ST 3189 Coursework Project

Analysis of Gender Discrimination at Houston College of Medicine

Candidate Number : 210500000

Table of Contents:

Introduction:

Although there has been significant advancement in women's rights at the workplace over the years, women continue to face the challenge of gender discrimination. While it is widely understood that inclusivity and diversity are instrumental in creating a strong workforce, to this day, we see numerous examples of gender discrimination at the workplace. Inequality in salary scales is one of the most evident forms of gender discrimination.

This report aims to explore the issue of gender discrimination through the analysis of the data presented as evidence to support the allegations made by the female doctors at the Houston College of Medicine. They claimed that female doctors were often paid less than their male counterparts, and were less likely to be made full-time professors. This case was argued in the United States District Court of Houston and was filed under Title VII of the Civil Rights Act of 1964, 42 U.S.C. 200e et seq.

The objective of this analysis is to determine whether the disparity in salary and positions (ranks) across the medical faculty at the Houston College of Medicine is rooted in gender discrimination. Thus two specific aspects need to be examined; whether salaries can be predicted based on gender; and whether gender has a bearing on the position or rank obtained.

The accompanying dataset was used to conduct the analysis through simple visualisations, exploratory data analysis and more complex, Supervised and Unsupervised Machine Learning models. The results and conclusions drawn from the data are presented using suitable graphs and tables, and then are compared to existing relevant literature.

Literature Review:

Title VII of the Civil Rights Act of 1964 prohibits employment discrimination based on race, religion, sex and national origin [1]. Since then, the law has been amended multiple times to provide greater protection against discrimination. Despite this there are still countless instances of gender discrimination in the workplace.

Gender discrimination results when there is an unfavourable bias towards a person or persons of a particular gender. While gender discrimination is an issue that may affect both men and women, reports, surveys and articles from around the world indicate that women are far more likely to be victims of discrimination at the workplace. Studies show that women are four times more likely than men to be treated as if they were incompetent due to their gender and about five times as likely to feel as though they have been paid less than their male counterparts for doing the same job [2].

Gender discrimination is not always easily apparent. It can manifest itself in various forms beyond instances of unequal pay or harassment. In fact, discriminatory practices could be present in hiring of employees based on their gender, or even granting promotions based on gender. Another form of discrimination is when female employees are held to stricter standards

or are evaluated more harshly. Oftentimes, these forms of gender discrimination are not explicit or even intentional. This can potentially be attributed to unconscious biases that women may face at the workplace: for instance, Performance Bias - where men are judged on their potential whereas women are judged on their past performance; or Maternity Bias - where women are thought to be inferior employees due the possibility of becoming mothers.

Workplace gender discrimination can have far - reaching consequences not only on the individual but on the entire organisation. Gender discrimination can cultivate a toxic work environment with increased workplace conflict. It can lead to fear or distrust in 'management' due to their discriminatory actions or lack of action against discrimination. This, in turn, can lead to lower employee morale and productivity [3].

The first step towards preventing gender discrimination is identifying/ acknowledging the problem and raising awareness. In the past few decades, there has been growing awareness around issues such as the pay gap, sexual harassment and lack of representation of women in leadership positions. While significant progress has been made, there is still a long way to go in the journey towards true gender equality – at the current rate, it will take 132 years to achieve gender parity, as stated in the Global Gender Gap Report [4].

Data:

The dataset consists of information on 261 faculty members of which 155 were male (59.3%) and 106 were female (40.7%). The dataset includes 10 variable columns - ID, Dept (Department of faculty member : Biochemistry/ Molecular Biology, Physiology, Genetics, Paediatrics, Medicine or Surgery), Gender - (Male or Female), Clin1 (Primary clinical emphasis), Cert1 (Board Certified), Prate (Publication Rate), Exp (Experience since earning MD), Rank (Assistant professor, Associate professor or full professor), Sal94 (Salary in 1994) and Sal95 (Salary after increment in 1994). The data was clean with no null or missing values present.

