Exploring Fairness of Multi-class Classifiers in Job Salary Prediction

Anonymous

1 Introduction

The job posting can provide valuable insights associate with salary range. With the huge volume of online job description, decision makers can discover the linkage between the description and corresponding job salary, and further provide a competitive salary. The machine learning techniques (ML) are widely used for that scope due to their prediction capacity by fully utilizing open data. However, these prediction can cause bias and unfairness such as gender unfairness in salary levels which can be learnt by models, bringing up the risk of unfair outcomes for decision makers.

In this research, we will investigate the potential gender unfairness on the task of salary prediction under three major supervised learning classifiers: Logistic Regression, Random Forest and K-Nearest Neighbour(KNN). Besides, we will use zero-R method as the baseline. Instead of using regression methods, we define our prediction task as a multiclass classification problem, since we are more interested in the fairness across different categorical salary range bins instead of absolute quantify number in order to better investigate the fairness. To more specific, our research question is: Are common classifiers can fairly predict salary level based on job description and not affected by gender information?

2 Literature Review

The fairness evaluation has been widely researched and well defined. There are three mainstream fairness metrics include Demographic parity (Calders et al., 2009), Equalized odds and Equal opportunity (Hardt et al., 2016). These metrics involve dividing the data into groups with sensitive members (e.g., a vulnerable group suffered from unfairness) and groups with non-sensitive members. These metrics primary use various classification performance measurements such as true negative rate, false negative rate, and more as a basis for fairness assessment. For example, giving a high positive prediction rate to a group implies that it gives "benefit" to that group and may further result in unfairness to other groups. Most previously work on fairness evaluation are for binary classification problem. Recent works on fairness matrices for multi-class problems are extended from these three major matrices by evaluating them in a group-wise manner (Putzel and Lee, 2022) (Blakeney et al., 2022).

2.1 Demographic Parity

The Demographic Parity holds when the ML prediction is independent with the sensitive group. In other words, a classifier will achieve the demographic parity when it brings the positive prediction to each group equally. For example, the proportion of positive prediction outcome for "getting a job" is at equal rate in each gender group (sensitive group).

2.2 Equalized Odds

The Equalized odds takes the True Positive Rate (TPR) and False Positive Rate (FPR) into accounts and expects that each of the matrices is at the same level across different sensitive. It considers the groups' FPR which measures a group are wrongfully benefited from the model. In other words, the Equalized odds considers the fairness in terms of prediction fairness and instance's fairness associated with qualification or not for benefits.

2.3 Equal opportunity

The Equal opportunity evaluates the true positive rates (TPR) across different sensitive groups. In other words, it measures whether a model or classifier will give an equal chance to correctly predict as positive case across different groups. This metric can be seen as the proportion of people who truly have the qualification for the positive label such as a competitive candidate received an offer.

For Multi-class classification tasks, either fairness metric will hold when the distribution of their assessment outcome is equal across different sensitive groups.

3 The Data Set

We will investigate the dataset derived from (Bhola et al., 2020). The features we use are from the job description, which are transformed as TFIDF with 500 dimension and embedding format with 384 dimensions. The target label we would like to classify is salary bin, which includes binned salary categorical ranged from 0-9. Moreover, based on our defined problem, we only focus on the labeled instance. Besides, we also use the provided demographic labels which contains gender code: 0 (gender-balanced occupation), 1 (female-dominated occupation), 2 (male-dominated occupation) for fairness assessment. Table 1 is a brief summary about these data:

Dataset	No. Instance	Features Dimensions	
		TFIDF	Embedding
Training	8000	8000*500	8000*384
Validation	1737	1737*500	1737*384
Test	1738	1738*500	1738*384

Table 1: The Dataset

As shown in Figure 1, the data set is balanced in terms of salary bins distribution. However, when looking at the distribution dividing by the gender code categories (Figure 2), we can see the frequency distribution is unbalanced with Female-dominated job primary located at the salary bin from 0 to 4, and Male-dominated job which are primary located at the salary bin from 3 to 9.

