GENDER WAGE GAP AND SALARY PREDICTION

Problem Statement:

In today's society, gender pay gap remains a significant issue across various industries and professions worldwide, reflecting disparities in earnings among men and women. Understanding the factors influencing salary discrepancies between genders is crucial for promoting fairness and equality in the workplace. This project aims to investigate the factors influencing salary discrepancies between genders using data from the 2018 Kaggle Machine Learning & Data Science Survey. By analyzing demographic and professional attributes such as age, education, profession, industry, and experience, the goal is to understand the extent of the gender pay gap and develop predictive models to estimate salaries based on these factors.

Data and Methods

Data Set: The dataset used in this analysis is sourced from the 2018 Kaggle Machine Learning & Data Science Survey, which comprises responses from data professionals worldwide. It contains information on various demographic and professional features, including gender, age, country, education, profession, industry, experience, and annual salary.

Methodology

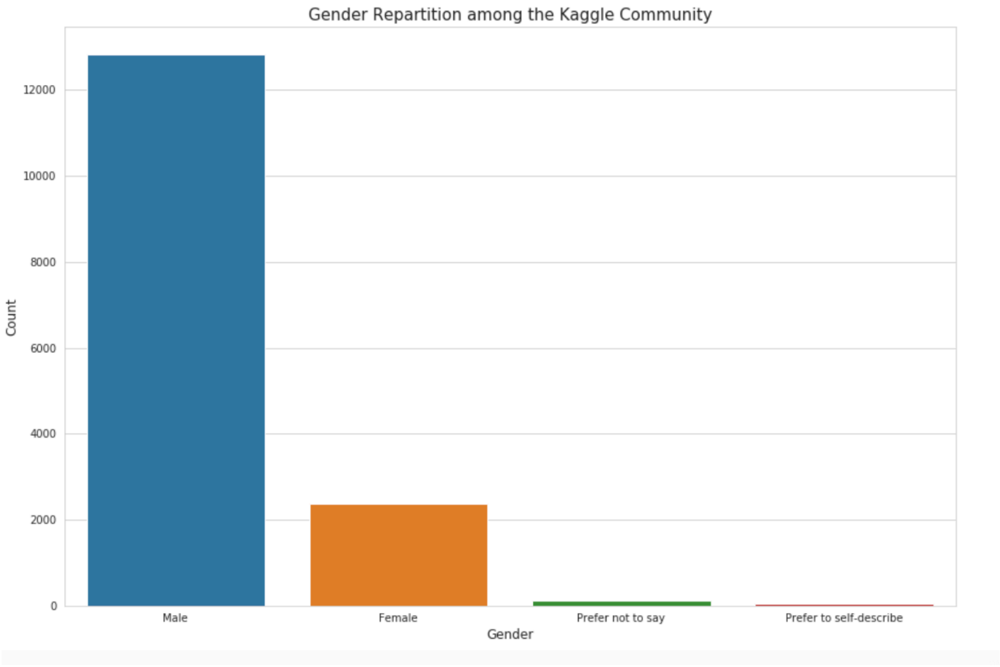
a. Data Cleaning

For this study, I only focused on the multiple choice questions, since they are much easier to construct a machine learning model. For this reason, I removed unnecessary columns and focus on 9 columns which are “Gender”, “Age”, “Country”, “Education”, “Major”, “Profession”, “Industry”, “Experience”, and “Annual Salary”. I removed text values from those column, since we only need numerical values for training a machine learning model. Some potential inconsistent responses were excluded such as students who earn more than half million in a year.

b. Exploratory Data Analysis (EDA)

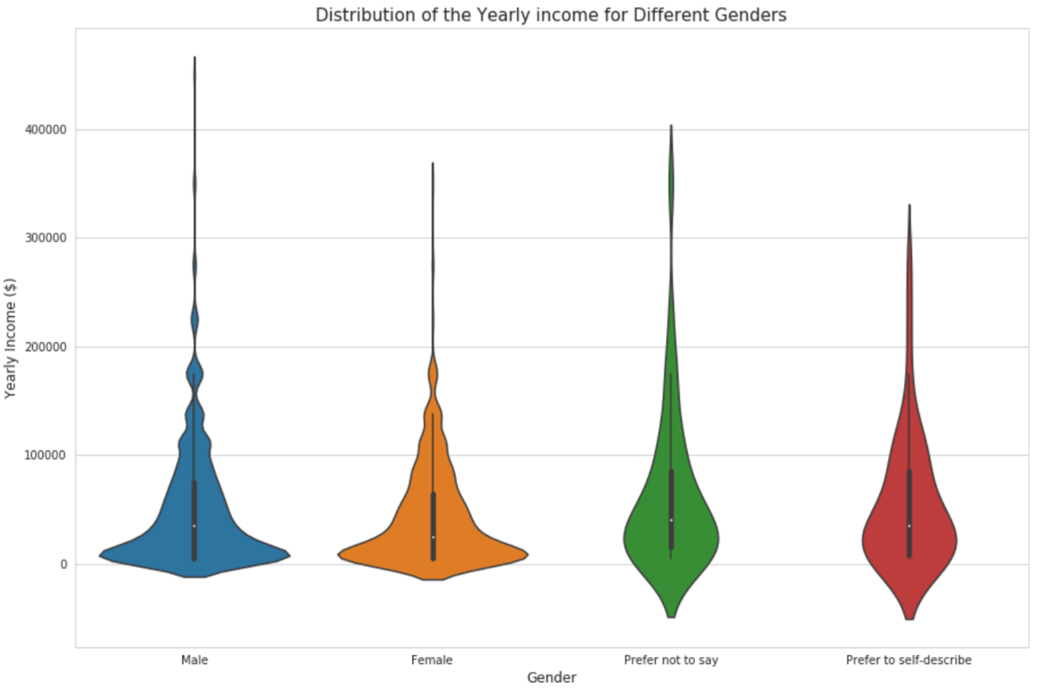
Exploratory Data Analysis is conducted to gain insights into the distribution of gender and salary within the dataset. Visualizations such as histograms, box plots, and scatter plots are utilized to analyze the relationships between gender and salary, as well as other demographic and professional attributes. This analysis provides a foundation for understanding the gender pay gap and identifying potential contributing factors.