Since the primary focus of the report is assessing the impact gender has on the prediction of rank and salary, the 'ID' column was dropped as it has no influence on the predictions of the model. In addition to this, the predictions for salary were made on the 'Sal94' column and since the correlation between the 'Sal94' and 'Sal95' columns were extremely high, it was assumed that predictions on one of the salary variables would be nearly identical to the other.

Methodology:

As explained in the introduction, the objective of the report is to determine the impact of Gender on salary and rank. The data was first explored with simple plots and tables to observe the distribution of gender. Following this, a correlation heatmap was created to assess the correlation of the features with one another. To better understand the distributions of variables with each other, a scatterplot matrix was plotted and inferences and conclusions were drawn.

For the unsupervised learning section, K Means Clustering was chosen in order to create clusters of data with similar characteristics. Since the report's focus is on the influence of gender, 2 clusters were chosen and their respective gender ratios were observed. The mean salaries of both clusters were calculated and a chi-square test for significance was performed to show that there is a statistical difference in the proportion of gender in both clusters. The significance test validates the hypothesis that there is a significant difference in the proportion of gender between the two clusters. However, while this tells us that gender is a significant differentiator, it is not the only factor that is impacting the difference in salary between the clusters.
In order to conduct a more conclusive analysis on whether gender is an accurate predictor of salary and rank of faculty members, regression and classification models were created to understand the importance of the various features in determining both salary and rank. However, before creating the models, the categorical variables such as department and rank had to be converted to dummy variables which not only allows us to use categorical data in the models but also to better understand the relationship of each of the classes of the categorical variables on the target variable. In addition to this the continuous variables were scaled - a process where numerical data is standardised to have a mean of zero and a standard deviation of 1.

The models were used to determine feature importance. Therefore, train test split was not performed on the models as the primary focus was determining which features were the best predictors of the target variables. In addition to this, due to the small sample size, splitting the data in training and testing data can reduce model accuracy. The issue being highlighted by the lawsuit is gender discrimination with respect to salary and rank. The regression and classification models attempt to show the features on the prediction of salary and rank. Thus by analysing the importance of gender and the other features on salary and rank, we can attempt to draw some conclusions on if there is evidence of gender discrimination.

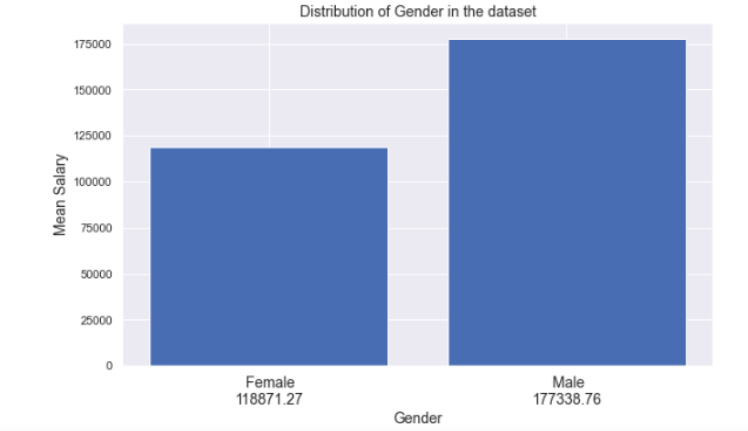
Now that the data is prepared, 2 regression models - Linear regression, Random forest regressor. - were used to analyse the impact of gender on salary. These models were chosen as they are of varying complexity and due to the fact that they have different approaches to determining feature importance. Linear regression is the simplest model and is therefore easy to understand. The importance of features is determined by interpreting the p values of the features. Random Forest is significantly more complex and can be difficult to understand but was used due to its ability to determine feature importance through mean decrease accuracy.

The classification models were used to answer the second question - Is gender a significant predictor of rank. The three models used were multinomial logistic regression, Random Forest Classifier and Decision Tree Classifier (CART). Three models were used as the models showed varying results. Logistic regression is similar to linear regression in that it is simple and easy to interpret. In this case a multinomial logistic regression model was chosen as the target variable - Rank has three classes. CART is a binary tree model which is intuitive and its results are easy to interpret. Random Forest is also a decision tree model.

