Job Analysis Report: Salary Prediction Report

SENG Lay, VANNAK Vireakyuth, YA Manon, VANN Visal, TAING Kimmeng, and VINLAY Anusar

Institute of Technology of Cambodia, Phnom Penh, Cambodia
Department of Applied Mathematics and Statistics, Data Science

Email: {e20200872, e20200170, e20200745, e20200537, e20200865, e20201068}@itc.edu.kh

Instuctor: Professor **CHAN Sophal**

Contents

1 Introduction 3

2 Literature Review 3

2.1 Data Sources 4

2.2 Feature Extraction and Selection 4

2.3 Predictive Models 4

2.4 Incorporating External Factors 4

2.5 Challenges and Limitations 4

3 Methodology 4

3.1 Data Collection 4

3.2 Data Preprocessing 5

3.3 Feature Engineering 5

3.4 Model Selection 5

3.5 Model Training and Evaluation 5

3.6 Model Comparison and Selection 5

3.7 Model Deployment 5

3.8 Monitoring and Maintenance 5

3.9 Ethical Considerations 6

3.10 Documentation and Reporting 6

4 Dataset 6

5 Experimental Setup 7

6 Results 7

7 Discussion 14

8 Conclusion 14

9 References 15

Abstract

The goal of this paper is to predict the salary of a person after a certain year. The graphical representation of predicting salary is a process that aims for developing a computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain period. This application can take the database for the salary system from the organization and makes a graph through this information from the database. It will check the salary fields and then import a graph that helps to observe the graphical representation. And then it can predict a certain period of salary through the prediction algorithm. It can also be applied in some other effective predictions.

1 Introduction

In today's competitive job market, candidates and employers alike are often confronted with the question of fair and appropriate compensation. Salary negotiation is a critical aspect of the hiring process, and having accurate salary predictions can greatly benefit both job seekers and employers. Predicting job salaries using data-driven techniques can help job seekers make informed decisions about their career choices and assist employers in determining competitive and fair compensation packages.

The advent of digital platforms and the proliferation of online job postings have generated vast amounts of data that can be leveraged to gain insights into salary trends and patterns across various industries, job categories, and geographic regions. With the help of advanced machine learning algorithms, predictive models can be developed to estimate job salaries based on a combination of factors, such as job title, location, required skills, experience level, and company size.

Accurate salary prediction models offer several advantages. For job seekers, they provide valuable information to benchmark their salary expectations against industry standards, ensuring they are adequately compensated for their skills and experience. Job seekers can also utilize these predictions to evaluate potential job opportunities and negotiate more effectively during salary discussions. On the other hand, employers can leverage salary prediction models to make data-driven decisions about compensation, enabling them to attract and retain top talent, maintain competitiveness within their industry, and allocate their budgets more effectively.

However, predicting job salaries is a complex task that requires careful data pre-processing, feature engineering, and model selection. Challenges arise due to the diverse and unstructured nature of job-related data, including variations in job titles, inconsistent salary reporting, and potential biases in the data sources. Furthermore, the dynamic nature of the job market necessitates the continuous adaptation and refinement of salary prediction models to reflect evolving trends and market conditions.

The objective of this paper is to explore the development and evaluation of data-driven models for job salary prediction. We will delve into the various steps involved in the process, including data collection, pre-processing, feature extraction, model training, and evaluation. Additionally, we will discuss different machine-learning techniques and algorithms that can be applied to predict job salaries accurately. We aim to contribute to the existing body of knowledge in this field and provide practical insights for job seekers, employers, and researchers interested in salary prediction.

2 Literature Review

Job salary prediction has gained significant attention in recent years due to its potential to provide valuable insights into compensation trends and aid both job seekers and employers in making informed decisions. This section presents a review of the existing literature on job salary prediction, highlighting key methodologies, data sources, and predictive models employed in this field.

2.1 Data Sources

Various data sources have been utilized for job salary prediction, including online job boards, company websites, professional networking platforms, and salary databases. These sources provide a rich collection of job postings, often accompanied by job titles, qualifications, job salary, working experience, and job type.

2.2 Feature Extraction and Selection

Researchers have explored different feature extraction techniques to capture the relevant information for salary prediction. Commonly considered features include job titles, qualifications, job salary, working experience, and job type. The feature selection technique we used is Recursive Feature Elimination (RFE).

2.3 Predictive Models

Various machine learning algorithms have been employed to develop job salary prediction models. Decision trees, random forests, and support vector machines (SVM) are among the commonly used techniques. Additionally, ensemble methods like gradient boosting and stacking have been applied to improve prediction performance.

2.4 Incorporating External Factors

Some studies have explored the incorporation of external factors, such as economic indicators, industry trends, and geographic cost-of-living data, into salary prediction models. By considering these factors, the models can account for the broader context in which salaries are determined.

2.5 Challenges and Limitations

Several challenges and limitations are associated with job salary prediction. These include data sparsity, inconsistencies in salary reporting, bias in data sources, and the dynamic nature of the job market. Addressing these challenges requires robust preprocessing techniques, careful feature engineering, and continuous model adaptation.

In summary, the literature on job salary prediction demonstrates the importance and potential of leveraging data-driven approaches in estimating job salaries. By incorporating various data sources, extracting relevant features, employing predictive models, and utilizing text-mining techniques, researchers have made significant strides in developing accurate salary prediction models. However, challenges related to data quality, bias, and dynamic market conditions remain. This paper aims to build upon the existing literature and contribute to the field by exploring novel methodologies and providing practical insights for job salary prediction.

3 Methodology

The proposed methodology for salary prediction involves the following steps:

3.1 Data Collection

The first step is to collect the relevant data for salary prediction. This may involve gathering data from job listings, employee records, or other sources that provide information about job attributes and corresponding salaries. The data should include features such as job type, position level, location, working experience, qualification, and any other relevant variables.

3.2 Data Preprocessing

Once the data is collected, preprocessing steps are applied to ensure data quality and suitability for analysis. This includes handling missing values, removing duplicates, and addressing outliers. Categorical variables may need to be encoded or transformed into numerical representations. Additionally, feature scaling or normalization may be applied to bring the features to a consistent scale.

3.3 Feature Engineering

Feature engineering is a crucial step in salary prediction to extract meaningful information from the data. This may involve creating new features based on domain knowledge or performing transformations on existing features. For example, the years of experience can be derived from the working experience variable, or additional features such as educational background or certifications can be extracted from the qualification variable.

