**Problem Statement**

Solving this problem will help you to gain a basic understanding of the HR industry

and also how analytics and data science are applied in practice in this domain.

**A brief understanding of the HR industry**

Human resources (**HR**) is the division of an organization that is responsible for finding, screening, recruiting, and training job applicants, and administering employee-benefit programs.

This management is the strategic approach to the efficient management of people in any organization in a way that can help their business gain a competitive advantage. It is designed to improve employee performance in service of an employer's strategic objectives.

**Understanding the Problem Statement**

In this project, you will have to put yourself in the shoes of an HR analytics expert and by leveraging the data provided, you are supposed to identify **how salary is varying with respect to other considerable factors**.

The dataset should be divided into test and train and you would have to use the training data to build the model and thereby apply it to test data to measure the performance and robustness of the model.

**General steps to be followed and tentative timelines:**

1. Understand the problem statement: 1 hour
2. Understand the data, make a valid business sense and write about it: 4 hours
3. Data analysis and interpretations: 4 hours
4. Data Cleaning and justification : 5 hours
5. Feature engineering(If needed) : 4 hours
6. Model building : 5 hours

(Justify why are you choosing a particular technique over others)

1. Cross-validation, performance check, and its explanation : 4 hours
2. Rework on different models, if needed : 4 hours
3. Visual interpretation : 2 hours
4. Results and Recommendations: 2 hours

**Steps to be performed for successful completion of capstone project:**

1. Understand the data set - Feature description provided below.
2. Check the number of variables that are normally distributed & explain
3. Find out the correlation between variables using spearman rank correlation & explain
4. Examine the relationship between categorical variables and the dependent variable & explain
5. Clean and pre-process the data with other operations required, like, missing values, handling outliers, etc. as and when necessary & explain
6. Apply ML algorithm to the pre-processed columns and find out, which of these algorithms is best suitable for the same & explain

Briefly write your inference for every step, justifying the logic for the same.

**Feature Description**

1. Gender - talks of the gender - Male or female

2. Business - if the person has a business or no

3. Age - age of the person

4. Salary - CTC of the employee

5. Dependants - number of people dependant on the person

6. Months - duration of the person employed with the company

7. Calls - if the person has authority to make calls or not

8. Type - salary settlement type

9. Billing - Subscribed to billing plans or no

10. Unit sales - unit sale made by the person

11. Total sales - total sales made by the person

12. Rating - If he has been given a rating by a superior or no

13. Bonus - amount received by a person as bonus for sales

14. Base pay - Base pay of the employee

15. Unit price - The Unit price of a sale

16. Volume - volume allotted to a person

17. Opening balance - The opening balance of an employee

18. Low - lowest balance allotted to a person.

19. Closing Balance- The closing balance of an employee

20. Education- Educational background of an employee

**Grading system**

The final submission should have the following components:

1) All codes (properly documented) - 10%

2) The code explaining what and why they did in each step.

- Steps 3 to 5 : 30%

- Steps 6 to 7 : 30%

- Step 8 to 10 : 30%

**How are you contributing to the organization**

Based on your inputs the hiring manager would understand how different features considered are affecting the salary component. This analysis can greatly help in making huge structural changes to the financial wing of the HR department.