Finiata TakeHome Task – Data Science

Dear Data Scientist and aspiring Fini,

Welcome to Finiata's take home assessment for data scientists! This assessment is part of our functional screening part of our hiring process. Finiata is a fully data science centered company. Our business model settles completely on our ability to personalize lending offers to our customers that seek to optimize a fine but challenging balance between risk (of default) and reward (i.e. revenues from fees and interests). Therefore, we are very careful on finding the best data science and engineering talent out there to help us on building this...together! #oneTeam

This task is divided in three parts and it is designed to be solved all-in-all in less than 4 hours. Please use Python for all the tasks below. Python Notebook is a preferable form of sharing the results and code. Good luck and have fun!

Part 1 Machine Learning Foundations

Please answer to the question below using your own words and thoughts as you may need to elaborate over those during later interview stages. Justify your answers with clear line of thoughts and sound justifications (e.g.: technical, empirical; avoiding colloquial words such as "typically"). Please, do not exceed 8 lines/200 words per answer/question.

1 . Imagine that you need to learn a **grouping structure** from a dataset with 8000 features and 10M examples using a regular laptop (assuming that you would not be able to load the entire dataset into its volatile memory). Please, describe the sequence of actions that you would carry out in order to obtain an easily interpretable result in less than 2 hours time.

Part 2 SQL Syntax

Given the below subset of Finiata's schema, write executable SQL queries to answer the questions below. Please answer in a single query for each question and assume *read-only* access to the database (i.e. do not use CREATE TABLE).

1. For each week of the present year, product type and interest_rate, show the sum of the expected profit per week for that loan up to the current week assuming profit=0 for weeks after the *week_first_time_past_due*.

Assume a PostgreSQL database, server timezone is UTC.

Table Name: **open_loans**

Column Name	**Datatype**
id	integer
client_id	Integer (foreign key to client data)
product_id	Integer (foreign key to product data)
amount_granted	integer
duration	integer
interest_rate	Enum('5%,'10%','15%')
expected_profit_per_week	integer
requested_at	timestamp with timezone
ever_went_pastdue	boolean
status	Enum('ok','past_due','defaulted')
week_first_time_past_due	integer

Table Name: **clientdata**

Column Name	**Datatype**
id	integer
city	string
country	string
postcode	string
stated_turnover	integer
company_name	string
director_name	string
director_age	integer
director_contact	string

Table Name: **Product**

Column Name	**Datatype**
id	integer
type	Enum('RCF','Factoring','LaaS')
country	string
name	string

Part 3 Modelling

In this part, we provide you a dataset with 18 features and one target variable [expert_opinion] that we would like to model. For your convenience, we already split the dataset between training and test, respectively. Additionally, provide a written answer to the questions describing the intuition for all the choices that you have made.

1 [Data Visualization] Exploratory Data Analysis is a must-have tool for any data scientist. In this task, we need to explore the presented dataset throughout visual and/or numerical tools and methods order to discover some insights that may help you (and us) to understand better the data and the supervised learning problem at hand.

2 [Modelling] Please, do train a model that maximizes the Area Under the ROC curve on unknown data using only data from the file “train.csv” as part of the training set. Return also at least 1 benchmark of your model’s performance using only the data from the file “test.csv”. Feel free to put forward the pipeline/sequence of tasks that you feel necessary to maximize your result. However, please also note that this model will need to be deployed in production.