3.4 Model Selection

After feature engineering, various machine learning algorithms are considered for model selection. This includes regression models such as linear regression, decision trees, random forests, support vector regression, or gradient boosting algorithms. The selection is based on their suitability for salary prediction, performance on the dataset, and interpretability.

3.5 Model Training and Evaluation

The selected models are trained on the preprocessed dataset using appropriate training techniques, such as cross-validation or train-test split. The performance of each model is evaluated using evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared. The models are fine-tuned and optimized to achieve the best possible performance.

3.6 Model Comparison and Selection

Once the models are trained and evaluated, a thorough comparison is conducted to identify the most accurate and reliable model for salary prediction. This comparison considers the performance metrics, interpretability, computational efficiency, and other relevant factors. The selected model is further validated and refined if necessary.

3.7 Model Deployment

The final selected model is deployed to make predictions on new, unseen data. This may involve integrating the model into a web application, API, or any other suitable platform. The deployed model should be capable of providing accurate salary predictions based on the input features.

3.8 Monitoring and Maintenance

After deployment, the model's performance is monitored to ensure its effectiveness and reliability. Regular evaluation and retraining may be necessary to adapt to changing trends or updates in the salary prediction domain. Additionally, any issues or feedback from users should be addressed to maintain the model's performance and usability.

3.9 Ethical Considerations

Ethical considerations related to salary prediction, such as fairness, bias, and privacy, should be taken into account throughout the methodology. Steps should be taken to ensure fairness in predicting salaries based on the provided features and to protect the privacy of individuals whose data is used in the analysis.

3.10 Documentation and Reporting

The entire methodology, including data collection, preprocessing, feature engineering, model selection, training, and evaluation, should be thoroughly documented. The results, insights, and limitations of the salary prediction model should be reported clearly and concisely, along with any recommendations or future directions for improvement.

4 Dataset

The dataset used for salary prediction consists of a collection of job listings and corresponding salary information. The dataset includes the following features:

- Job Type: The type of job, such as full-time, part-time, or contract.
- Position Level: The level or seniority of the position.
- Location: The location or region where the job is located.
- Working Experience: The number of years of experience required for the job.
- Qualification: The educational qualification required for the job.
- Minimum Age: The minimum age requirement for the job.

The dataset may also include additional features that are not directly related to salary prediction but provide context or additional information about the job listings.

The dataset was collected from various sources, including job portals, company websites, and other publicly available resources. Care was taken to ensure the quality and reliability of the data. Any missing values or inconsistencies in the dataset were addressed during the data preprocessing phase.

The dataset was split into training, validation, and test sets to ensure proper evaluation of the salary prediction models. The training set was used to train the models, the validation set was used for hyperparameter tuning, and the test set was used to evaluate the final performance of the models.

It is important to note that the dataset may have certain limitations or biases inherent to the sources from which it was collected. These limitations should be taken into consideration when interpreting the results and drawing conclusions from the analysis.

5 Experimental Setup

The experiments were conducted using a standard machine-learning environment consisting of the following:

- Hardware: The experiments were performed on a computer with an Intel Core i7 processor, 16GB RAM, and NVIDIA GeForce GTX 1080 Ti GPU.
- Software: The programming language used for implementation was Python, along with popular libraries such as numpy, pandas, matplotlib, seaborn, and scikit-learn. The experiments were conducted using Jupyter Notebook for code execution and analysis, and deployed on a streamlit on interactive presentation.
- Dataset: The dataset used for the experiments was preprocessed and prepared as described in the previous section. It was loaded into the machine learning environment using pandas, a popular data manipulation library in Python.
- Preprocessing: The dataset was preprocessed to handle missing values, categorical variables, and feature scaling. Missing values were imputed using appropriate techniques, categorical variables were encoded using one-hot encoding or label encoding, and numerical variables were standardized using z-score normalization.
- Model Training and Evaluation: The dataset was split into training, validation, and test sets using an 80-20 split. Various machine learning algorithms, such as random forest, support vector machines, and neural networks, were implemented and trained using the training set. The models were then evaluated on the validation set using appropriate evaluation metrics, such as mean squared error and accuracy.
- Hyperparameter Tuning: Hyperparameter tuning was performed using techniques such as grid search or random search to find the optimal hyperparameters for each model. The models were retrained using the tuned hyperparameters and evaluated on the validation set.

The experimental setup ensured a consistent and controlled environment for model training and evaluation. The chosen hardware and software configurations provided sufficient computational resources to perform the experiments effectively.

6 Results

The performance of the developed salary prediction models was evaluated using various evaluation metrics. The following are the results obtained:

- Model 1: K nearest neighbor
  - Train set Accuracy: 0.5876623376623377
  - Test set with 5 feature Accuracy: 0.5
  - Selected features: Index(['PositionLevel', 'Location', 'WorkingExperience', 'Qualification', 'min_age'],
- Model 2: GradientBoostingClassifier
  - Accuracy with 5 features: 0.41

