Capstone Project – Employee Salary Compensation

DS106 | Lee Yi Ping

Problem Statement

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

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

Goals To Solve The Problem

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

Dataset – Employee Salary Compensation

- 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.

- It contains 678,524 rows and 22 attributes.

Sample of 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

Important attributes from the dataset

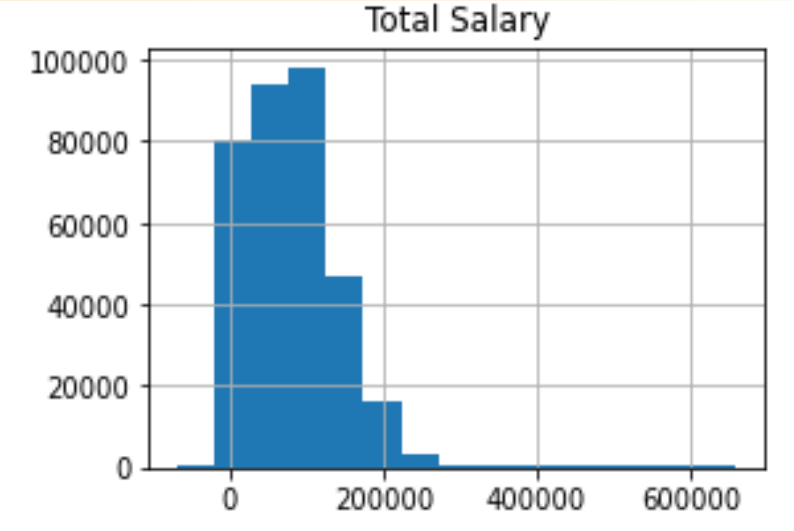
1. **Organization Group:** Org Group is a group of Departments
2. **Department:** Departments are the primary organizational unit
3. **Job:** Jobs are defined by the Human Resources classification unit
4. **Total Salary:** Consists of Salary, Overtime and Other Salary paid
5. **Total Benefits:** Covers Health and Dental, Benefits and Other Benefits
6. **Total Compensation:** Comprises of Total Salary and Total Benefits

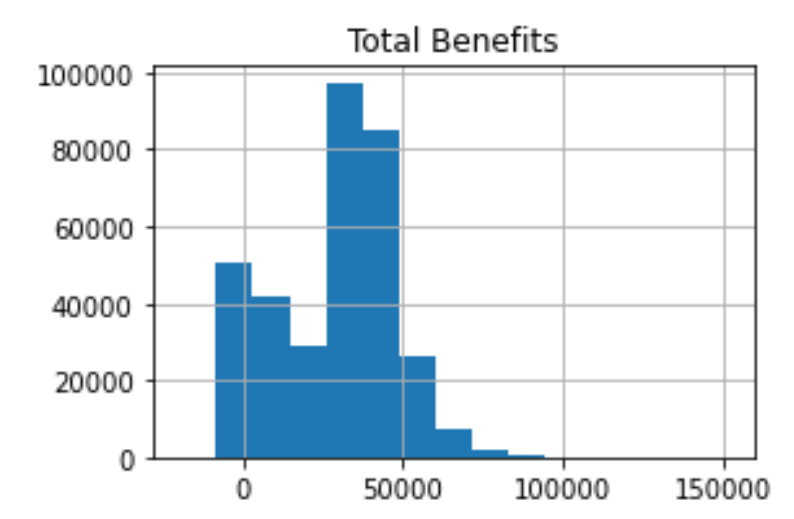
Challenges and Difficulties

- Data Cleaning:
  - To replace missing values from both similar feature, ie. Code and name.
  - Same code with multiple similar values for 4 features to be standardize.
  - To decide which features to drop that is not needed for modelling.
- Data Quality Assurance:
  - To justify whether negative value is needed for model training.
- Which scaling would best fit to the training models.
- Selection of encoding method for categorical feature.
- How to have best hyperparameter tuning results on the selected model to improve the model performance.

