Welcome to the Data Science Technical Test

Being hands on is one of the key requirements for this role. We ask you to complete the following tasks within 7 to 10 days and submit the results through Greenhouse. We want to see how you approach the problem described below and also that the rationale you follow makes sense in Glovo's business context. The model you submit needs to be at least better than chance but its actual performance is less important than the explanations you provide along the way.

Therefore, it's important that you write down concisely but with sufficient level of detail the insights you discover, any decision you take during the analysis and to give a proper interpretation of any data visualization you create. Please avoid including in your submission any unjustified data prints, uninterpreted plots or any unused or uncommented code. Consider also explaining any additional steps you think it makes sense to try, even if you don't have time to implement them.

The problem

Imagine you already are a Data Scientist at Glovo. The manager of an important city needs to size the fleet of couriers in order to properly serve the orders we expect to get in the next few weeks. Since couriers (aka glovers) are free to work or to stop working at any moment, they approach you and ask which couriers in the city fleet we can expect to work in the next 3 weeks (the churn window) and which couriers are not likely to come back, a classic churn problem. In order to act on the insights you provide, the city manager will need to have the predictions of your model available at least a week in advance.

We provide some raw data to answer this question using an ML model that you will build following the steps below.

About the Data Set

Attached to this task you will find 2 CSV files. The features are renamed for confidentiality purposes and a data dictionary will NOT be provided. However, in both CSV files, the courier ID's represent the same courier. The first data set consists of the courier's lifetime features, i.e. intrinsic courier-level characteristics. The second one consists of the courier's weekly variant features, so features at a courier-week level for each week a courier has delivered at least one order.

Task 1: Exploratory Analysis and Data Munging

In this task, you are expected to explore and clean the data, treat missing values, find out related features and create your target variable. Before starting, note that each courier doesn't work every week. Thus, some of the courier-week data combinations are not provided.

Now you must create the target variable for this churn problem, so follow this procedure: for each courier, if data for any of the weeks 9, 10 or 11 is provided (the churn window), it means that the courier delivered at least 1 order in the last 3 weeks of the dataset, i.e. they did not churn, so you can label those couriers as "0" (*not churned*); otherwise, label them as "1" (*churned*), since they did not show up in any of the last 3 weeks. After labeling, remove the data for weeks 8, 9, 10 and 11 to avoid a bias in your next task. We ask you to remove also week 8, due to the business requirements of having model predictions available 1 week before the churn window, so week 8 information will not be available at prediction time.

Then perform a complete Exploratory Data Analysis: univariate, bivariate and timing analysis, explore correlations between features and between features and labels and explain your findings. You should also come up with a way to treat missing values and outliers in the data.

Task 2: Create a Predictive Algorithm

For the second task, you are expected to create a model that classifies every courier into churned/not churned just once, using the labels that were created in the first task. First, choose what you consider is the best metrics for the model to optimize during training, which is suitable for the problem at hand and explain your selection. Then create a model using Python or R. You are free to choose any external algorithms and libraries / packages. Last, tune the hyper-parameters of your model using any search method you think is the best and explain your reasoning for this choice.

Task 3: Evaluating the Model

Once the model is trained, as a final task you are asked to assess its performance. First, get an unbiased measurement of the optimization metrics you selected in Task 2 and share your score. Compute any other metrics that you think are relevant to understand the strong and weak spots of your model and use any plot or visualisation necessary for the type of ML problem and the business case we are trying to solve.

The submission

Please submit your solution in a zipped file without passphrase through Greenhouse. This file must include any code you used to solve the business case (Jupyter notebook, R script...), a list of requirements to run that code if any (software versions, packages installed...) and, optionally, a document or slide deck summarizing your results. You don't need to submit any of the data we provided or the data generated during the analysis or training of the model.

If you have any technical questions, please email them to takehometest.ds@glovoapp.com and we will get back to you as soon as we can.

Good luck!