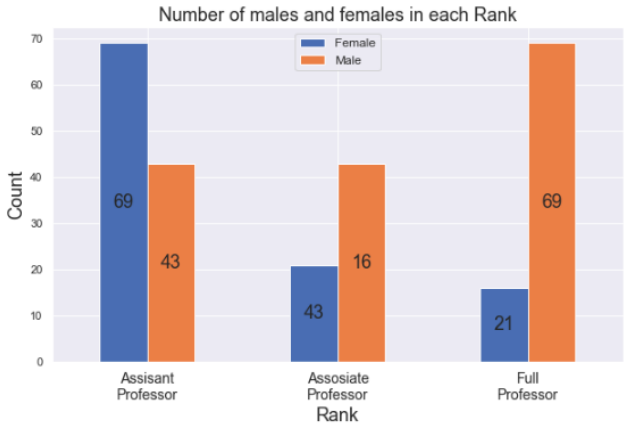
5. Data Analysis and Model Comparison:

5.1 Exploratory Data Analysis:

As mentioned earlier the data consisted of 106 (40.7%) females and 155 males (59.3%). First, the mean salary of males and females was calculated and visualised:



Following this the distribution of gender in each rank was visualised:



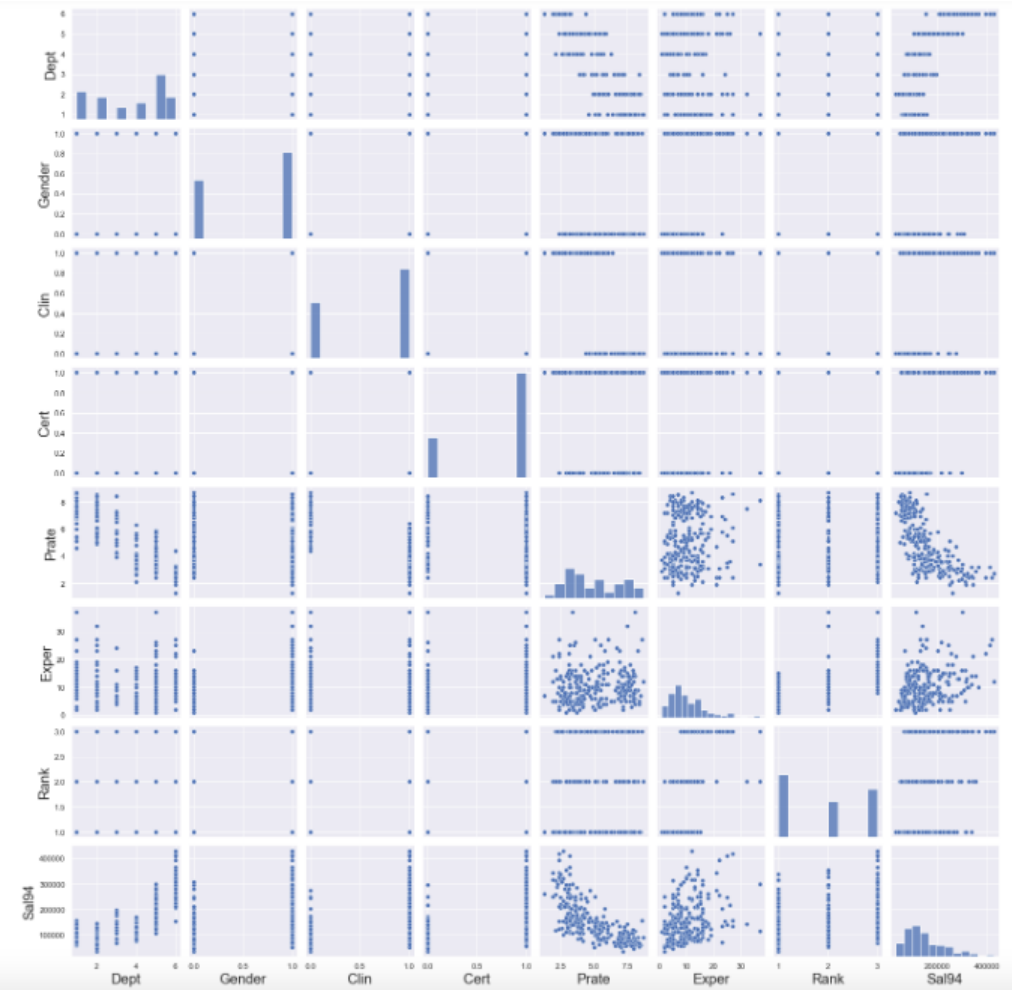
The above graphs support the claims made by the plaintiffs. The first graph clearly shows that female faculty members make considerably less than male faculty members and the second graph reveals that the ratio of female to male associate and full professors are drastically lower as compared to assistant professors. However, it is important to note that there are many factors in the dataset that could potentially affect salary and rank. Therefore more robust methods of analysis were used that take into account all relevant features of the dataset to try and determine if gender is a significant differentiator.

