Tech Salaries and Job Title as Accurate Salary Predictors

Harrison Barnes, Jabari Thomas, Medha Modini, Bennett Shearin, Dylan Wilson

School of Data Science, University of North Carolina – Charlotte

DTSC 1302

Dr. Scipioni & Dr. Ageenko

08 December 2024

Introduction

In this paper, we discuss the findings that our model concluded regarding how various factors affect salary differences among employees within the tech industry. Based on the dataset given to us, we had a multitude of variables that would affect the total yearly compensation. These variables include location, race, gender, years of experience, job level, company, job title, tenure, highest level of education, and the presence of bonuses. We narrowed our focus to two primary factors: years of experience and job title. We selected these variables due to their relationship with career progression and the natural expectation of higher salaries with more experience. With all these factors in mind, the research question we formulated is “Are years of experience and job title accurate predictors of salary?”. With this question, we aimed to determine if years of experience and one’s job title are statistically significant predictors of an individual's salary. By creating a program to predict the variables that impact salary statistically, we were able to perform a comprehensive analysis. Our model incorporated regression techniques to quantify the relationship between the independent variables (years of experience and job title) and the dependent variable (salary).

Context and Implications

In recent years, a growing number of companies have been administering Payment Benchmark Algorithms (PBA). As stated by Tim Decker and Ben Hanowell from ADPResearch, these specialized large language models (or LLMs for short) utilize transformers in order to predict pay based on the context in which people work (Hanowell, et al., 2023). While this contextual framework is crucial and should be prioritized, years of experience and job title remain important supplementary factors in a comprehensive approach to wage determination and organizational compensation strategies. Based on the dataset provided, our research holds

significant implications for both employers and employees**.** For employers, this study offers insight into enabling them to make fair and unbiased decisions when setting salaries for new hires and providing raises to existing employees. This data can contribute to fostering a non-toxic, happy work environment. The findings of this research will help employees assess whether their current salary is fair, considering their years of experience and job title. It also aids in evaluating if a potential employer's offer is reasonable.

Measurement

The measurement aimed to examine what variables contribute to the highest pay among employees in the technology industry. To answer the research question “Are years of experience and job title significant predictors of salary?” we have to find the correlations between all of the variables chosen. Each variable of an employee's years of experience, job title, and total yearly compensation plays a role in determining what leads to the most compensated employees. We coded a correlation heatmap, a pair plot, and a histogram to see how the three variables are related.

Our heat map revealed that years of experience were more related than the job title. The years of experience variable was 42% correlated and the job title was only 25% correlated with yearly compensation. The statistical models helped narrow down the data to get a pin down on what it all means. Years of experience is the total time employees have been working within the technology field. The employees' job title is their designated role within the company or organization. Total yearly compensation is the amount of money that each employee makes in the year of work. This includes any bonuses, stock investments, and salary.

Data

When first examining our dataset, it was evident that there were many instances of outliers and missing data that interfered with the interpretation of our dataset. In order to allow ourselves to move along with analyzing our data and answering our question of whether an employee's experience and job title were statistically significant in determining their yearly compensation, we opted to remove the rows with missing data. Given that the provided dataset contained over sixty thousand employees, removing these rows with missing values still left a sufficient number to analyze.

Removing the rows with missing values allowed us to improve the quality and reliability of the dataset that we would analyze, leading to a more confident conclusion regarding the correlation between these factors on an employee's salary. While we could have opted to use more advanced methods of imputation to fill in the missing values, the quantity of data provided didn't limit our analysis. Approaching our dataset in this way allowed us to find and determine the potentially meaningful correlation between an employee's experience, job title, and salary.

After having manipulated the data through deletion, we created multiple data visualizations in order to showcase the relationship between our variables. Of these, there were three primary visualizations that best did so when used together: a correlation heatmap, a pair plot of variables, and a histogram.

