**Predictive Analysis of Salary Categories: Leveraging Machine Learning for Strategic Compensation Management**

RACHANON CHOMPHOO (*Author*)

Student Number: 22080474

**ABSTRACT**

**In this study, we harness machine learning to predict salary categories based on various employee attributes. Using a dataset comprising features like job roles, experience, and educational qualifications, we evaluated multiple algorithms, with XGBoost emerging as a top performer. The model provided insights into key determinants of salary classifications, aiding stakeholders in making informed compensation decisions. The findings underscore the potential of predictive analytics in streamlining HR processes, setting competitive salary benchmarks, and enhancing organizational strategies.**

**Keywords— Predictive analysis; Salary categories; Machine learning; Employee attributes; XGBoost; Compensation decisions; HR processes; Organizational strategies.**

I. INTRODUCTION

Data science is one of the fastest-growing occupations, with a 36% increase expected between 2021 and 2031 [1]. In modern business, human resource management has transcended traditional administrative roles, evolving into a strategic function pivotal to organizational success. Central to this function is the challenge of determining compensation packages that attract and retain top talent and align with industry standards and organizational capabilities. With the proliferation of data and advancements in machine learning, there is an increasing opportunity to harness these technologies to address this challenge.

Historically reliant on human judgment, market surveys, and broad industry benchmarks, salary determination can now be augmented with predictive analytics. Machine learning algorithms can offer data-driven insights by analyzing patterns in employee profiles, job roles, and compensation data, enabling HR professionals to make informed decisions.

By employing a machine learning algorithms, ranging from Logistic Regression, Random Forests, Decision Tree, Support Vector Machine and Extreme Gradient Boosting, to predict salary categories based on a range of employee attributes. Through systematically exploring model performance and its implications, the study aims to elucidate the benefits, limitations, and practical applications of machine learning in compensation management.

II. DATASET OVERVIEW

The dataset use in this project was obtained from Kaggle [2], data originally from the aijobs.net in 2020 to 2023. The dataset contains 3,755 observations and 11 attributes includes:

* Job information including job title, employment type, employment residence, experience level and work year.
* Salary information including salary, salary in USD and salary currency.
* Company location and company size.
* Remote ratio - overall amount of work done remotely.

Exploratory Data Analysis (EDA) has been implemented to perform initial investigations on data to discover patterns and test hypotheses through visualization [3].

*Figure 1. Top 5 job designations as a Percentage*

A graph of different colored bars

Description automatically generated

Figure 1 shows the top 5 job designations. The data engineer role is the most popular (27.70% of the dataset), indicating higher demand than data scientist, with a 5.33% gap.

Figure 2. Percentage of Job Categories Designations

A graph with numbers and text

Description automatically generated with medium confidence

Figure 2 illustrates a considerable 22.05% difference between data engineering and data science, possibly due to increased funding for data teams as their importance is recognized [4].

Figure 3. Experience Level

A screenshot of a computer screen

Description automatically generated

Senior-level employees earn the most, followed by mid-level, entry-level, and executive-level employees. Most employees are mid-level or higher.

Figure 4. Distribution & QQ-plot of Salary in US Dollars

A graph of a graph

Description automatically generated with medium confidence

The salary distribution has an average of $137,570.39, but there's variability with some earning less or more. The standard deviation is $63,047.23, with a slightly positively skewed distribution. The data is normally distributed 68% of the time with a kurtosis value of 0.83. It's not excessively outlier prone [5].

Figure 5. median salary across different job categories based on the number of works years.

A graph of a salary

Description automatically generated with medium confidence

Figure 5 shows median salaries tend to increase with work years, with "Management" and "Data Science" having higher median salaries than other roles even with fewer work years.

Figure 6. salary comparison between employee residence and company location

A graph of a salary comparison between employees residence and company location

Description automatically generated

Figure 6 shows that some areas, like the United State (US), have more significant concentrations, indicating popular combinations of employee residences and company locations.

Supervised learning was the appropriate approach for predicting target variable. The dataset provided records with inputs and corresponding salaries, allowing the model to learn the features' relationship.

III. PROBLEM DEFINITION

Given the multifaceted nature of salary determination and the plethora of influencing variables, there is a pressing need for a systematic, data-driven approach to predict salary categories. This would aid stakeholders in making informed decisions, setting realistic expectations, and ensuring alignment with industry standards.

