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.

Deliverables:

Using a notebook / a code repository, return to us the following:

1) a trained model
2) features + featurization code based off of the provided data
3) an ability to score new customers as they come in
4) a brief analysis of the performance of the model
5) an explanation of
    a) your features and why you chose them
    b) how you evaluated the model
    c) qualitative analysis of model performance

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.

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.

• Feature Engineering Challenge

The credit bureau report is one of the additional external source data used for Credit model and ML decision making. 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 features that you think are important 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 which 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 is received 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