SPICEMONEY Data SCIENCE INTERVIEW

PROBLEM

The attached data files are for understanding customer risk of business.

We want to predict Business risk.

There are two data files in the zip file attached: they are "assignment_train.csv" and "assignment_test.csv".

The last column in assignment_train data set is 'business_risk' - it measures the riskiness of the customer dropping off. Higher this 'business_risk' coefficient, higher are the chances that customer will abandon using our platform.

Your task is to predict the business risk on test dataset, by making use of "assignment_train.csv" dataset to train the machine learning model.

The rows are customer ids(known as 'agent_id') and columns are features observed about their business. The knowledge of column names is not important, however, it should be noted that same prefix of column name implies same family of observation (for eg. '3M_all_min' is 3month minimum sales in all products, and '3M_all_max' is 3 month maximum sales in all products)

Your Output would be predicted values of 'business_risk' for all agent_id in "assignment_test.csv" dataset.

All the features are numerical in nature.

Response:

- Please send a running code (in .ipynb or .py or .R file) accompanied by a text file explaining all the steps and assumptions you made while creating the model
- Your model should predict the business risk coefficient for "assignment_test.csv" dataset
- Bonus points for neatly summarizing the results