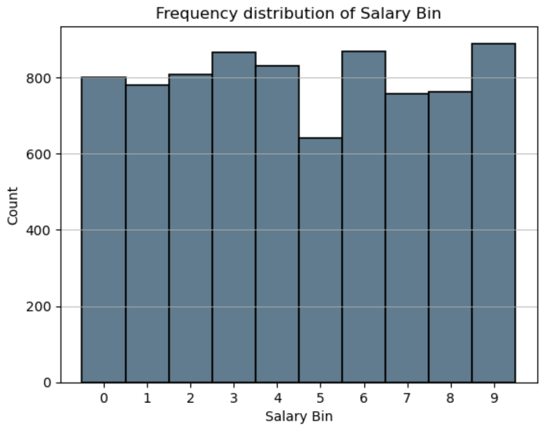


Figure 1: Frequency Distribution

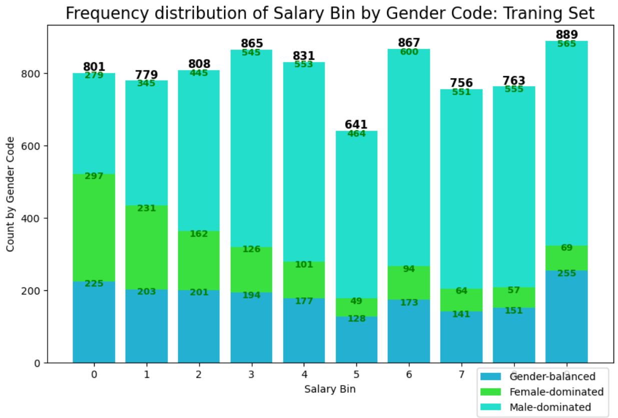


Figure 2: Frequency Distribution by Gender Code

4 Method

4.1 Salary Bin Classification

We aim to investigate the fairness levels in terms of different distinct classifier types. We choose common classifiers to see their ability of performing classification task fairly. We apply one linear classifier: Multinomial Logistic Regression, one non-linear classifier: K-Nearest Neighbors, and one tree-based ensemble classifier Random Forest. We use the Zero-R method as our evaluation baseline. The reason for not using the common method- Naive Bayes, is that it assumes the features are conditional independent each other. However, it apparently not suitable for text features since the texts in a description are relevant. Moreover, we also not adopt the Decision Tree Method because the input features are in high dimensions. A single decision tree may not be able to split the tree effectively. Therefore, we adopt the Random Forests to tackle this issue which consists of lots number of decision trees performing selecting subset features randomly.

We then used 3-fold grid search to optimize the parameters of these models. Table 2 is the summary parameter settings.

Classifier	Parameter Setting	
	Embedding	TF-IDF
Multinomial Logistic Regression	C=0.7	C=0.2
Random Forest	Max depth=20 No. trees=1000	Max depth=20 No. trees=1000
KNN	K=30	K=30

Table 2: Parameter Setting

For model evaluation, we use the F1 score, Recall, and Precision matrices to better know the performance comprehensively. Moreover, since we do not examine the different levels of contribution for certain salary bin classes, we adopt the Macro version for each metric which treat the nine bins equally by averaging the per-class metric. Finally, we will evaluate the classifier performance in terms of the two different types of features: Embedding and TF-IDF.

4.2 Gender Fairness Investigation

For gender fairness assessment, we firstly define our sensitive groups which is the labels column named "gender code". Then, we will use the Demographic Parity for fairness evaluation. We define a fair classifier is able to label salary bins distribution equally across each sensitive group without preferring to certain groups. We adopt the Demographic Parity stated in (Putzel and Lee, 2022) which works for the multiclass classification task fairness. It states that a multiclass classifier will achieve demographic parity if the classification label outcome distribution is equal across sensitive groups.

5 Results and Discussion

5.1 Classification Result

Classifiers	Precision		Recall		F1-Score	
	Embedding	TF-IDF	Embedding	TF-IDF	Embedding	TF-IDF
Logistic Regression	0.217	0.188	0.241	0.218	0.218	0.190
Random Forest	0.265	0.250	0.259	0.255	0.235	0.219
KNN	0.223	0.195	0.229	0.211	0.219	0.196
Zero-R	0.010	0.010	0.100	0.100	0.018	0.018

Table 3: Result Summary

Overall, the Random Forest Method performance outperforms other classifiers in terms of Precision, Recall and F1-Score on the salary bin classification. The reason is that the Random Forest Method is ensembled with several decision trees performing random bootstrapping sampling from the training data set and then adopt the majority voting for final label prediction, which can reduce variance of prediction on unseen data and thus has a better prediction ability generally. For Zero-R performance, since the label bin 9 in the training dataset has the highest proportion, so for the validation dataset it classifies all of the instance as 9 and thus get only 0.01 precision, which is the true frequency proportion of salary bin label as 9 in the validation set.