Data Exploration/Insights

- ***Analysis – Numerical Attributes***





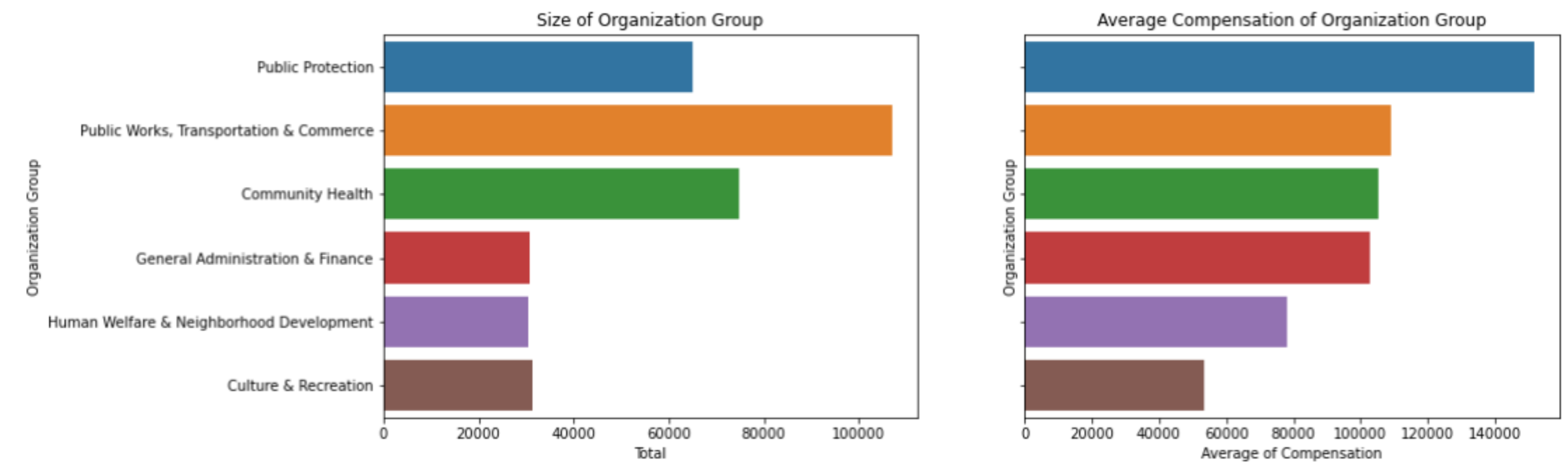
1. **Zero and negative dollar from "Total Salary" and "Total Benefits" and this was only 0.3% of the data.**

2. **Negative values may be due to adjustment for being overpaid or employees have left the organization but had overclaimed their compensation like flexi benefits.**

3. **For zero amount from "Salaries", it could be an employee only receiving a one-time payout, ie. Temp staff.**

