**Home Assignment for Data Scientist**

**[Data Science and Analytics Team Prague]**

Q1. PySpark assessment:

You are given data in the bigdata environment, and you need to train a model in the cluster environment. Imagine you chose PySpark to train the model and make prediction on bigdata. (Please add code snippets and the outputs)

1. Setup a local environment to run PySpark.
2. Download the Census Income data at <http://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)>
3. Print the data Schema, Summary, # of columns and # of rows .
4. Print a table that distinct values of all columns.
5. Make exploratory data analysis and visualize your findings from data.
6. Create a binary target variable to predict income above 50k and below 50k.
7. Select the most promising features without using any ML model (assume that dataset contains many features, and you cannot train a model on all of them). Hint: you may use some heuristics, significance tests etc. to select the predictive variables. You may not rely on model-based feature reduction techniques like Stepwise selection, Regularization. The goal here is to not rely on modeling algorithms to select features, as this may be very costly.
8. Should the raw model dataset be randomly distributed to Train/Test before or after identifying the most predictive features?
9. Perform any feature engineering and explain the reasons behind. e.g. New features from existing, re-grouping categorical variable categories, meaningful encoding, binning etc.
10. Choose a ML algorithm to build an acceptable predictive model in PySpark and explain the reasons for the selection of such algorithm.
11. Print the model performance measurements such as ROC, AUC, Confusion Matrix for both train and test data.
12. Explain the steps taken to avoid model overfitting.
13. What changes do you propose to make this modeling process more effective, if you are to train this in a cluster with 1,000 nodes, 300 million rows and 5,000 features in AWS?

Q2. SQL skills assessment:

1. Create your own employee table in the following format: EMPLOYEE\_ID, NAME, PROFESSION. How do you generate a new table with number of employees in each of the following categories of professions: IT, SALES and OTHER? Please provide both your table and the query. Hint: based on the string match in the profession.
2. You have the following definition and the transactional tables. How do you construct the output table with the columns “date, #new\_users, #active\_users, #churn\_users, #reactivated\_users” in SQL? In the definition table (user\_status\_table), “x” represents activity, “-“ represents no activity and “NA” represents when it’s not important.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| user\_status\_table | |  |  |  | activity\_table | |
| **user\_status** | **Activity before 60 days** | **Activity 30-60 days ago** | **Activity in the last 30 days** |  | **date** | **user\_id** |
| **New** | - | - | x |  |  |  |
| **Active** | NA | x | x |  |  |  |
| **Churned** | NA | x | - |  |  |  |
| **Reactivated** | x | - | x |  |  |  |

1. You have two tables as shown below. Someone has accidently deleted a column from the second table and rows from the first table. Please help the analyst to evaluate the data quality issue with SQL queries to:
   1. Identify the users that are not in the first table but are in the second table.
   2. Join the two tables with only one record for duplicates.
   3. Explain your assumptions and approach.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| users\_table |  |  |  | transaction\_table | |  |  |
| user\_id | name | gender |  | index | user\_id | age | weight |
| 1 | A | M |  | 1 | 1 | 23 | 60 |
| 2 | B | F |  | 2 | 2 | 34 | 75 |
| 3 | C | F |  | 3 | 3 | 55 | 80 |
| 4 | D | M |  | 4 | 3 | 43 | 90 |
| 5 | E | F |  | 5 | 4 | 54 | 66 |
|  |  |  |  | 6 | 7 | 23 | 57 |
|  |  |  |  | 7 | 4 | 65 | 64 |
|  |  |  |  | 8 | 7 | 44 | 80 |

* 1. Find how many unique names for each gender category do we have in the users\_table?

Q3: Modelling Assessment:

For this task use the Bank Marketing dataset available on the following address:

<http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

Build 2 binary classification models using any 2 of the following methods (in R or Python):

1. Logistic Regression
2. Random Forest
3. GBM
4. Xgboost
5. Neural Network

Try and minimalize overfitting. Compare the performance of both models using ROC graphs, AUCs, confusion matrices.

Provide also the full source code and description of any variable transformations or balancing performed.