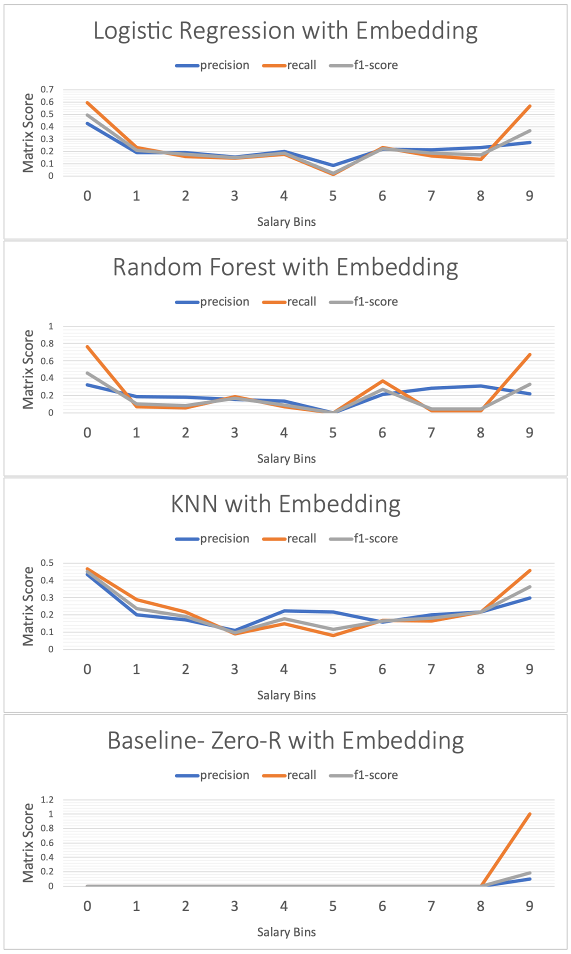


Figure 3: Classifier Performance by Salary Bin

If we further observe these performance matrices across different salary bins for our three Classifiers (see Figure 3), we can see that the predictors always get high performance score in the low-salary bin ranging from 0 to 1 and high-salary bin ranging from 8 to 9, especially for the Recall score. Bin ranging at 3 to 6 get the lower performance score on average. The potential reason that it is hard for classifiers to capture the feature pat-terns associate with salary information for medium-level salary bins. However, for jobs in the low and high level salary bins may contains distinct information or word combination enabling classifiers can better capture the patterns. This adequately reflect to the high Recall score, meaning that some useful feature information associated with low/high salary bin enable classifiers to label instance correctly with low false-negative.

We further process the TF-IDF word feature from "tfidf words" file and select the top 5 TF-IDF score words for each labelled instance. Then we group the processed instance by salary

Salary Bin	Top 10 Frequency Word for Each Bin (Retrieved from top 5 scores of TF-IDF)	Distinct Word Combination
0	customers, duties, customer, assist, sales, stock, perform, orders, food, handle	duties, assist, stock, food
1	sales, customers, customer, clients, day, assist, service, project, duties, prepare	service, day, prepare, duties
2	sales, project, customers, prepare, clients, customer, assist, design, site, marketing	design, site, marketing
3	project, sales, design, customers, site, research, customer, drawings, clients, day	sales, research, drawings, design
4	design, project, sales, experience, clients, customer, business, research, data, software	design, experience, research, data, software
5	design, project, data, business, sales, research, software, experience, support, client	design, experience, research, data, software
6	project, design, business, sales, data, support, software, development, experience, research	software, development, experience, research
7	project, business, experience, design, software, sales, technical, data, security, application	software, security, application, technical
8	business, project, sales, design, development, management, product, data, client, technical	management, technical, design, data
9	business, project, team, product, sales, global, marketing, technical, data, regional	team, global, technical, regional

