Analysis of Salaries in Tech Companies

Daniel Vazquez, Hiza Mvuendy, Natasha Kroll, Samy Babikerali

The University of North Carolina at Charlotte

DTSC 2302

Marco Scipioni, Christopher Dong

May 3, 2024

Uncovering the Variables that Truly Impact Your Salary

Introduction

The dynamic cultural norms within the tech industry and education can be confusing when determining the optimal career pathway. The common belief suggests that early entry jobs in tech companies require a degree, experience, and networking. Despite this, many big personalities in tech such as Elon Musk, Brandon Dawson, and Matt Schuldt all argue against pursuing a professional degree for jobs in tech (The Oracles, 2020). If college isn't always the answer, what is the most effective pathway to succeed in the tech industry? This report centers on the most impactful factors that affect the total compensation in tech companies. To achieve this, the study will first focus on the context and implications of the research. This includes a literature synthesis of content related to the research. Following that, it will focus on the potential stakeholders, potential conflicts, and our ethical considerations and frameworks. This will explain how the research question is measured. This includes conceptualizing and operationalizing the key terms and variables. Finally, the research will provide an overview of how data manipulation was accomplished. This will include a summary of the statistics and data findings.. Therefore, the study will answer the following research question: What variables have the highest impact on total yearly compensation for STEM jobs? Research shows that certain variables can have a positive or negative impact on total yearly compensation for STEM jobs due to a myriad of factors.

Context and Implication

The importance of location during the decision-making process of finding a job or university institution cannot be overlooked. Several geographical variables can have a large impact on a person's financial outlook in the future according to various researchers. The proximity of post-secondary education institutions to major tech hubs such as San Jose, Seattle, Austin, and Raleigh-Durham can provide real advantages to job-seeking graduates in these areas. Tiffany Chow's research reveals that alumni from universities who are close to major tech hubs tend to earn higher wages in sought-after markets compared to those from universities located far from these dense tech hubs (Chow, 2022). Not only does this reveal the geographic correlation between secondary education institutions and higher yearly compensation, but it leads to the question of STEM qualifications and how these may translate into the real world. Research has shown that the matching of STEM credentials, such as a Bachelor's degree in Computer Science, with STEM employment is more likely to occur in STEM agglomerations (previously mentioned tech hubs), as opposed to larger metropolitan areas (New York, Los Angeles, etc) with a greater number of STEM jobs. This can mainly be attributed to the higher concentration of STEM positions in these tech hubs (Wright et al., 2016). While location plays a pivotal role in influencing income and employment within the tech industry, this is not the only factor that has been researched.

Diversity in the workplace is something that has been a pressing issue for years, and research on this variable is crucial to this project's objective. The prevalence of tech jobs in Silicon Valley allows it to be a great setting for an inside look into the tech industry's diversity. In a private sector survey conducted by the U.S. Equal Employment Opportunity Commission in

2014, the research found that the participation rates of whites, Asian Americans, and males in high-tech industries were disproportionately higher, notably in the Silicon Valley area. This was especially noted in the high-tech leadership positions such as executives/senior level officials and managers nationwide. At the same time, African Americans and Hispanics were disproportionately fewer in these leadership positions and in technology jobs throughout the nation overall (EEOC, 2014).

Furthermore, businesses in the high-tech industry have claimed that they promote a culture of "...flexibility, collaboration, teamwork, clearer and fewer rules, less internal politics, and flatter job ladders" that some argue can benefit women more than bureaucratized alternatives. Despite this, these benefits have favored mainly white women in these spaces in the last decade, with women of color not experiencing similar progress, despite these companies significantly expanding their workforces in this period. Not only this, but with women only making up 20% of executives, senior and managerial positions, the prevalence of gender discrimination and sexism can only exacerbate salary discrepancies between men and women in the high-tech industry (Neely et al, 2023).

