We are an AutoML product, we are looking for a data scientist, who can innovate, maintain code.

This is a perfect test for us to know if the candidate can start working from day 1. This test can be done in a few hours if you have been dealing with these sorts of problems. If not, it will take some time.

We are providing you a small dataset. There are 2 target variables in it. One is a categorical variable (Target Satisfaction) and another is a numeric variable (Target Salary).

The task is -

You have to clean this data, prepare it for ML models, build more features and ultimately build a random forest (scikit).

Split the dataset into 80:20, train:test.

Print the Weighted F1 on test dataset (target\_satisfaction) and RMSE (Target\_salary)

Please note -

If you do this problem without any cleaning or building any additional features, you won't get the RMSE or good F1. There are lots of variables which need feature engineering.

Follow the usual data science process.

Please if you plan to submit any random value, we won’t be giving you a call , since even before running your code in google colab, we know the possible range of answers.

Please don’t print graphs, visualizations, it makes readability of code lower.

Do well comment on the code. Commenting the code well is very important from a readability point of view.

Before attempting this, few more things to let you know, so the expectations are right and you attempt only when you are sure.

1. You have to be very good with python and data science concepts.
2. Please do not attempt it if you are not very good with data science, as this is not a training internship and we expect you to contribute from day 1, with our minimal assistance.

Dataset -

<https://drive.google.com/file/d/16plggcGFOBEvjiYK65--Z6SmRwdeDbUy/view?usp=sharing>