IT5006 Milestone 4 - Final Report

Gender Pay Gap in Tech Industry

Group 10:

Huang Yunqi (A0275689A)

Liu Chang (A0275717U)

Wong Kit Long (A0275793L)

1 Dataset and Preprocessing

1. Dataset:

- The dataset comprises a survey collecting demographic, occupational, coding, and machine-learning skill information from Kaggle users. It spans a duration of 3 years.

2. Import Libraries:

- Utilize pandas (pd), numpy (np), matplotlib and shap.

3. Dataset Consolidation:

- Extract data from three distinct CSV files (2020, 2021, and 2022) and generate three pandas DataFrames (DF).
- Retrieve question titles from the first row of each DataFrame.
- Select columns from the three DataFrames and concatenate them into a single DataFrame using pd.concat.

4. Data Preprocessing:

- Eliminate irrelevant data not pertaining to our primary focus, which is gender wage differences.
- Address missing gender information, students, and unemployed individuals.
- Filter out data lacking salary information or completed in less than one minute to prevent non-response, selection, and sampling biases.
- Transform feature values by mapping, ensuring consistency across the three years.
- Convert categorical columns to integers (e.g., age, salary).
- Combine and remove certain categories, grouping low-occurrence categories into ‘other'.
- Exclude data with null fields.
- Convert category columns into dummy features to streamline visualization and analysis.
- Standardize salary by region to account for varying salary levels across countries.

5. Feature Selection:

- Conduct correlation analysis and eliminate features with a correlation greater than 0.7 (Figure 1).
- Employ Shap Value: Train an XGBoost model and use Shap values to identify the importance of each feature. Less important features like "use other ML model" and "use other programming language" are subsequently removed (Figure 2).

2 Exploratory Analysis Insights

As the tech industry rapidly evolves, the topic of gender equality in the workplace is becoming increasingly important so we embark on our data analytics project on a comprehensive examination of the gender pay gap within the tech industry. Through in-depth analysis of Kaggle survey datasets, we aim to uncover the multifaceted factors that contribute to income inequality between men and women.

2.1 Feature Trend

Our investigation begins with a meticulous time series analysis, charting the evolution of median salaries across different dimensions(regions, education levels, etc.) from 2020 to 2022. In terms of the overarching trend, the gender pay gap within the tech industry exhibited a notable shift, exceeding 250% in 2020, but showing a substantial reduction to approximately 120% by 2022 (Figure 3). This suggests a positive trajectory over the span of three years. Nevertheless, a considerable pay disparity persists, and the extent of the gap remains substantial.

In terms of **educational qualifications**, the gender pay gap for individuals with a Bachelor's degree and a Master's degree exhibited an upward trend from 2020 to 2021. However, there is a positive development as the pay gap significantly decreased from 2021 to 2022, dropping below 100%. This encouraging shift suggests a potential improvement in pay equity within the two educational categories. Nevertheless, the gender pay disparity within the category of Doctoral degrees has shown a persistent upward trajectory, escalating consistently each year from 150% in 2020 to nearly 350% in 2022(Figure 4). This alarming trend underscores a profound and concerning gender inequality issue within the realm of higher education. When we deep dive into the **regions**, we found that Russia witnessed an expanding gender pay gap, however, in other countries, there was a gradual narrowing of the pay gap (Figure 5). This indicates positive results from efforts to address pay disparities, although challenges persist in specific regions.

2.2 Statistical Insights

Furthermore, we scrutinize the intricate interplay between salary and gender, dissecting it by various aspects. Initially, with the accumulation of **coding experience**, there is a tendency for the gender pay gap to expand over three years, peaking between 3-5 years of coding experience(Figure 6). Subsequently, a gradual decline in the pay gap becomes apparent. This shift could be attributed to the increasing visibility of an individual's actual contributions, which are less influenced by gender factors. This observed trend is consistent with the patterns observed in individuals gaining **experience in machine learning** (Figure 7).

The gender pay gap demonstrates variations across different roles. **Occupations** with more advanced and demanding skill requirements, such as machine learning engineers, teachers/professors, and research scientists, consistently exhibit higher pay gaps in comparison to roles with fewer technological skill prerequisites, like managers and business analysts(Figure 8).

The choice of **programming languages** also impacts the pay gap, with higher gaps observed among individuals using C, C++ and Java, commonly associated with software engineering roles(Figure 9). Moreover, for professionals using **deep learning packages** such as JAX, Fastai and Keras, a notable gender pay gap is evident, whereas individuals relying on libraries like XGBoost and scikit-learn experience smaller gaps(Figure 10). This indicates that within the realm of top-tier data scientists and machine learning experts, there persists a substantial gender pay gap, aligning with the prevailing trends associated with these roles.