Before using machine learning models for analysis a scatterplot matrix was created to reveal the relationships between pairs of variables in the dataset from which the following inferences were made :

1. Certain departments (Surgeon, Medicine) have significantly higher salaries.
2. There is a visible difference in the highest earners for men and women, with the male top earners earning much higher salaries than women.
3. Employees with Primary Clinical Emphasis have high top salaries.
4. Employees with Board Certification have high top salaries.
5. It appears that publication rate has a negative correlation with salary. The faculty members with the lowest publication rates have the highest salaries

6. It appears that experience does not seem to be a determining factor in predicting salary.

Scatterplot Matrix:



5.2 Unsupervised Learning:

For the unsupervised learning section, K mean clustering was used to group the dataset into clusters of similar features where K is the user defined number of clusters. In this case, 2

clusters were chosen. After K mean clustering was run, the data was split into two dataframes based on the assigned clusters. The proportions of gender in both clusters were analysed and the results showed that the first cluster was 52% males to 48% females. The second cluster was made up of 66% males and 34% females.

The clusters show a significant difference in the proportion of gender.. Since the clusters showed such a difference in their proportions of gender, they were then analysed to see if this difference in gender proportion would reflect in the average salary of each cluster.

```
Salary Statistics for Cluster 1:
count       118.000000
mean      93347.059322
std       28776.478417
min       34514.000000
25%       70194.000000
50%       89955.500000
75%      110702.500000
max      172793.000000
Name: Salary94, dtype: float64


Salary Statistics for Cluster 2:
count       143.000000
mean     203307.062937
std       75292.715683
min       77087.000000
25%      146804.500000
50%      185413.000000
75%      248820.000000
max      428876.000000
Name: Salary94, dtype: float64
```

The two clusters did show significant differences in the mean salary. However in order to investigate if the differences between the gender proportions of the clusters was statistically different, a Chi - Square test was performed. The test yielded a p-value of 0.029 which means that the null hypothesis (there is no difference between the clusters) can be rejected and that there is evidence to suggest that the proportions of gender in the two clusters is statistically different.

From the results of clustering and the Chi - square test we can conclude that there is a difference in the proportions of gender and that the cluster with a notably higher proportion of male faculty members, has much higher average salary. This would suggest that gender is an important differentiator of salary. However, whale clustering splits the data into similar clusters, it does not help us understand the relationship each of the features has on salary or rank. Moreover, salary is dependent on numerous factors, hence, further analysis is required in order to conclusively say that there was gender based salary discrimination at the college.

Supervised Learning :

In order to answer the two questions three regression and three classification models were used for their varying complexity and to allow us to arrive at a more conclusive answer to the questions.

Q1 : Is gender a significant differentiator of salary?

