**Senior Data Scientist Challenge**

**This challenge has three case tasks, they are created to test your skills set in the areas of**

* **Credit risk modelling – ML development**
* **Credit Bureau data extraction**
* **SQL code challenge**

**The required attachment to complete the challenge can be found in the body of this email.**

* **Credit Risk Modelling Challenge**

You will be given a small dataset **German*Credit.csv*** with data of customers applying for a loan. The goal

is to predict their defaulting behavior. The data structure described in the Data section of this

Document.

* Train a model to predict defaulting behavior.
* The choice of scoring function(s) is free and must be justified.
* Evaluate your model and discuss the results.

We will evaluate the quality of the implementation, the score of the final model, how you got there, and every additional comment / analysis on the data / model training.

You can either use the tools you want (notebooks or code repository).

**Data Dictionary**

The target is the variable **default**.

The data has the following structure:

* Observation\_id: unique observation id
* Checking\_balance: Status of existing checking account. (German currency)
* Savings\_balance: Savings account/bonds (German currency)
* Installment\_rate: Installment rate in percentage of disposable income
* Personal\_status: Personal status and sex
* Residence\_history: Present residence since
* Installment\_plan: Other installment plans
* Existing\_credits: Number of existing credits at this bank
* Dependents: Number of people being liable to provide maintenance for
* Default: 1 is a good loan, 2 is a defaulting one.
* **Features Engineering Challenge**

The credit bureau report is one of the additional external source data used by Credit Direct Limited for Credit model and ML decision making. Our main Credit Bureau is Credit registry, In this section, we will test your ability to extract and transform features from such json format file.

The expectation is that you

* Extract create all the feature that you think are important to you in the attached file ***Credit\_bureau\_sample\_data.json***
* For each constructed variable, tell us how it could be relevant to improve the risk scoring model.
* The output should be a Python function/class who takes one or more credit reports as input and returns the features as they can be used by a model.

We will evaluate your ability to manipulate semi-structured data and your business sense.

**Data**

The file ***Credit\_bureau\_sample\_data.json***  is a json file that is structured in this way:

[

{

“application\_id”: id,

“data”: credit\_report

},

...

]

This is a list of dictionaries with two elements, the ***application\_id*** which is equivalent to an observation and the ***credit\_report*** which is a json full response that **Credit Direct Limited** recovers from the external provider.

**SQL Code Challenge**

You have been given a data ***BikerDatav2.csv*** which contains information on bike hailing business, write and attach simple SQL query to perform the following basic operations.

* On which day of the week do we on average have the longest trip?
* What month/year has the most bike trips and what is the count of the trips?
* In the same table, return which particular trip has longest duration and the trip that has the shortest duration (return all the information(columns) on the table for this record)

If more than 1 record has the same duration, return the earliest trip [start time]

*NB: Exclude 'Missing' and 'Stolen' as values in the end\_station\_name column.*

*Exclude trips that start and end at the same station.*

*Your final output will be 2 rows*

**Data dictionary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field name | Type | Mode | Key | Description |
| trip\_id | INTEGER | NULLABLE | Primary Key | Numeric ID of bike trip |
| subscriber\_type | STRING | NULLABLE |  | Type of the Subscriber |
| bikeid | STRING | NULLABLE |  | ID of bike used |
| start\_time | TIMESTAMP | NULLABLE |  | Start timestamp of trip |
| start\_station\_id | INTEGER | NULLABLE |  | Numeric reference for start station |
| start\_station\_name | STRING | NULLABLE |  | Station name for start station |
| end\_station\_id | STRING | NULLABLE |  | Numeric reference for end station |
| end\_station\_name | STRING | NULLABLE |  | Station name for end station |
| duration\_minutes | INTEGER | NULLABLE |  | Time of trip in minutes |