New York City Current Job Postings

Data Preparation Project Report by Amruta Katke

Introduction

Salary is important in most basic sense. Majority of people would not do their jobs if they are not getting paid for it. When people look for a job, the major deciding factor for most of the people is salary. Fair salary for a particular job is also important. The actual salary number is important but its relative sometimes. For example, many people would be wondering if they are getting paid sufficiently or if they are getting paid above or below the market. Salary of any particular job is dependent on factors like job title, job level, job category, job location, etc. There are various job sites which provide salary related information which is only limited to some factors or some geographic areas. These job sites provide information only for some specific category of the job or profiles. Such information does not consider important factors like job location, job category, the level of the job, etc.

Now a days, there are multiple sites which are using data science to predict the salary using various predictors related to the job profile. These sites predict salary for various job profiles within certain geographic locations. These predictions bring transparency to the murky side the job market.

The objective of this project is to leverage machine learning algorithms to predict the salary based on independent variables like job level, job category and other job-related information. The project aims at visualizing information and understanding the relation between various independent variables. The project also focuses on choosing best predictive model for the given data and projecting new salaries for new jobs. The goal of the project is to apply appropriate methods for handling missing values, methods for dimension reduction and model selection prior to the use of machine learning model.

Data Exploration

Dataset: New York City Current Job Postings

The dataset New York City Current Job posting contains current job postings in the city of New York. It has 3527 data samples and 28 columns/variables. The columns are all related to the job postings in New York area.

Dataset Source: Kaggle

This dataset is maintained using Socrata's API and Kaggle's API. Socrata is assisting countless organizations with hosting their open data and has been an integral part of the process of bringing more data to the public.

Link: https://www.kaggle.com/new-york-city/new-york-city-current-job-postings#nyc-jobs.csv

About the Dataset

Size: 3527 x 28

Number of numerical variables: 4

Number of Categorical variables: 24

1. **Job.ID**: This is a job opening ID. This predictor is a numerical variable.
2. **Agency**: Name of the City agency where the vacancy exists. It's a categorical variable with 54 different values.
3. **Posting.Type**: This attribute tells about the type of the job posting like Internal or External job posting.
4. **X.Of.Positions**: This is the number of vacancies to be filled. It's a numerical variable.
5. **Business.Title**: This attribute gives the business title i.e. the title of the job position. It has 1551 unique categorical values.
6. **Civil.Service.Title**: Civil service title has 350 different categories.
7. **Title.Code.No**: Civil Service title Code gives the code for those 350 Civil Service title categories. This is a numerical variable.
8. **Level**: This categorical attribute provides the Civil Service title level like 0, 1, 2... M1, M2 etc. It has 14 different levels.
9. **Job.Category**: This attribute has 1872 different job categories or fields of the jobs.
10. **Full.Time.Part.Time.indicator**: This is giving the type of the job opening as Full time or Part time.
11. **Salary.Range.From**: This is the starting salary value for that particular job. This is a continuous numerical variable.
12. **Salary.Range.To**: This is the proposed salary to value. This variable and Salary.Range.From variables are giving the salary range for a particular job posting.
13. **Salary.Frequency**: This gives the proposed salary frequency in terms of the categorical variable like Annual and Hourly and Daily.
14. **Work.Location**: This is the agency location in New York for that particular job and this column contains 216 different locations that mean 216 categorical variables.
15. **Division.Work.Unit**: Department/Division within the hiring agency. This has 715 different categories.
16. **Job.Description**: Description of the job responsibilities for this position. It has 1971 unique values.
17. **Minimum.Qual.Requirements**: Minimum qualifications required for position. It has 351 different values.
18. **Preferred.Skills**: Preferred skills for this position.
19. **Additional.Information**: Additional information provided by the hiring agency.
20. **To.Apply**: Instructions on how to apply for this position.
21. **Hours.Shift**: Working hours and shift information.
22. **Work.Location.1**: Specific work location for this opening.
23. **Recruitment.Contact**: Recruitment contact information.
24. **Residency.Requirement**: Residency requirements for this position.
25. **Posting.Date**: Date that the position was posted.
26. **Post.Until**: Date Posting will be removed. (Blank means post until filled)
27. **Posting.Updated**: Last Modification Date.
28. **Process.Date**: Dataset created date.