Graph 1. Distribution of Gender



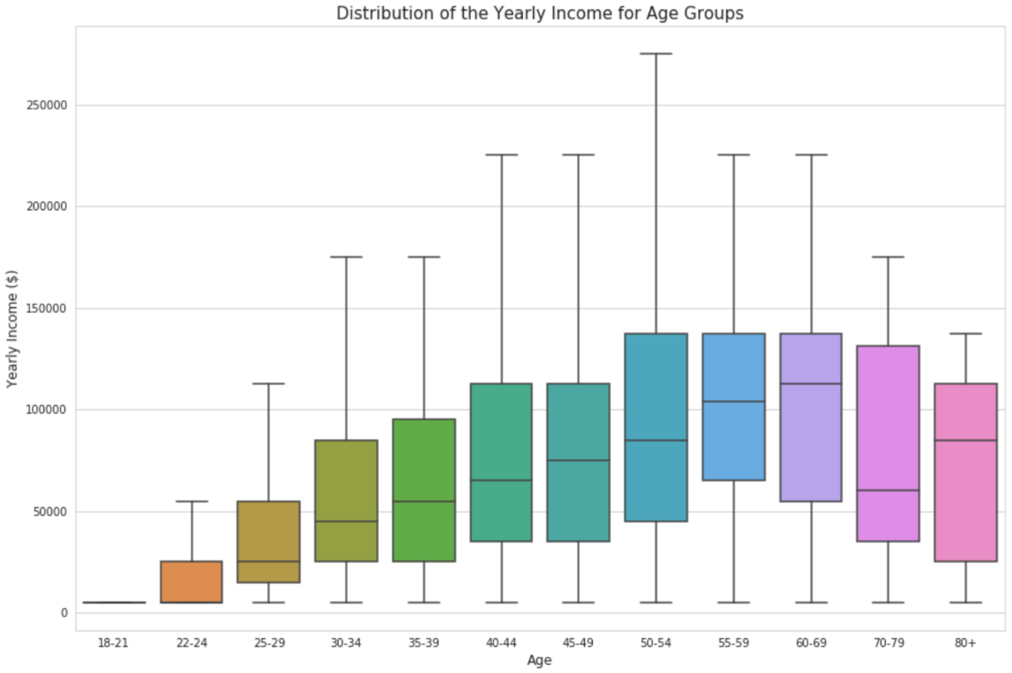
According to the survey, there are more men than women among the respondents. Men consist %84 all respondents, where as %16 of respondents are women (Graph 1.). When we look at the distribution of salaries among male and female, it is quite similar (Graph 2.).

Graph 2. Distribution of Yearly Income Among Different Genders



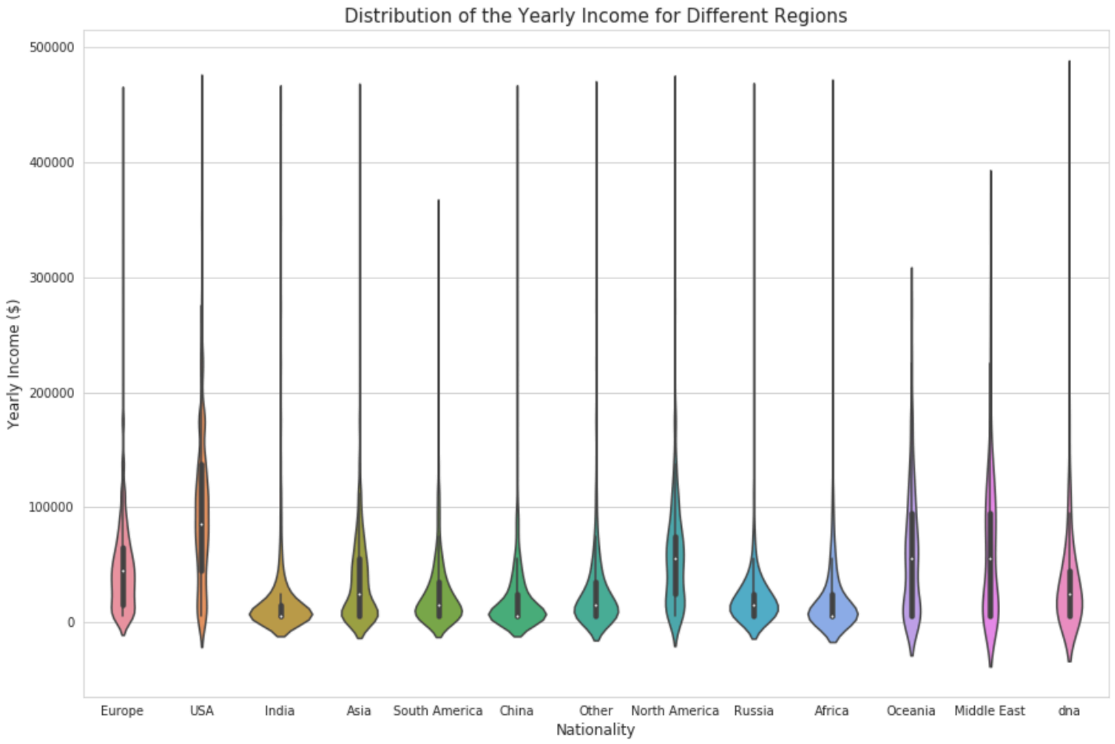
When it comes to distribution of salaries among different age groups, the positive correlation can be seen. The older the respondents are, the more they earn until the retirement.