Correlation Heatmap

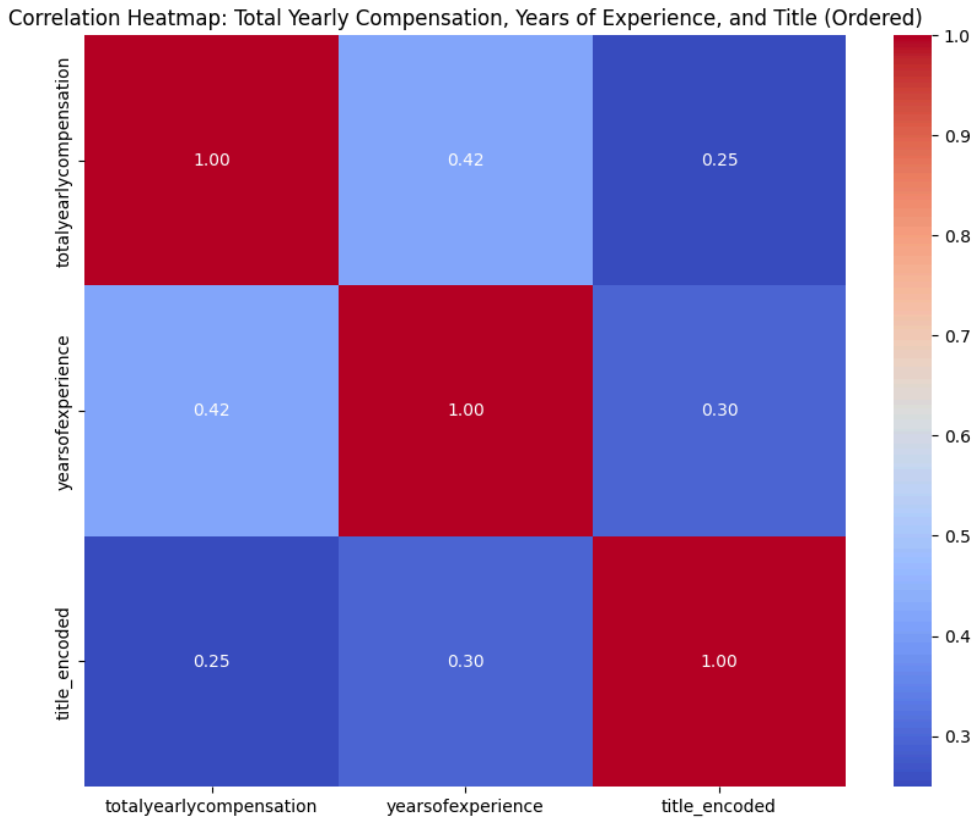

Figure 1.

A correlation heatmap presents the correlation between each set of variables. In our case, that is an employee's total yearly compensation, years of experience, and job title, which has been converted into numerical values for calculation. The closer the number is to 1.00, and, visually, the more red its box is, the stronger the correlation between the two variables. As a result of all correlations we analyzed being positive, we opted to show these numbers on a scale from 0 to 1, rather than -1 to 1. Along the diagonal, every box has a correlation of 1.00 because it is being compared to itself.

Based on the correlation heatmap shown in Figure 1, we discovered that in general, an employee's job title held no significant correlation to either their yearly compensation or experience, with scores of just 0.25 and 0.30, respectively. This low correlation communicates that while one's title might play a role in determining one's yearly compensation or experience, it

is not a primary factor. Shown later, other visualizations will help break down the relationship based on each job title. One's annual compensation and years of experience, however, showed a correlation of 0.42. While also not a strong correlation, there is a moderate positive correlation between the variables. As an employee stays in that field of work for a longer period of time and gains experience, their yearly compensation will begin to increase, however, it is very likely that there are other factors of strong influence, similar to one's job title.

Pairplot of Variables

The second primary visualization we used to help analyze the relationship between our variables was a pair plot. A pair plot shows the relationship between 2 variables as the independent variable changes. In Figure 1, the correlation heatmap only displayed the correlation coefficient between the two variables. In Figure 2, however, the pair plot of the same variables displays the patterns, trends, and outliers.

Through Figure 2, it is evident that, as the correlation heatmap communicated, there is no significant relationship between any of the variables. As one's title improves, their yearly compensation increases slightly (shown in the top right grid), but only for the last two titles. These two titles are product manager (left) and software engineering manager (right). Outside of these two job titles, along with software engineering (listed 6th from the left), the remaining jobs are compensated very similarly. This means that a job title can be an important communicator of one's yearly compensation for a few specific titles, but in general, doesn't play a significant role.