The main stakeholders relevant to our research project are the companies paying employees' wages, employees, universities, and students. The goal of the companies is to maximize profits, which can be done by cutting down on expenses. By decreasing the amount of profits utilized for paying employee salaries, they would be able to consequently decrease their expenses and increase their profit margins. This research project will show which regions can expect to have lower pay and which will have higher pay. Companies can implement this by hiring more employees in locations within these lower-paying regions to meet their goal of minimizing costs. This way, they would still have the same number of employees working the same number of hours, but wouldn’t be required to pay the employees as much due to the expectations of salaries within these regions. The companies could also hire more employees based on experience levels to again minimize expenses. Employees with less experience will not be expecting as much pay compared to other employees with more experience, even if they have equal levels of skills. These companies can use the research data to purposefully select employees based on their skill level, ignoring their years of experience, to get the same competency of employees while minimizing pay.

The goals of the companies mentioned above, directly conflict with the goals of their employees, another main stakeholder in this research project. Employees aim to maximize their pay while doing the smallest amount of work. This research study can be implemented by the employees to analyze the regions with the highest pay for the least amount of work experience. As an example, there is a company with a location that has a very high pay that expects only a few years of experience. This same company has another location with lower pay that expects more years of experience. The only difference between these two positions is the DMA ID sections they are in. Employees could utilize this research project by applying to jobs in these DMA ID sectors with higher-paying jobs with less experience. Employees can also utilize the study by increasing their work experience to consequently receive a higher salary.

Another main stakeholder in this project is universities. This project examines the correlation between degrees and salary, when both the university and occupation are local. Universities could implement this project for marketing purposes. Universities often have connections to companies within their area, an example being UNC-Charlotte’s connections to financial companies such as Bank of America. Advertising these connections and the percent of graduates who are employed by these companies is already a huge marketing strategy, but it could be even more effective if the universities implement the findings for DMA ID and total yearly compensation. The universities located in the DMA ID sections with increased total yearly compensation could advertise that jobs in that area, which they have connections with, have a higher total yearly compensation. For example, if UNC-Charlotte was within a DMA ID sector with this higher yearly compensation, their advertisement could look something like this: "_% of our graduates land jobs with companies in our network, boasting higher average yearly compensation than similar roles across the U.S. Come to UNC-Charlotte today!"

This aligns with the goals of the future college students. When selecting which university to attend, there are many factors to take into account. One of these factors is the likelihood of getting a high-paying job after graduating from that university. If there are two universities with the same tuition and level of prestige, then a tipping point for the decision of where to attend may come down to the connection between the university and jobs after graduation. This being said, future students could utilize the correlation between DMA ID and total yearly compensation to determine the university they plan to attend.

This research topic holds significant importance and can be implemented in several different ways. Firstly, this topic can highlight equity in compensation. With this data, we will be able to understand how experience and geographical factors influence compensation. By identifying discrepancies, stakeholders will be able to highlight potential pay disparities and take steps to correct them. Another potential benefit of focusing on the research question regarding the factors that are most influential on total yearly compensation is being well-versed in strategic workforce planning and informed decision-making. Stakeholders can use these insights to improve their resource allocation, talent recruitment strategies, and workforce planning and retention. Another potential benefit to highlight is stakeholder policy and regulation compliance. Additionally, many regions have regulations that govern fair pay.

A potential harm of this project is privacy concerns. Considering that real employee compensation data is used, making sure to protect and safeguard sensitive data and private information while also complying with data protection regulations is essential. This way, data breaches or privacy violations can be avoided. Manipulation for unethical purposes is another harm that was considered for this project. There is a potential for some stakeholders to manipulate the findings of this project to justify discrimination practices or unfair, unequal compensation policies. An example of this would be a company paying an employee’s wages. An organization/company could selectively interpret our findings to justify underpaying employees in certain regions, employees with less experience, or even unintentionally discriminating by race/gender.

There are several ways the research is set up to ensure it does more good than harm. Unintended consequences can happen if all of the context surrounding this project is not considered. Some of which include the prevention of privacy breaches, unintended consequences, and harmful exploitation of the data used in the research. An example would be adjusting pay/compensation based solely on a few factors such as region or years of experience without considering performance or market trends. This can result in inequitable outcomes. Additionally, this research included meticulous analysis to prior studies and current conditions to ensure the progression of the subject. Using a utilitarian approach, the impact on institutions, ecosystems, and cultural perspectives were examined.