In terms of **regions**, Russia and India experienced a widening gender pay gap, at 317% and 250% respectively from 2021 to 2022, whereas the disparity is comparatively smaller in developed countries such as Germany and the UK(Figure 11). This discrepancy can potentially be attributed to the heightened awareness of gender equality in developed nations, prompting employers to demonstrate a greater inclination toward equitable pay practices. Conversely, in some developing countries, cultural norms may adhere to more traditional gender role concepts, contributing to more pronounced salary differences.

3 Recommendation Task

3.1 Evaluation Metric

Before delving into the model selection process, it is crucial to establish the metrics by which the models will be evaluated. For the task of predicting job titles, we chose three primary metrics to assess the performance of our models: Accuracy, Recall, and F1 Score. **Accuracy** measures the overall correctness of the model, indicating the percentage of total correct predictions. **Recall** ensures that the model is capable of identifying the correct job titles as often as possible. **F1 Score** balances the precision and recall, providing a single metric for the model's accuracy concerning both false positives and negatives.

3.2 Model Selection

Our goal is to predict job titles from user input, modeling it as a Multiclass Classification problem. This supervised learning method involves training the model to forecast one among multiple discrete labels corresponding to various job titles. We first encode the labels (i.e., role titles) into integers and then divide the dataset into training and test sets. To navigate the complexities of predicting job roles—a task with numerous potential outcomes—we test a range of classifiers known for their efficacy in multiclass settings. Finally, we sort the results by accuracy and print them. The results of some classifier tests are presented in table 1.

LightGBM stands out due to its high accuracy and efficiency, especially in dealing with the categorical nature of our data. It is faster than many other gradient boosting frameworks because of its histogram-based optimization and leaf-wise growth strategy as opposed to level-wise growth. This makes it particularly useful for a responsive user experience in a web application. It also has better control over overfitting compared to other non-ensemble methods, even when dealing with data that has numerous features, which is common in user-profile-based predictions.

3.3 Model Fine-tuning with LightGBM

After selecting LightGBM as our model of choice for the job title prediction task, the next critical step is to fine-tune the model to optimize its performance. We employ three prominent hyperparameter optimization techniques to find the best set of parameters for LightGBM: RandomizedSearchCV, GridSearchCV, and Optuna.

RandomizedSearchCV offers a probabilistic approach to hyperparameter optimization. Instead of exhaustively searching over all possible combinations, it randomly samples a predefined number of parameter combinations from a specified distribution.

GridSearchCV takes a more exhaustive approach, systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. It is more likely to find the optimal parameters due to its exhaustive nature.

Optuna is a newer, open-source hyperparameter optimization framework that automates the process of finding the best hyperparameters. It differs from the aforementioned methods by employing a more efficient search through the parameter space using Bayesian optimization. Optuna uses past trial knowledge to predict which set of parameters might lead to better results. It can stop the evaluation of a hyperparameter set early if it determines that these parameters are unlikely to yield good results, saving computational resources.

Fine-tuning LightGBM with these advanced techniques allows us to harness the full potential of the model. After fine-tuning, the accuracy of the model can reach 0.4822. By methodically iterating over the hyperparameter space and rigorously validating the results, we ensure that our model is both accurate and robust, leading to a reliable and effective job recommendation system.

Model Finalization and Deployment

Upon fine-tuning the LightGBM model to achieve the best performance on our validation set, we proceed to save the model. This ensures that the exact state of the model can be preserved, allowing for consistent predictions and eliminating the need to retrain the model each time it's used.

When a user enters their information into the web application, the input data is preprocessed in the same way as the training data to ensure consistency. The preprocessed data is then passed to the LightGBM model through the API. The model predicts the most probable job titles based on the user's information and returns these predictions to the front end of the web application. Job titles are ranked by the probability of match, giving users clear guidance on their best-fit roles.

4 Reflection and Conclusion

Reflection

Throughout this project, our journey encompasses a multifaceted learning experience. In the realm of data analysis, we delve into comprehensive data preparation, involving intricate processes such as data cleaning, feature engineering, and judicious feature selection. Integrating learning visualization and plotting, we convey insights from data analysis and machine learning, enhancing our multifaceted journey. Navigating the landscape of machine learning, our focus extends to model selection, meticulous parameter tuning, and thorough model evaluation. The practical application of these skills is further demonstrated in the development of a web app using Streamlit. We gain proficiency in utilizing the ChatGPT API, honing skills in prompt engineering, and fine-tuning parameters such as temperature to enhance model responses. Alongside our newfound knowledge, we conscientiously recognize the limitations inherent in our approach (text 1). These acknowledgments serve as crucial signposts for refining and expanding our methodologies in future endeavors.