Regression Models:
The three regression models used to answer the question were Linear regression and Random forest regressor. The main outcome of interest is the salary of the faculty members. The models will include the other features in the dataset to predict salary. Analysing the results of the models will tell us which features are the most important predictors of salary and therefore try to evaluate if there is gender discrimination in salary.

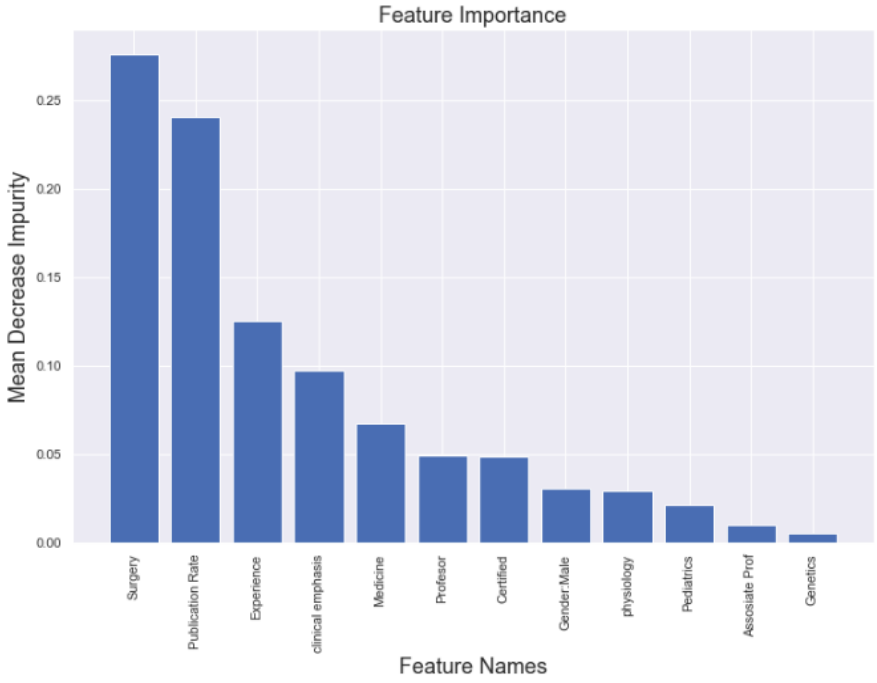
Linear regression:
A linear regression model was made with 'sal94' as the predictor. The R^2 value is 0.904 which indicates that the model is a good fit and explains 90.4% of variability in the data. The model shows that gender and Department-2 (Physiology) showed p-values of 0.185 and 0.381 respectively. This means that at a 5% significance level, Gender is not a significant predictor of salary. The model shows that the other features such as rank, publication rate etc. show are significant predictors of salary with p-values lower than 0.0001.

Model Summary

	coef	std err	t	P>|t|	[0.025	0.975]
Prate	2.867e+04	3299.350	8.691	0.000	2.22e+04	3.52e+04
Exper	1.528e+04	2246.929	6.799	0.000	1.09e+04	1.97e+04
Dept_2	-5127.6384	5843.895	-0.877	0.381	-1.66e+04	6382.129
Dept_3	4.435e+04	6982.528	6.351	0.000	3.06e+04	5.81e+04
Dept_4	6.761e+04	7262.571	9.310	0.000	5.33e+04	8.19e+04
Dept_5	1.144e+05	5843.769	19.573	0.000	1.03e+05	1.26e+05
Dept_6	2.264e+05	7864.120	28.788	0.000	2.11e+05	2.42e+05
Gender_1	5193.8141	3907.486	1.329	0.185	-2502.124	1.29e+04
Clin_1	5.277e+04	5503.632	9.588	0.000	4.19e+04	6.36e+04
Cert_1	2.331e+04	4304.568	5.415	0.000	1.48e+04	3.18e+04
Rank_2	2.427e+04	4679.520	5.187	0.000	1.51e+04	3.35e+04
Rank_3	4.191e+04	5211.381	8.043	0.000	3.17e+04	5.22e+04

Random Forest Regressor:
Similar to the previous model, the Random Forest Regressor was fitted with the features to predict Sal94 with the number of decision trees = 500 and a maximum feature selection of 3. The model has an R^2 value of 0.986. Which shows that the model is a very good fit. The importance of the features was determined using the mean decrease impurity which in random forest is calculated by seeing how much the mean squared error (MSE) changes when permutations of features are changed. Higher increase in MSE indicates that the feature is more important.
The graph clearly shows that Experience and Publication rate are amongst the best predictors of salary. Departments such as Surgery and Medicine are also important predictors of salary. Gender however, appears to not be very significant in predicting salary.