Considering the problem statement's objectives and the inherent characteristics of bivariate and multivariate analyses, it is academically sound to conclude that multivariate analysis is more suitable for this analysis. While bivariate analysis can offer preliminary insights, a multivariate approach is essential to delve deeper into the complex web of relationships among the multiple attributes and their collective influence on salary categorization.

*Algorithm selection:*

Algorithms were chosen based on the problem and dataset. Salary ranges were used to classify. One initial algorithm and two ensemble methods were selected for evaluation:

* *Logistic Regression* - estimates the probability of an instance belonging to a category using the logistic function. It is simple to interpret when features have a linear relationship with the response variable, making it popular for binary classification [6].
* *Random Forest* - is a machine learning algorithm that uses decision trees to make accurate predictions and efficiently handle large datasets. It also reduces overfitting by randomly selecting features and can rank features based on importance.
* *Decision Trees* - segment datasets based on input features, creating visual representations of output values and input feature conjunctions.
* *Extreme Gradient Boosting (XGBoost)* - is a powerful library for gradient boosting, handling missing data and parallel processing. It is excellent for classification and regression tasks and provides feature importance scores.
* *Support Vector Machine (SVM)* - is a robust supervised learning algorithm that can identify the best hyperplane for separating classes in high-dimensional spaces using kernel functions.

*Algorithm not selected*:

* *Linear Regression* - is not a viable approach for our problem domain, as it is intended to predict continuous variables. Our current issue involves classification, wherein the target variable (salary range) is categorical. Therefore, we must explore alternative methods better suited for the task.

Confusion Matrix is essential for evaluating classification algorithms with accurate performance measures, primarily when a class imbalance exists. It provides metrics such as actual outcomes and model predictions breakdown [7].

A high precision highlights the model's ability to minimize false optimistic predictions. This analysis demonstrates the model's proficiency in accurately classifying salary ranges without over-assigning classes.

A model with high recall captures most positive instances without omission. It signifies the model's ability to classify salary ranges without under-representation correctly.

A high F1-Score suggests a harmonious combination of precision and recall for the salary range prediction task, indicating strong model performance.

When predicting salaries, we balance accuracy and completeness. Precision measures accurate predictions, while recall captures all relevant instances. We use the F1-Score and confusion matrix to minimize misclassifications.

IV. ANALYSIS & EVALUATION

Start preprocessing the data by cleaning and checking missing values. Remove irrelevant columns like work\_year, salary, and salary\_currency for better analysis.

*Feature Transformation & Encoding*

We transformed the salary\_in\_usd column into categories: 'Low', 'Medium', and 'High', using defined salary ranges. This is an example of binning a continuous variable [8]. Converted categorical variables to numerical values using label encoding. Data was divided into training and testing sets to assess the model's predictive abilities accurately.

*Model Selection*

Five classification algorithms were selected: Random Forest, Extreme Gradient Boosting (XGBoost), Decision Tree, Logistic Regression, and Support Vector Machine (SVM). The data was split into training, and test sets with 80% and 20%, respectively.

Table I. Table I. Accuracy Score of the selected model

| Model | Accuracy Score |
| --- | --- |
| Random Forest | 77.63% |
| XGBoost | 79.63% |
| Decision Tree | 78.03% |
| Logistic Regression | 77.23% |
| Support Vector Machine | 76.03% |

Ensemble methods such as XGBoost, Decision tree and Random Forest tend to outperform other models in classification tasks due to their ability to capture complex patterns as shown in Table I.

Table II. Summary of model performance

| Model | Salary Category | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- |
| Random Forest | Low | 0.68 | 0.61 | 0.64 |
| Medium | 0.52 | 0.31 | 0.39 |
| High | 0.83 | 0.94 | 0.88 |
| XGBoost | Low | 0.69 | 0.60 | 0.64 |
| Medium | 0.53 | 0.32 | 0.40 |
| High | 0.83 | 0.94 | 0.88 |
| Decision Tree | Low | 0.71 | 0.58 | 0.64 |
| Medium | 0.50 | 0.32 | 0.39 |
| High | 0.82 | 0.93 | 0.87 |
| Logistic Regression | Low | 0.65 | 0.27 | 0.39 |
| Medium | 0.49 | 0.37 | 0.42 |
| High | 0.80 | 0.92 | 0.86 |
| Support Vector Machine | Low | 0.67 | 0.19 | 0.30 |
| Medium | 0.56 | 0.21 | 0.30 |
| High | 0.78 | 0.99 | 0.87 |