Measurements

The original goal of this research project is to determine and calculate which variables had the highest correlation to success in STEM jobs. It is important to outline the terms used throughout this research project. STEM jobs are any jobs within the Science, Technology, Engineering, or Mathematical industries. With this research project, the dataset contains information pertaining to tech companies, so there is a focus on the tech sector within STEM jobs. The general definition of success given by Merriam-Webster Dictionary is "the attainment of wealth, favor, or eminence", but in the context of this research project, there is a focus on the attainment of wealth as a measure of success. By focusing on the wealth aspect of success, total yearly compensation will be a good representation. Knowledge is another concept focused on in this research project. Knowledge can be defined as the level of expertise a person possesses in a particular field. Location, or region, has a large impact on how much your salary is, an example being increased pay for the same job position in Los Angeles, California due to the cost of living in the area. Due to this, there will likely be a high correlation between region and total yearly compensation. Hence, the region is a concept focused on for this project.

The terms used in the dataset that are implemented in this project are total yearly compensation, locations, years of experience, years at the company, tag, base salary, stock grant value, bonus, gender, city ID, DMA ID, education, and race. Total yearly compensation in the dataset is measured by the sum of base salary, stock grants, and bonuses. Location is the specific city and state in the United States where the data was collected from. DMA ID represents the Designated Marketing Area ID, which breaks the US into 210 different sections for marketing purposes. Figure 1 is a map displaying the individual sectors within the contiguous United States (United States without Alaska and Hawaii). City ID is a number representing every city in the United States. Location, DMA ID, and City ID are methods to measure region in the conceptualization of this research project. Gender is if the person the data was collected from is male or female. Education is determined by the degree of the employee, including a Master's Degree, a Bachelor's Degree, a Doctorate Degree, a high school degree, and some college degree. Years of experience is the number of years that a person has worked in that particular field. Years at the company can be defined as the number of years a person has spent working for that particular company. All three of these factors (education, years of experience, and years at the company) represent the concept of knowledge and are how we operationally define knowledge.

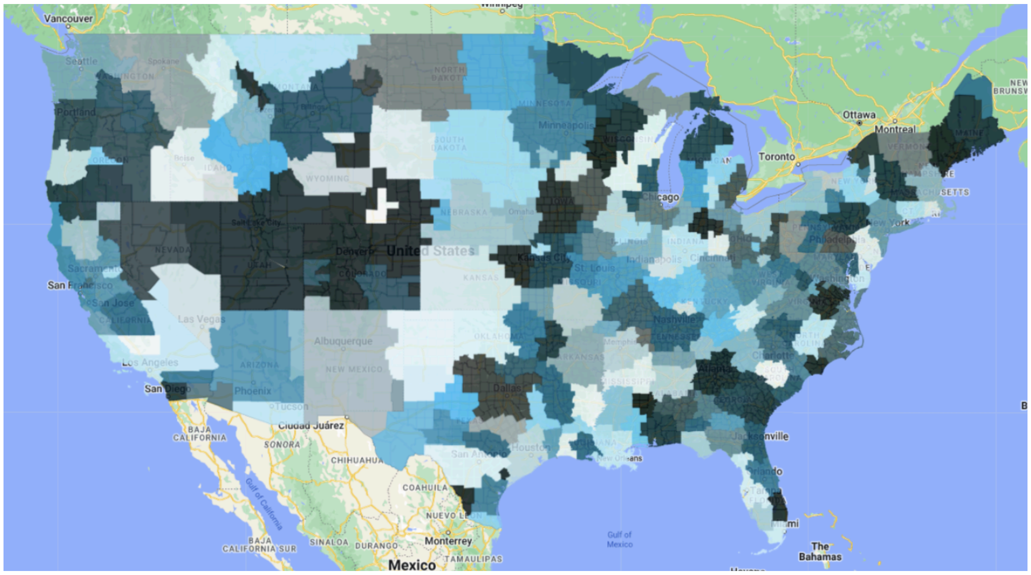

Figure 1 - DMA ID sectors in the contiguous United States.

To answer the research question, the project uses multiple regression models to display the current correlation between different variables and is used as a predictor for total yearly compensation based on given independent variables. Before using the multiple regression models, a correlation heat map is used to determine which variables have the highest impact on success. The variables with the highest correlation to total yearly compensation are then plugged into different multiple linear regression models. Using the r-squared value, the most accurate model will be selected for predicting total yearly compensation. The r-squared value determines the variation of the dependent variable based on the independent variable, with a larger r-squared value representing a more accurate model. Hence, the best predictor is determined by the model with the highest r-squared value.

