Exploring the Factors That Influence Base Salary: A Comparative Analysis Of Regression Techniques

Saurav Upadhyaya

CS 588-01

Department of Computer Science

The University of Alabama in Huntsville

su0010@uah.edu

22th April, 2023

Table of Contents

List of Figures

1. Introduction

The "**Exploring the Factors That Influence Base Salary: A Comparative Analysis of Regression Techniques**" project intends to look into the numerous elements that affect pay and evaluate their effects using regression techniques. The Citywide Payroll Data (Fiscal Year) dataset, which has 5.11 million rows and 17 columns of data on payroll numbers, agency names, employee names, start dates, work locations, job titles, base salaries, pay bases, regular and overtime hours, and total pay, will be used for this project[1]. The project aims to compare several regression algorithms to analyse the factors that influence base salary based on these factors by studying this enormous dataset in order to acquire useful insights into the factors that influence base salary in the public sector.

2. Data

The Citywide Payroll Data (Fiscal Year) data is big data in terms of the 5Vs model. The 5Vs model refers to the five key characteristics of big data: volume, velocity, variety, veracity, and value. Analysing the data in the context of each of these characteristics;

1. Volume: The dataset contains information on salaries and pay for tens of thousands of employees over multiple years, which means it has a large volume of data.
2. Velocity: The data is collected and updated annually, which means it is continuously growing at a fast pace. Additionally, the data is generated and processed in real-time, which requires fast processing and analysis.
3. Variety: The dataset contains a variety of different data types, including text, numerical and categorical data.
4. Veracity: The data comes from a reliable source, Office of Payroll Administration, New York So, it can be trusted and is consistent.
5. Value: The data has significant value for understanding government salaries and pay structures, and can be used for a wide range of purposes such as budgeting, policy-making, and transparency in government.

Therefore, this dataset can be considered big data based on the 5Vs model, as it has a large volume, a moderate velocity, a variety of data types, high veracity, and significant value for various applications.

3. Methods

Before analysing the data, some data preprocessing steps are performed. LabelEncoder is used for converting categorical variables into numerical values aiming to assign a unique integer to each category, thereby transforming them into numerical labels that the model can understand. It seems there are certain columns that are not much needed. So, it has been removed. To Standardise the features of a dataset, StandardScalar has been used, which transforms the data so that it has zero mean and unit variance[4]. Also, to make the data more interpretable and to ensure all features are on the same scale, it is used. To

reduce the complexity of the model, improve its performance and to reduce the risk of overfitting, SelectKBest and Principal Component Analysis are used for feature selection[3]. SelectKBest selects top k features based on their statistical significance. It does this by evaluating the relationship between each input feature and the output variable and selecting the k features with the highest scores. Actually, it identifies the k most important input features that have a high significant impact on the target variable. This helps to reduce the dimensionality of the data and improve the performance of the model. KBest selects Pay Basis, Regular Hours and Agency Name as best features. PCA is also used to ensure that the most relevant and informative features are selected for the model[2]. In PCA, the original features are transformed into a new set of orthogonal features which captures the most important features in the data. It does this by identifying the directions of maximum variance in the data and projecting the data into these directions. This helps to reduce dimensionality, where its goal is to reduce the number of features while preserving its information. After applying PCA, the reduced feature set comprised the columns Regular Gross Paid, Pay Basis, Total Other Pay. Regular hours and Agency Name.

Regression techniques are used to analyze the factors that influence base salary. In this case, the input variables are the factors that influence base salary and the output variable is the actual base salary. The performance of several regression techniques such as Random Forest, Gradient Boosting, Linear and two voting regressors that combine several regression techniques are compared.[3]

1. Random Forest Regressor: As part of ensemble learning, the RandomForestRegressor combines different decision trees to produce predictions that are more precise. It is famous for its capacity to handle non-linear correlations between variables and is frequently used for regression issues.
2. Gradient Boosting Regressor: It is an additional ensemble learning method that combines a number of insufficient predictors (in this case, decision trees) to produce an adequate predictor. It is frequently employed for regression issues and is renowned for its capacity to manage complex interactions between variables.
3. Linear Regression: Simple yet effective, linear regression presupposes a linear connection between the input factors and the output variable. It is simple to understand and can be used to pinpoint the most crucial output variable predictors.
4. voting regressor: An approach called the voting regressor combines different regression algorithms to provide a predictor that is more precise. In this instance, two voting regressors are used. Voting Regressor I includes Linear Regressor, Decision Tree Regressor, and Random Forest Regressor. Voting Regressor II includes Random Forest Regressor, K Neighbors Regressor, GradientBoostingRegressor, and Extra Trees Regressor[5].