Graph showing mean decrease impurity:



Q2 : Is gender a significant predictor of Rank?

Classification Models:
The two models used to determine feature importance were Logistic regression, Random forest classifier and CART. The models used to predict Rank and analyse the importance to the features in predicting rank if a faculty member.
The models do not use salary in their predictions as Rank will show clear differences in salary.

Logistic Regression:
The model was fitted with the features in the dataset except ‘Sal94’ and was used to predict ‘Rank’. The significant predictors or rank could be determined in multiple ways. First, the coefficients and p-values were observed. At a 5% significance level, features with p-values lower than 0.05 are considered to be statistically significant. Gender and Experience are both found to be statistically significant. However, to check if these features are significant in predicting Rank, the odds ratio confidence interval for all the features was computed. If 1 does not lie in the odds ratio confidence interval for a feature, it is considered to be a significant predictor. The table shows Gender and Experience to be significant predictors of Rank.

Model Summary:

Rank=2	coef	std err	z	P>|z|	[0.025	0.975]
Prate	0.0047	0.352	0.013	0.989	-0.684	0.694
Exper	2.8202	0.434	6.502	0.000	1.970	3.670
Dept_2	1.1822	0.707	1.673	0.094	-0.203	2.567
Dept_3	0.4111	0.752	0.547	0.584	-1.062	1.884
Dept_4	1.4578	0.853	1.709	0.088	-0.215	3.130
Dept_5	0.4533	0.709	0.639	0.523	-0.937	1.843
Dept_6	0.0438	0.934	0.047	0.963	-1.786	1.873
Gender_1	1.2512	0.419	2.986	0.003	0.430	2.072
Clin_1	-0.1452	0.647	-0.225	0.822	-1.413	1.122
Cert_1	-1.0850	0.528	-2.056	0.040	-2.119	-0.051

Rank=3	coef	std err	z	P>|z|	[0.025	0.975]
Prate	-0.4157	0.387	-1.075	0.283	-1.174	0.342
Exper	3.6582	0.460	7.945	0.000	2.756	4.561
Dept_2	1.2828	0.734	1.747	0.081	-0.156	2.722
Dept_3	0.3061	0.817	0.375	0.708	-1.295	1.907
Dept_4	-0.4060	0.993	-0.409	0.683	-2.353	1.541
Dept_5	-0.5543	0.761	-0.728	0.466	-2.046	0.937
Dept_6	-1.4846	1.015	-1.463	0.144	-3.474	0.505
Gender_1	1.6374	0.460	3.559	0.000	0.736	2.539
Clin_1	-0.3997	0.688	-0.581	0.561	-1.747	0.948
Cert_1	-0.3607	0.563	-0.641	0.522	-1.464	0.742

Odds Ratio Confidence Interval:

Rank		Upper	Lower
2	Prate	0.504394	2.001127
	Exper	7.171400	39.260340
	Dept_2	0.816406	13.029997
	Dept_3	0.345672	6.582874
	Dept_4	0.806929	22.878933
	Dept_5	0.391849	6.318240
	Dept_6	0.167656	6.510742
	Gender_1	1.537162	7.944378
	Clin_1	0.243472	3.071877
	Cert_1	0.120128	0.950414
3	Prate	0.309185	1.408462
	Exper	15.733890	95.644343
	Dept_2	0.855233	15.209522
	Dept_3	0.273893	6.734414
	Dept_4	0.095112	4.668251
	Dept_5	0.129243	2.553566
	Dept_6	0.030987	1.656883
	Gender_1	2.087065	12.667740
	Clin_1	0.174229	2.580358
	Cert_1	0.231419	2.100565

CART and Random Forest Classifier:

In order to support the finding from the logistic regression model two additional models were used. CART and Random forest are both decision tree models which determine feature importance by calculating the effect each feature has on the Gini Impurity - a measure to determine how the nodes should be split.