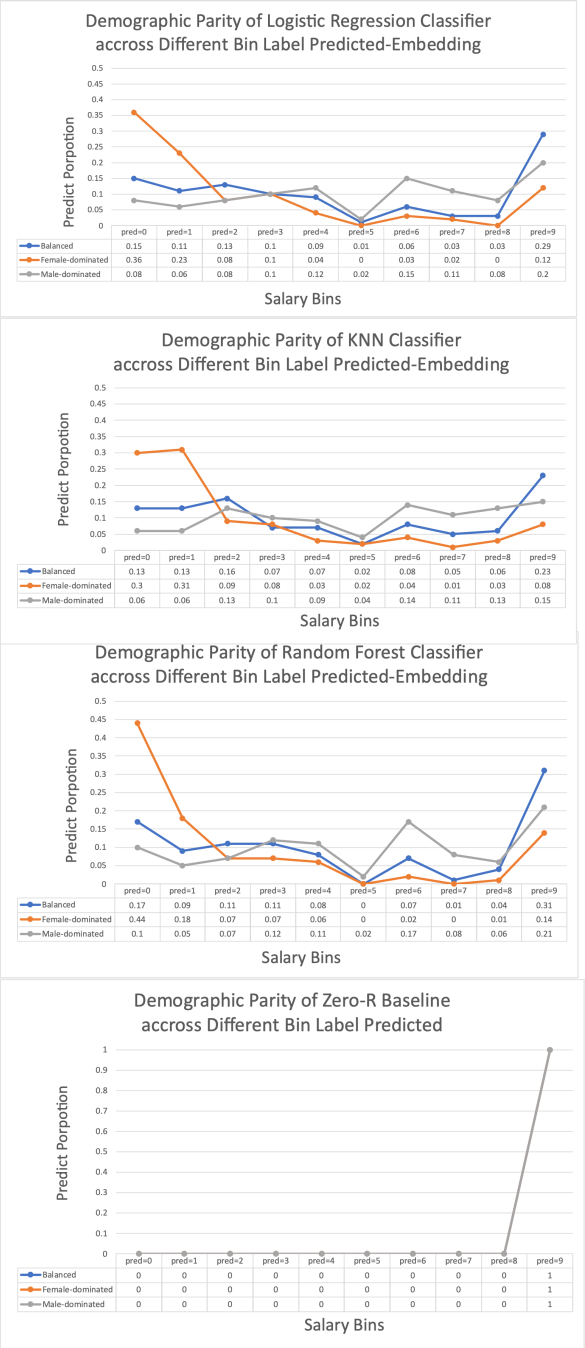
Table 4: Important Word Feature from TF-IDF for Each Bin

bins, choosing top 10 TF-IDF score words for each bin (see Table 4). We then roughly identify the distinct word combination for each bin which can be a representative feature word bags for that bin's corresponding job. For example, for Bin 0, there are distinct representative words are "duties", "assist", "stock", "food". For Bin 9, the distinct representative words are "team", "global", "technical", regional". These distinct representative words can be highly relevant factors to their corresponding salary bin and thus learnt by classifiers, leading high performance for each classifiers as shown in Figure 3. However, when looking the word combination at the Bin 4, Bin 5 and Bin 6, we can see these words are highly overlap such as "design", experience, "research", "data", "software". As a result, it makes classifiers hard to find a pattern for prediction. Only the Random Forest method performs well at Bin 6 with F-1 score around 0.3, achieving by its advantage of bootstrapping mechanism.

Finally, we can see that the outcome with Embedding feature performs well than the TF-IDF feature. The reason is that the TF-IDF does not take the order of job description text into account. It only considers the single word importance. The word Embedding, on the other hand, takes the contexts into account which can capture more information and results a better outcome.

5.2 Fairness Investigation

Figure 4 shows that the assessed gender fairness or Demographic Parity by different classifiers predicted with embedding features. A classifier will be seen as fair when the proportion of positive prediction for a certain bin across three groups are the same (e.g., three lines will be as close as possible in each bin). We can see that the low-salary bin 1 and bin 2 yield the highest unfairness gaps for the three classifiers. This shows that the three classifiers always give positive prediction with instances in "Female-dominated" for these low-salary bins. On the other hand, the medium-salary bin 6 also yields a high unfairness gap, showing that the three classifiers always give positive prediction with instances in "Male-dominated". The high-



	pred=0	pred=1	pred=2	pred=3	pred=4	pred=5	pred=6	pred=7	pred=8	pred=9
Balanced	0.15	0.11	0.13	0.1	0.09	0.01	0.06	0.03	0.03	0.29
Female-dominated	0.36	0.23	0.08	0.1	0.04	0	0.03	0.02	0	0.12
Male-dominated	0.08	0.06	0.08	0.1	0.12	0.02	0.15	0.11	0.08	0.2

	pred=0	pred=1	pred=2	pred=3	pred=4	pred=5	pred=6	pred=7	pred=8	pred=9
Balanced	0.13	0.13	0.16	0.07	0.07	0.02	0.08	0.05	0.06	0.23
Female-dominated	0.3	0.31	0.09	0.08	0.03	0.02	0.04	0.01	0.03	0.08
Male-dominated	0.06	0.06	0.13	0.1	0.09	0.04	0.14	0.11	0.13	0.15