Data

The dataset used for this research is 'Tech Companies Salary". Due to missing data and outliers, before running any models, some manipulation of the data set was necessary. The variables gender, education, and race were all missing a considerable amount of data, 31.19% for gender, 51.2% for education, and 64.2% for race. The missing data in these variables is filled with the most frequent result, which would be a Master's Degree for education, Asian for race, and male for gender. Additionally, categorical variables such as gender, education, race, and tag (title of position) are converted to binary data using "One Hot, and Label Encoding". To achieve the most accurate results, outliers in total yearly compensation are removed. This was accomplished by computing the interquartile range, lower bound, and upper bound. Followed by the removal of data rows that are above the upper bound and below the lower bound.

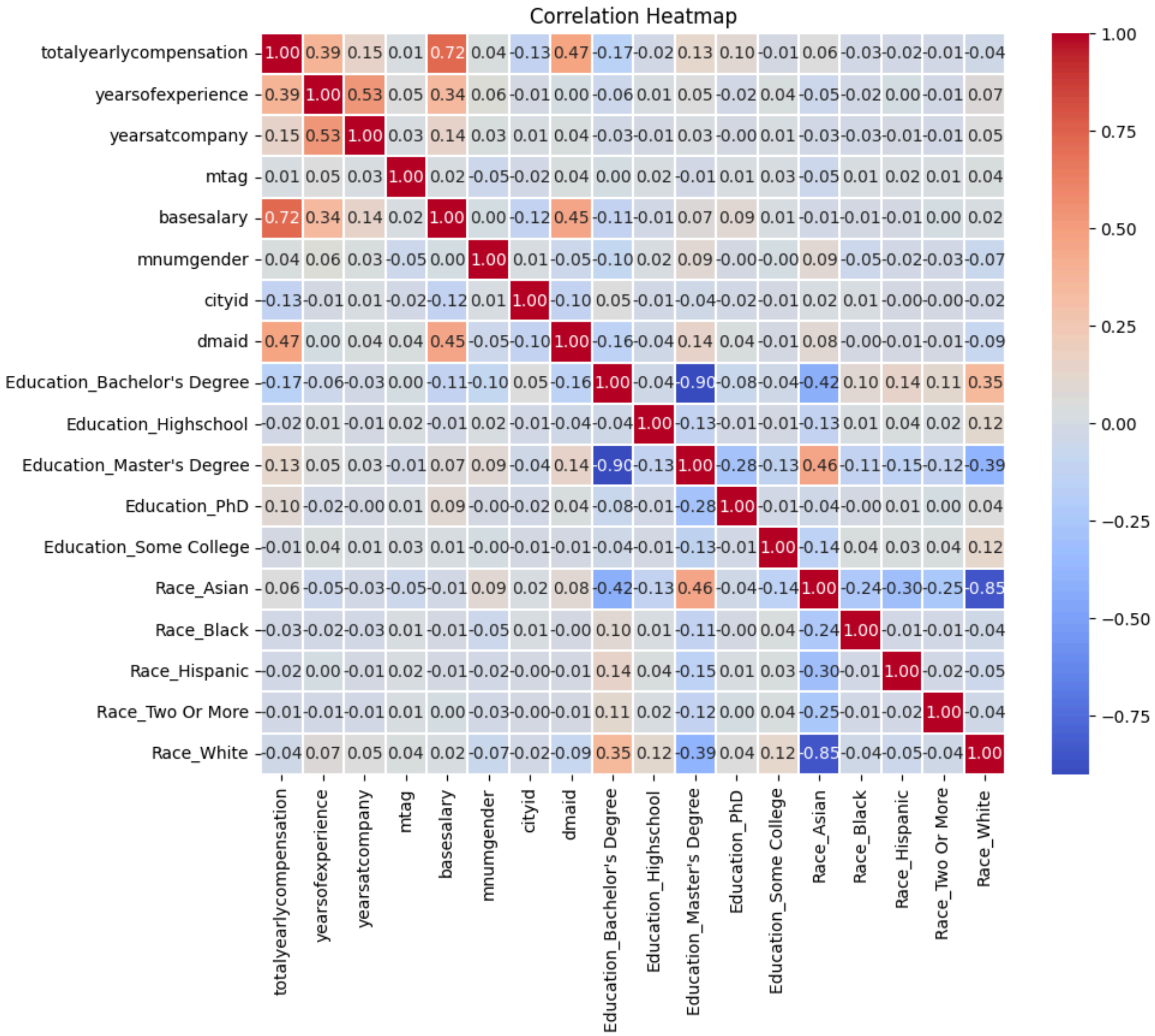

Figure 2 - Correlation Heat Map using the factors within the “Tech Companies Salary” dataset.