The research revealed that XGBoost, Random Forest, and Decision Tree models exhibited comparable performance. However, XGBoost demonstrated a slight edge in flexibility and tended to yield marginally superior outcomes with hyperparameter tuning [9]. Once hyperparameters were tuned, we analyzed performance metrics for the top model, XGBoost.

A grid search over a selected range of hyperparameters for the XGBoost model has been performed, giving accuracy results of approximately 80.93%.

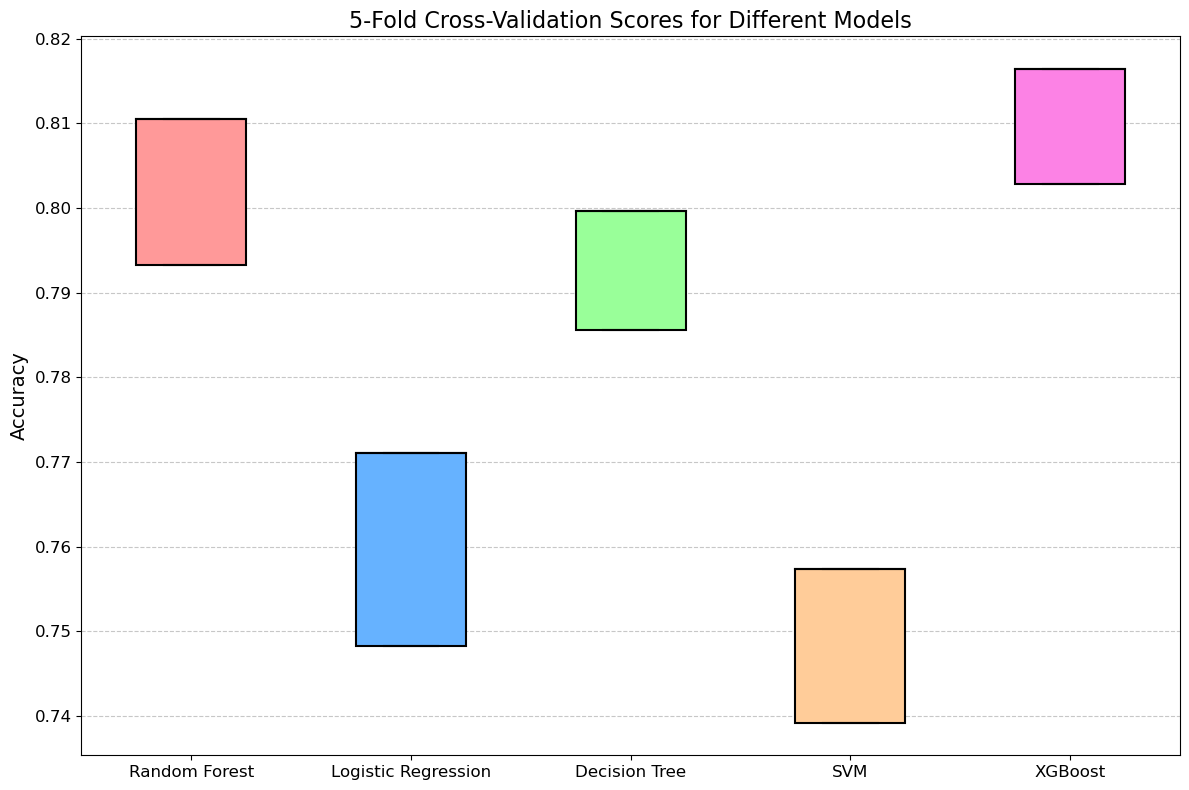
Table III. Model performance of XGBoost

| Model | Salary Category | Precision | Recall | F1-Score |
| --- | --- | --- | --- | --- |
| XGBoost | Low | 0.71 | 0.6 | 0.65 |
| Medium | 0.53 | 0.35 | 0.41 |
| High | 0.83 | 0.94 | 0.88 |

The model's performance is consistent with our previous evaluations and shows a good performance, especially for the "High" salary category as shown in the Table III.