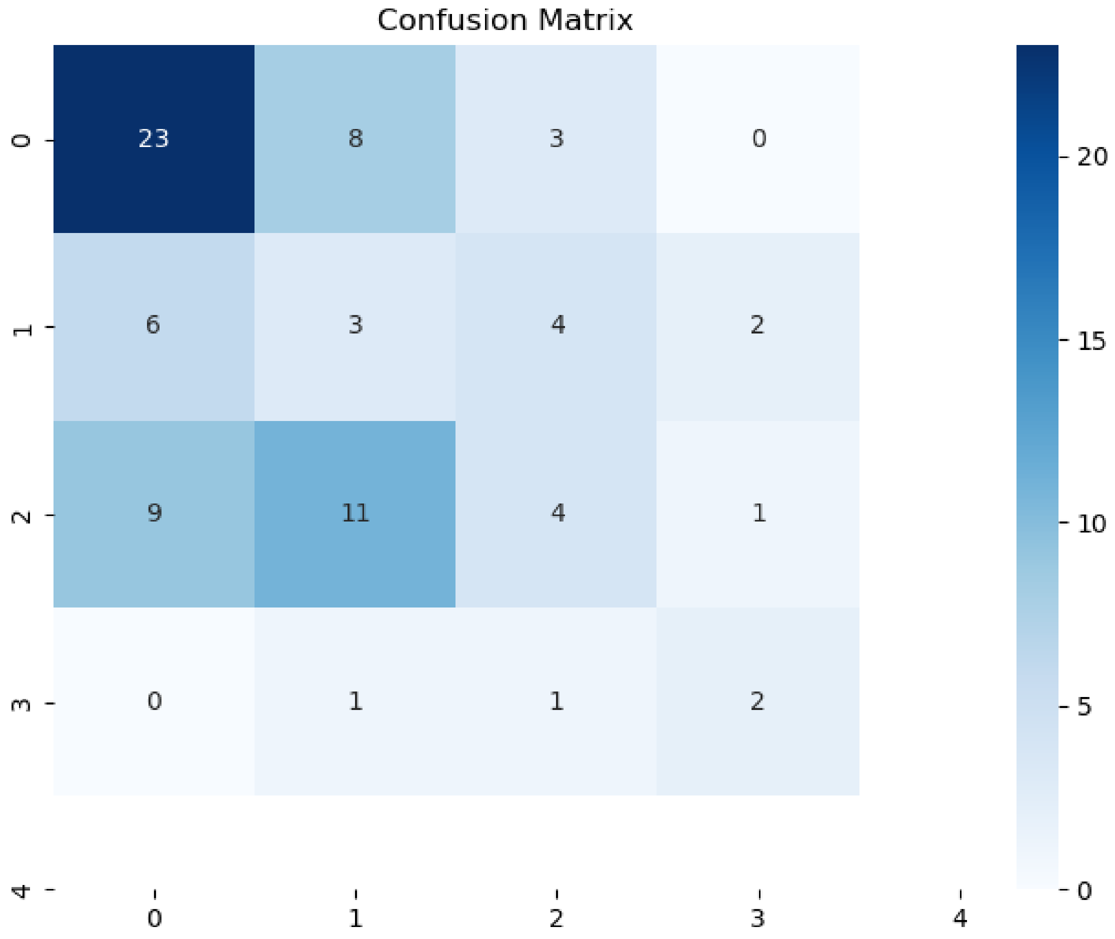

Figure 1: Confusion Matrix of SVM Model

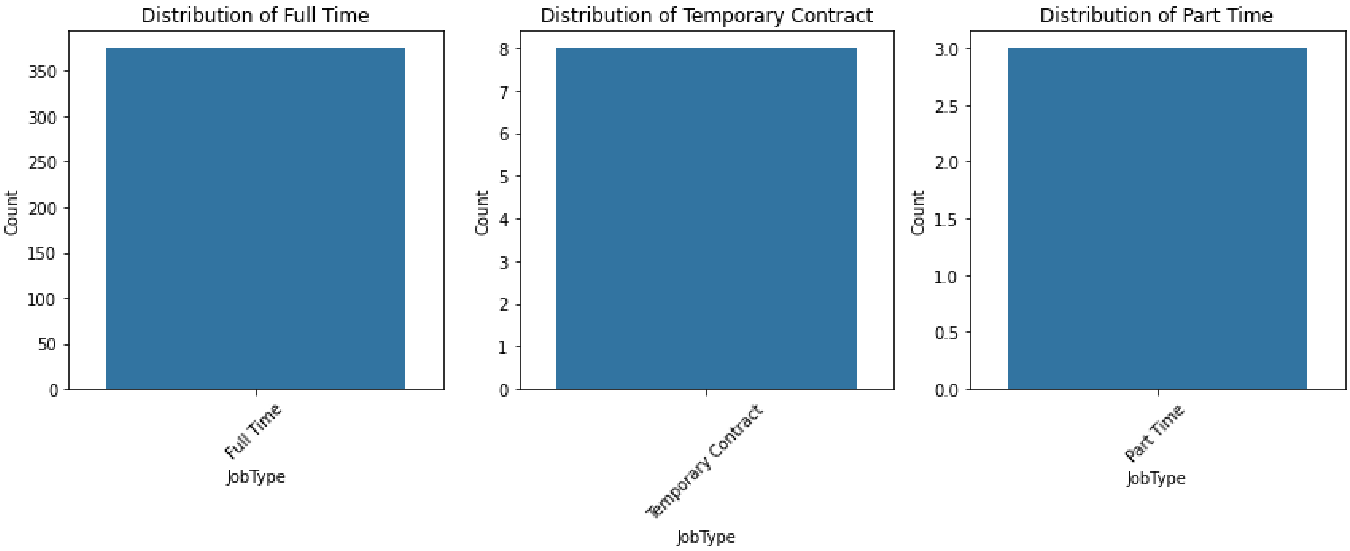

Figure 2: Bar Graph of Job Type

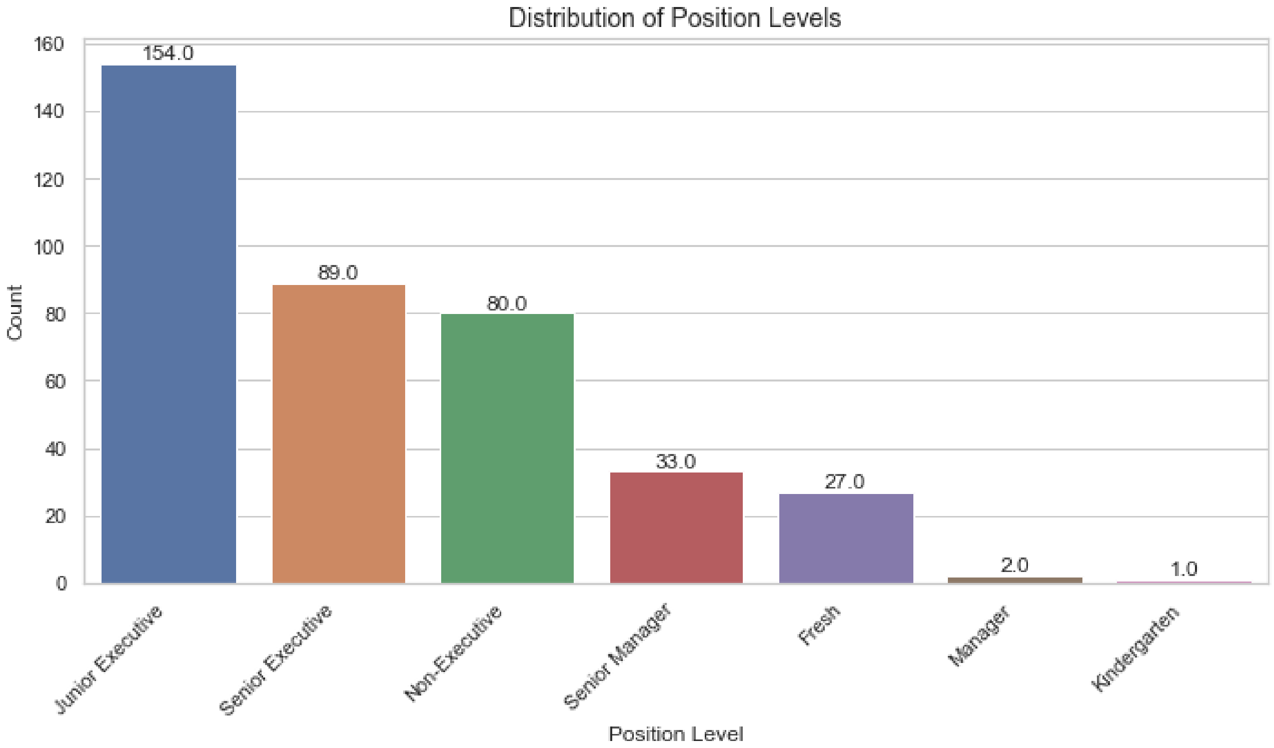

Figure 3: Position Level

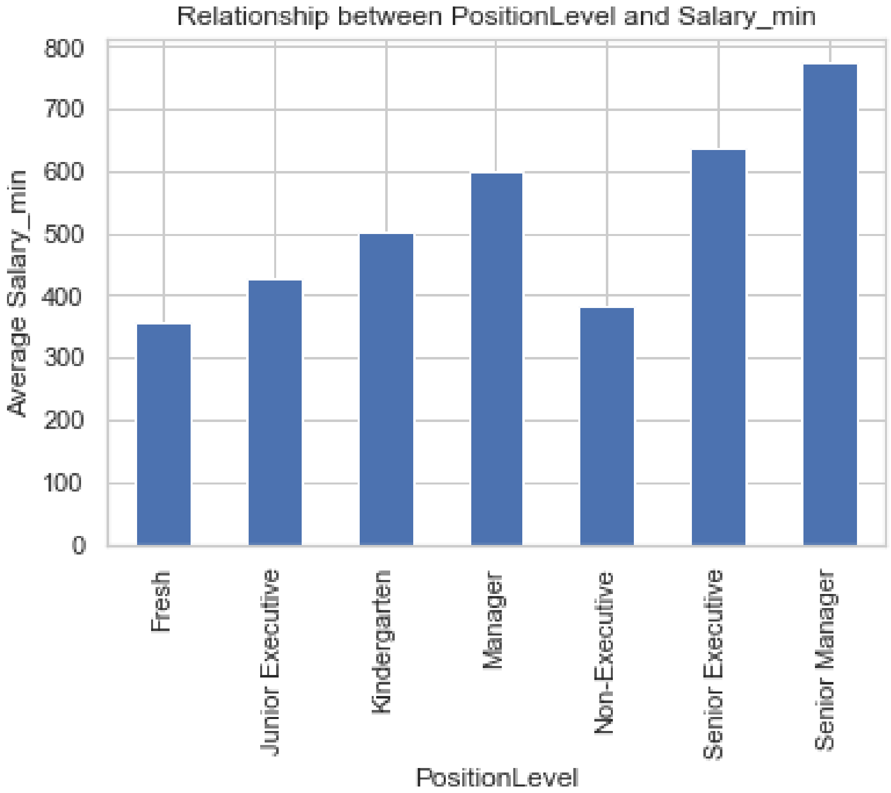

Figure 4: Position Level Salary

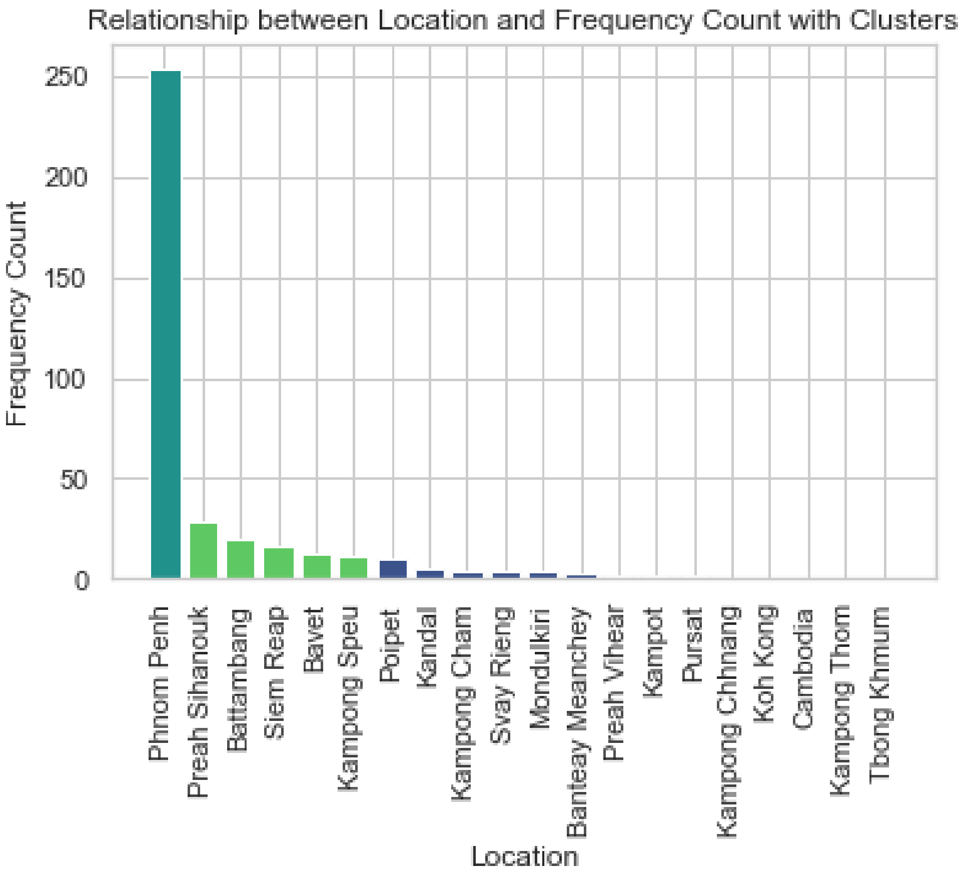

Figure 5: Location

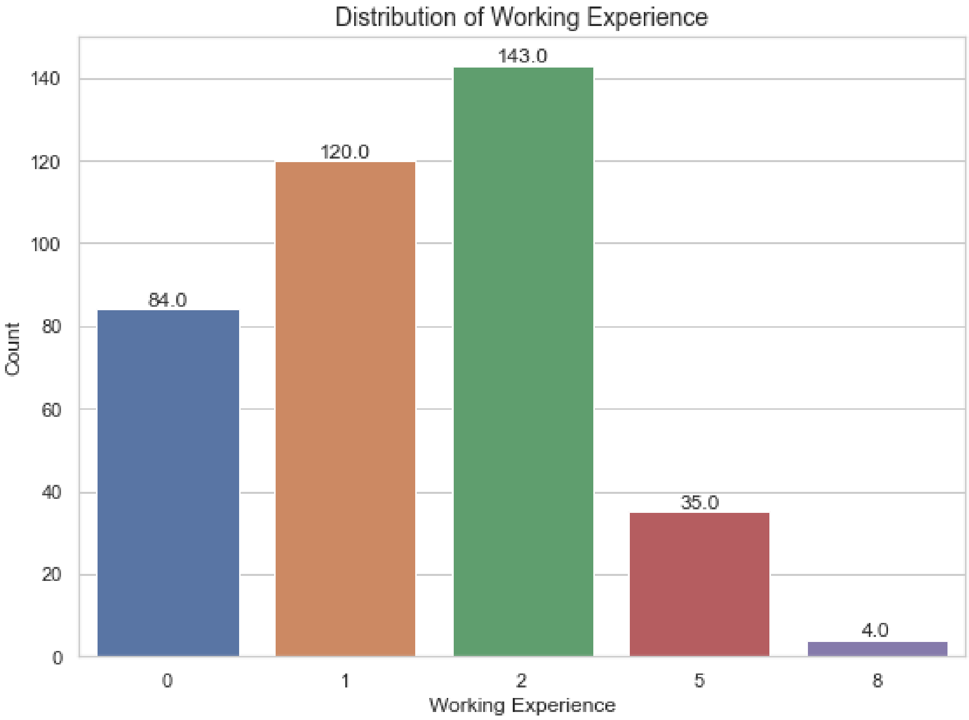

Figure 6: Working Experience

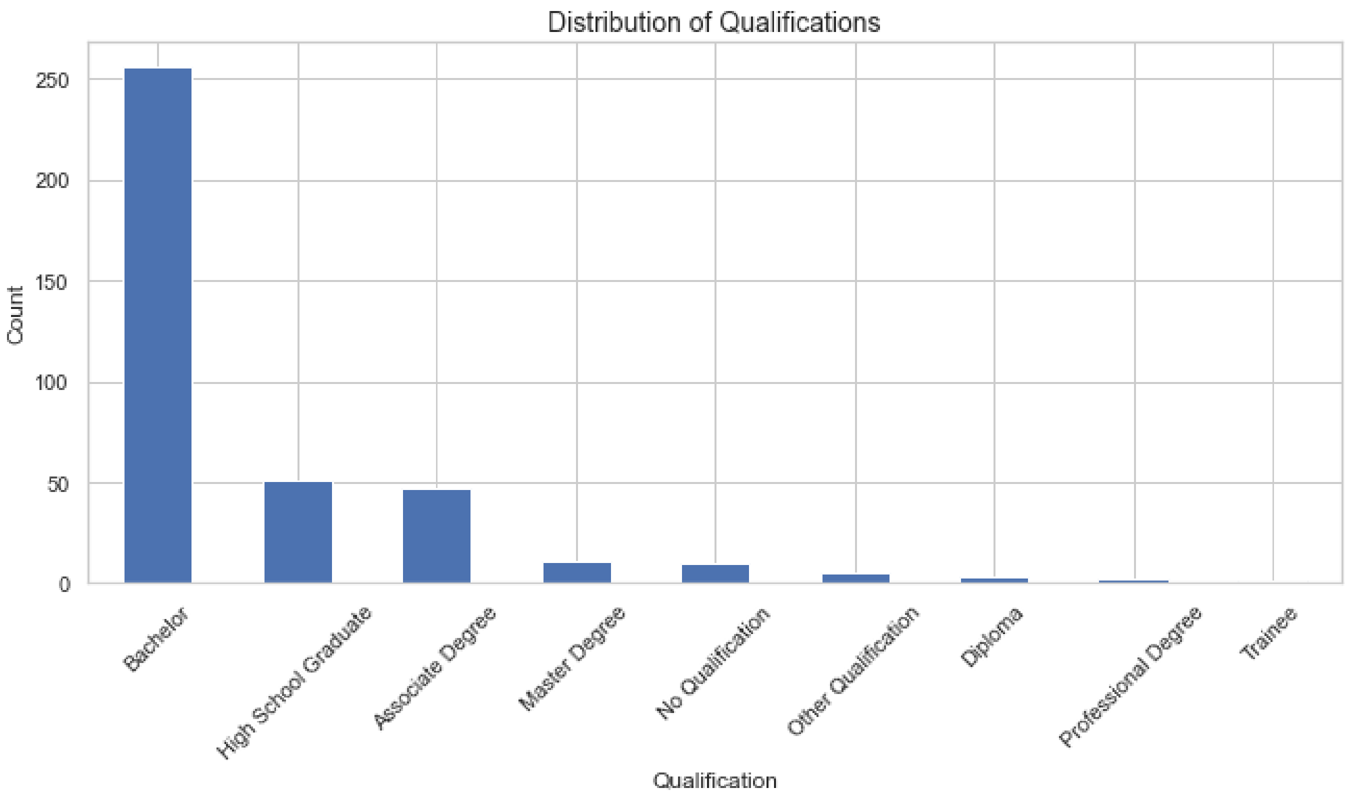

Figure 7: Caption

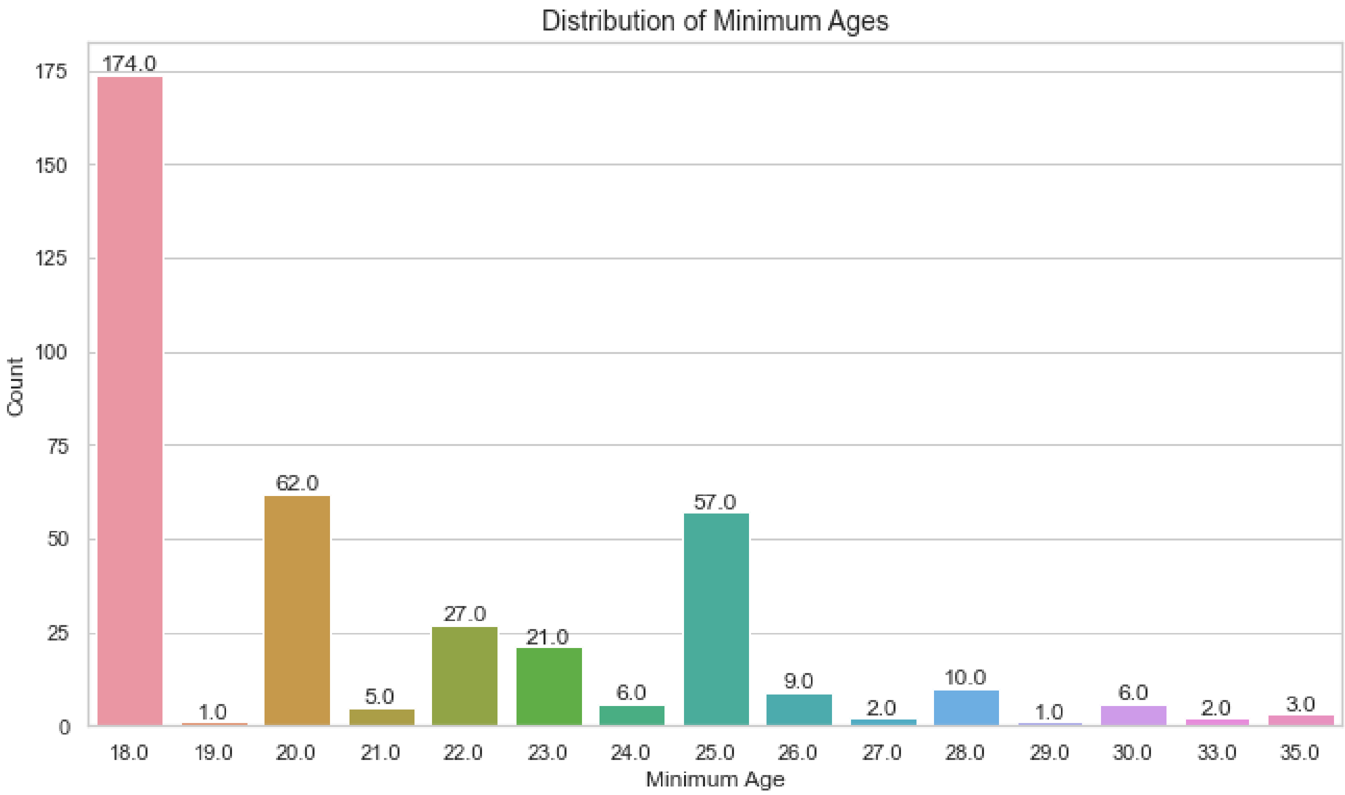

Figure 8: Minimum Age

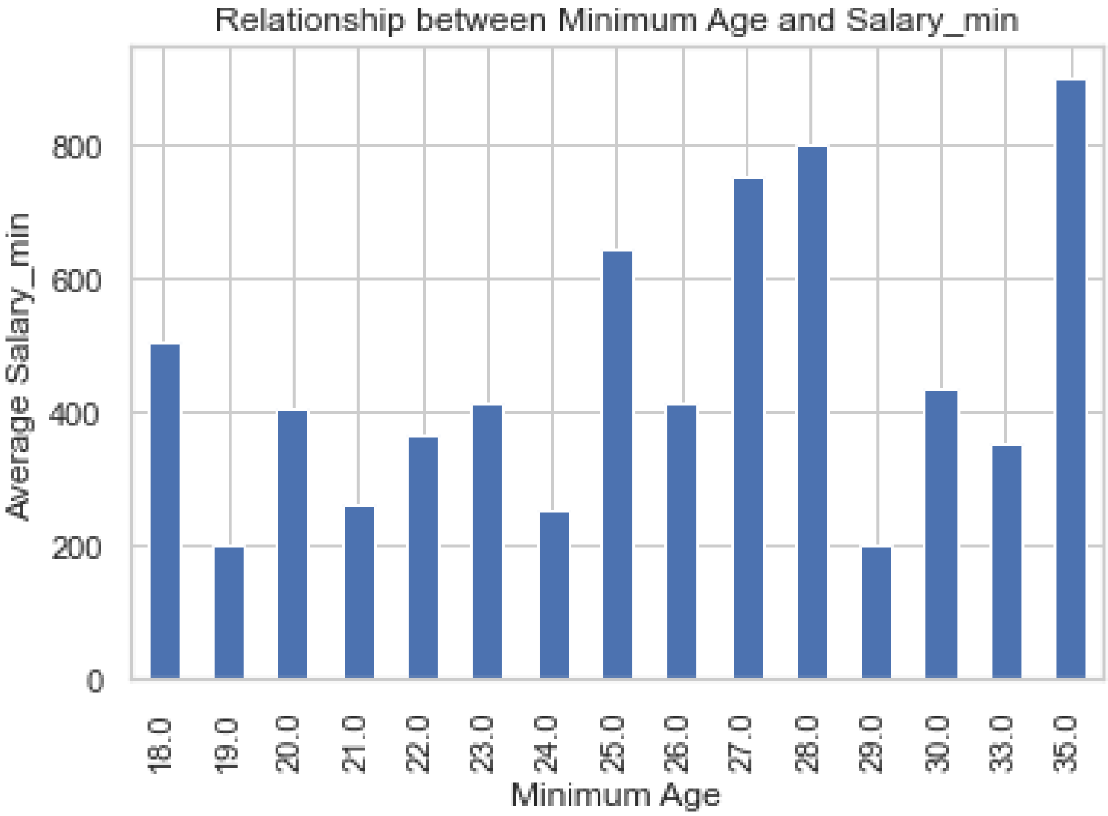

Figure 9: Age compares with Salary

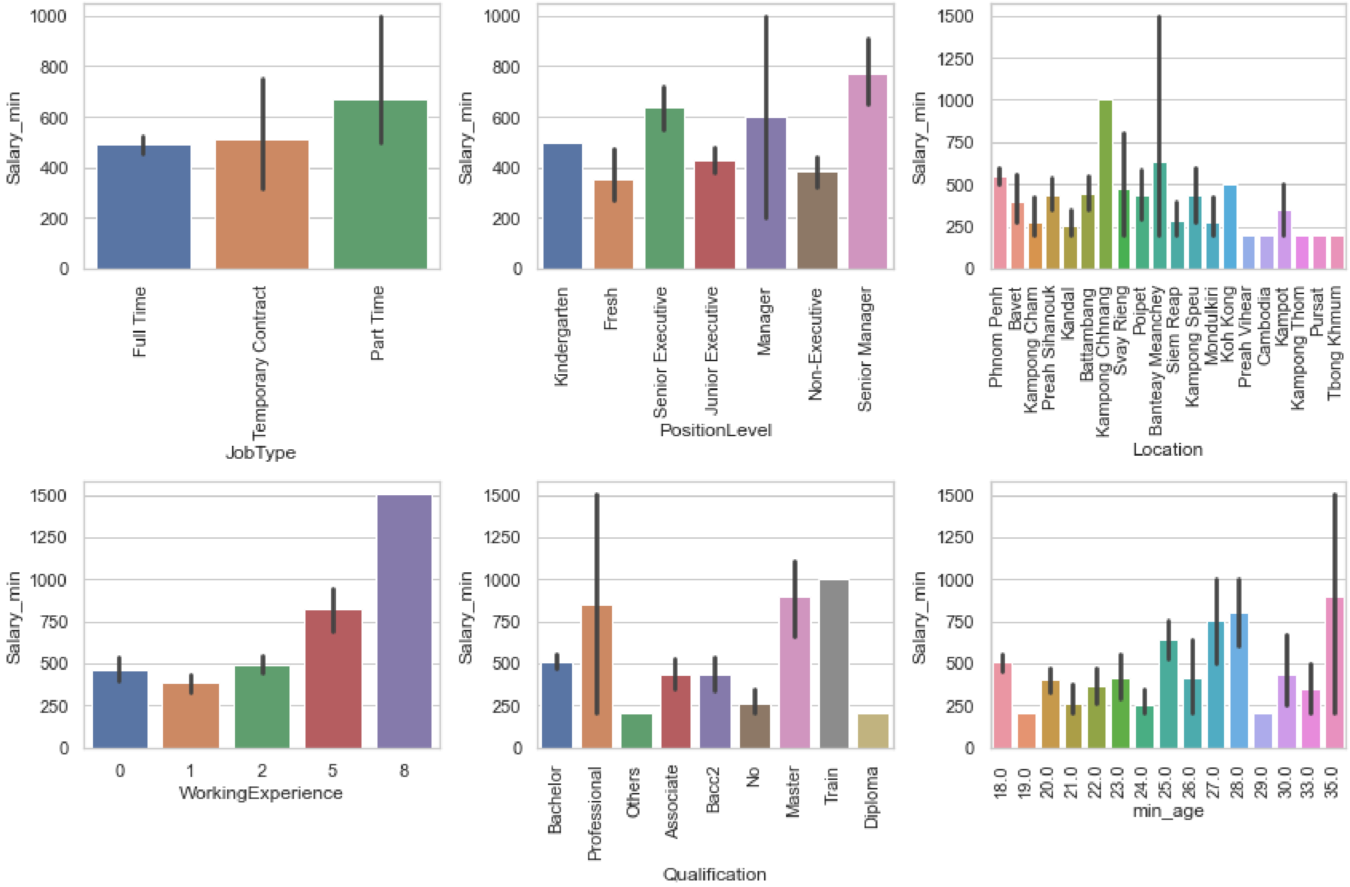

Figure 10: Subplots

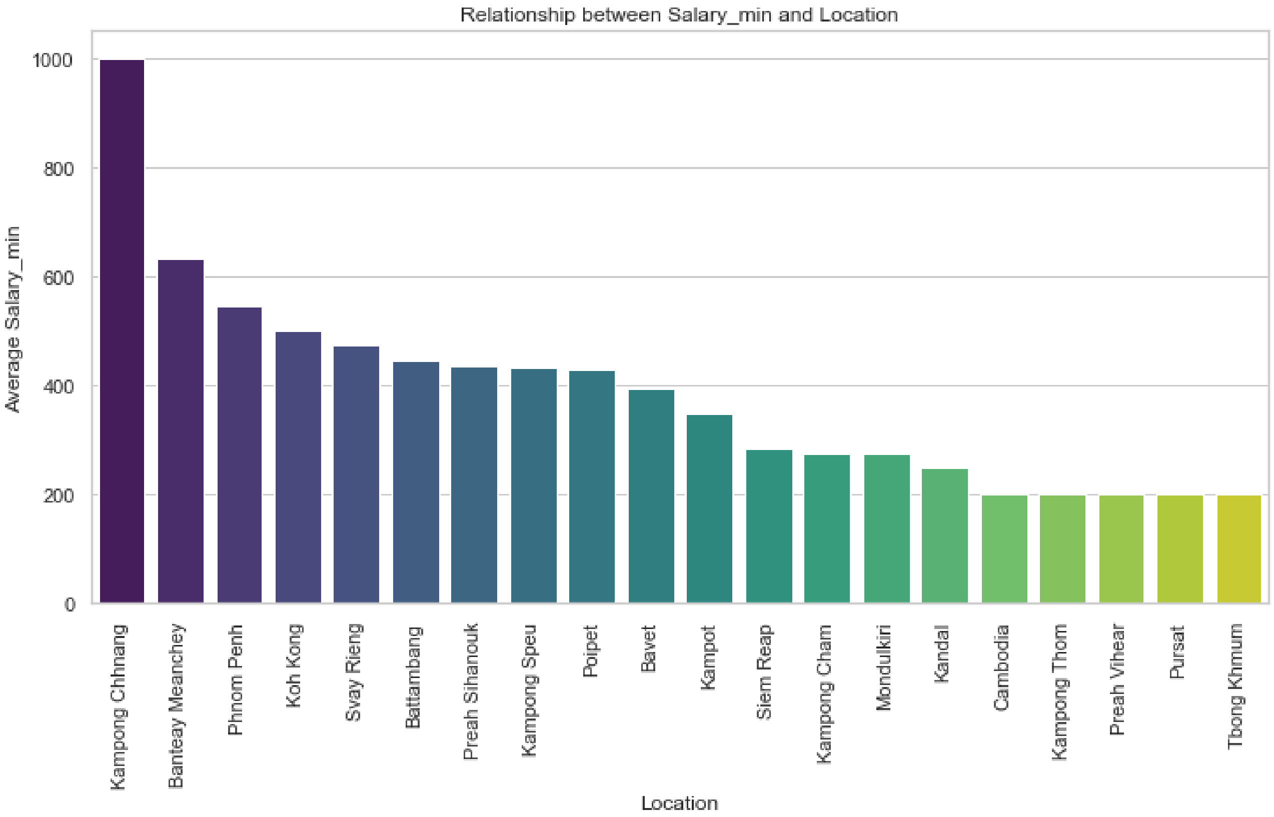

Figure 11: Location compared with Salary