Demographic Parity of Random Forest Classifier
accross Different Bin Label Predicted-Embedding
Predict Porpotion
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Salary Bins

	pred=0	pred=1	pred=2	pred=3	pred=4	pred=5	pred=6	pred=7	pred=8	pred=9
Balanced	0.17	0.09	0.11	0.11	0.08	0	0.07	0.01	0.04	0.31
Female-dominated	0.44	0.18	0.07	0.07	0.06	0	0.02	0	0.01	0.14
Male-dominated	0.1	0.05	0.07	0.12	0.11	0.02	0.17	0.08	0.06	0.21

Demographic Parity of Zero-R Baseline
accross Different Bin Label Predicted
Predict Porpotion
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Salary Bins

	pred=0	pred=1	pred=2	pred=3	pred=4	pred=5	pred=6	pred=7	pred=8	pred=9
Balanced	0	0	0	0	0	0	0	0	0	1
Female-dominated	0	0	0	0	0	0	0	0	0	1
Male-dominated	0	0	0	0	0	0	0	0	0	1

Figure 4: Fairness Investigation

salary bin 8 and bin 9 also yield high unfairness gaps, but the gap lies in ”Gender-balanced” group.

Accordingly, we can see there exists gender-unfairness associated with job salary Bin in 0, 1, and 6 based on our classification task outcomes. If we look at the distribution for each group across salary bin (see Figure 5), we can see that in Bin 0 and Bin 1, the Female-dominated job accounts for the highest proportion around 24% and 19%, respectively. For Bin 6, the Male-dominated job accounts for the highest proportion around 12%. For Bin 9, the Gender-balanced job accounts for the highest proportion around 14%. Based on these trends, we posit the unfairness aroused from our classification is due to the unbalanced of our training data, or historical bias (Mehrabi et al., 2022). This kind of bias reflect the past hiring characteristics and further reflect in the training data. Since the ML models learn patterns from the historical data, what they learnt will bring the relevant outcomes to the unseen data, causing the unfairness, even accentuating the unfairness in an accumulate way by performing ML technique again and again to more unseen data.

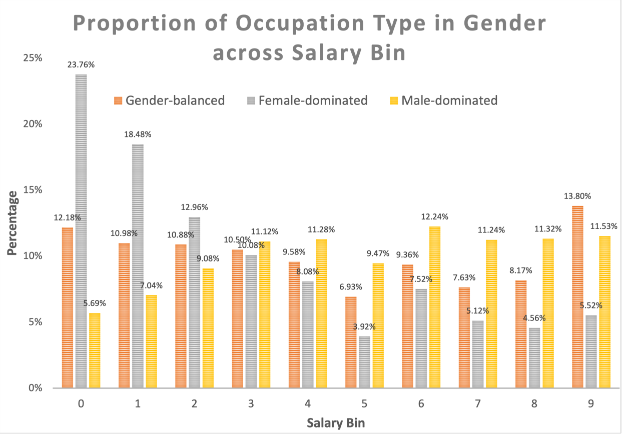


Figure 5: Occupation Distribution

6 Conclusion

In this research, we investigate the potential gender unfairness on the task of salary prediction under three major supervised learning classifiers: Logistic Regression, Random Forest and K-Nearest Neighbour. For the prediction performance, the Random Forest classifier outperforms other classifiers due to its bootstrapping strategy suitable for high-dimension features. For unfairness measurement in Demographic Parity, we found that the low-salary bin 1 and bin 2 yield the highest unfairness gaps for ”Female-dominated” group for all classifiers. The medium-salary bin 6 also yields high unfairness gaps for ”Male-dominated” group. We posit these gaps are aroused by the historical bias existing in the training data with unbalanced distribution of job type by gender. These unbalanced pattern will be learnt by moderns classifiers and thus exacerbates the unfairness gaps for further classification tasks.

References

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org.

Bhola, A., Halder, K., Prasad, A., and Kan, M.-Y. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5832–5842, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Blakeney, C., Atkinson, G., Huish, N., Yan, Y., Metsis, V., and Zong, Z. (2022). Measuring Bias and Fairness in Multiclass Classification. In *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*, pages 1–6.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. ISSN: 2375-9259.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs].

Putzel, P. and Lee, S. (2022). Blackbox Post-Processing for Multiclass Fairness. arXiv:2201.04461 [cs].