4. Data Visualization and Analysis

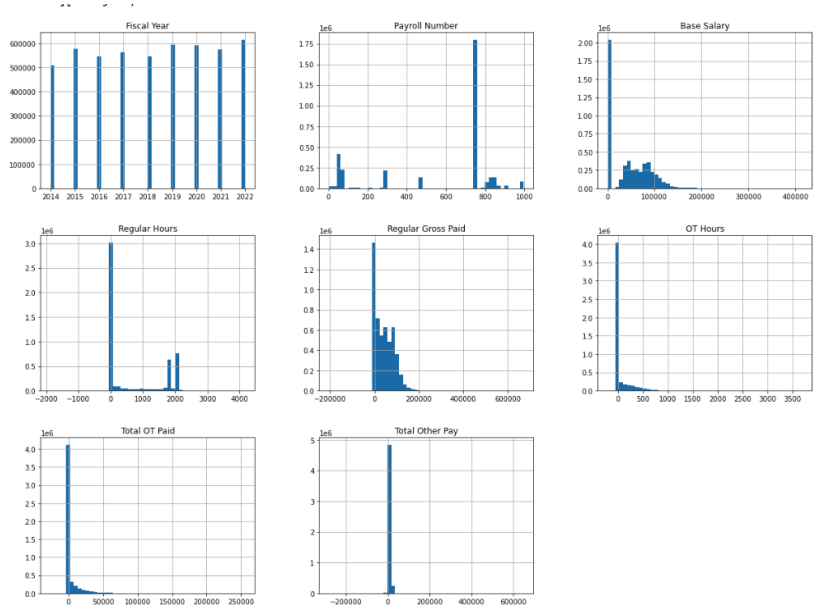

Fig. 1. Histogram plot of all features

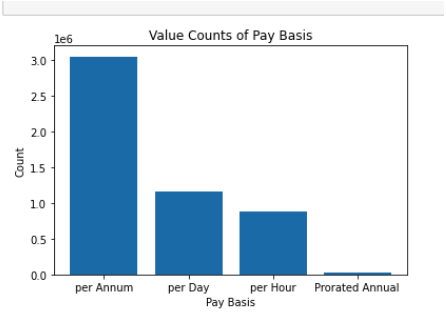

Fig. 2. Distribution of Data across Pay Basis

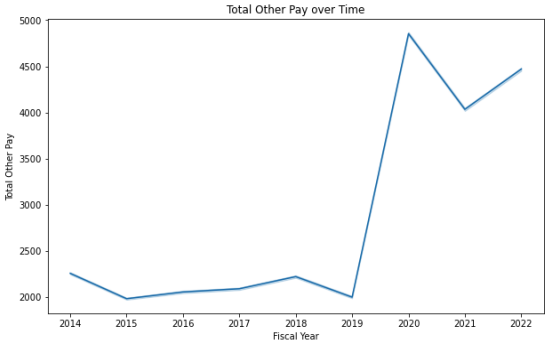

Fig 3. Line chart of Total Other Pay Over Time

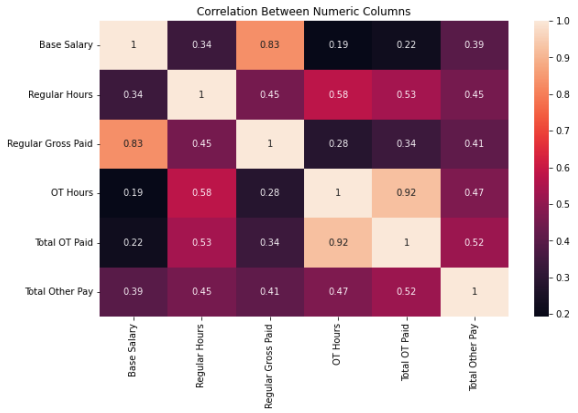

Fig. 4. Heatmap of Correlation Between Numerical Columns

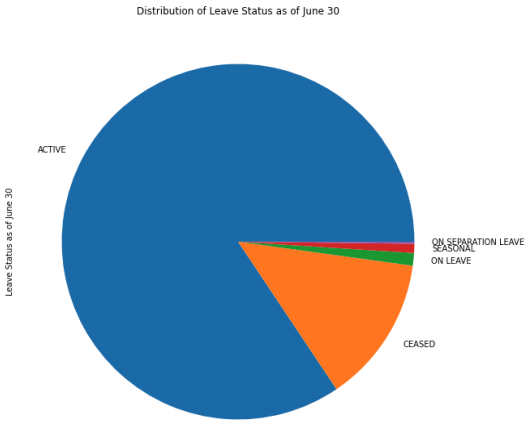

Fig. 5. Analysing Leave Status as of June 30

5. Results and Conclusions

a) Using Kbest features

Regressor	RMSE Train	RMSE Validation	RMSE Test	Mean Absolute Error (MAE)	R-squared (R2)
1. Random Forest	0.42	0.44	0.44	0.24	0.80
2 Gradient Boosting	0.45	0.45	0.45	0.26	0.79
3 LinearRegression	0.43	0.48	0.43	0.54	0.57
4 Voting Regressor:I	0.47	0.47	0.47	0.33	0.77
5. Voting Regressor:II	0.45	0.46	0.46	0.26	0.78