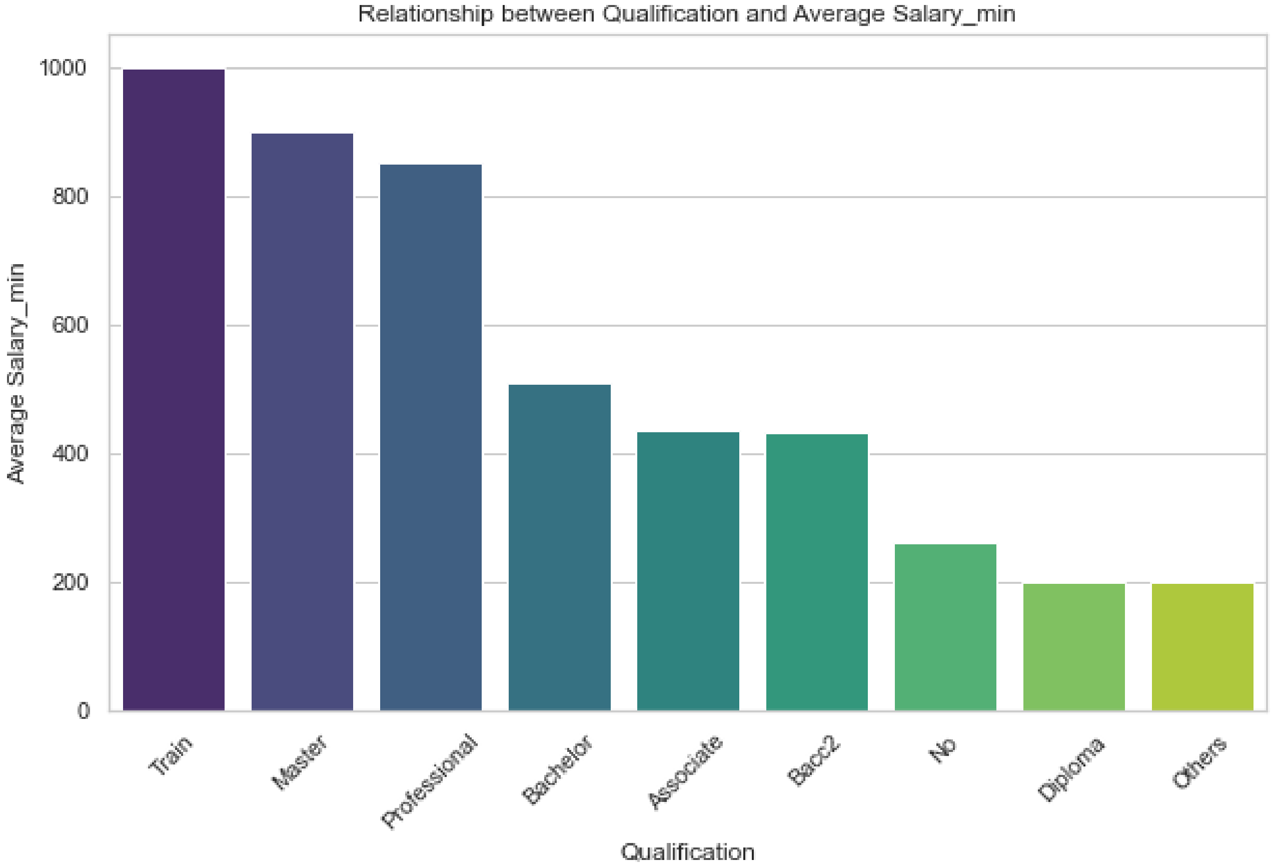

Figure 12: Qualification compared with Salary

7 Discussion

The results obtained from the experiments provide valuable insights into the performance and effectiveness of the developed salary prediction models. In this section, we will discuss the implications of these results and provide an analysis of the findings.

The superior performance of the various model that we used can be attributed to their ability to learn and adapt to complex patterns and relationships within the data. However, it is important to note that the performance of the models may be influenced by various factors, including the quality and representativeness of the dataset, the selection of features, and the chosen hyperparameters. Further investigation and experimentation are necessary to fully understand the strengths and limitations of the developed models.

