**Overdose AI**

**Data Scientist Intern Practical Exercise**

Hello there! Congratulations on being selected for the next round of the hiring process!! In this round, you will be provided with a practical exercise that will test your programming, analytical and report writing skills.

**Note:** *The exercise will involve 2 parts, Part A and Part B.*

* *Part A is mandatory for all applicants.*
* *Part B is optional and includes a more challenging task.*

*Although Part B is not mandatory but completion of it significantly increases the likelihood of the candidate being called for the next rounds.*

**Instructions (Please read carefully):**

* The programming must be done on Jupyter Notebook using Python language. (IDEs such as VSCode may be used to boost productivity.)
* Google Colab is also allowed for the coding but ensure to download it as a .ipynb file for submission.
* Taking references from the internet is accepted but the candidate must ensure that they understand the concept. They may be asked questions regarding the code.
* A thoughtful, clean & commented code is expected during submission.
* Please do not upload the submission in any public sites such as GitHub. Doing so immediately disqualifies the candidate.

**Deliverables:**

1. The Jupyter Notebook (.ipynb) file
2. A word document (Report)

If Part B is attempted,

1. Additional Jupyter Notebook (.ipynb) file

After you are done with the assignment below, please upload the deliverables to Google Drive and fill the submission form below:

[Submission Form Link](https://forms.gle/vmSXsq87SZSPv7TFA)

**Note:** *Don’t use public sites such as GitHub as that immediately disqualifies the candidate.*

**Part A**

Dataset: [Link](https://drive.google.com/file/d/1fWUG__B-11mV2td-eoqS7eF1eADnsdRs/view?usp=sharing)

The attached dataset “daily\_offers.xlsx” is a representation of steel items offered to customers. Each row represents an item, and has an offered/selling price of that item.

You as the Data Scientist is tasked to:

* Analyze all the features, using statistical and visualization methods of your choice and prepare the data for a regression model. (You may demonstrate your creativity and skills here by involving techniques like data summarization, correlation, data cleaning, feature engineering, etc.)
* Split the data into training and testing sets in the ratio of 90:10 respectively. Build multiple regression models to predict the price of an item (At least 4 models).
* Hyperparameter tune the best regression model.
* Evaluate the best model using the metric r2 score on the testing set.

Finally, you are required to create a detailed report to document your work where you may showcase your analysis, visualizations, results, findings, etc.

**Remarks:** *We are more interested in understanding your thought process, creativity and analytical skills. Use this as an opportunity to extract as many insights as possible. Feel free to use any type of visualizations and also any techniques that you feel is justified.*

**Part B (Optional)**

Dataset: [Link](https://drive.google.com/file/d/1e9dOqKxIBaRdfFBVhiPOeN2CrOQZRYOs/view?usp=sharing)

You are given the dataset “movieReplicationSet.csv” which features ratings data of 400 movies from 1097 research participants. The dataset details are as:

*1st row: Headers (Movie titles/questions)*

*Columns 1-400: These columns contain the ratings for the 400 movies (0 to 4, and missing)*

*Columns 401-420: These columns contain self-assessments on sensation seeking behaviors (1-*

*5)*

*Columns 421-464: These columns contain responses to personality questions (1-5)*

*Columns 465-474: These columns contain self-reported movie experience ratings (1-5)*

*Column 475: Gender identity (1 = female, 2 = male, 3 = self-described)*

*Column 476: Only child (1 = yes, 0 = no, -1 = no response)*

*Column 477: Movies are best enjoyed alone (1 = yes, 0 = no, -1 = no response)*

You as the Data Scientist is tasked with specific instructions:

* Impute the missing ratings with a blend (50:50) of the arithmetic mean of each column and each row. (Let’s say that the rating of user 350 for movie 200 is missing and that the average rating of this user for other movies is 4 and the average rating (by other users) for this movie is 3, then the to-be-imputed rating would be 3.5, using this method.)
* For each of the 400 movies, use a simple linear regression model to predict the ratings. Use the ratings of the \*other\* 399 movies in the dataset to predict the ratings of each movie (that means you’ll have to build 399 models for each of the 400 movies). For each of the 400 movies, find the movie that predicts ratings the best. Then report the average COD (Coefficient of determination) of those 400 simple linear regression models. Please include a histogram of these 400 COD (Coefficient of determination) values and a table with the 10 movies that are most easily predicted from the ratings of a single other movie and the 10 movies that are hardest to predict from the ratings of a single other movie (and their associated COD (Coefficient of determination) values, as well as which movie ratings are the best predictor, so this table should have 3 columns).

**Remarks:** *We are more interested in understanding how well equipped are you in understanding extremely detailed and specific instructions.*