Objective: To predict salary for the job postings

Target Variable: Salary

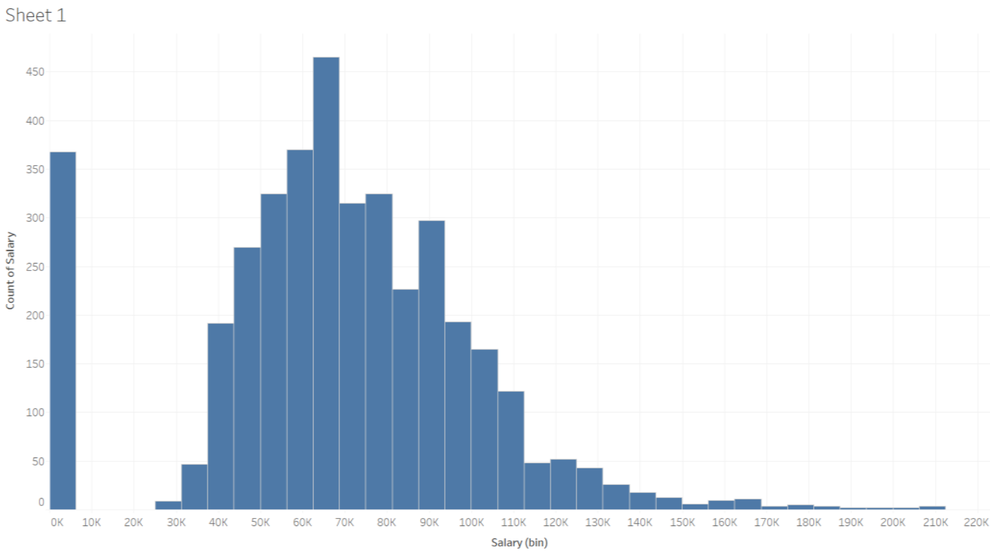
The dataset contains two columns related to salary which are Salary.Range.From and Salary.Range.To. In order to use machine learning algorithm to predict salary we need only one continuous numerical variable. The final target variable was created by combining both the variables. The average of Salary.Range.From and Salary.Range.To was taken as the target variable.

$$Salary = (Salary.Range.From\ + Salary.Range.To)/2$$

Data Visualization

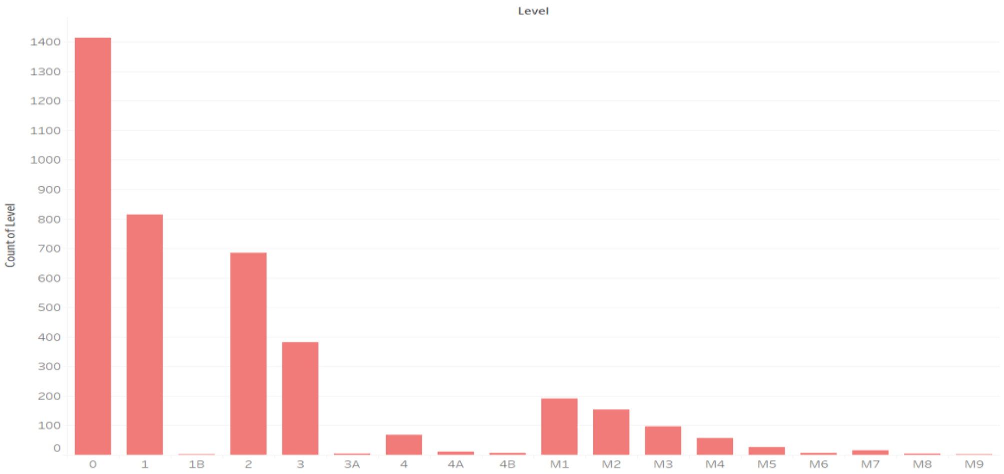
The plots are created using Tableau software for relevant variables towards predicting the salary.