Fig. 6. Results of all regression models when used KBest Features

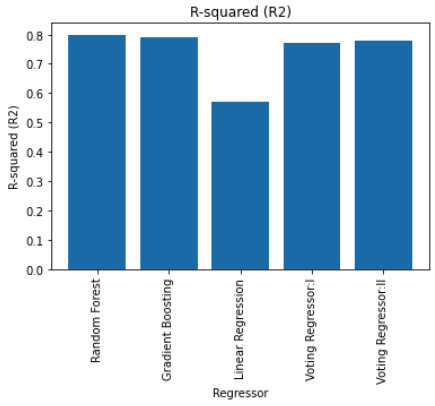

Fig. 7. Plot of Regression models vs R-Squared

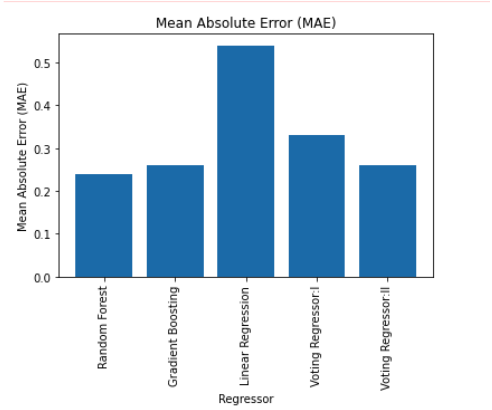

Fig. 8. Plot of Regression models vs MAE

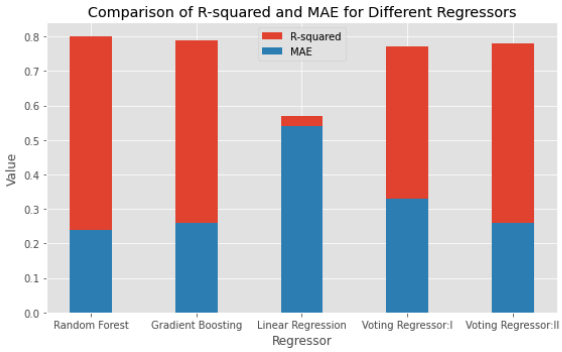

Fig. 9. Comparison of R-squared and MAE for different Regressors

From the plot, Fig.9., showing R-squared and MAE values for different regressors, we can infer the performance of each regressor on the given dataset. The R-squared value indicates how well the model fits the data and a higher value indicates a better fit. On the other hand, MAE measures the average magnitude of the errors in the predictions made by the model, with lower values indicating better performance. Looking at the plot, we can see that Random Forest and Gradient Boosting perform similarly in terms of R-squared, but Gradient Boosting has a slightly better MAE. Linear Regression has a lower R-squared value compared to the other two, indicating a poorer fit to the data. Voting Regressor I has the worst performance with a very low R-squared value and a relatively high MAE. Voting Regressor II performs similarly to the other two with respect to R-squared and has a relatively lower MAE.

b) Using PCA to select features

This means that after applying PCA to an original dataset that contained multiple features, the new feature set was obtained by selecting a subset of these features based on their contribution to the principal components. And below regression models were used. The result is shown below.

Regressor	RMSE Train	RMSE Validation	RMSE Test	Mean Absolute Error (MAE)	R-squared (R2)
1. Random Forest	0.06	0.15	0.15	0.04	0.97
2 Gradient Boosting	0.21	0.21	0.21	0.09	0.95
3 LinearRegression	0.18	0.18	0.18	0.29	0.81
4 Voting Regressor:I	0.22	0.23	0.23	0.14	0.94
5. Voting Regressor:II	0.12	0.15	0.15	0.06	0.97