Conclusion

In summary, this comprehensive exploration into the gender pay gap within the tech industry illuminates key facets through data analysis and machine learning. Leveraging Kaggle survey datasets spanning 2020 to 2022, we navigate the intricacies of data collection, preprocessing, and modeling, employing various techniques and tools. Our findings unveil nuanced insights and trends surrounding the gender pay gap, dissecting its impact across educational backgrounds, professional experience, job roles, programming languages, and geographical regions.

The application of LightGBM as a machine learning model to predict job titles showcases the potential of such tools in addressing real-world challenges. The fine-tuning process, incorporating different hyperparameter optimization methods, underscores the significance of model refinement for enhanced performance. Our evaluation metrics, encompassing accuracy, recall, and F1 score, provide a robust measure of the model's efficacy.

The significance of the gender pay gap in the tech industry is undeniable. As we navigate the intersection of technology and societal challenges, data analysis and machine learning emerge as powerful tools for positive change. Recognizing the persistent challenges, we acknowledge their potential to meaningfully contribute to a more equitable future in the tech workforce.

Appendix

Figure 1: Feature Correlation Plot

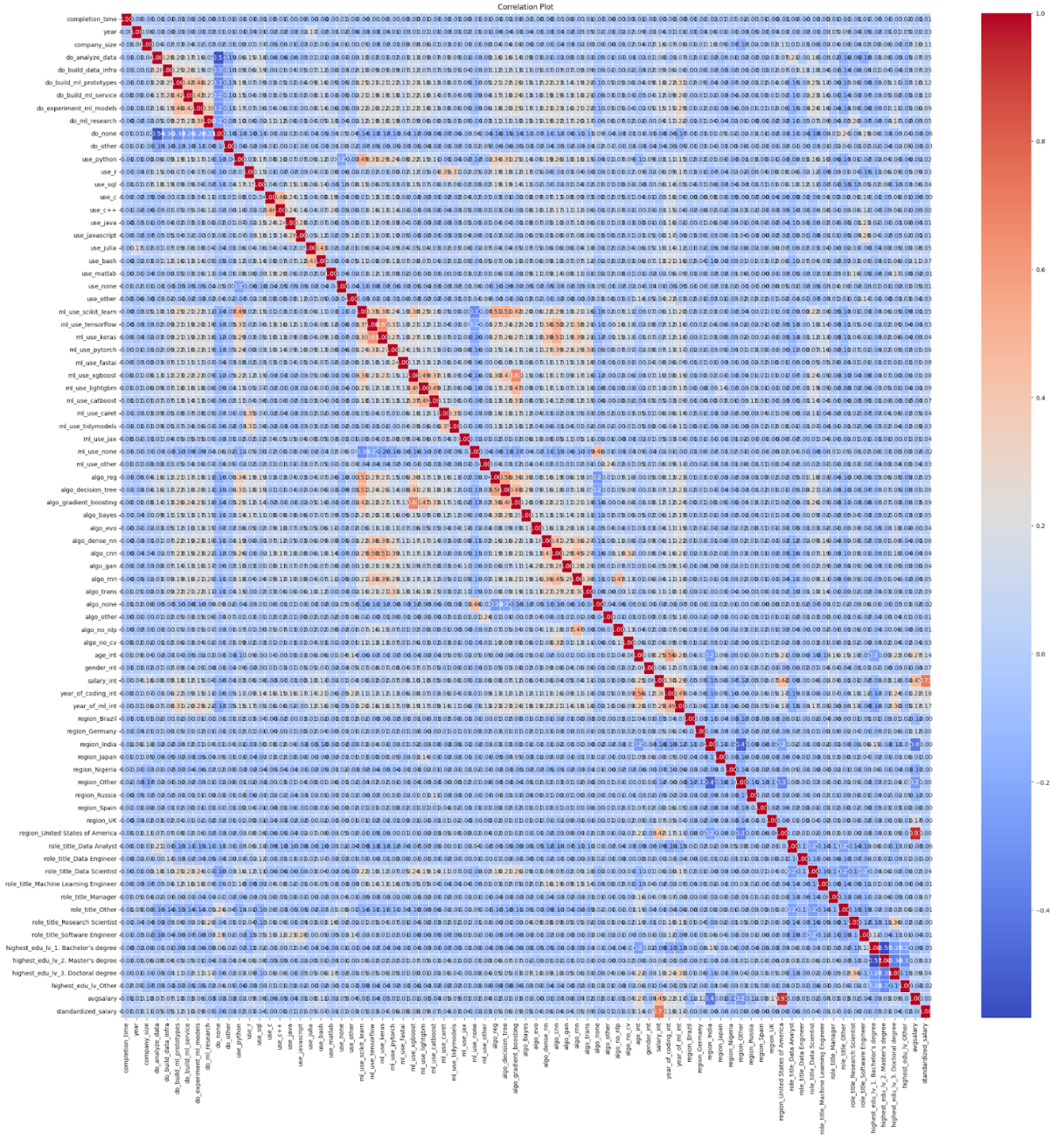
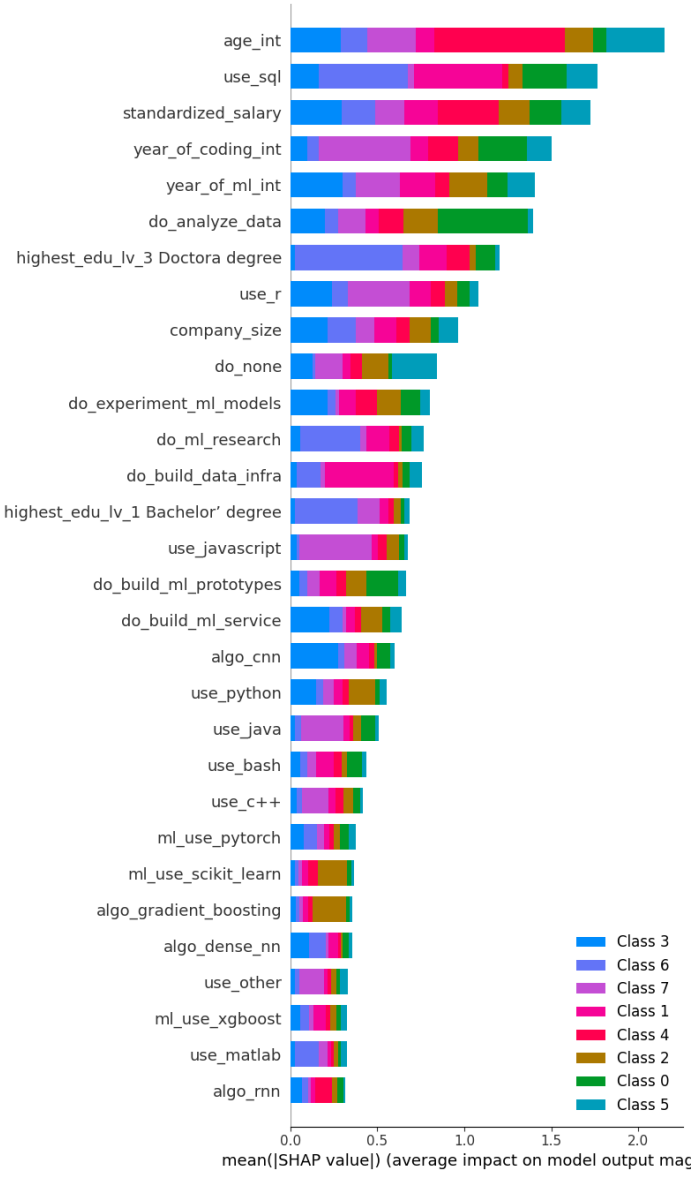


Figure 2: Feature Shap Value Plot



{0: 'Data Analyst', 1: 'Data Engineer', 2: 'Data Scientist', 3: 'Machine Learning Engineer', 4: 'Manager', 5: 'Other', 6: 'Research Scientist', 7: 'Software Engineer'}