1.Histogram of Target variable

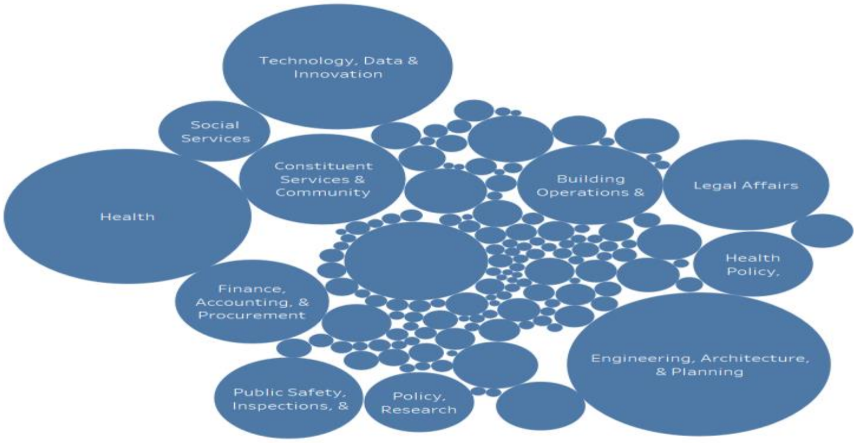


The trend of count of Salary for Salary (bin).

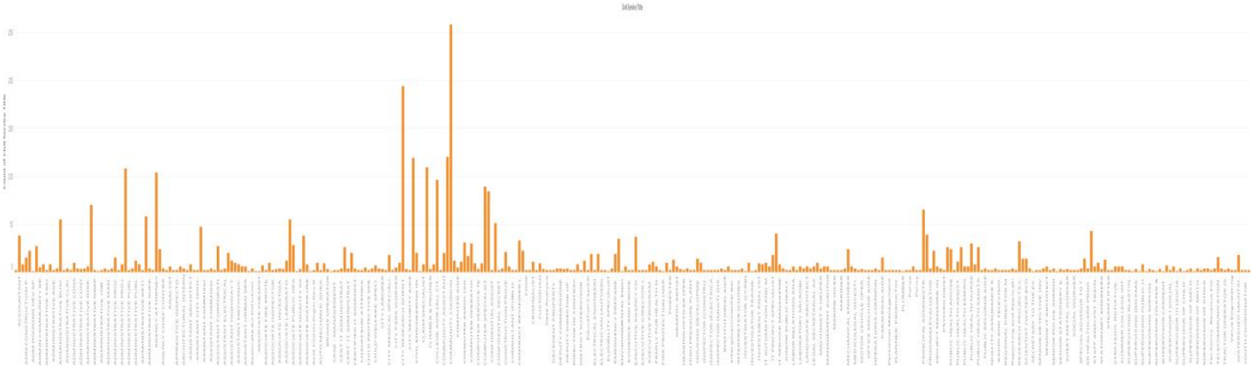
2.Histogram of Level variable



3.Packed Bubble Graph of Job Categories: The categories with higher frequency are visible in the named bubbles.