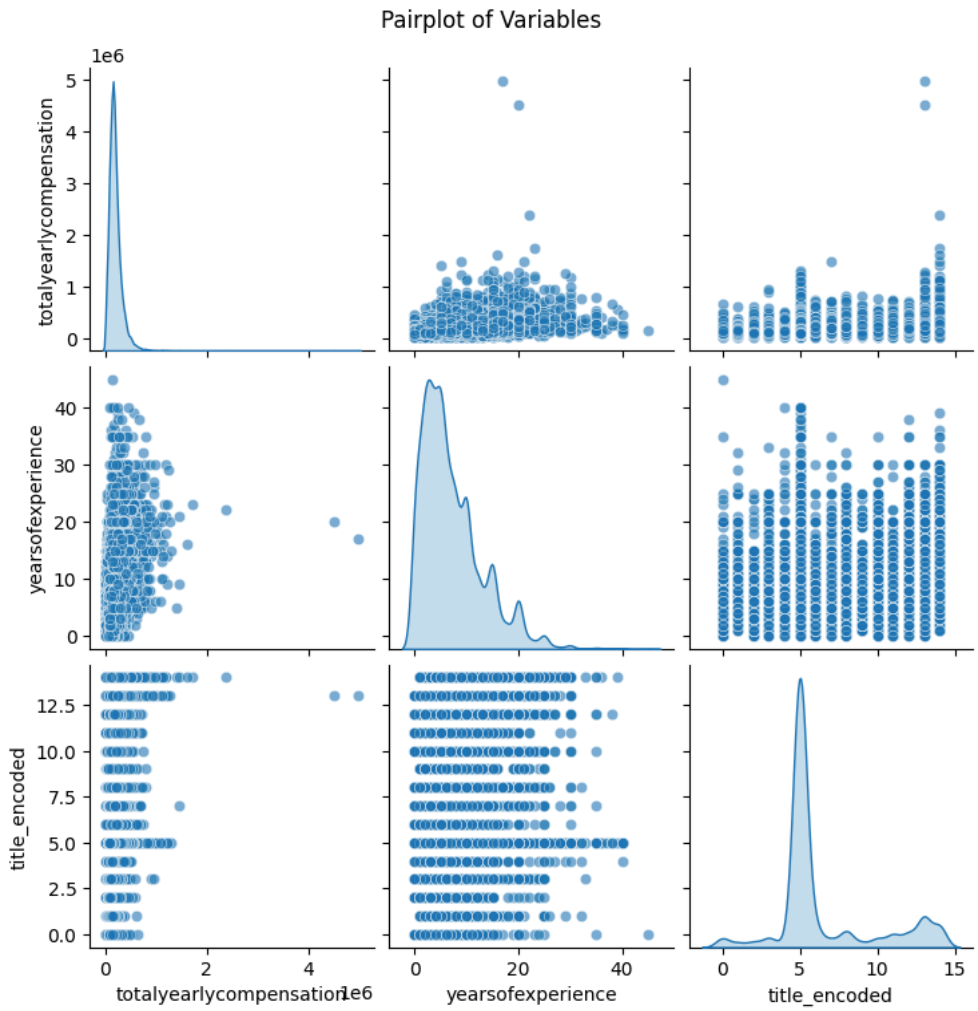

Figure 2.

In the grid showing the trends between experience and yearly compensation, it is evident that there is a moderate, positive correlation, but it is difficult to truly see the extent of the correlation. To fix this, we created another pair plot in which the total yearly compensation was the log of itself, as seen in Figure 3 below. Doing so helped compress the effect of the outliers and better show the correlation.