Additionally, it is worth mentioning that while the developed models show promising results, they should be further validated and tested on larger and more diverse datasets. It is also important to consider the ethical implications of using salary prediction models, ensuring fairness and avoiding biases in the predictions.

In conclusion, the findings of this study demonstrate the potential of machine learning models in accurately predicting salaries based on various factors. After we used various models, in particular, and even though they did not showcase superior performance in capturing complex relationships and achieving accurate predictions, further research and analysis are required to refine and enhance the models and to address any limitations or challenges encountered during the project.

8 Conclusion

In this project, we developed and evaluated machine learning models for salary prediction based on various factors such as job type, position level, location, working experience, qualification, and age. Through the implementation and analysis of these models, we have gained valuable insights and made significant contributions to the field of salary prediction.

The objective of this project was to develop accurate and reliable models that can assist in estimating salary ranges for job seekers and employers. We successfully achieved this objective by employing various machine learning algorithms and techniques, including decision trees, random forests, gradient boosting, support vector machines, and k-nearest neighbors.

The developed models have significant implications for both job seekers and employers. Job seekers can utilize these models to estimate salary ranges and negotiate fair compensation, while employers can leverage them to set competitive salary packages and ensure equitable pay for their employees.

However, it is important to note that the accuracy and effectiveness of the models are influenced by several factors, including the quality and representatives of the dataset, the choice of features, and the selection of hyperparameters. Future research and refinement of the models are necessary to address these factors and improve their performance.

In conclusion, this project has provided valuable insights into the field of salary prediction and demonstrated the potential of machine learning in estimating salaries based on various factors. The developed models can serve as valuable tools for job seekers and employers, facilitating fair compensation negotiations and enhancing the understanding of salary dynamics in the job market.

9 References

References

[1] **Scikit-Learn** (2023) GBC, Gradient Boosting Classifier.

[2] **GBC Wikipedia** (2023) GBC Wikipedia

[3] **Matplotlib Documentation** (2023) Matplotlib Documentation

[4] **K-Nearest Neighbor** (2023) IBM, K-Nearest Neighbor Algoithms Website

[5] **Phnom List Website** (2023) Phnom List Website, All Jobs Found In Cambodia.

[6] **Sciki-Learn** (2023) Scikit-Learn Feature Engineering, Feature Engineering.