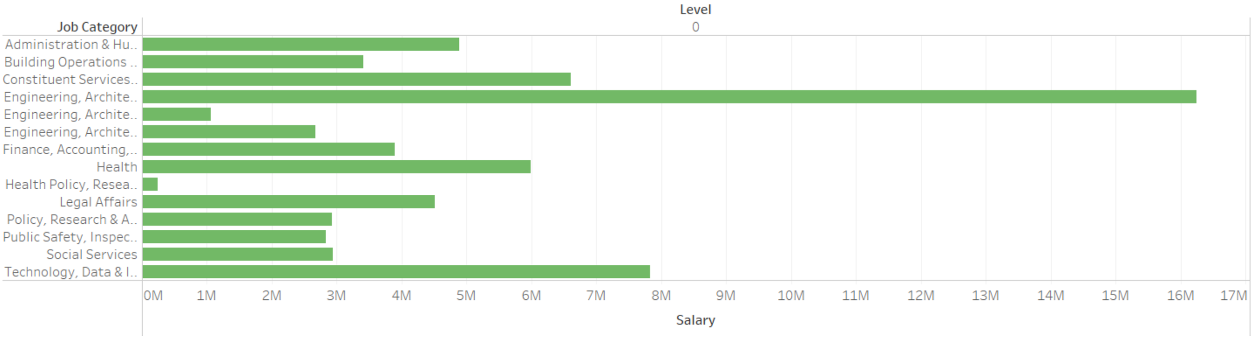


4.Histogram of Civil Service Title



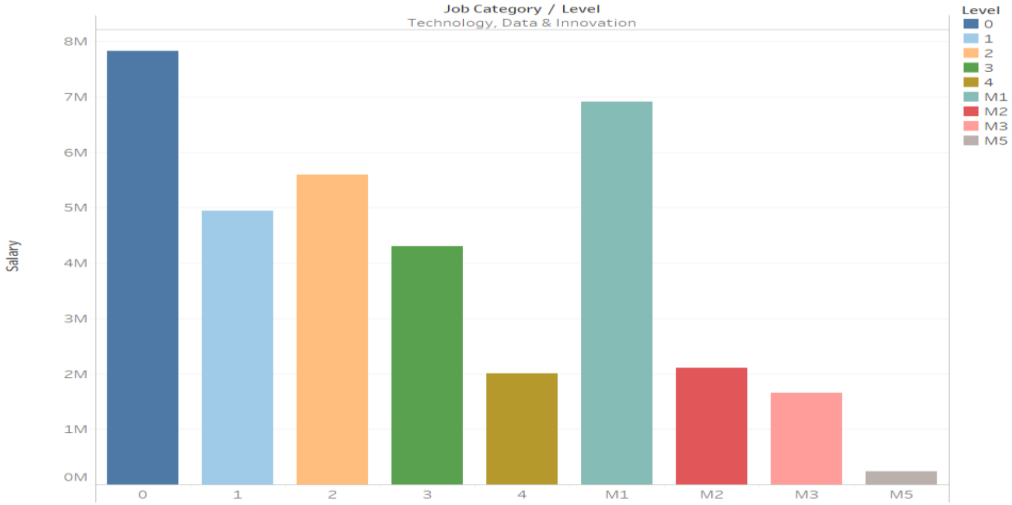
5. Job Category Vs Level (0)

The following chart shows that the salary for the job categories differ for the same job level-0. Later it was checked for other job levels.



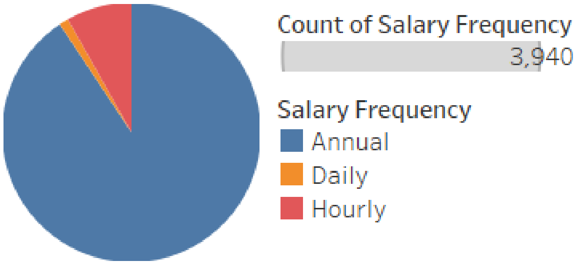
6. Bar Chart: Job Category (Technology, Data & Innovation) Vs Job Level

The following bar chart shows that for any job category the salary differs based on the its job level.



7. Pie Chart of Salary Frquency

Annual salary has highest number of data samples. Hourly salary frequency has smaller number of data samples than Annual type but greater than Daily salary frequency.



