STEM Major
Salary Prediction

IEOR 4523 -- Final Project

Professor Uday Menon

Section 001

Data Another Day

Zhuoyang Han (zh2454), Xiao Huang (xh2488), Weishan Lu (wl2778), Wenhao Yang (wy2378), Xiaodie Ni (xn2139)

Intro & Background

From 2008 to 2018, the number of STEM jobs has grown about 9 percent, where SDE jobs specifically increased 22 percent. As technology rapidly develops and the information age expands, an increasing number of people are pursuing a degree in STEM-related disciplines, including ourselves, or are exploring a career switch into STEM-related sectors. Our project is based on the analysis of a dataset on STEM occupation salary downloaded from Kaggle. We aim to use predictive modeling to help people get a better understanding of how much compensation they may receive based on demographic information. With our model, we intend to empower STEM students in planning their future career path, selecting their preferred working location, comparing company cultures and policies, and negotiating salaries.

Data Overview

The dataset contains 62643 rows of salary records for people working in STEM-related fields collected from 2017 to 2021. A total of 29 columns including: 'company', 'level', 'title', 'totalyearlycompensation', 'location', 'yearsofexperience', 'gender', 'Education', etc.
Referring to Figure 1.1, there is a significant amount of data missing in ‘gender’ (31.19%), ‘Race’ (64.2%), and ‘Education’(51.52%) at earlier stages. This can be explained by a non-mandatory fill for the above fields in initial surveys.

Data Processing

Selecting Columns

- ['cityid', 'dmaid','rowNumber','tag','otherdetails'] are dropped as they are not informative without further supporting contents and cannot directly benefit later analyses.
- [‘level’] is excluded because different companies have very different decomposition of job levels and it is hard to unify into a fair standard.
- ['basesalary','stockgrantvalue', 'bonus'] are not being evaluated despite the fact that they are highly correlated with the dependent variable (‘totalyearlycompenstation’) since they are the building blocks of the salary package, and thus may cause overfitting problems.

Handling Outliers & Missing Data

From Figure 1.2, we observe that some records have a greater number of years at the company than total years of experience, thus are illogical and were removed. In addition, we excluded outliers that are more than 3 standard deviations away from the mean for the dependent variable. Although [‘gender’,’Race’,’Education’] have a large proportion of data missing, our group believes the existing records can still provide partial insights for the exploratory analysis. Therefore, our later findings on those subjects are only based on known data and are not concerned with filling in the unknowns.

Transforming & Filtering Data

In order to handle the right skewness in 'totalyearlycompensation', we applied log transformation and obtained an approximately Normal distribution (Figure 1.3). We also extrapolated 'Year' from 'timestamp' as an additional feature because salary packages might differ from time to time due to different market conditions.

As the vast majority of the records are within the U.S, we narrowed our dataset to the top 10 American states. Similarly, among 1600+ companies, our group put a focus on the top 7 giants, and we also filtered out the 5 most popular job positions for further analysis.

After filtering, our team transformed categorical data (['company','title','State','Year','gender']) into dummies using one hot encoding.

Exploratory Data Analysis

We first take a look at the basic data statistics across the selected seven companies. As we can see from figure 2.1, the 25th percentile value of facebook salary is roughly the same as Microsoft's 75th percentile value. Comparing statistics of the two as in figure 2.2, there are more entry-level records from Microsoft while Microsoft has the least standard deviation among the companies, which means the salary discrepancy is low.

Title:

Since Software Development Engineer and Product Manager are two of the most common job titles in the record, we used the seaborn.violinplot to take a further look at these two titles. We can see from figure 2.3 that Product Managers usually earn more in top-level positions, while Software Development Engineers appear to receive higher average salaries than Product Managers in the same level.

Gender:

Within the indicated records of Male (85.4%) and Female (13.9%), from figure 2.4, it is clear that there is a big difference between the salary earned by males and females in the STEM field. While males dominate the positions, they receive higher average salaries and more opportunities to be promoted, as the highest income bar is much higher than that of females.

Race:

As shown in figure 2.5, the major races in STEM fields are Asain (56.7%) and White (31.8%). When there are more Asians than Whites, Asians have a lower upper bound of income, which indicates a lower top-level salary. On the other hand, Black employees earn the least average salary and also the least upper bound, implying a glass ceiling. It is also worth noticing that Product Managers with 2+ race backgrounds have an unusually high 75th percentile value.

Location:

After limiting the location to the United States, we group the data by states and count by population. From figure 2.6, we can see from the heat map that most employees working for top technology companies are based in Seattle and California.