Evaluate the model's performance using k-fold cross-validation to get a more robust measure of its accuracy and generalization capability.

Figure 7. 5-fold cross validation scores



According to our analysis, the XGBoost model outperforms all other models and achieves the highest median accuracy score. This is further supported by the relatively narrow distribution of scores, as illustrated in Figure 7.

Table IV. five-fold cross-validation model results

| Model | Mean | Std. |
| --- | --- | --- |
| XGBoost | 80.96% | 0.68% |
| Random Forest | 80.19% | 0.86% |
| Logistic Regression | 75.97% | 1.14% |
| Decision Tree | 79.26% | 0.70% |
| Support Vector Machine | 74.83% | 0.91% |

The cross-validation results indicate that the model is stable across different subsets of data and provides a robust performance estimate.

Figure 8. Feature Importance using XGBoost.

A graph with blue bars

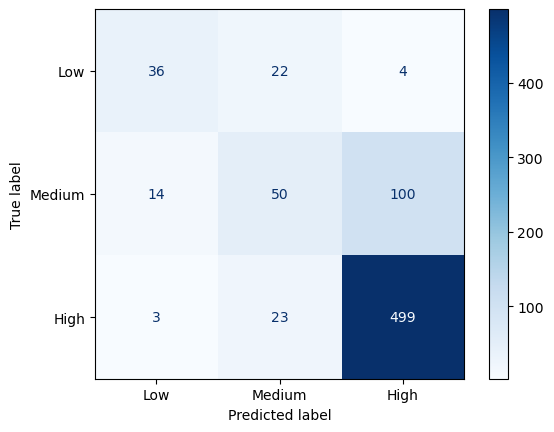
Description automatically generated

From Figure 8, *'experience\_level\_x\_employment\_type*' is the most contributed in this model prediction, followed by *'company\_location\_x\_employee\_residence*' and ‘*employ residence’*. These tell us that the experience level of each employment type has become a critical factor in determining the salary. Also, the distance between company location and employee residence can influence the salary, too.

*Model Evaluation*

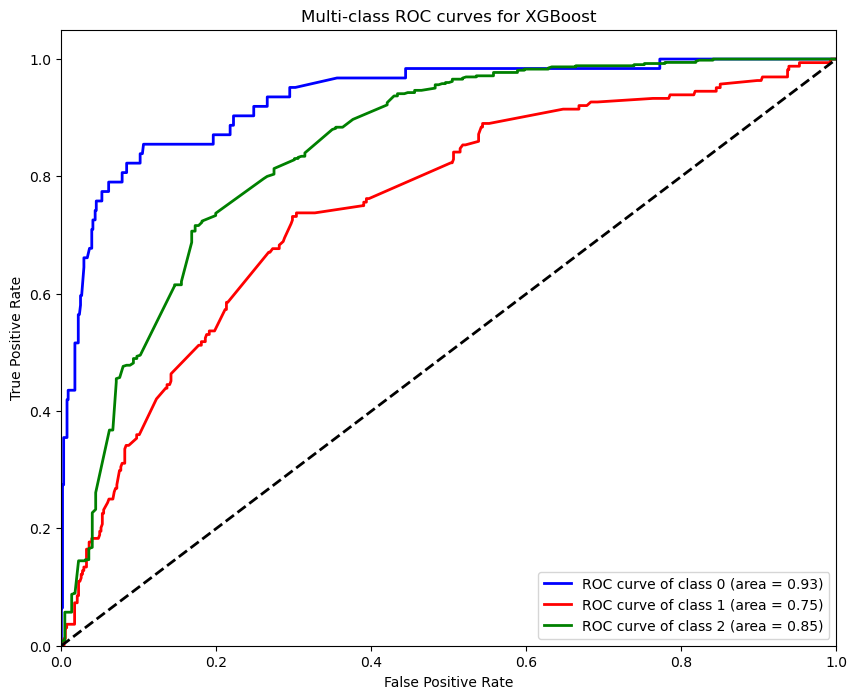
The confusion matrix displays accurate and inaccurate salary predictions for each category: "Low", "Medium", and "High". Actual "Low" salaries mainly were correctly classified, but some were misclassified as "Medium" or "High". Similarly, some "Medium" salaries were misclassified as "Low" or "High". However, the model performed well for the "High" salary category, with most actual "High" salaries correctly classified (Figure 9).