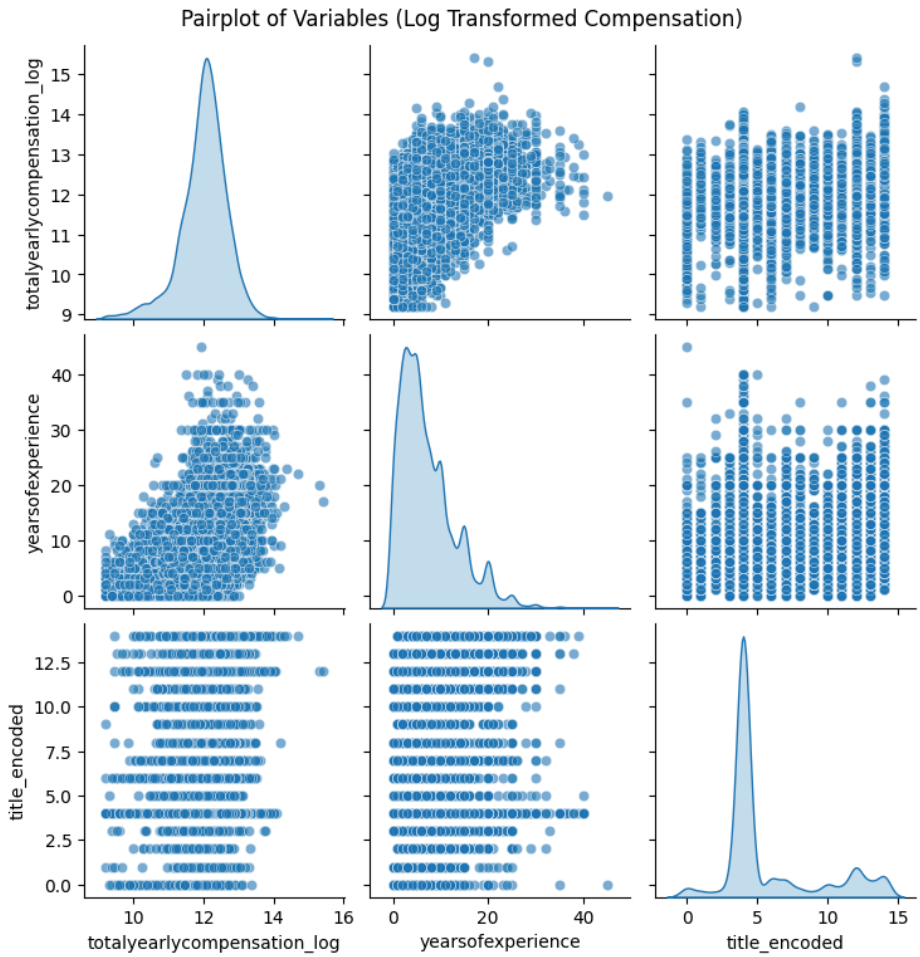

Figure 3.

As a result of using the log of each employee's total yearly compensation, the correlation between the compensation and years of experience is more evident. As one becomes more experienced in their field, their compensation begins to increase to a certain extent. Eventually, however, it seems that the compensation levels off and no longer consistently improves through more experience. Like in the correlation heatmap shown in Figure 1, the pair plots of Figure 2 and Figure 3 display that there is a slight correlation between experience and job title to compensation. The pairplots, however, provided a deeper insight into the trends within these correlations.

