Integrity
Above
All

ICEA LION Group
Chiromo Road
Nairobi

Nairobi, 8 September 2022

Dear Data Scientist Candidate

Re: Test

After our initial reviews we are happy to invite you to an aptitude assessment for the position of Data Scientist. The assessment has two parts, four general questions and a specialized analytics case study. The assessment is designed to score your ability to align with our corporate strategy and needs. Please find the case study below. You are required to submit it by EOD Monday 12th September 2022.

Sincerely
ICEA LION Group

General Questions:

1. Mention the palette of tools which you would recommend to be used for analysis of our data, and write for each of them how you would be using each tool.

2. What metrics would you track on a regular basis, and how do you use the information to adjust your approach?

3. How would you go about identifying the unique needs for new products from our customers? You have all departments at your disposal, and you need those data, telling management what our customers needs are - so that we can start delivering! What would you do?

4. Analysis: The good old joke persists;

   "20% of all traffic accidents are caused by people who have drunk alcohol. This means that 80% of all traffic accidents are caused by drivers who are sober. Conclusion: Drive drunk!".

   Now from a data analytics perspective, explain step by step with arguments presented in good order, whether this above conclusion is right or wrong, and how you do the analysis to reach to the conclusion.

Case study

This dataset contains columns simulating credit bureau data. Credit default risk is the risk that a lender takes the chance that a borrower fails to make required payments of the loan. The main purpose of this analysis is to predict whether a new customer can be a reliable customer. It's a way to avoid default and increase the bank's revenue. This can be used to automate approving and declining loan applications more accurately.

Case Study Tasks:

Please prepare a presentation with around **5 to 10 slides** that provides an insight driven analysis of the problem. The analysis should include the answers to the tasks described below. You'll find more information for the tasks attached.:

1. Using the attached dataset (credit_risk_dataset_training.csv) analyse the data and visualize the most important aspects using your preferred method. Furthermore,

share three ideas on how to increase the % of successful loan applicants. Document your steps where needed.

2. Predict the outcome of a loan: is a customer likely to satisfy or default on the loan obligations? Use credit_risk_dataset_training.csv to train your model and credit_risk_dataset_test.csv to predict the missing value ('loan_status'). Please document your steps and method used. Include the accuracy or evaluation metric used for calibrating your model

Key success factors:

- Be precise and structure your answers in a clear manner. Don't beat around the bush.
- Presentation of the answers is key. Show us what you did.
- Showcase your creativity and have fun doing it!

About the Dataset

Detailed data description of Credit Risk dataset:

Feature Name	Description
person_age	Age
person_income	Annual Income
person*home*ownership	Home ownership
person*emp*length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan*int*rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan*percent*income	Percent income
cb*person*default*on*file	Historical default
cb*preson*cred*hist*length	Credit history length

The objective of this challenge is to create a machine learning model that will predict whether a customer will satisfy or default on their loan obligations.

Files available for download:

- **credit_risk_dataset_training.csv** - contains the target variable “loan_status”. This is the dataset that you will use to train your model.
- **credit_risk_dataset_test.csv-** resembles the training dataset but without the target-related columns. This is the dataset on which you will apply your model to for submissions and evaluation.