Figure 9. Confusion Matrix for XGBoost.



Using the ROC curve (Figure 10), The XGBoost model demonstrates strong performance across all salary categories, with exceptionally high discriminatory power for the Low (AUC = 0.93) and High (AUC = 0.85) salary classes. However, the performance for the Medium (AUC = 0.75) salary class suggests distinguishing this class from the others may be more challenging.

Figure 10. ROC curve for XGBoost.



*Model Interpretation*

After evaluating several models, including Random Forest, Logistic Regression, Decision Tree, SVM, and XGBoost, it was found that Random Forest had the highest initial accuracy. However, XGBoost, after hyperparameter tuning, was selected for detailed evaluation due to its versatile nature and other advantages. The analysis revealed that factors like experience level and job title played significant roles in determining salary levels. The XGBoost model demonstrated strong performance in distinguishing between Low and High-salary classes, but there might be challenges in distinguishing the Medium-salary class. The model can be used by employers, job seekers, and career consultants for various purposes, including setting salary expectations and benchmarking salaries.

V. Conclusion

Utilizing the XGBoost algorithm, this analysis aimed to predict salary categories based on various input features. The model exhibited notable efficacy, particularly in identifying Low and High-salary classes, as reflected by ROC-AUC values of 0.93 and 0.85, respectively. However, its performance for the medium salary class was moderate, with an AUC of 0.75.

These insights have significant real-world implications:

* *Employers & HR Professionals* - The model's findings can inform strategic compensation decisions, ensuring competitive and industry-aligned salary packages. Key features, such as experience level and job title, emerged as pivotal in influencing salary predictions.
* *Job Seekers & Career Consultants* - The model provides a framework for setting realistic salary expectations, aiding in career planning and transitions.
* *Academics & Researchers* - This analysis adds to the discourse on machine learning applications in HR, highlighting the nuances and complexities of salary determinants.

While the XGBoost model offers valuable predictive insights, it should be viewed as a complementary tool, supplementing human expertise and industry knowledge in HR decision-making.

*Limitations and Further Work*

The present analysis is subject to certain limitations, which stem from the quality and completeness of the dataset, the risk of overfitting, the need for transparency, and the computational demands of some of the models employed. These limitations may affect the results' accuracy and reliability and should be considered when interpreting the findings. Therefore, it is essential to exercise caution and prudence when drawing conclusions based on this analysis.

To enhance predictive power, it is recommended to implement feature engineering, ensemble approaches, regularization, and continuous feedback. Additionally, incorporating deep learning with neural networks can further improve results. These techniques can improve accuracy and effectiveness in predictive modelling for various business and academic applications.

REFERENCES

1. Luna, J. (2023) Data Science Salary Expectations in 2023. [online]. [Accessed 11 Aug 2023].
2. Randomarnab, (2023) Data Science Salaries 2023. [online]. Available from: [Accessed 11 August 2023].
3. Patil, P. (2018) What Is Exploratory Data Analysis. [online]. [Accessed 11 August 2023].
4. Mizrahi, E. (2023) How Will Data Engineering Change Over the Next 5 Years. [online]. [Accessed 11 August 2023].
5. Turney, S. (2022) What Is Kurtosis? | Definition, Examples & Formula. [online]. [Accessed 12 August 2023].
6. Shrivallabh, (2023) Logistic Regression: A Powerful Tool for Binary Classification. [online]. [Accessed 13 August 2023].
7. Markham, K. (2014) Simple Guide to Confusion Matrix Terminology. [online]. [Accessed 13 August 2023].
8. Vaidya, D. (2023) What Is Data Binning. [online]. [Accessed 14 August 2023].
9. Martin, D. (2021) XGboost: A Complete Guide to Fine-tune and Optimize Your Model. [online]. [Accessed 15 August 2023].
10. Pannmie, D. (2023) Ds Salary: Full Eda | Geo | Cluster + Xgboost. [online]. [Accessed 15 August 2023].
11. Joshi, V. (2023) Eda and Prediction (TensorFlow). [online]. [Accessed 15 August 2023].