Histogram

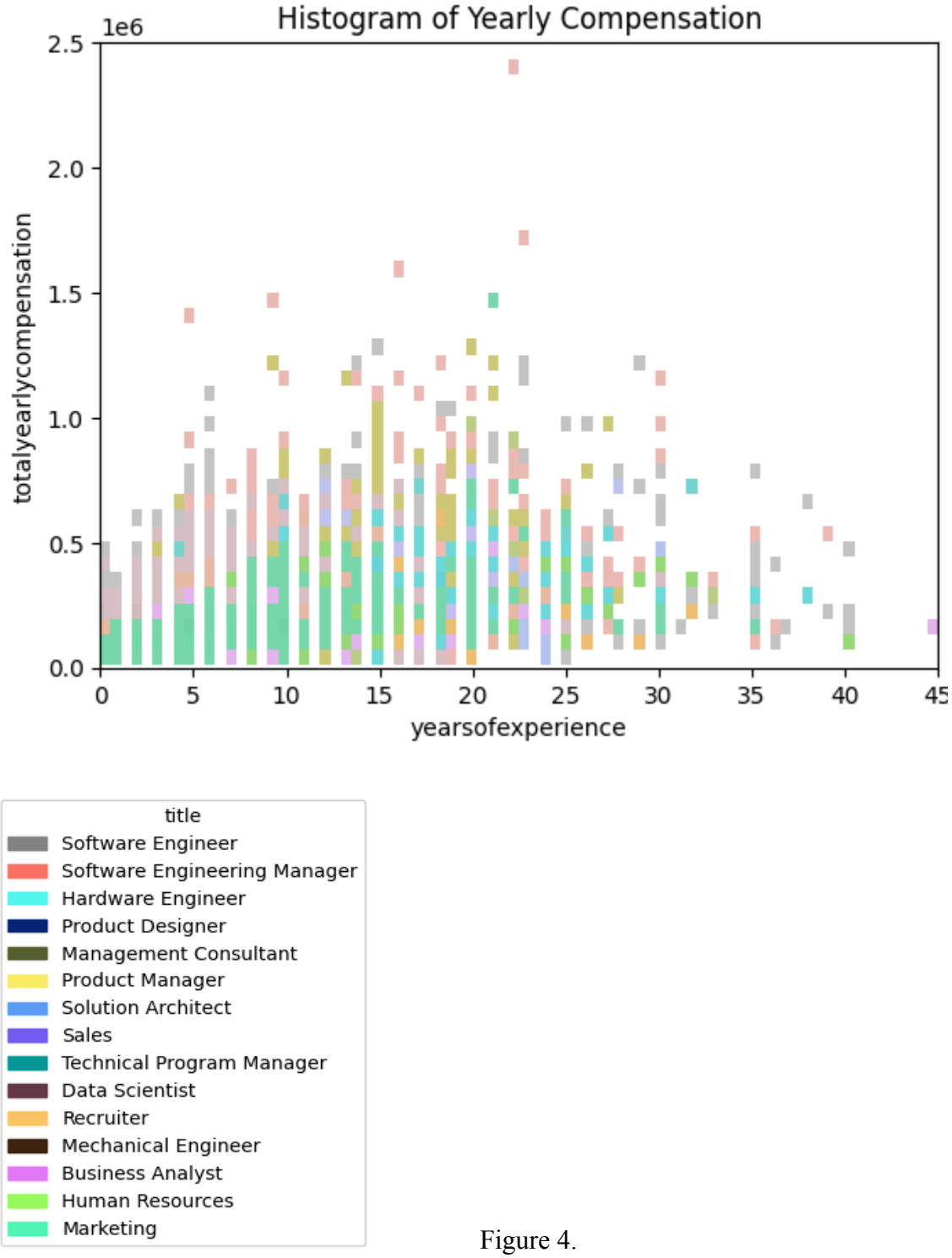

Figure 4.

In an effort to further supplement the heatmap and pair plots, we created a histogram displaying the relationship between compensation and experience, with coloring based on job

title. This histogram, shown above in Figure 4, is the same visualization in the upper middle grid of Figure 2, except colorized. In order to better view the changes in color throughout the histogram, we also created bins so that there aren't thousands of squares displayed.

Through Figure 4, it is clear that certain colors, and therefore job titles, consistently have a higher total yearly compensation than others. For instance, a Software Engineering Manager is often alone at the top of the histogram, meaning that if someone is a Software Engineering Manager, then they should be paid a fairly substantial amount more than their colleagues. Product managers also seem to be compensated more than the majority of their colleagues. This supports the information shared by the pair plots. Like in the pair plots, the remaining titles are primarily clustered near the bottom of the histogram. If an employee holds one of those job titles, then it will not be a great indicator of their pay, so someone looking to predict their pay would have to look at other variables.

Statistics Summary

OLS Regression Results

Dep. Variable:	totalyearlycompensation_log	R-squared:	0.162
Model:	OLS	Adj. R-squared:	0.162
Method:	Least Squares	F-statistic:	2084.
Date:	Sat, 07 Dec 2024	Prob (F-statistic):	0.00
Time:	17:25:27	Log-Likelihood:	-19598.
No. Observations:	21515	AIC:	3.920e+04
Df Residuals:	21512	BIC:	3.923e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.5661	0.009	1218.756	0.000	11.547	11.585
yearsofexperience	0.0391	0.001	53.089	0.000	0.038	0.041
title_encoded	0.0247	0.001	18.559	0.000	0.022	0.027

Omnibus:	3241.389	Durbin-Watson:	1.845
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6402.085
Skew:	-0.933	Prob(JB):	0.00
Kurtosis:	4.912	Cond. No.	25.8

Figure 5.