Fig. 10. Results of all regression models when used new dataset

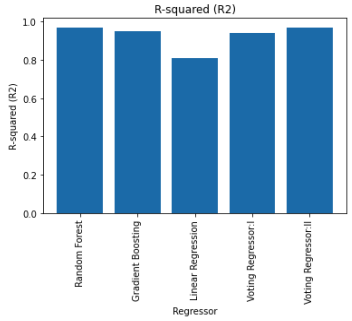

Fig. 11. Plot of Regression models vs R-Squared

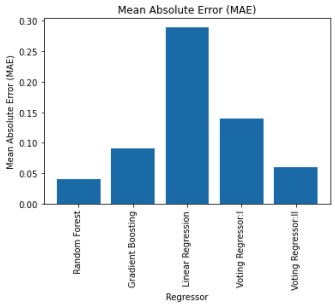

Fig. 12. Plot of Regression models vs MAE

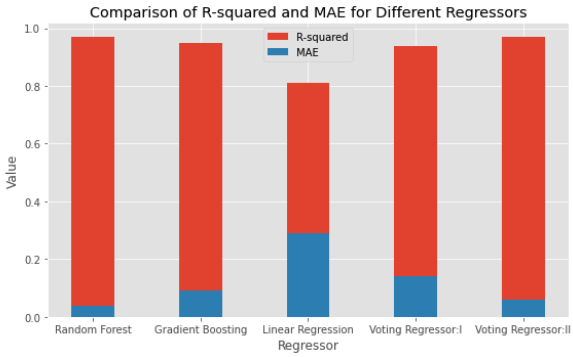

Fig. 13. Comparison of R-squared and MAE for different Regressors

Based on the table Fig.10, among different regression techniques, Random Forest performed the best with the lowest RMSE on the training, validation, and test sets. It also has a high R-squared value of 0.97, indicating that it explained most of the variability in the data. Linear Regression and Gradient Boosting techniques also performed relatively well, with comparable RMSE value, and R-squared values of 0.81 and 0.95 respectively. The two Voting Regressor models did not perform as well as the individual models. Voting Regressor I had the highest RMSE value on all sets, while Voting Regressor II had lower RMSE value, but still not as good as Random Forest.

Based on the table Fig.6, the performance of the regression models has decreased significantly when the input features are reduced to only 'Pay Basis', 'Regular Hours', and 'Agency Name'. The RMSE values have increased compared to the previous results. Once again, Random Forest and Gradient Boosting are the top-performing models, with similar RMSE value, and R-squared values around 0.8. Linear Regression also performs well, with lower RMSE value than Gradient Boosting. The Voting Regressor models once again did not perform as well as the individual models, with higher RMSE value. So, the results suggest that predicting Base Salary based on only 'Pay Basis', 'Regular Hours', and 'Agency Name' is more challenging and less accurate than when more features are considered. Random Forest, Gradient Boosting, and Linear Regression models are still suitable for this task, but their performance has decreased.

In conclusion, the comparison of the two tables, Figure 6 and Figure 10, indicates that the use of additional features, namely 'Regular Gross Paid' and 'Total Other Pay', significantly improves the performance of regression models in predicting Base Salary. The Random Forest model appears to be the most suitable for this task, achieving the lowest RMSE value, followed by Linear Regression and Gradient Boosting models. The reduced feature set consisting of 'Pay Basis', 'Regular Hours', and 'Agency Name' resulted in decreased model performance across all regression techniques. The Voting Regression models did not perform as well, which may be attributed to the combination of different models with varying performance. Overall, these findings suggest that the inclusion of relevant features is crucial for accurate base salary prediction, and the Random Forest model is a strong candidate for this task.

References

1. Office of Payroll Administration (OPA). (n.d.). Citywide Payroll Data (Fiscal Year). Retrieved from https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e
2. Scikit-learn. (n.d.). Principal component analysis. Retrieved from https://scikit-learn.org/stable/modules/decomposition.html#pca
3. Scikit-learn. (n.d.). Feature selection. Retrieved from https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection
4. Sklearn.ensemble.VotingRegressor. (n.d.). Scikit-learn 1.0 documentation. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingRegressor.html
5. Scikit-learn. (n.d.). Preprocessing data. Retrieved from https://scikit-learn.org/stable/modules/preprocessing.html

Appendix

1. Data
   Link:
   https://data.cityofnewyork.us/City-Government/Citywide-Payroll-Data-Fiscal-Year-/k397-673e
   Description: The Citywide Payroll Data (Fiscal Year) dataset, which has 5.11 million rows and 17 columns of data on payroll numbers, agency names, employee names, start dates, work locations, job titles, base salaries, pay bases, regular and overtime hours, and total pay. The dataset contains information on salaries and pay for tens of thousands of employees over multiple years, which means it has a large volume of data.

2. SauravUpadhyaya.ipynb/SauravUpadhyaya.html
   Link:
   https://drive.google.com/file/d/1UxWIjIMLTvp39fUVkF0EcnazApdiiaPq/view?usp=share_link

   Description: In this file, dataset[1] is explored, each features are visualized and then feature selection is done using KBest and PCA. The new feature set was obtained by selecting a subset of these features based on their contribution to the principal components. Once features are selected, regression models are used and their performance is evaluated. At the end, best features and models are identified.