Graph 3. Distribution of Yearly Income Among Age Groups



Salaries highly vary depending on the country that people work in. To inspect the distrubiton of salaries among countries, I regroup most country by region and continent except the most represented ones which are USA, India, China, Russia, and Brazil.

Graph 4. Distribution of Yearly Income Among Different Regions



Due to the economic instability in North America, Oceania, and Middle East, the salaries are higher in those areas. The fact that the life conditions are not cheap, their economic system permits high wages (Graph 4.).

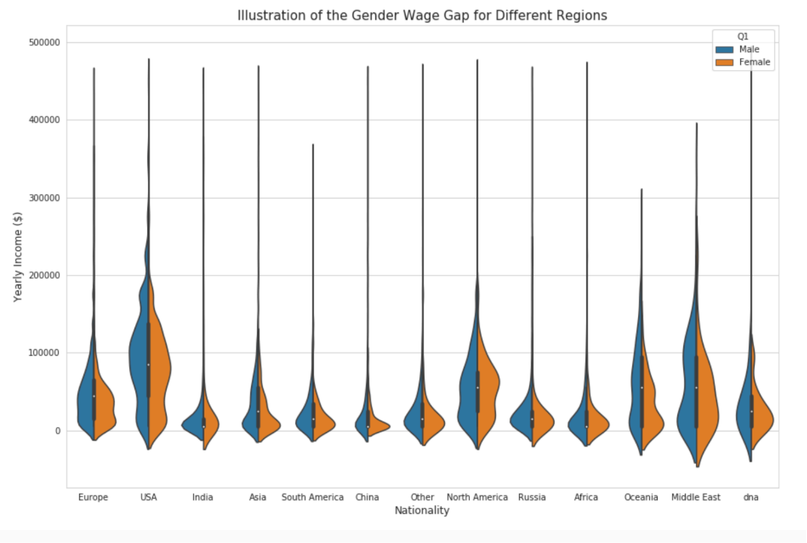
c. Gender Pay Gap Analysis

The gender pay gap is quantified by comparing the median and mean salaries of men and women within the dataset. Statistical tests are performed to assess the significance of the differences in salary distributions between genders. Additionally, the gender pay gap is examined across different demographic and professional categories to identify potential disparities.

Gender Wage Gap in Different Countries

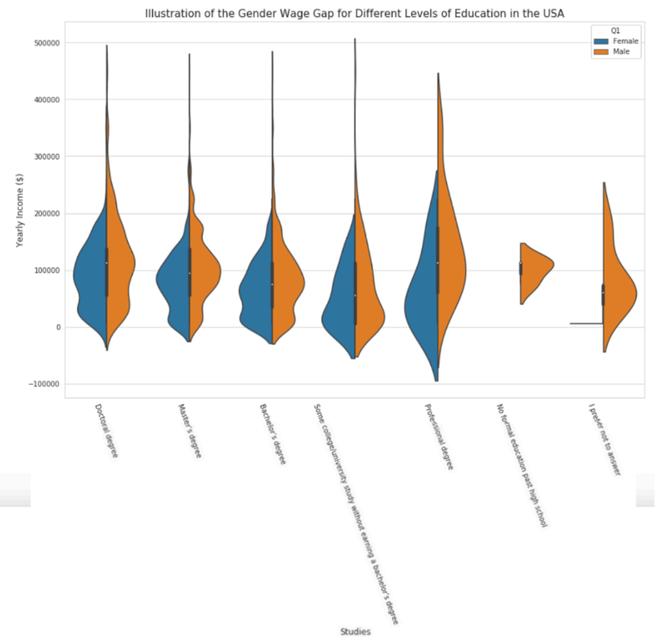
The wage gap is quite visible in the graph below. The wage gap appears to be higher in Europe and in North America compared to Asia (Graph 5).