Modeling

Linear Regression

Linear Regression captures a mathematical linear relationship between input and output values, and we used it as a base model. With this commonly used model, the variation of outcome total annual compensation as the dependent variable could be explained by other factors easily. R-square and Mean Square Error were used to evaluate the model performance. After substituting all the features, we obtained the testing R-square value of 0.535 and MSE of 0.069. And Company_facebook shown in *Figure 3.1* was identified as the most important feature based on the coefficient and P-value.

Lasso Feature Selection

Lasso places an L1 regularization into the objective function. And all the beta coefficients are penalized equally regardless of their values. As a result, it is possible that some beta values will be pushed to zero. Therefore, Lasso regression could derive a certain beta to zero, which works as a model selection tool. Starting with standardizing each independent variable and re-center the dependent variable using mean value, it would allow us to tease out the impact of bata zero in the regulation. After standardizing the dataset and performing the feature selection based on the R-square metric, 36 features were selected from a total 43 features with an optimal lambda of 0.001, which was shown in *Figure 3.2 and 3.3*. We ended up with a testing R-square value of 0.557, which displayed an improvement in the regression performance.

Random Forest

Random Forest, an ensemble learning method with a collection of individual decision trees, was used to construct a multitude of decision trees for predicting the total annual compensation. Since we have both categorical and numerical variables, Random Forest would be able to account for both types. As shown in *Figure 3.4*, year of experience was selected as the most important feature. For optimizing the Random Forest model, several parameters could be tuned using grid search or hyperopt. After performing the Random Forest model, we obtain an R-squared value of 0.608 and MSE of 0.058.

XGBoost

Similar to Random Forest, XGBoost is also a tree-based ensemble learning method. However, random forest builds trees in parallel, while XGBoost, as a gradient boosted model, builds trees sequentially. Each of the trees in XGBoost is built based on the information from previous trees, so it generally performs better than random forest. From our output, XGBoost is the best one among all models we used. The advantages of dealing with missing values and consisting of many hyperparameters that can be tuned may also account for the best performance. After tuning hyperparameters, we got 0.629 for the R-square value and 0.054 for MSE on the testing set (see Figure 3.6). Besides years of experience, XGBoost also selects software engineering manager and doctorate degree as important features just like Lasso does, which shows in *Figure 3.5*.

Support Vector Regression (SVR)
SVR uses the same principle as Support Vector Machine (SVM) to do a regression model. Unlike linear regression, SVR fits a best line within a threshold of values. It sets a margin of tolerance that allows the error within that margin while it minimizes error. It also has a regularization parameter to control over-fitting problem. We got an R-square score of 0.608 and MSE of 0.058 for SVR (shown in Figure 3.6), which is not bad. But it took a significantly longer time to train the SVR model and do hyperparameter tuning compared to other models.

K-Nearest Neighbors
KNN regression model learns from neighbors and calculates the mean of K nearest data points as output. Thus, the accuracy of KNN models will rely on the similarity of data points in the neighborhood. Also, a large number of predictors may cause bad performance of KNN. We used grid-search cross-validation to find the best K value: K=10, but the result of the KNN regression model for our data is still not good. We obtained a testing R-square of 0.53 and MSE of 0.074 (see Figure 3.6). One explanation is that we have many dummy variables in our dataset, which may affect the performance of KNN.

Conclusion and Future Improvements

Overall, our model focused on key factors that impact salaries of STEM-related jobs. With various data analytics tools and machine learning techniques, our models in general offer good predictions. Based on feature importance and coefficient, our model suggests that working for Facebook may increase your chance of being paid higher. Additionally, if you work in the states of New York or California, you could potentially earn more. Furthermore, choosing a career path as a software engineer manager and having a doctorate degree would also help.

Admittedly, we believe there is still room for improvement in our model. For the time being, we only consider some large technology firms, but we believe that our model would be more valuable if we could include additional data from medium and small firms, and build features based on company size rather than specific companies. We may also include data from technical professions in other industries, such as banking or consulting, and develop industry-specific features. As for data source, instead of utilizing Kaggle, we believe we could web scrape various websites, such as Glassdoor and Levels.fyi, to obtain more comprehensive and up-to-date data.

Furthermore, we feel it is critical to include additional qualitative criteria while utilizing our model to make decisions. For instance, company culture, future career prospects, work-life balance, and even the city's weather. Those are difficult to quantify but crucial in determining quality of life, and we should combine them with salary to make an informed decision.