Data Exploration/Insights

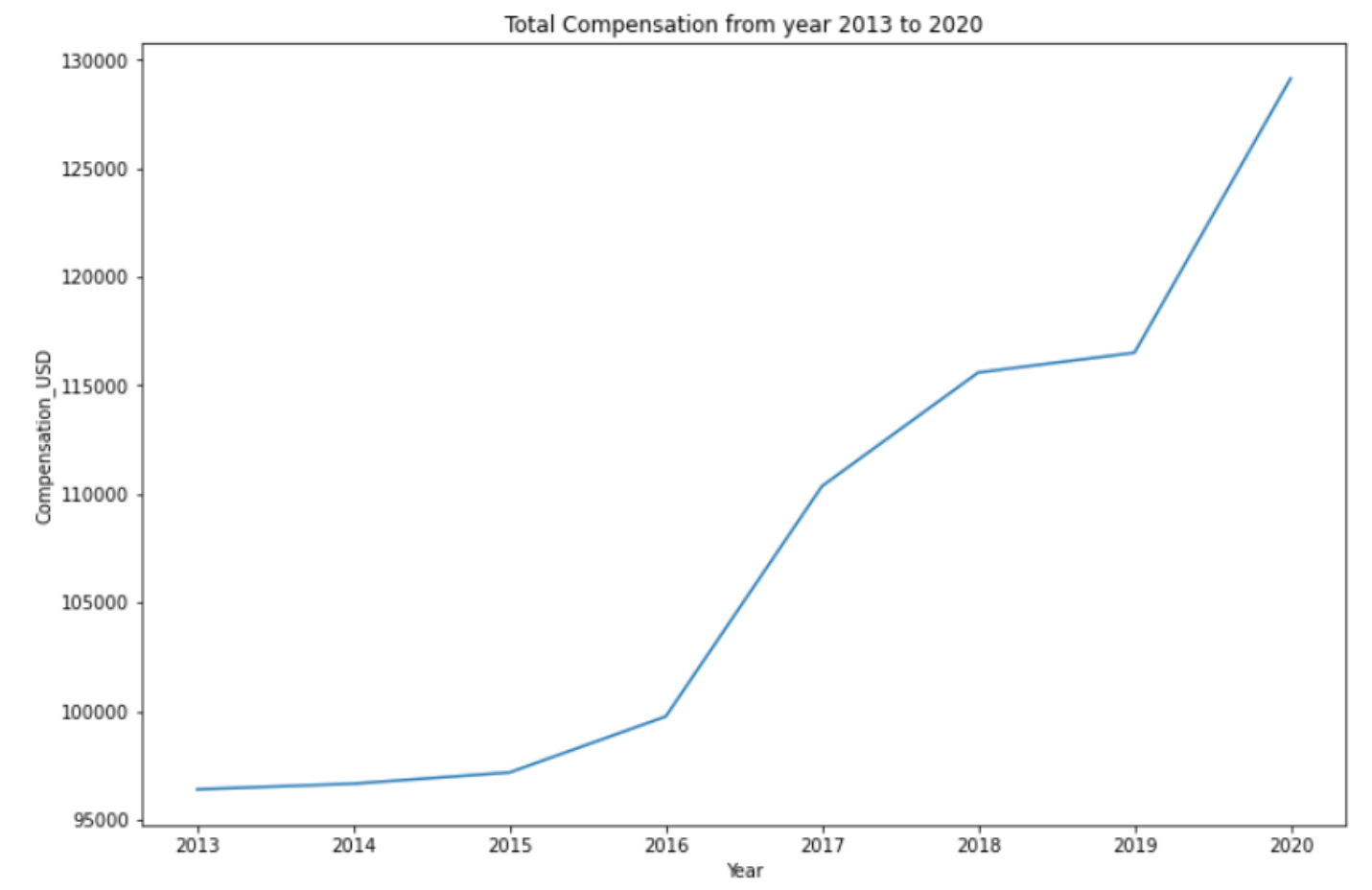
- ***Comparing the size against average compensation of Organization Group***



- From the chart, we can see that Public Protection have the highest compensation among all the other organization even though it did not have the highest count.
- It could be due to the nature of their job as protecting the public as fire fighter or police.
- Although the count of Community Health is double from General Admin and Finance, their average compensation is the same.

Data Exploration/Insights

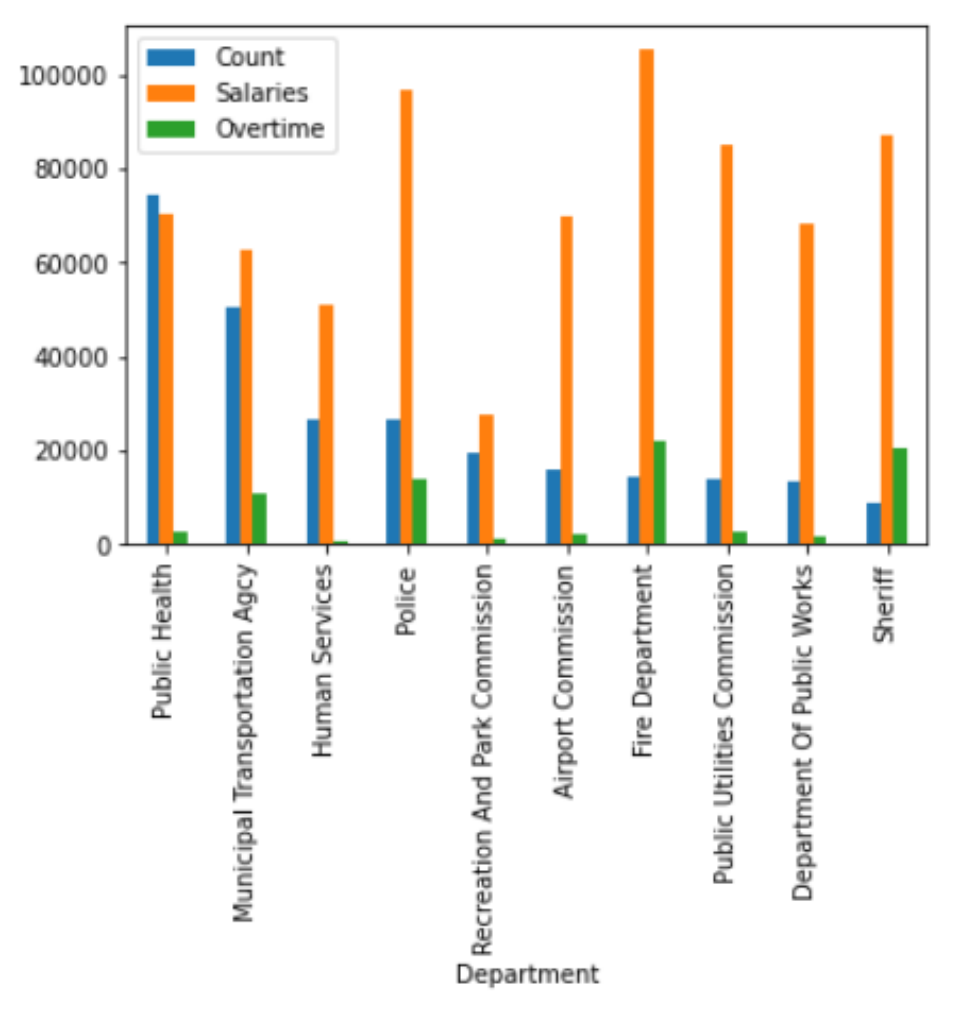
- ***How did the compensation change over the year?***