Graph 5. Gender Wage Gap Among Different Regions

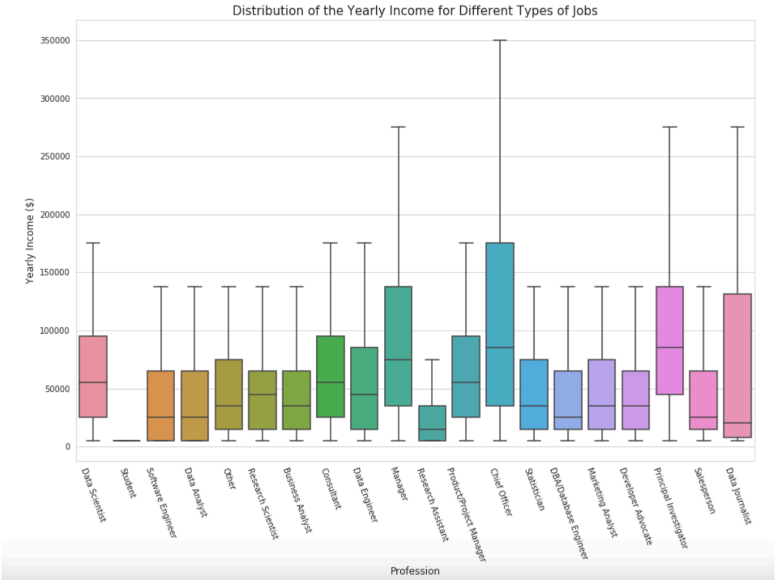


When we examine the case for USA (Graph 6.), gender wage gap is not only caused by education level differences, as there is differences inside each type of studies. According to many studies, women tend to study more than men in the USA, however, this is not the case in this study since women are underrepresented. It can be said that the length of the studies is not the cause to the gap in salaries among genders.

Graph 6. Gender Wage Gap for Different Levels of Education in USA



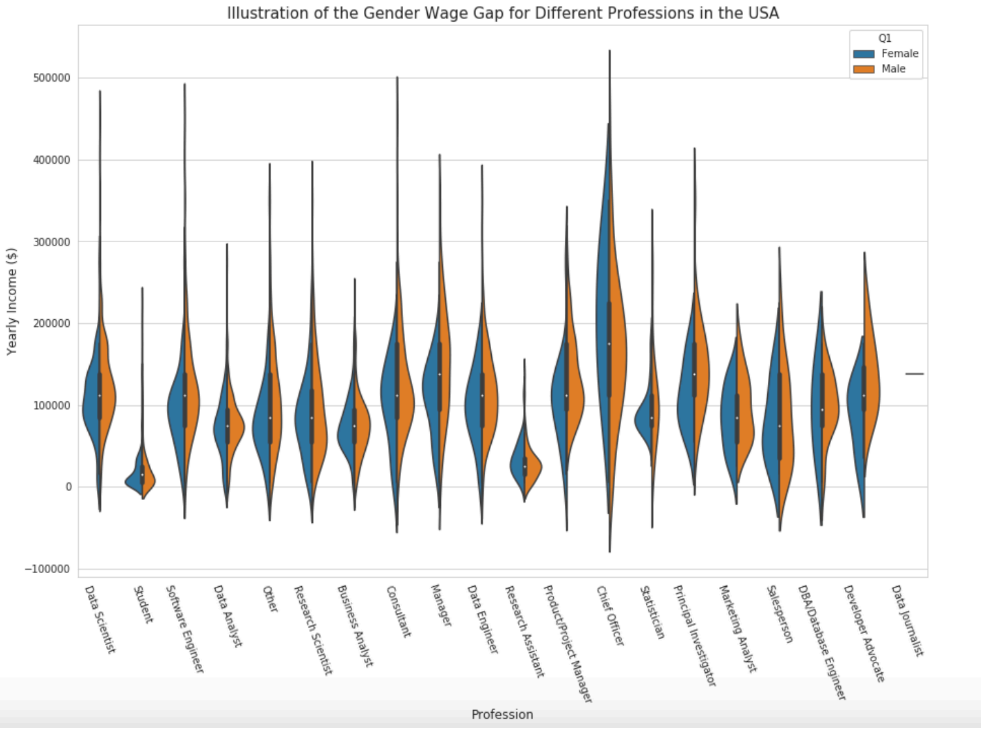
Graph 7. Distribution of Yearly Income for Different Types of Jobs



According to Graph 7., top earning jobs are “Chief Officer”, “Manager”, and “Principal Investigator” which is not surprising since those kinds of jobs generally held by older people.

