## Capstone Project Proposal (DS106)

Problem Statement:

The project objective is to develop a compensation prediction model, which calculates the total sum of additional benefits to your salary.

The problem here is a regression problem to predict the total compensation of city's employees based on his or her overtime, other salaries, retirement, health and dental and other benefits.

Introduction of dataset:

Dataset: Employee_Salary_Compensation.csv

The San Francisco Controller's Office maintains a database of the salary and benefits paid to City employees since the fiscal year 2013. This dataset contains more than 650k employee records found in San Francisco from 2013 to 2020. The dataset consists of 678,524 rows and 22 columns.

There are 22 attributes of the dataset have as per below:

1. Organization Group Code: Org Group is a group of Departments. For example, the Public Protection Org Group includes departments such as the Police, Fire, Adult Probation, District Attorney, and Sheriff.
2. Job Family Code: Job Family combines similar Jobs into meaningful groups.
3. Job Code: Jobs are defined by the Human Resources classification unit. Examples include gardeners, police officers, and accountants.
4. Year Type: Fiscal (July through June) or Calendar (January through December).
5. Year: An accounting period of 12 months. The City and County of San Francisco operate on a fiscal year that begins on July 1 and ends on June 30 the following year. The Fiscal Year ending June 30, 2012, is represented as FY2011-2012.
6. Organization Group: Org Group is a group of Departments. For example, the Public Protection Org Group includes departments such as the Police, Fire, Adult Probation, District Attorney, and Sheriff.
7. Department Code: Departments are the primary organizational unit used by the City and County of San Francisco. Examples include Recreation and Parks, Public Works, and the Police Department.
8. Department: Departments are the primary organizational unit used by the City and County of San Francisco. Examples include Recreation and Parks, Public Works, and the Police Department.
9. Union Code: Unions represent employees in collective bargaining agreements. A job belongs to one union, although some jobs are unrepresented (usually temporarily).
10. Union: Unions represent employees in collective bargaining agreements. A job belongs to one union, although some jobs are unrepresented (usually temporarily).
11. Job Family: Job Family combines similar Jobs into meaningful groups.

12. Job: Jobs are defined by the Human Resources classification unit. Examples include gardeners, police officers, and accountants.
13. Employee Identifier: Each distinct number in the “Employee Identifier” column represents one employee. These identifying numbers are not meaningful but rather are randomly assigned for the purpose of building this dataset. Employee ID has been included here to allow users to reconstruct the original report. Note that each employee’s identifier will change each time this dataset is updated, so comparisons by employee across multiple versions of the dataset are not possible.
14. Salaries: Normal salaries paid to permanent or temporary City employees.
15. Overtime: Amounts paid to City employees working in excess of 40 hours per week.
16. Other Salaries: Various irregular payments made to City employees including premium pay, incentive pay, or other one-time payments.
17. Total Salary: The sum of all salaries paid to City employees.
18. Retirement: City contributions to employee retirement plans.
19. Health and Dental: City-paid premiums to health and dental insurance plans covering City employees. To protect confidentiality as legally required, pro-rated citywide averages are presented in lieu of employee-specific health and dental benefits.
20. Other Benefits: Mandatory benefits paid on behalf of employees, such as Social Security (FICA and Medicare) contributions, unemployment insurance premiums, and minor discretionary benefits not included in the above categories.
21. Total Benefits: The sum of all benefits paid to City employees.
22. Total Compensation: The sum of all salaries and benefits paid to City employees.

The dataset was obtained from the website of Kaggle as per the link below.

https://www.kaggle.com/siddheshera/san-francisco-employee-salary-compensation

Sample of the dataset

| | Organization Group Code | Job Family Code | Job Code | Year Type | Year | Organization Group | Department Code | Department | Union Code | Union | ... | Employee Identifier | Salaries | Overtime | Other Salaries |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1000 | 1021 | Calendar | 2013 | Public Protection | ADP | ADP Adult Probation | 21.0 | Prof & Tech Engineers - Miscellaneous, Local 21 | ... | 37730 | 57534.65 | 0.0 | 0.00 |
| **1** | 1 | 1000 | 1023 | Calendar | 2013 | Public Protection | ADP | ADP Adult Probation | 21.0 | Prof & Tech Engineers - Miscellaneous, Local 21 | ... | 26574 | 57678.50 | 0.0 | 0.00 |
| **2** | 1 | 1000 | 1031 | Calendar | 2013 | Public Protection | ADP | ADP Adult Probation | 21.0 | Prof & Tech Engineers - Miscellaneous, Local 21 | ... | 8148 | 63532.93 | 0.0 | 0.00 |
| **3** | 1 | 1000 | 1054 | Calendar | 2013 | Public Protection | ADP | ADP Adult Probation | 21.0 | Prof & Tech Engineers - Miscellaneous, Local 21 | ... | 27436 | 101274.51 | 0.0 | -7058.59 |
| **4** | 1 | 1000 | 1062 | Calendar | 2013 | Public Protection | ADP | ADP Adult Probation | 21.0 | Prof & Tech Engineers - Miscellaneous, Local 21 | ... | 37730 | 5084.00 | 0.0 | 0.00 |

5 rows × 22 columns

Challenge or Difficulties:

The challenges that were anticipated during the preparation of the data for modelling were to identify the missing values and decide whether to drop similar or unimportant features, which could cause discrepancy of the prediction model.

Another challenge would be for data processing, in order to prevent the training dataset from having an excessive number of columns, we need to select the best fit encoding method. With several numerical attributes, proper scaling is needed.

Lastly, we need to decide the best hyperparameter tuning results on the selected model to improve the model performance.

Some questions that can lead to the goal from this dataset:

Analysis:

1. How are base pay, overtime pay, and benefits allocated between different department?
2. What is the percentage that allocated to retirement based on the job?
3. How does different jobs influence the salary given?
4. Which department have the most overtime claims?
5. When was the highest compensation?
6. Which organization group have the most department with the largest difference in compensation?
7. Which Union has the highest average health and dental?

Goals:

1. To find the top 3 baseline models to test, we will be using LayzRegressor from lazypredict library to compare across all the models available. After which, we will select the top 3 models and do hyperparameter tuning.
2. Find suitable performance metrics for the model.
3. Explore tuning parameters needed for hyperparameter tuning.

Hope this prediction is able to help The San Francisco Controller's Office to have better planning of their budget and resources to help prevent under or over estimations and maintain a stable calculation through the years to come.