Correlation Heatmap results	
Variables	**Correlation**
Total Yearly Compensation	1
Years of experience	0.3883
Years at company	0.1513
Base salary	0.7229
City ID	0.1286

DMA ID	0.4672
Bachelor's Degree	0.1658
Masters Degree	0.1256

To find the best variables to predict total yearly compensation, a correlation heat map is used. The 7 highest variables that explain total yearly compensation are presented at the table above, which have been extracted from the heatmap. Out of these variables, the 2 that are best suited for a model are years of experience and Designated Market Area ID (DMA ID). Hence, years of experience and DMA ID, the two independent variables, are used to predict the total yearly compensation, the dependent variable. Coincidently, the missing data did not impact the independent variables, which assisted in the creation of a robust and reliable model.

The dataset is trained on multiple different regression models (Linear regression, Random Forest, KNN, XGBoost, LightGBM, and SVM). To find the best-performing model, statistical analysis is performed to test the utility of each model.

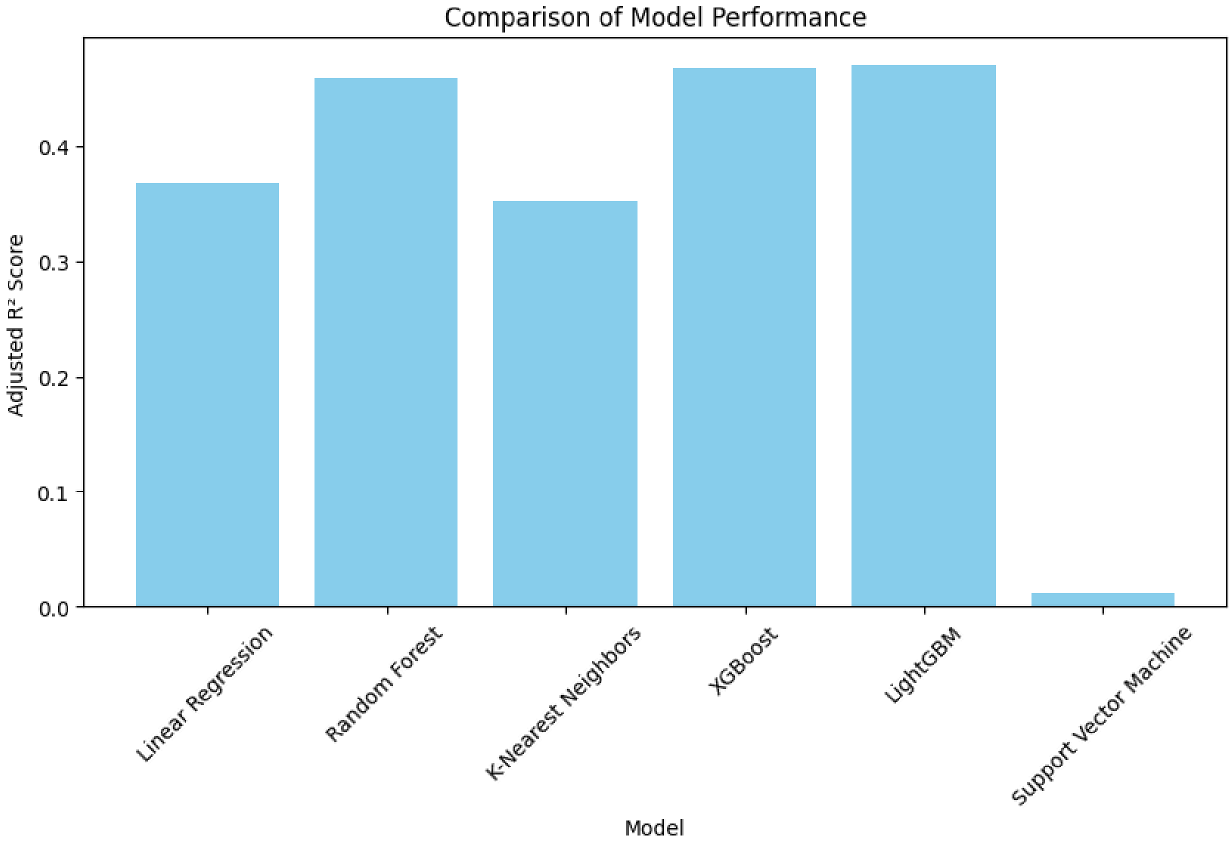