Figure 3

Figure 4

Figure 5

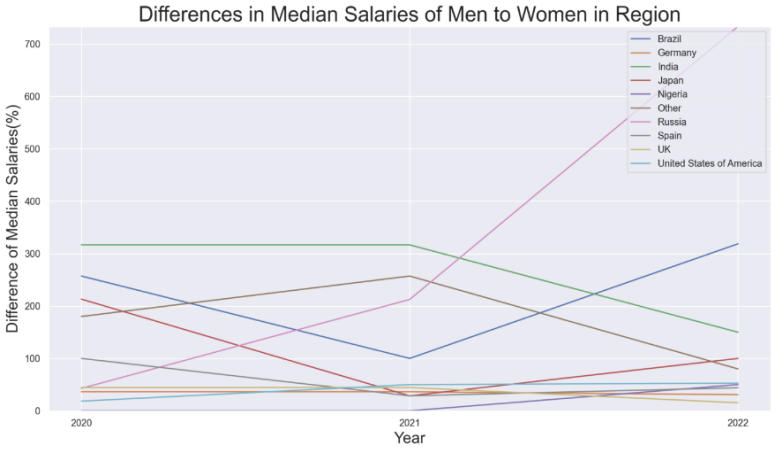


Figure 6

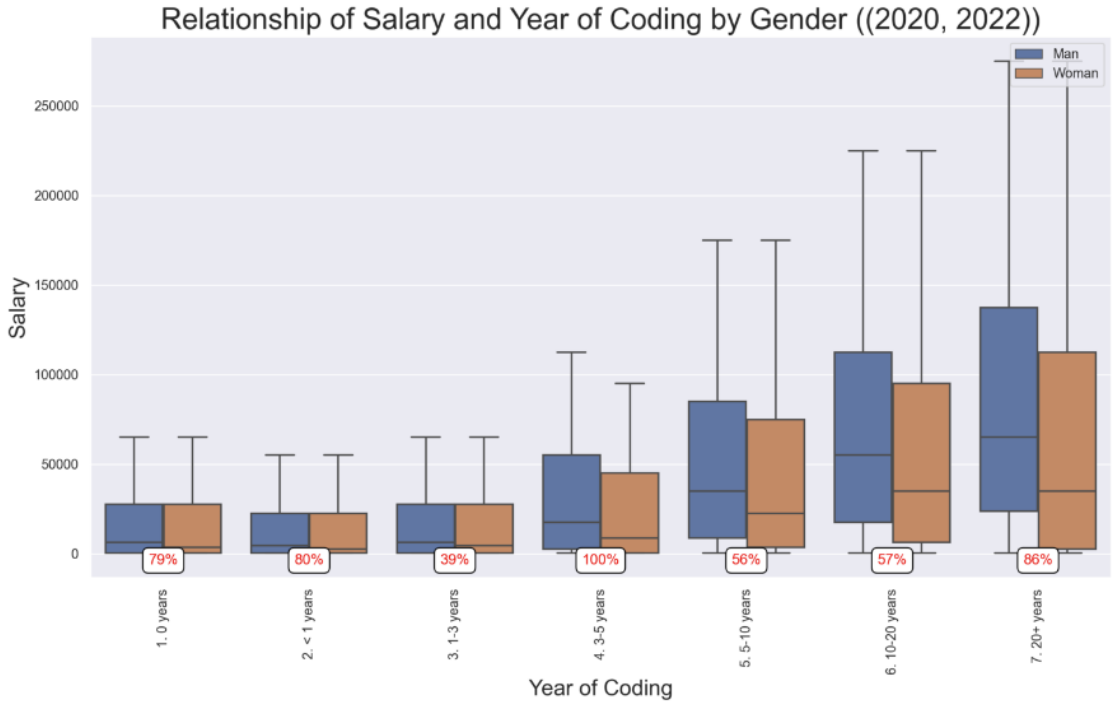


Figure 7

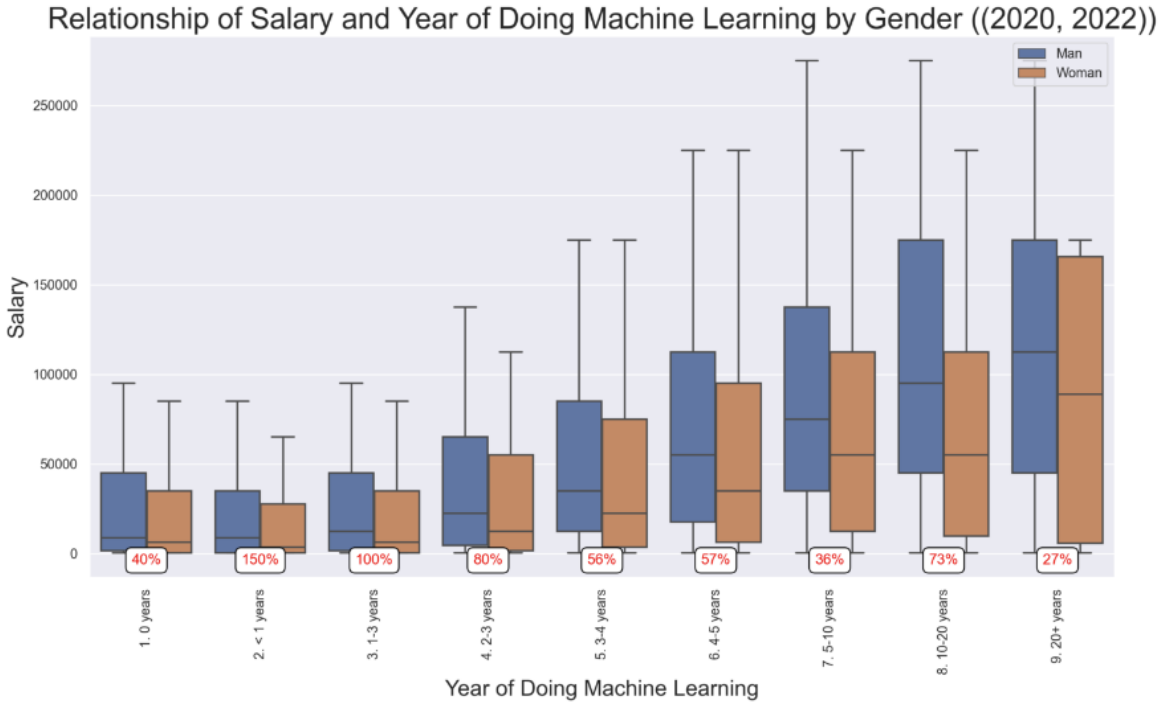


Figure 8

Figure 9

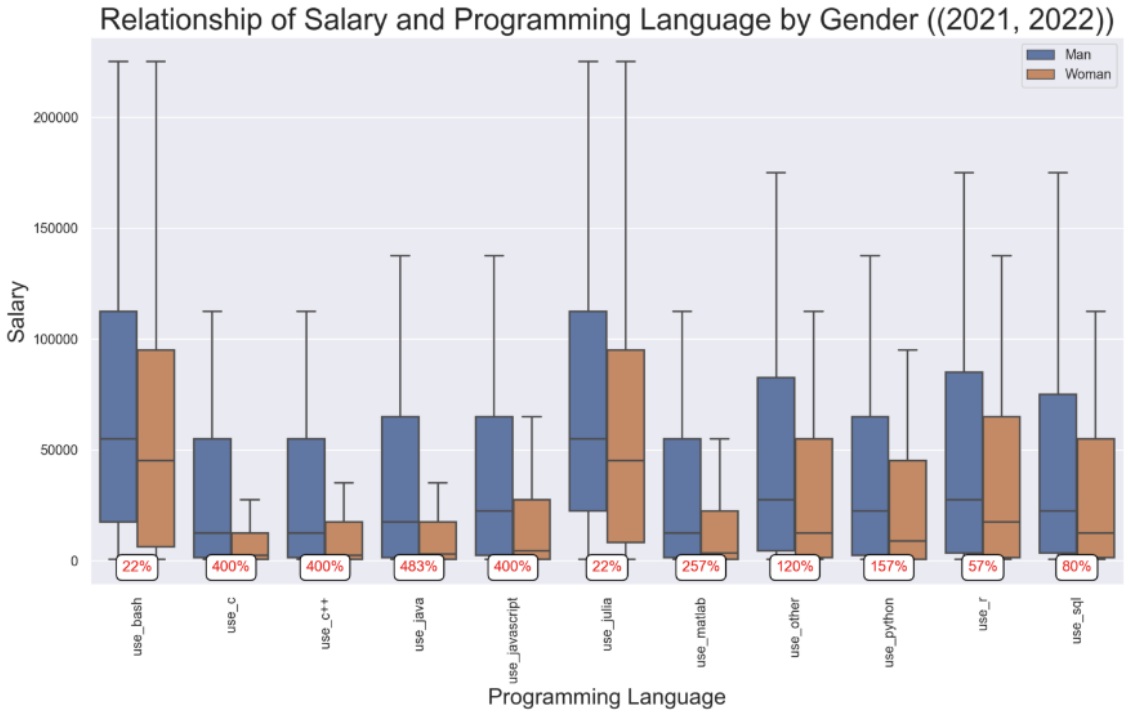