- ❑ We can see from the chart, year 2013 to 2015 are more consistent as compared to year 2016 to 2017 and year 2019 to 2020 that have more incremental amount.
- ❑ After 2-3 years of consistency, there is a rise in the compensation of at least 10%.
- ❑ The surge in 2017 and 2020 could be low turnover rate which will have more increment of salary and benefits.

Data Exploration/Insights

- ***Which is the top 10 department with the highest count and what is their salaries and overtime claims?***



- ❑ **Based on the chart, we can see that the number in Public Health is inline with the Salaries paid out but OT claims is lesser. It could be due to their role not require OT and also refer to contract that certain threshold of salary range there will be no overtime claims.**

- ❑ **Fire Department have the highest salaries and overtime claims followed by Police among the top 10 department, this could be their work nature is more high risk and greater responsibilities compared to others. Hence the City attract people to take up the job by giving more remuneration as it is essential and this is a job that required passion.**

Data Exploration/Insights

- ***Is the nature of the job affect their health and dental or salaries?***

	Job	Count	Average of Compensation	Average of Salaries	Average of Health and Dental
686	Transit Fare Inspector	21440	96366.785885	52985.094664	11444.607673
516	Public Health Nurse	12548	45648.925017	35634.193819	810.054410
132	Clinical Nurse Specialist	11034	166224.617600	109706.573817	12109.601439
225	EMT/Paramedic/Firefighter	6982	174649.207004	98895.884610	14780.140126
140	Communications Line Wrk Sprv 2	6963	11425.460409	7762.933193	2016.117336
591	Sergeant 2	6818	169770.841049	106502.620034	11917.858282
183	Custodial Assistant Supervisor	6745	77328.325081	47325.127663	11139.674463
531	Recreation Specialist	5838	9304.817622	7341.643407	1084.786257
64	Attorney (Civil/Criminal)	5833	19066.088411	12656.177619	3543.739834
501	Principal Account Clerk	5661	93396.715231	61706.106317	11666.038253

- ❑ **Transit Fare Inspector has the most headcount in San Francisco City but their salaries and compensation is not the highest as it may due to blue collar nature of work.**
- ❑ **Public Health Nurse has the lowest Health and Dental as all the medical expenses is under the hospital cost.**
- ❑ **Among the top 10 of Job, we know that EMT/Paramedic/Firefighter is the high risk job hence their health and dental coverage is the highest.**

Key Steps in Data Pre-processing

Missing Observation Analysis

There is null values and it has been all replaced without dropping any rows.

Outlier Observation Analysis

There is outlier in numerical feature however we are not removing it.

Drop Unwanted Column

Drop features not needed for machine learning: “Year Type”, “Employee Identifier”.