Figure 3 - Comparison of Model Performance bar chart.

In Figure 3 (pictured above) the independent variables are the different regression models, including linear regression, Random Forest, K-Nearest Neighbors, XGBoost, LightGBM, and Support Vector Machine. The dependent variable is the adjusted r-squared score, ranging from 0.0 to 0.5. While Random Forest, LightGBM, and XGBoost are relatively close in performance, LightGBM performed slightly better with an adjusted R square value of 0.47. This means that the independent variables (years of experience and DMA ID) explained 47% of the dependent variable (total yearly compensation). Additionally, the Global F test, which is a statistical analysis that tests the utility of the model is performed. The test computed a P value < 0.0001 which strongly indicates the rejection of the null hypothesis at any reasonable level of significance. This very low P-value suggests that the model predictors collectively provide a significantly better fit to the data than a model without any predictors (just the mean of the dependent variable).

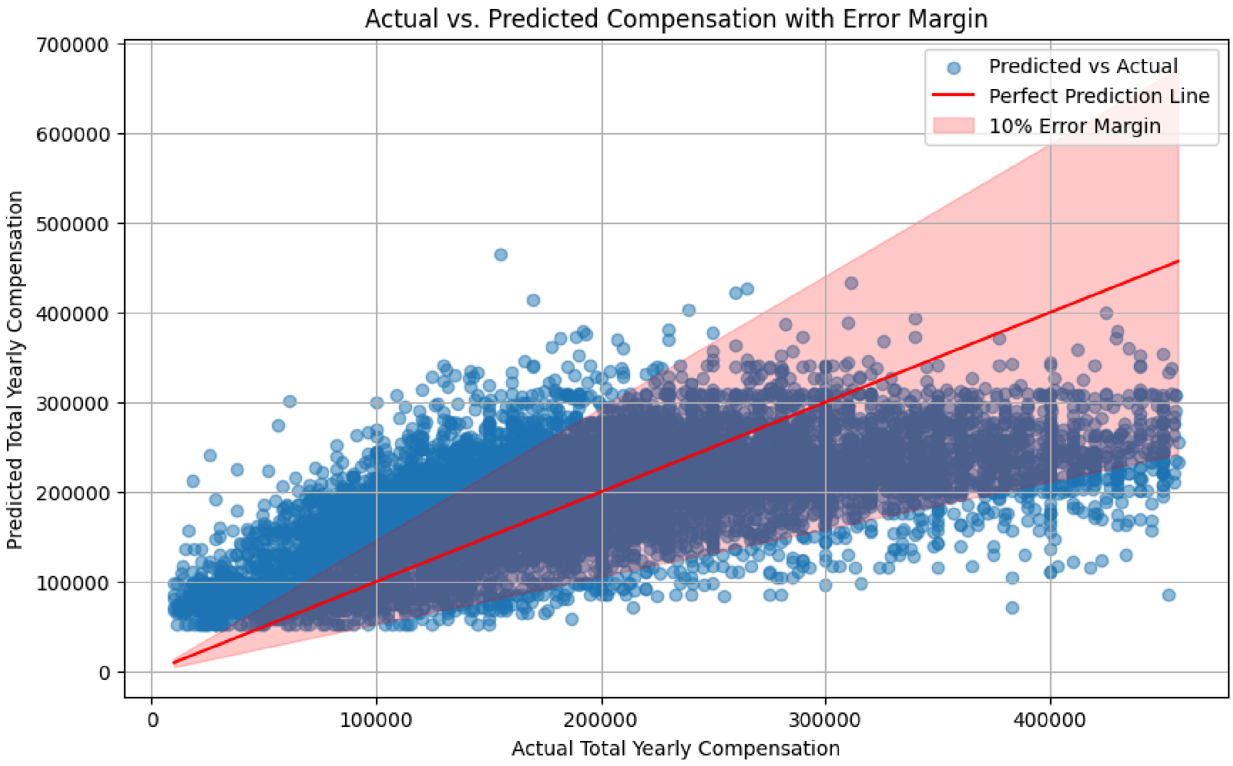

Figure 4 - Actual vs Predicted Total Yearly Compensation Graph.