Figure 10

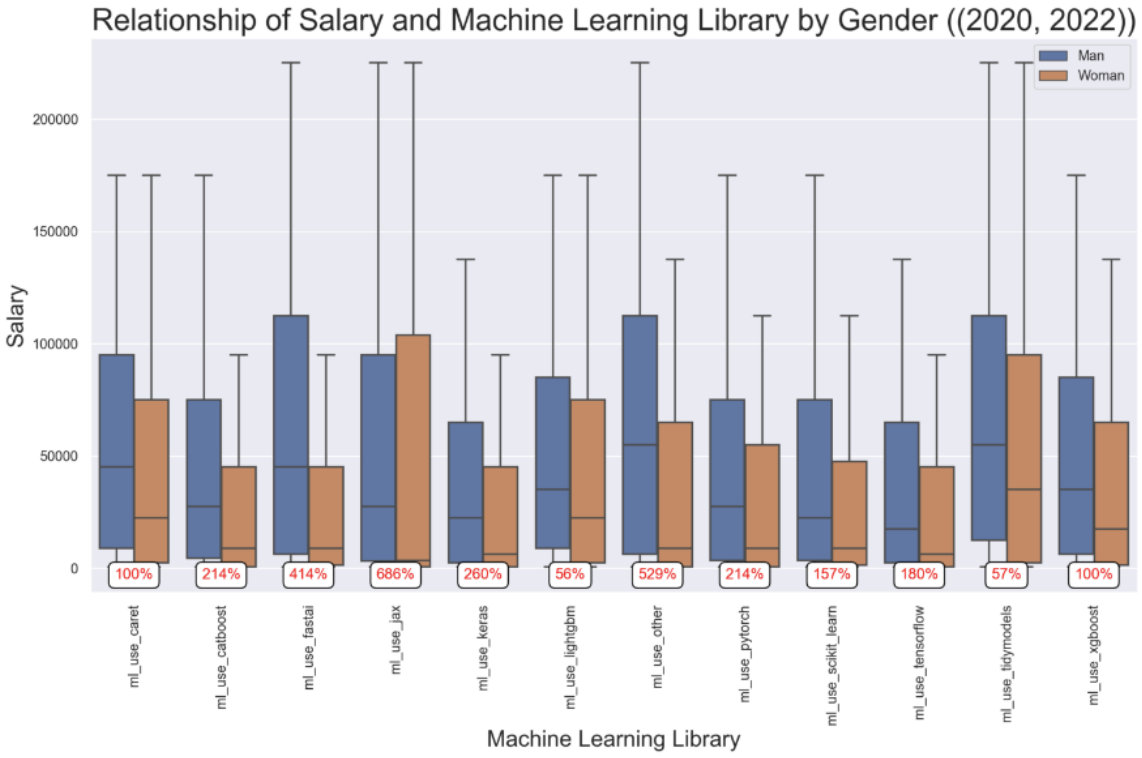

Relationship of Salary and Machine Learning Library by Gender ((2020, 2022))
Man
Woman
Salary
200000
150000
100000
50000
0
100%
214%
414%
686%
260%
56%
529%
214%
157%
180%
57%
100%
ml_use_caret
ml_use_catboost
ml_use_fastai
ml_use_jax
ml_use_keras
ml_use_lightgbm
ml_use_other
ml_use_pytorch
ml_use_scikit_learn
ml_use_tensorflow
ml_use_tidymodels
ml_use_xgboost
Machine Learning Library