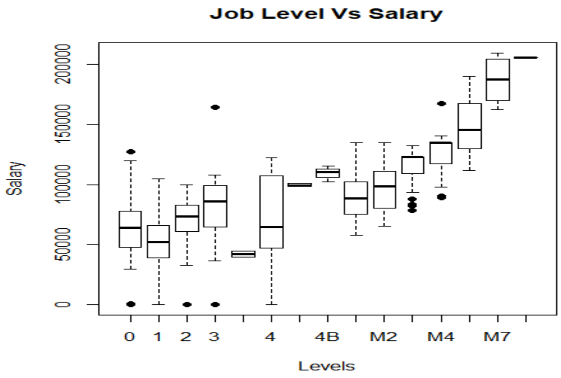
Data Preparation

Missing value Imputation

Missing values in the columns Salary.Range.From or Salary.Range.To were handled. If any value was missing in the Salary.Range.From, then it was imputed with the corresponding Salary.Range.To value. Similarly, the missing values in Salary.Range.To were imputed with Salary.Range.From value. The imputation was done this way as the target variable is the average of Salary.Range.From and Salary.Range.To.

Outlier Detection

The dataset does not have many numerical values, it mostly consists of categorical predictors. In existing numerical values i.e. Salary, there were no outliers. The outliers Detection was done by plotting box-plots and by checking points greater and smaller than upper quartile and lower quartile. Scatter plot also helped in detecting outliers. For example, some of the data samples were removed with level 0, 2, 3, etc.



Variable Selection

Domain Knowledge: Domain knowledge played an important role in deciding the data to be used for prediction. Contextual information of the problem at hand helped to make the project faster and to yield a useful answer. In prediction of salary, the possible predictors were chosen based on their relevancy towards the salary prediction.

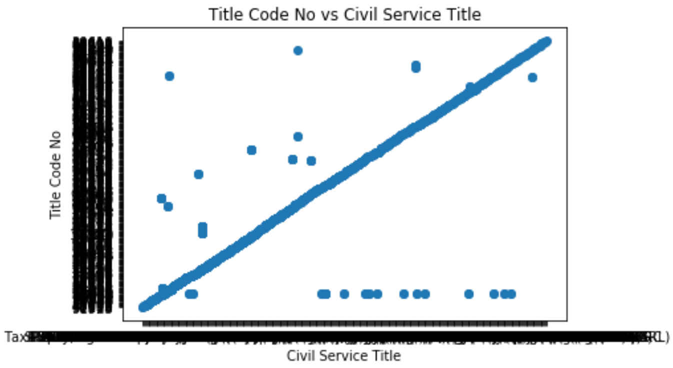
Following columns were dropped as they are insignificant for salary prediction.

1. Job Description
2. Number of positions
3. To Apply
4. Recruitment Contact
5. Residency Requirement
6. Posting Date
7. Post Until
8. Posting Updated
9. Minimum Qual Requirements
10. Preferred Skills
11. Work.Location
12. Division.Work.Unit
13. Full-Time/Part-Time indicator
14. Work Location 1
15. Hours.Shift
16. Posting.Type
18. Process Date
19. Job ID
20. Posting Type
21. Agency
23. Business Title

Model Selection

For model selection along with domain knowledge methods like correlation analysis was used. Civil.service.Title and Title.Code.No are highly correlated which we can see from the correlation analysis of these categorical variables.

The Correlation value is 0.916456404, hence one of the variables can be removed.



Title.Code.No was derived from Civil.Service.Title variables hence they are highly correlated to each other. Adding both the variables to the model will affect the model performance due to multicollinearity problem.

Later, through model selection method of backward elimination the Title.Code.No was dropped as it was not statistically significant to the salary prediction. Also, there are large number of job tittles and each containing its unique Title.Code.No. The final model should be selected such that it can accommodate new job postings in future. If Title.Code.NO is one of the predictors, then the model will not perform if new job postings are added later. Level and Job.Category are the variables which can accommodate new job postings added in future, as thousands of new job profiles are being added each year. Each new job profile will have its category and level, hence Job.Category and Level are significant to for the salary prediction.