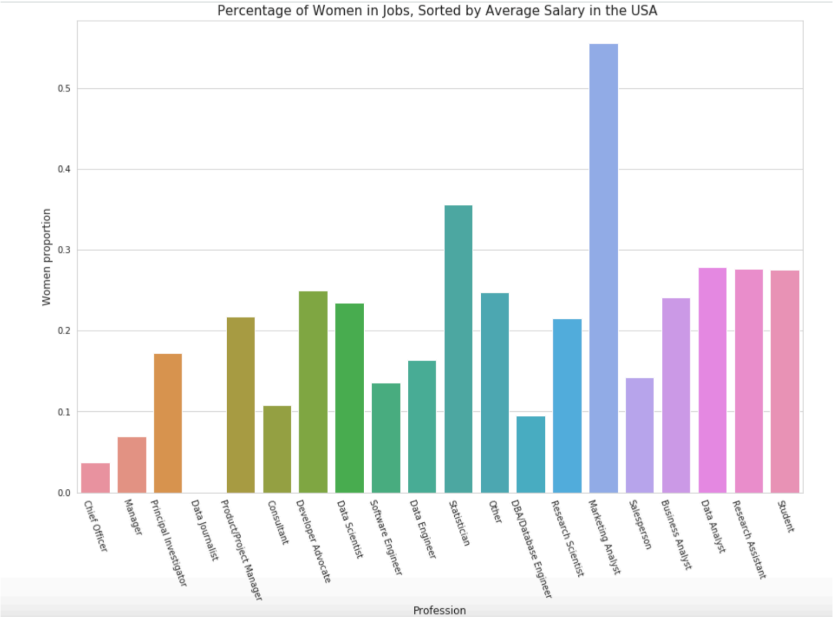
Gender Wage Gap in the Same Profession

Graph 8. Gender Wage Gap for Different Professions in the USA

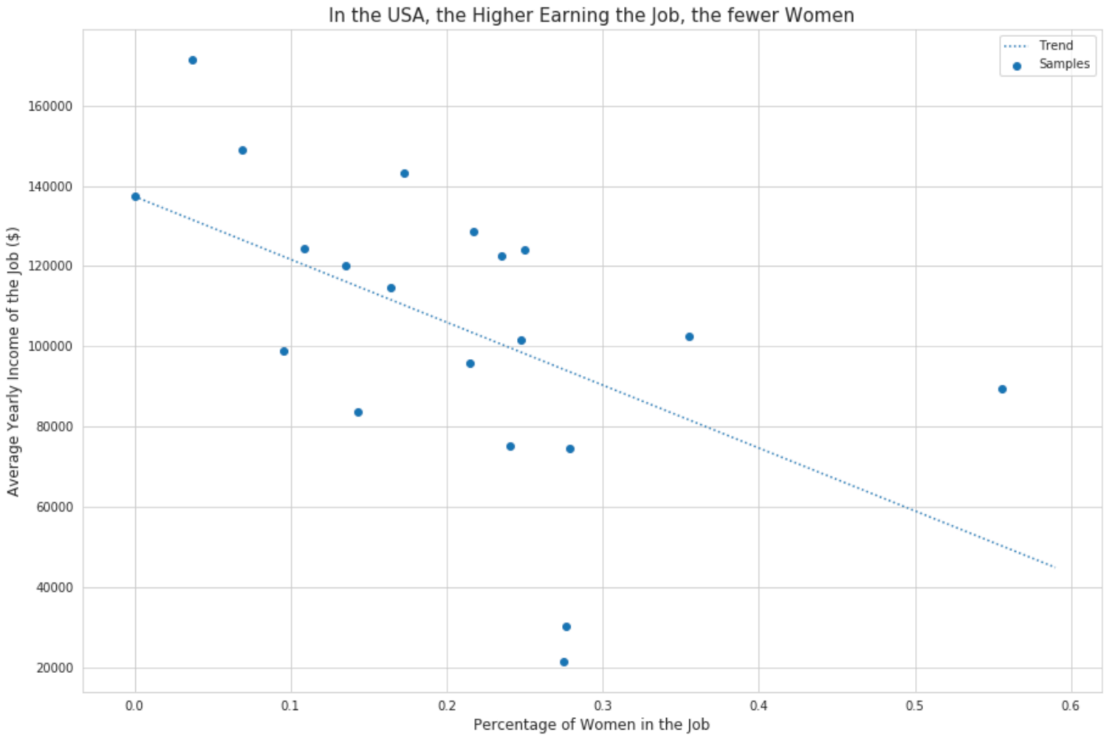


For the same job, the wage gap seems to be smaller. Overall, me seem to gain more salary then men, however women do better as Chief Officer and approximately as much as men in most jobs (Graph 8.). However, it can be said that men tend to occupy higher earning jobs. To be able to show this, linear regression performed to analyze the relationship between the percentage of women in a job and the average yearly income of that job. The jobs with higher salaries recruit fewer women (Graph 10.).

Graph 9. Percentage of Women in Jobs in the USA (sorted by Average Salary)



Graph 10. Gender Diversity and Income in US Jobs: Linear Regression Analysis



d. Preprocessing and Salary Prediction Modeling

Predictive models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression, are developed to estimate salaries based on demographic and professional features. The dataset used in this study contains categorical variables such as 'Gender', 'Age', 'Country', 'Education', 'Major', 'Profession', 'Industry', 'Experience', and 'Annual_Salary'. Dummy variables were created for these categorical features in the preprocessing pipeline using OneHotEncoder. Then, the data set was split into training and testing sets.