Key Steps in Data Pre-processing

Duplicate Values Checking

After dropping the feature "Employee Id" there are 3,275 rows duplicated and we drop and keep the last row.

Encode Categorical Features

1. Using OHE to encode "Organization Group" and "Job Family"
2. Encode "Department" using Frequency Encoding.

Feature Scaling

Scale numerical feature using Standard Scaler except for Target, ie. Salaries and Benefits. There are total 6 numerical features.

Training Process of the machine learning model

Model Selection and Baseline Model Training Results

1.1 Decision Tree Regressor
1.2 Random Forest Regressor
1.3 Linear Regression
1.4 Gradient Boosting Regressor
1.5 XGB Regressor
1.6 Extra Trees Regressor
1.7 Histogram Gradient Boosting Regressor
1.8 Elastic Net
1.9 AdaBoost Regressor
1.10 KNeighbors Regressor

	model	MAE	MSE	RMSE	R2
0	DT	1589.005005	12811814.619121	3579.359526	0.997567
1	RF	954.097765	5449582.376056	2334.434059	0.998965
2	LR	812.674334	3514773.755877	1874.772988	0.999333
3	GB	2162.101329	11129592.078538	3336.104327	0.997886
4	XGB	1339.204868	5042895.325087	2245.63918	0.999042
5	ET	736.096118	3951594.743749	1987.861852	0.99925
6	HGB	1518.409661	13219835.271078	3635.90914	0.997489
7	EL	9243.999993	160319247.846479	12661.723731	0.969554
8	AB	21832.722324	724092652.969227	26908.969749	0.862488
9	KN	2823.047665	35540737.834126	5961.605307	0.993251

Hyperparameter Optimization – GridSearchCV

Based on baseline model results, we are comparing by the metric "MAE" to evaluate the model as we are keeping the outliers and the metric of mean_absolute_error can robust the outliers. Hence, we had chosen the 3 models with the lowest values of "MAE" for hyperparameter tuning.

1. Random Forest Regressor

- Best parameters after 5 x attempts :
  - "n_estimators": [100,110,120,130,140] **– 140 the best**
  - "max_depth": [29,30,33,35] **– 33 the best**

2. Extra Trees Regressor

- Best parameters after 5 x attempts :
  - "n_estimators": [300,320,350] **– 300 the best**
  - "max_depth": [27,33,38] **– 33 the best**

3. Linear Regression

- There is no hyperparameter tuning on this algorithms.