The significance of Level and Job.Category was verified by the model selection methods and Pvalue analysis. Salary.Frequency is an important variable as it is a differentiating factor for salaries of particular range. If the Salary.Frequency is Annual then it has higher salary value in the range of thousands, whereas if the Salary.Frequency is Daily then the salary is in the range of hundreds and its in the range of 0-60 if the Salary.Frequency is Hourly.

Predictors for salary prediction

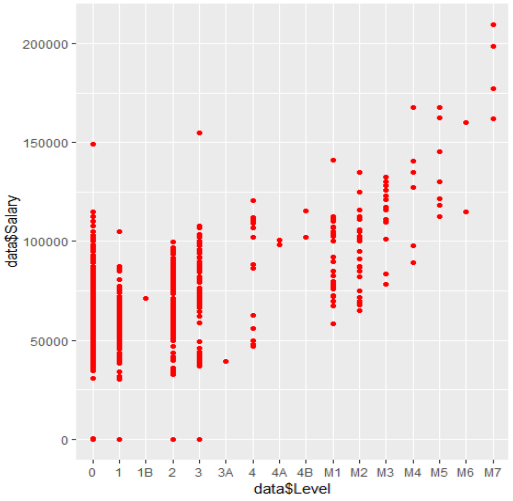
Following predictors were chosen for the final model.

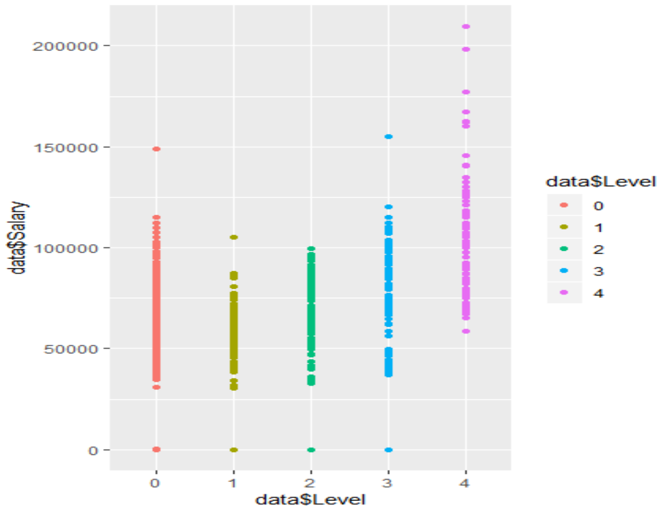
1. Level
2. Job.Category
3. Salary.Frequency

Data Preprocessing

Data preprocessing was performed on the final model for better performance and for dimension reduction of categories in Level and Job.Category. Conversion of each category into dummy variable is not possible in regression models. Some categories need to be removed as every category cannot be used in the model and not every category is statistically significant for the prediction.

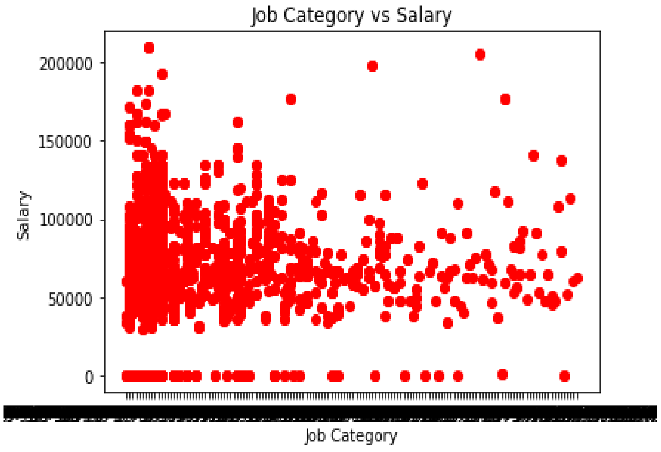
Level:

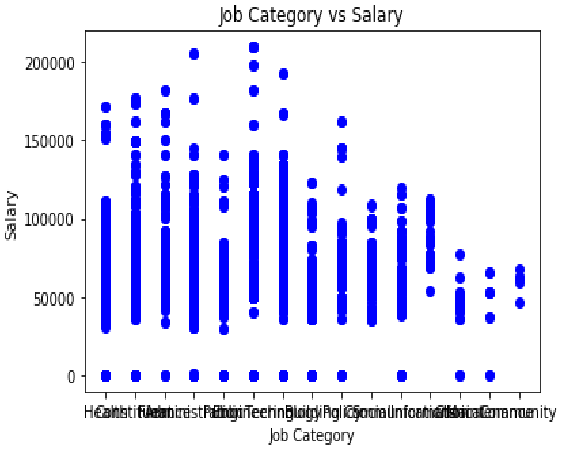




Levels with low numbers of data samples were combined with its adjacent levels. For example, level 1 and 1B we combined to get level 1, level 3 and 3A to 3, level 4, 4A and 4B to 4, etc.

Job.Category:





1. Categories like Health Public Safety, Inspections, & Enforcement + Health Policy, Research & Analysis, + etc. were combined to get Health category
2. Finance, Accounting, & Procurement + Finance, Accounting, & Procurement Health, + etc. were combined to Finance

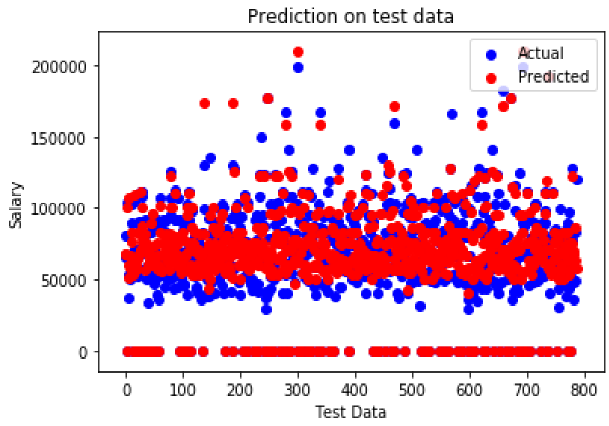
3. Engineering, Architecture, & Planning + Engineering, Architecture, & Planning Finance, Accounting, & Procurement + etc. were grouped together to get Engineering category. And many such categories.

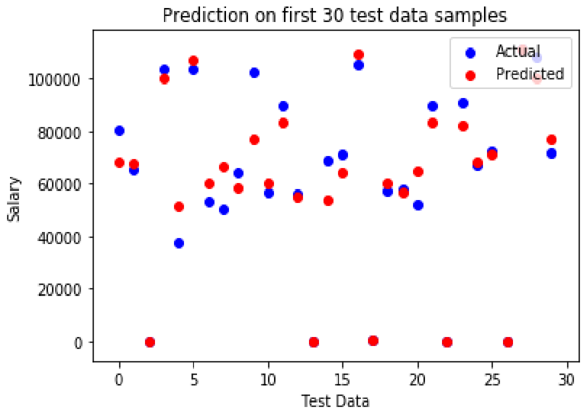
Results

The model is trying to predict a continuous numerical variable (Salary) hence this requires regression analysis which is supervised learning task.