The aim of this study is to predict the annual income in thousand of USD. The problem was tackled as a classification one rather a regression one since regression might give bad accuracy on lower salaries. That's why, 6 categories are created for the salaries which are:

1. less than 10k
2. between 10k and 30k
3. between 30k and 50k
4. between 50k and 80k
5. between 80k and 125k
6. more than 100k

Four different models were initialized which are Linear Regression, Decision Tree, Random Forest, and Gradient Boosting. The Gradient Boosting and Decision Tree were achieved perfect performance on the test dataset, however, the Gradient Boosting model has been selected since Decision Tree model may be prone to overfitting. The Gradient Boosting model, on the other hand, tends to generalize better by combining multiple weak learners. The Gradient Boosting model achieved strong performance with a relatively low MSE and a high R-squared score. While it didn't achieve perfect performance like the Decision Tree model, it's a more robust option that tends to generalize well to unseen data.

e. Hyperparameter Tuning Using GridSearchCV

Additionally, hyperparameter tuning was performed to optimize the performance of the models. Grid search or random search techniques were applied to search the hyperparameter space and identify the optimal combination of hyperparameters for each model. This process helped improve the predictive performance of the models and ensure robustness in the predictions.

f. Model Interpretation

Model performance is evaluated using metrics such as Mean Squared Error and R-squared to assess predictive accuracy. The trained models are interpreted to understand the relative importance of different features in predicting salary. Feature importance scores are analyzed to identify the key factors influencing salary estimation.

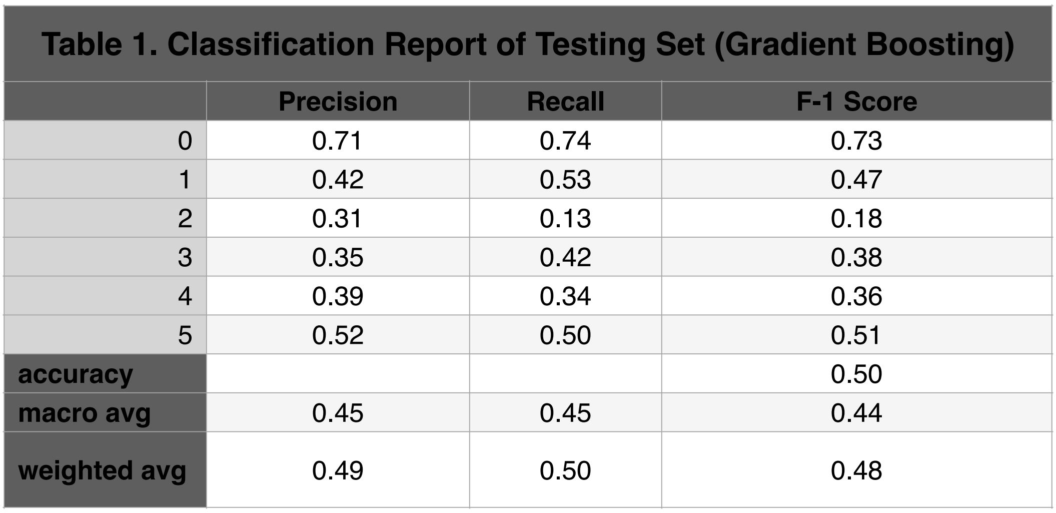
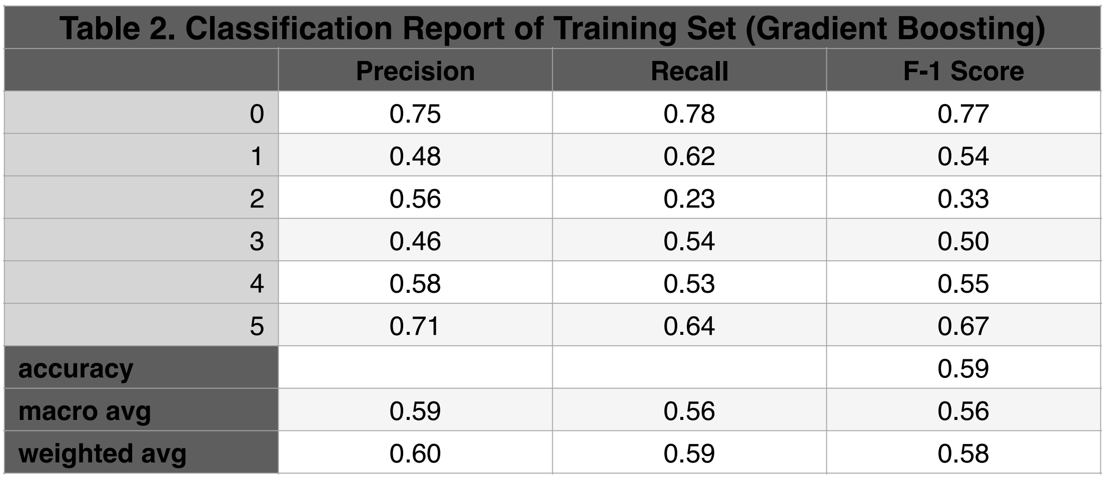


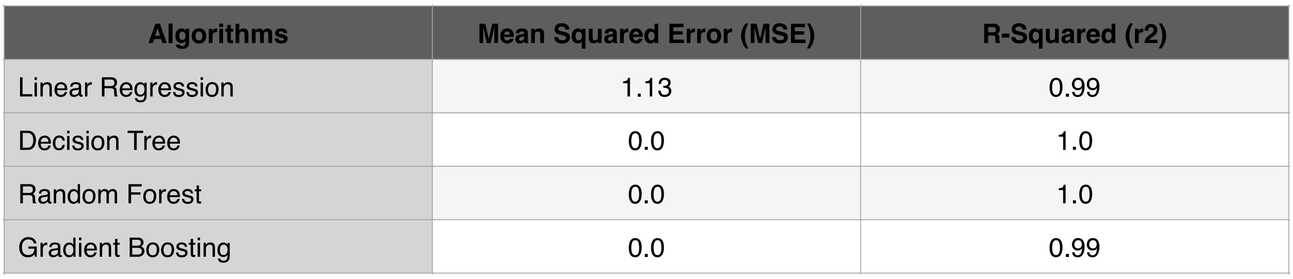
Table 1. Classification Report of Testing Set (Gradient Boosting)