Figure 11

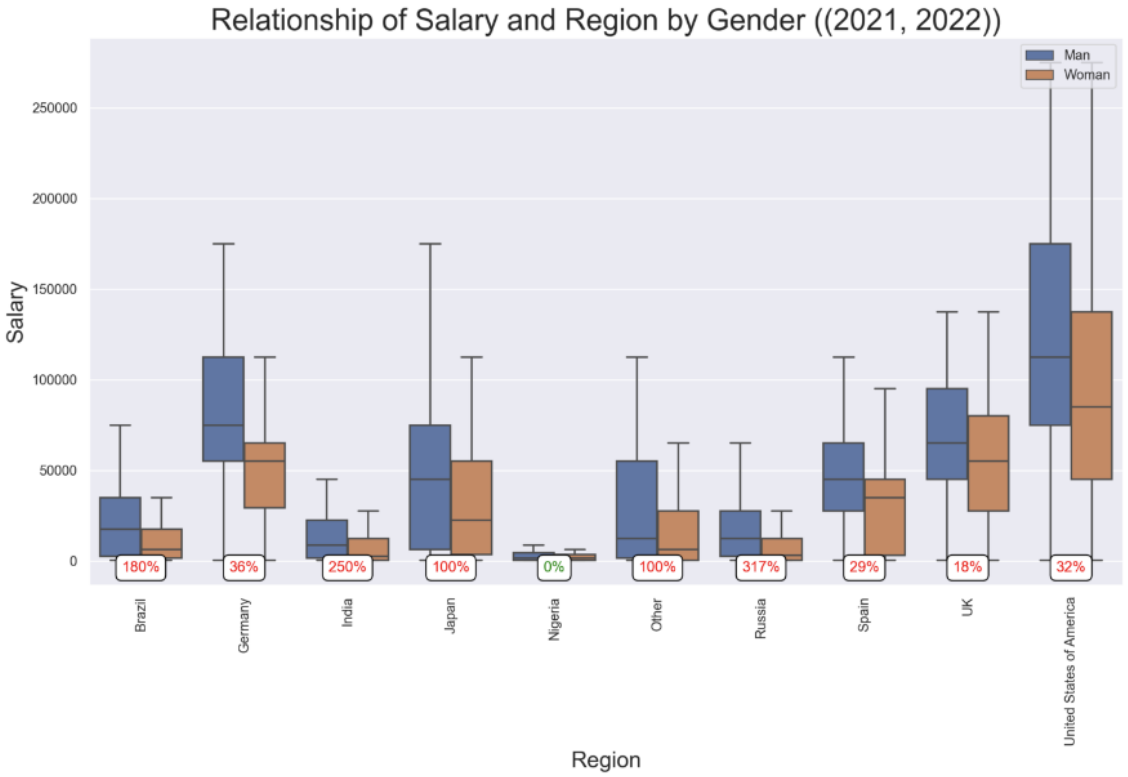

Relationship of Salary and Region by Gender ((2021, 2022))
Man
Woman
Salary
250000
200000
150000
100000
50000
0
180%
36%
250%
100%
0%
100%
317%
29%
18%
32%
Brazil
Germany
India
Japan
Nigeria
Other
Russia
Spain
UK
United States of America
Region

Table 1: Classifier Performance Comparison

Rank	Classifier	Accuracy	Rank	Classifier	Accuracy
1	LightGBM	0.4293	6	K-Nearest Neighbors	0.2429
2	Gradiant Boosting	0.4228	7	Naive Bayes	0.2259
3	XGBoost	0.4196	8	Logistic Regression	0.2125
4	Random Forest	0.4118	9	Support Vector Machine	0.2119
5	Decision Tree	0.2724	10	Neural Network	0.1803

Table 2: LightGBM Classification Performance

Top-1 Accuracy: 0.48221
Top-2 Accuracy: 0.69548
Recall: 0.31500
F1 Score: 0.32299

Confusion Matrix:

	Data Analyst	Data Engineer	Data Scientist	Engineer (non-software)	Machine Learning Engineer	Manager	Research Scientist	Software Engineer	Statistician	Teacher / professor
Data Analyst	734	15	245	1	17	34	47	116	5	6
Data Engineer	86	39	71	0	14	9	8	70	1	0
Data Scientist	237	8	893	0	94	13	93	63	9	4
Engineer (non-software)	44	0	6	2	1	10	6	15	0	2
Machine Learning Engineer	45	3	213	0	136	4	56	71	0	1
Manager	160	5	94	2	11	80	28	108	0	2
Research Scientist	56	2	119	0	28	18	301	31	8	14
Software Engineer	98	11	110	0	53	29	32	596	0	6
Statistician	60	0	34	0	1	3	19	3	9	0
Teacher / professor	23	1	21	0	2	3	60	14	1	16

Classification Report:

	precision	recall	f1-score	support
Data Analyst	0.48	0.6	0.53	1220
Data Engineer	0.46	0.13	0.2	298
Data Scientist	0.49	0.63	0.55	1414
Engineer (non-software)	0.4	0.02	0.04	86
Machine Learning Engineer	0.38	0.26	0.31	529
Manager	0.39	0.16	0.23	490
Research Scientist	0.46	0.52	0.49	577
Software Engineer	0.55	0.64	0.59	935
Statistician	0.27	0.07	0.11	129
Teacher / professor	0.31	0.11	0.17	141
accuracy	0.48			5819
macro avg	0.42	0.32	0.32	5819
weighted avg	0.47	0.48	0.45	5819

Text Material 1: Recommendations System Limitations

Selection Bias: Skewed Insights and Recommendations

- The Kaggle user dataset introduces bias by predominantly featuring men and individuals in data science roles.
- Representation imbalances exist across job types, with fewer instances of managerial, office, and non-data science roles, potentially skewing results.

Survey and Data Collection Limitations

- Inherent human errors in respondents' survey submissions may introduce inaccuracies to the dataset.
- Errors by individuals involved in the data collection process can have implications for the overall performance of the model.

Model Accuracy & Methodology

- Limitation in the dependent variable, as the training dataset only includes the user's current job title, omitting their complete work history.
- Occupation determination is influenced by factors beyond skillset, not fully captured by the model.
- Some crucial workplace skill sets are not covered, impacting the model's comprehensiveness.