In order to best communicate the dataset and correlations between experience, job title, and compensation, we opted to display the statistics summary as a log relationship. Through a

linear regression model, the data doesn't take into account how changes in values are impacted as experience grows or one's job title gets near the top of the company. In the logarithmic regression model, shown in Figure 5, the primary statistics to focus on are r-squared, coefficients, and log-likelihood.

The r-squared is low, at just 0.162. While this isn't necessarily bad, it does mean that only roughly 16% of the variability of the dataset in determining compensation can be explained by years of experience and job title. The other 84% is likely to be found in the original dataset, such as level, gender, company, and education. Through the coefficients, how the compensation changes based on changes in experience and job title is shown. With 0 years of experience and an entry-level job, one's salary in the tech industry is roughly the exponential function – since this is a log regression summary – of 11.5661, or $105,000. With every additional year of experience, the employee's compensation increases by nearly 4%, and with every additional increase in job level, an increase of roughly 2.5%. These are found by taking the exponential function of the coefficient, as well. Finally, the log-likelihood displays how well this model fits the data. The greater the number, the better it fits. While -19598 may appear like a very poor number, because of the number of observations, that is not necessarily the case. In comparison to many other summaries created, -19598, while far from perfect, was a much better fit.

Conclusion

Our analysis ultimately shows how years of experience and job titles influence salary differences in the tech industry. We found that years of experience had a moderate correlation with yearly compensation of 42% while job titles showed only a weak correlation of 25%. This suggests that while both factors play a role in determining salaries, years of experience carry more weight than one's job title. Through using tools like correlation heatmaps, pair plots, and

histograms, we were able to visualize these relationships, identifying trends such as the consistently higher salaries of software engineering managers and product managers.

These findings somewhat answer our research question: years of experience and job titles are somewhat effective predictors of salary, but they don't tell the whole story. While experience does show a correlation to higher compensation over time, neither it nor job title alone is a strong indicator. This shows that determining the causes of salaries is more complex than it might seem and will involve other factors like education, company policies, and job level. Our results offer valuable insights for employers looking to build better compensation models and for employees evaluating their pay based on their roles and experience.

That said, our study contained a few limitations. We had to remove missing data and outliers, which could have affected the results, and the low r-squared value of 16% shows that most of the variability in salaries comes from factors we didn't analyze. Future research should take a broader approach by including variables like education, company size, and regional economic trends. Exploring these could provide a more accurate and comprehensive understanding of what drives salary differences in the tech industry.

Looking ahead, those researching should consider taking a more holistic approach to analyzing and determining salaries. By utilizing a wider range of factors and using more advanced methods, we can move closer to building clear and transparent compensation systems that benefit both organizations and their employee's satisfaction.

References

1. Hanowell, B. (2023, June 29). When using AI to predict pay, context matters - ADP Research. *ADP Research*. https://www.adpresearch.com/when-using-ai-to-predict-pay-context-matters/

2. Li, F., Majid, N. A., & Ding, S. (2024). Unlocking the potential of LSTM for accurate salary prediction with MLE, Jeffreys prior, and advanced risk functions. *PeerJ Computer Science, 10,* e1875–e1875. https://doi.org/10.7717/peerj-cs.1875

3. Martin, I., Mariello, A., Battiti, R., & Hernandez, J. A. (2018). Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study. *International Journal of Computational Intelligence Systems, 11*(1), 1192. https://doi.org/10.2991/ijcis.11.1.90

4. Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. *Information, 13*(10), 495. https://doi.org/10.3390/info13100495

5. Tee, Z., & Raheem, M. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits - A Literature Review. *ResearchGate*, 70–74. https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review

6. Words Matter: The Text of Online Job Postings Can Predict Salaries. (2022, January 10). *Stanford HAI; Stanford University*. https://hai.stanford.edu/news/words-matter-text-online-job-postings-can-predict-salaries