	Precision	Recall	F-1 Score
0	0.71	0.74	0.73
1	0.42	0.53	0.47
2	0.31	0.13	0.18
3	0.35	0.42	0.38
4	0.39	0.34	0.36
5	0.52	0.50	0.51
accuracy			0.50
macro avg	0.45	0.45	0.44
weighted avg	0.49	0.50	0.48

Table 2. Classification Report of Training Set (Gradient Boosting)



	Precision	Recall	F-1 Score
0	0.75	0.78	0.77
1	0.48	0.62	0.54
2	0.56	0.23	0.33
3	0.46	0.54	0.50
4	0.58	0.53	0.55
5	0.71	0.64	0.67
accuracy			0.59
macro avg	0.59	0.56	0.56
weighted avg	0.60	0.59	0.58

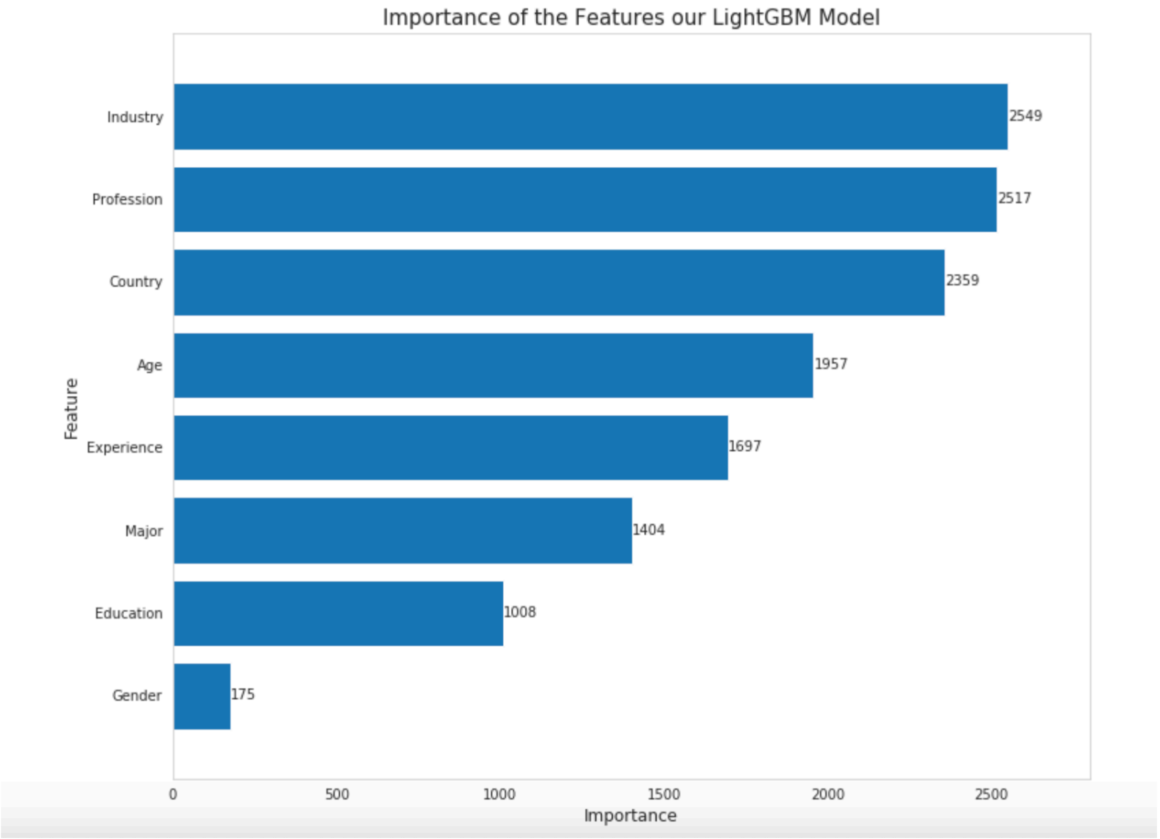


Algorithms	Mean Squared Error (MSE)	R-Squared (r2)
Linear Regression	1.13	0.99
Decision Tree	0.0	1.0
Random Forest	0.0	1.0
Gradient Boosting	0.0	0.99

