreover, in order to help you to gather some additional customer's information, to look at their personal rating history, we prepared some additional dataset trip rating dataset from the previous 30-days period (30 days before the training dataset collected):

customer_id, driver_id, rating, was_rated

Describe the model, feature engineering, parameter and feature selection process and recommend us a metric to estimate your work.
Send us your code and prediction file for the testing dataset in csv format:

entry_id, was_rated

Don't forget, we really appreciate comments

3***. Rating prediction for the most irresistible ones

If the previous task is not advanced enough for you or if you have some extra time, try to predict *rating* (0, 1, 2, 3, 4 or 5 stars) instead.
We bet you can't handle this one, can you? ;)

4. Feedback

As we mentioned, feedback is a really big and important part of XXXXXX culture and we hope you get to experience this on your own. For now, give us some feedback, we really want to find the best colleague for you:

- Did you have any issue understanding the assignment and accessing the data?
- How do you like the dataset? Was it easy for you to understand what all the columns mean? Which columns were difficult to interpret?
- From 0 to 10 estimate the difficulty of each of the tasks tackled above.
- How well defined were the tasks for you? Would you have preferred to have been given additional guidance?
- Additional comments.