Feature Importance using CART :

Feature Name	Experience	Publication Rate	Certified	Medicine	Physiology	Pediatrics	Gender:Male	Clinical Emphasis	Surgery	Genetics
Score	0.605485	0.226206	0.047007	0.030767	0.02386	0.019076	0.016513	0.013112	0.011571	0.006404

Feature Importance using Random Forest :

Feature Name	Experience	Publication Rate	Gender:Male	Certified	Clinical Emphasis	Medicine	Physiology	Pediatrics	Surgery	Genetics
Score	0.530279	0.245879	0.071055	0.03988	0.026427	0.020504	0.020367	0.019747	0.013578	0.012284

Conclusion:

The different supervised and unsupervised learning models were used to arrive at more concrete conclusions to the two questions.

As explained earlier, in order to explore gender discrimination in salary, the impact of gender on salary had to be studied. K Means Clustering showed that the two clusters had statistically different proportions of gender. Therefore when the average salary for each of the clusters was found to be different, it showed that gender is a differentiator of salary. To perform a more thorough analysis of gender's influence on salary, regression models were used. Linear regression Reveals that gender is not significant in predicting salary. The second model, Random Forest, again showed that gender is insignificant in predicting salary. Both models however show that features such as Experience, Publication rate and department play a much bigger role in determining salary of a faculty member. While clustering showed that gender is a differentiator of salary, when the effect of all the variables was taken into account, the regression model showed that gender is not a significant predictor of salary.
Therefore there is insufficient evidence to suggest that the difference in salary between male and female faculty members is a result of gender discrimination. The differences could instead be explained by factors such as Experience or Publication rate. Research suggests that while a gender pay gap exists, most factors contributing to the difference in pay can be explained by measurable factors that include education, qualification and work experience [4]. Pew research centre suggests that gender pay gap can also be attributed to factors such as role. In the Houston College of Medicine, more women are present in lower paying ranks. The difference could even be explained by Rank, which was shown to be more significant that gender.
This serves as a segue to the next claim - is the disparity in rank a result of gender discrimination. The classification models revealed mixed results. The multinomial logistic regression shows that gender and experience are the two variables with the highest impact on rank. But Experience is a more significant predictor of rank. The CART model shows that gender is not a significant predictor on rank whereas experience and publication rate are. Lastly the Random Forest model shows gender to be one of the significant predictors. The model's prediction of the importance of gender varies and hence we can not conclusively say that the disparity in rank is due to gender discrimination. However two or the three models show that gender is one of the significant predictors. Thereby suggesting that while gender may be a deciding factor in Rank, Experience and publication rate play a much larger role in determining the Rank of a faculty member. While the results suggest that there may be evidence of gender discrimination in allocation of Rank, the results also could point towards discriminatory practices in hiring. Numerous reports support these finding - That while possible explanations in the pay gap may be attributed to measurable factors, the gender biases, especially in hiring can further contribute to inequality in pay.[5]

References:

Dataset : https://www.kaggle.com/datasets/hjmjerry/gender-discrimination

1 - https://bohmlaw.com/areas/gender-discrimination/
2-https://www.ellevatenetwork.com/articles/8775-gender-discrimination-in-the-workplace-an-in-depth-look
3- https://smallbusiness.chron.com/effects-gender-discrimination-workplace-2860.html
4-https://www.pewresearch.org/fact-tank/2023/03/01/gender-pay-gap-facts/
5-
https://www.epi.org/publication/womens-work-and-the-gender-pay-gap-how-discrimination-societal-norms-and-other-forces-affect-womens-occupational-choices-and-their-pay/