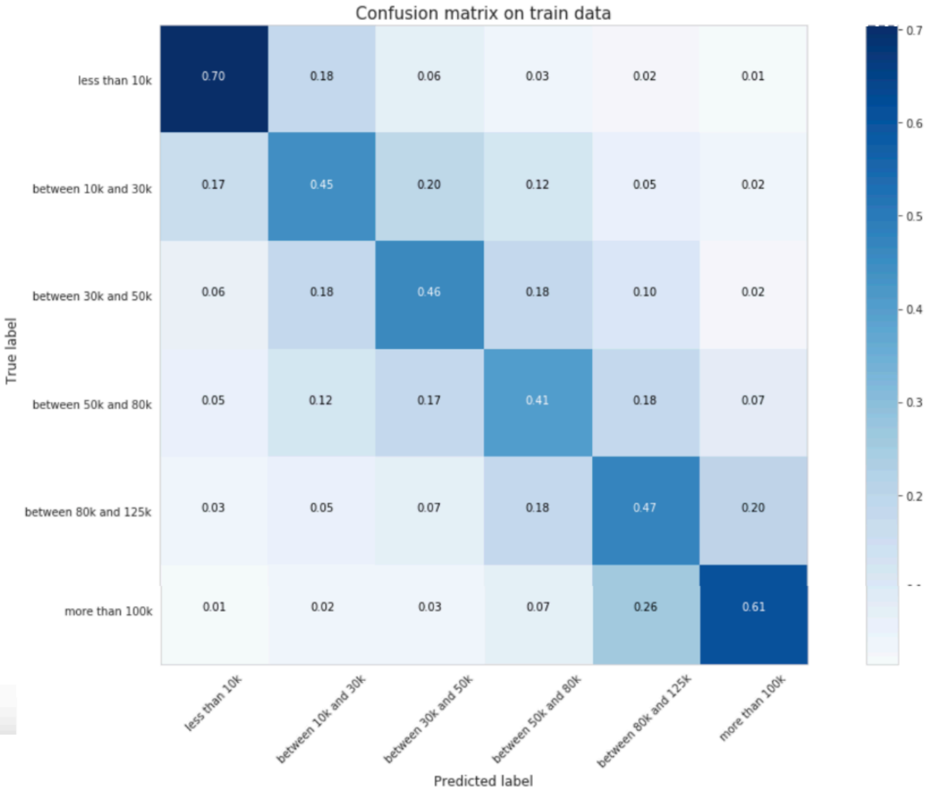
g. Feature Importance using Gradient Boosting with LightGBM

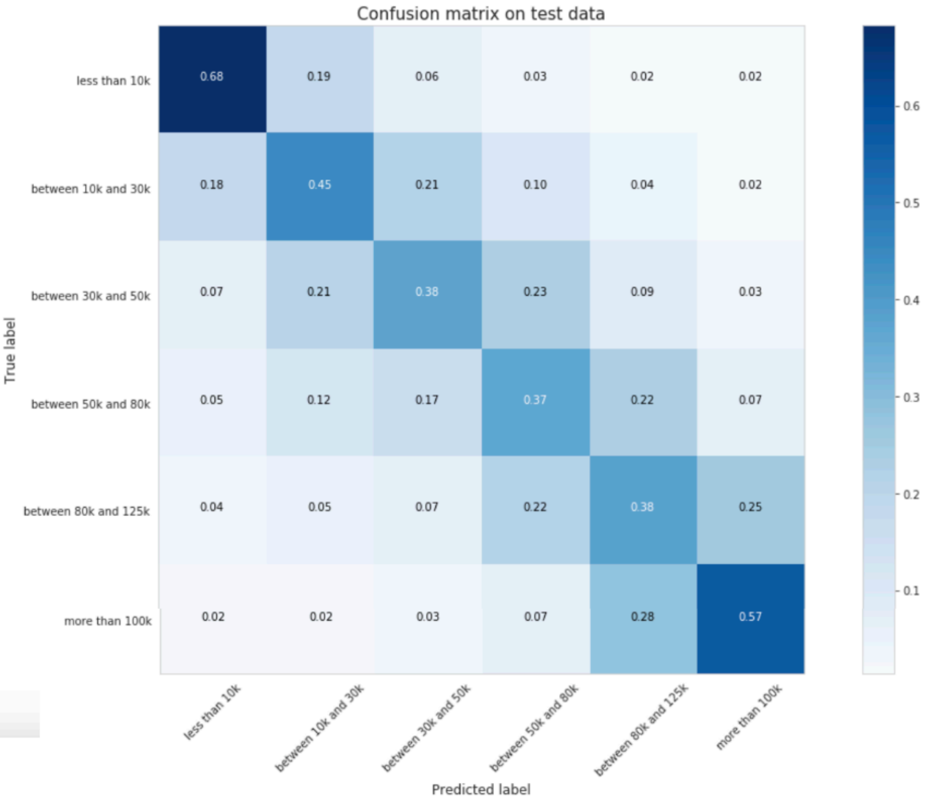
According to the result of the analysis, gender appears as the least important feature (Graph 11.). This does not mean that gender wage gap does not exist, however, it show that this is not a factors when determining the salary of respondents in this survey. The most important parameter is the profession and it can be seen that higher earning jobs are occupied by higher proportions of men.

Graph 11. Feature Importance (LightGBM Model)



According to our prediction after modeling, low paid and high paid scientist are the easier ones to detect. This is mostly due to the wide rang of high paid people and the fact that students are easy to detect.





Conclusion

In conclusion, this analysis highlights the existence of the gender pay gap and its implications for gender equality in the workplace. By leveraging data-driven approaches, we gain insights into the factors contributing to salary discrepancies between genders and develop predictive models to estimate salaries based on demographic and professional attributes. The findings from this analysis contribute to the ongoing efforts to promote fairness and equality in the workforce and provide valuable insights for policymakers, organizations, and individuals striving to address the gender pay gap.