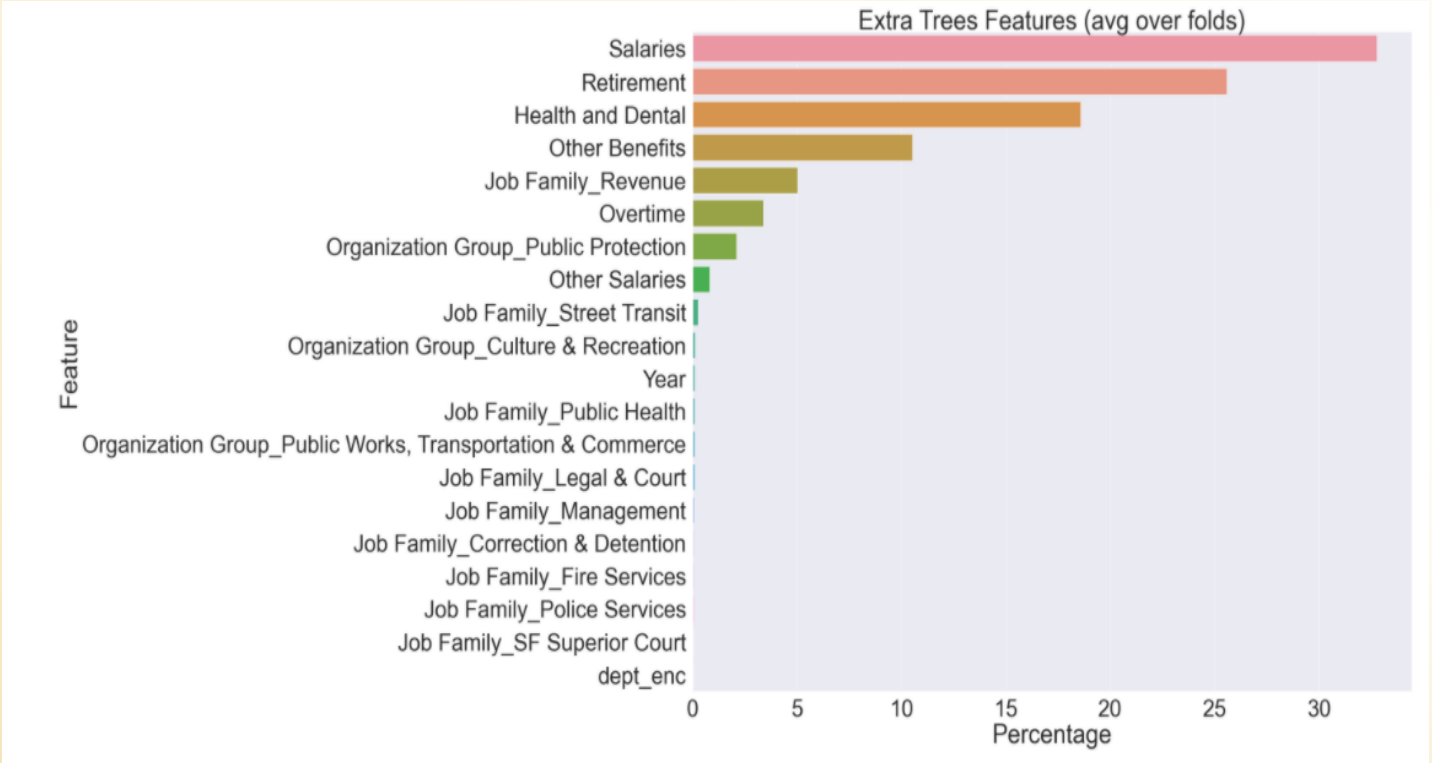
Model Selection and Training

- ***Comparison of metric results for 3 models before and after tuned.***

Metric	Random Forest (Baseline)	Random Forest (Tuned)	Extra Tree (Baseline)	Extra Tree (Tuned)	Linear Regression
MAE	954.097765	951.667821	736.096118	722.776508	812.674334
RMSE	2334.434059	2337.067295	1987.861852	1979.969633	1874.772988
R2	0.998965	0.998963	0.99925	0.999256	0.999333

By comparing the metric results of the table above, Extra Tree Regressor is the best model and we choose this model to forecast the employee salary compensation with the lowest average discrepancy between the actual and predicted value among the 3 models. The other 2 models are not chosen due to not much improvement after tuning. Extra Tree Regressor can forecast the compensation with mean absolute percentage error (MAPE) of 0.67% which are a very good results with a lower error rate.

Feature Importance (Extra Trees Regressor)



	feature	importance
1	Salaries	0.317464
4	Retirement	0.262234
5	Health and Dental	0.180690
6	Other Benefits	0.109664
31	Job Family_Revenue	0.052704
2	Overtime	0.032186
11	Organization Group_Public Protection	0.023092
3	Other Salaries	0.008669
35	Job Family_Street Transit	0.003152
8	Organization Group_Culture & Recreation	0.001367
12	Organization Group_Public Works, Transportation & Commerce	0.001201
0	Year	0.001009

Based on feature importance, we can see that salaries is 32% weightage among 38 features that we have included in the training model and followed by retirement of 26%. This 2 features have taken up half of the weightage and it is common to have "Salaries" holding the highest weightage as it is a basic requirement. San Francisco City is providing the employees with a good benefits of retirement to ease their worries after they retired. However, "Overtime" and "Other Salaries" were less than 5%. This shows that San Francisco City is taking care of their employees' work life balance because of the employees clock lesser overtime.

Conclusion

1. The prediction model 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. It can also help the city to request more manpower cost if they know in advance.

2. The City may have more time to requisite funds for essential stuff to make good of the city and provide maintenance of the public places after having sufficient fund for their manpower cost. They can take this opportunity to cut cost on certain department and provide adequate remuneration for the lower wage jobs.

- ***What would you have done differently if you were to do the project again?***
  - To have better handling on large dataset on training model
  - Train the feature separately and ensemble difference algorithms
  - To have more EDA on the dataset

Appendix

Dataset

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

GitHub

- https://github.com/aureliapy/SF_employee_salary_compensation

THANK YOU