Figure 4 is a visual for the LightGBM model, comparing the actual values to the predicted values. The independent variable is the actual total yearly compensation, ranging from 0 to 500000. The dependent variable is the predicted total yearly compensation, ranging from 0 to 700000. This is graphed using a scatterplot of predicted vs actual blue dots. There is a line in red representing the predicted values, with the red-shaded area representing a 10% margin of error. Generally, the model can predict tech companies' total yearly compensation by a 47% margin of error between $60,000 and $310,000.

Conclusion

Overall, this research investigated which variables had the highest impact on success. Multiple studies were evaluated and considered when creating the models. Some of the most relevant studies included the possible correlations between the independent variables which assisted in understanding the results of our model. The research used multiple regression models to display the correlation between different variables and is used as a predictor for total yearly compensation based on the given independent variables.

Before creating the models, a correlation heat map is used to determine which variables have the highest impact on success. The heat map analysis revealed that the largest factors correlating to success were location and knowledge, as indicated by DMA ID and years of experience. In the research project, DMA ID was used to represent these regions/locations. DMA ID represents the Designated Market Area ID, which breaks the US into 210 different sections for marketing purposes. An example of how location is related to total yearly compensation is increased pay for the same job position in Los Angeles compared to Charlotte due to the higher cost of living in Los Angeles.It is important to note that although there is an associated correlation between DMA ID and total yearly compensation, there may not be a correlation between DMA ID and disposable income. Our definition of success pertaining to wealth then may not be applicable to DMA ID when taking into account the cost of living. Finally, years of experience which is a variable representing knowledge was concluded as a logical correlator for compensation.

While the results of the model may seem unsatisfactory, it is important to note that the model achieves a 47% explainability by only using 2 variables. The model could have achieved better results if the data set had fewer missing values and more variables. Additionally, the research team could have utilized additional time to further study the time frame included in the dataset to possibly explain the type of missing data (MAR, MNAR, MCAR), and the state of the economy. Furthermore, future studies could provide insight into which factors truly impact salaries; If education is necessary for specific careers; If there is a disparity in fields that must be diverse.

References

Chow, T. (2022). The Geography of Jobs: How Proximity to a Prestige Labor Market Shapes Opportunity for Computer Science Degree Holders. *Social Sciences*, *11*(3), 116. https://doi.org/10.3390/socsci11030116

DMA Map. (n.d.). Media Market Map. Retrieved May 3, 2024, from https://www.mediamarketmap.com/market-map/

Merriam-Webster. “Definition of SUCCESS.” Merriam-Webster.com, 2009, www.merriam-webster.com/dictionary/success.

Neely, M. T., Sheehan, P., & Williams, C. L. (2023). Social Inequality in High Tech: How Gender, Race, and Ethnicity Structure the World’s Most Powerful Industry. *Annual Review of Sociology*, *49*(1), 319–338. https://doi.org/10.1146/annurev-soc-031021-034202

The Oracles. (2020, March 10). *8 business leaders explain why you don’t need a college degree to be successful*. Business Insider. https://www.businessinsider.com/business-leaders-dont-need-college-degree-be-successful#7-all-the-knowledge-you-need-is-already-at-your-fingertips-7

U.S. Equal Employment Opportunity Commission (EEOC). (2014). *Diversity in High Tech | U.S. Equal Employment Opportunity Commission*. Www.eeoc.gov. https://www.eeoc.gov/special-report/diversity-high-tech

Wright, R., Ellis, M., & Townley, M. (2016). The Matching of STEM Degree Holders with STEM Occupations in Large Metropolitan Labor Markets in the United States. *Economic Geography*, *93*(2), 185–201. https://doi.org/10.1080/00130095.2016.1220803