Model 1: Decision Regression Tree (Training data: 80%, Test data: 20%)

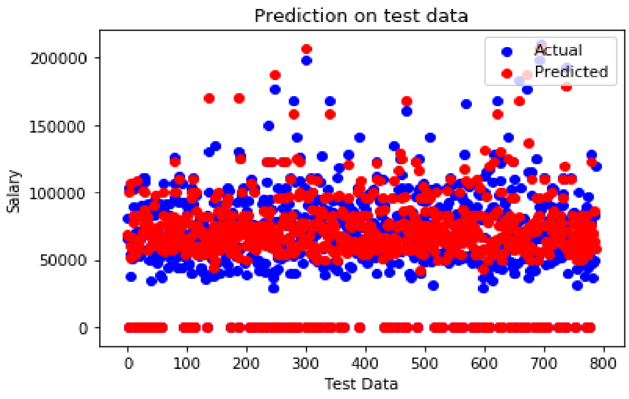
```
Mean Absolute Error: 10451.406009246966
Mean Squared Error: 210929973.66081384
Root Mean Squared Error: 14523.428440310292
```

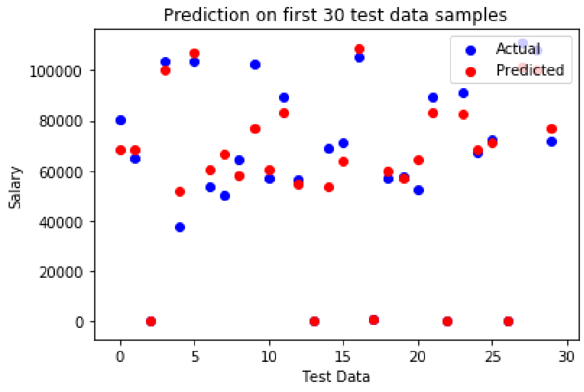




Model 2: Random Forest Regression (Training data: 80%, Test data: 20%, No. of Trees: 200)

```
Mean Absolute Error: 10579.580253247572
Mean Squared Error: 211804421.62367666
Root Mean Squared Error: 14553.502039841705
```





Summary

Model	MAE	MSE	RMSE
Decision Regression Tree	10451.406	210929973.660	14523.428
Random Forest Regression	10579.580	211804421.623	14553.502

After comparing the results of Decision Tree and Random Forest, it concludes that Decision Tree performed better in predicting the salary. There is no much difference in the RMSE, MSE, MAE values of two methods. The RMSE of the model with Decision Tree Regression is lesser than that of Random Forest Regression. Although Decision Tree performed better than Random Forest, the MSE, RMSE for decision tree model is high.
The reason behind the high values of errors in both the cases is that the target variable was not accurate. The target variable was generated by taking the average of two salary range variables. If there would have been an accurate measure of salary variable in the dataset then these two models would have performed even better.

Other methods like Multiple Linear Regression, Polynomial Regression were considered to carry out the prediction task. The level vs Salary plots were in curved shape hence Polynomial Regression method was being considered. Although these methods perform better in regression tasks, the model for salary prediction had large number of categorical variables. Multiple Linear Regression and Polynomial Regression need conversion of categorical variables to dummy variables which was not helpful in this project. Decision trees and Random Forest take categorically encoded variables as input without changing those variables to dummy variables/one-hot encoded variables. Hence Decision Tree Regression and Random Forest Regression were the best suitable methods to perform the salary prediction task with all the categorical predictors.

Conclusion

The project helped me in learning various data visualization methods and drawing conclusions from them. Multiple experiments were performed on visualizing the data and finding out the relation between the predictors and the target variable. Data visualization also helped in dimension reduction for correlation analysis of two or more predictors. It was taken as an opportunity to learn visualization tools like ggplot in R and tableau software. Dimension reduction is an important task in data processing, which was performed properly to reduce the number of predictors which are not statistically significant to the model. Although, the project did not employ PCA for dimension reduction as numerical predictors were not there, but the concept was understood during the lectures implemented in homework assignments. Model selection methods learned during the class were helpful in selecting most significant predictors. Machine learning algorithms were properly studied hence choosing appropriate models for such prediction tasks was easier. Large number of numerical predictors are not difficult to handle in the implementation of algorithms, but categorical variables are. Combining and grouping multiple categories which are similar, handling large number of categorical variables for better predictions and better performance of a model was accomplished through the project.

Links

GitHub Repository: https://github.com/amrutask/DataPreparationProject