Appendix

Link to dataset https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries

Figure 1.1

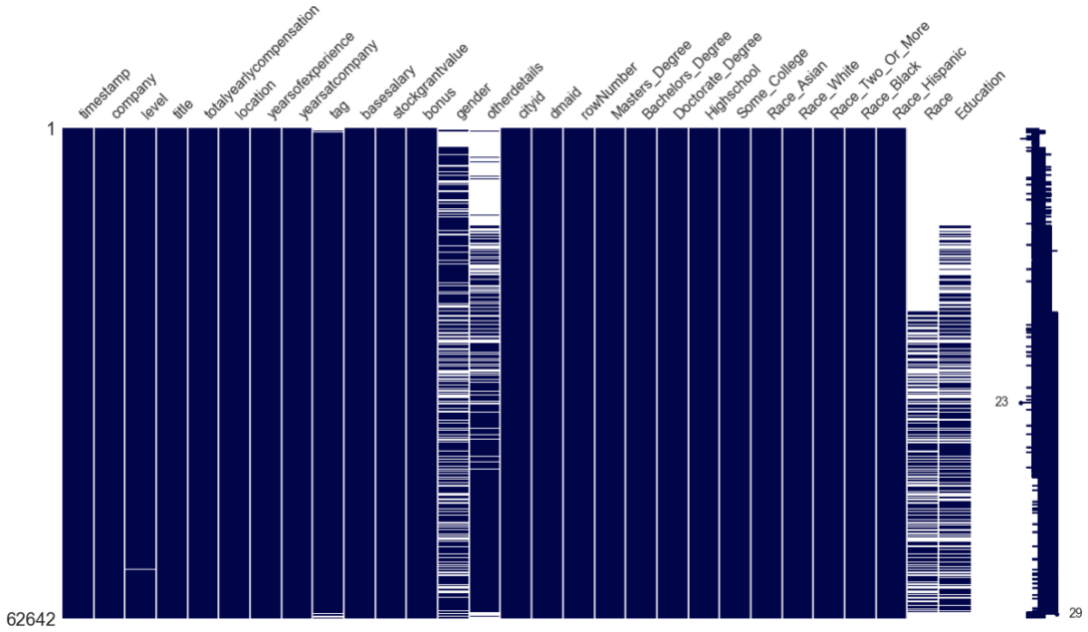

Figure 1.2

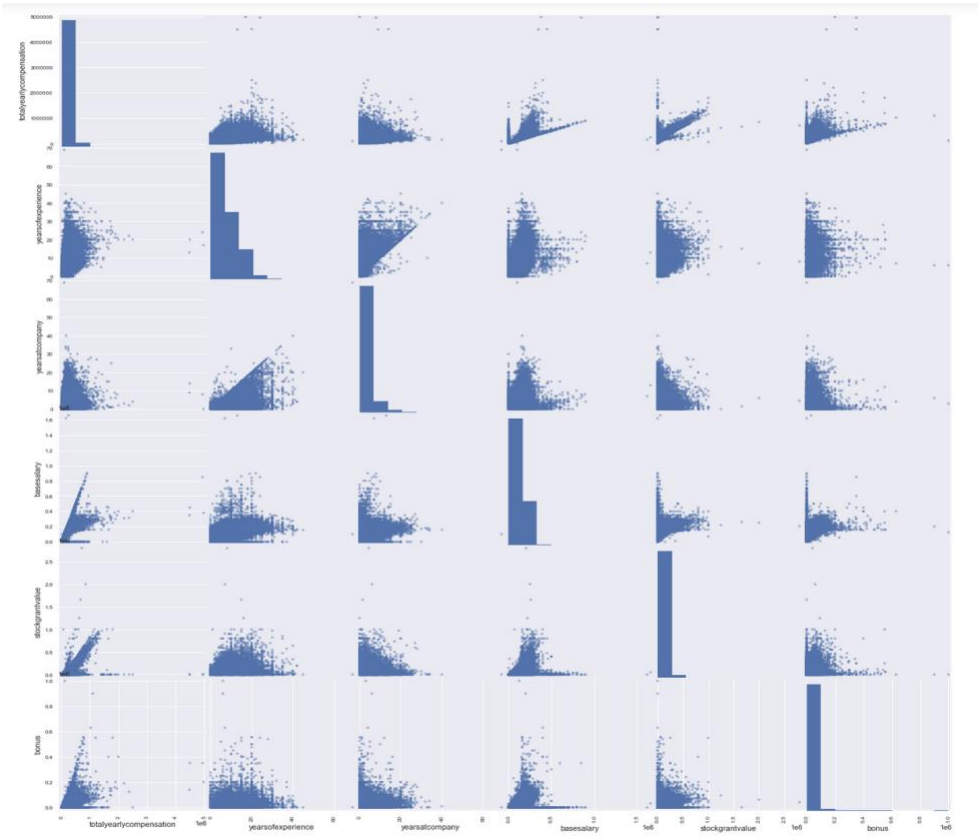

Figure 1.3

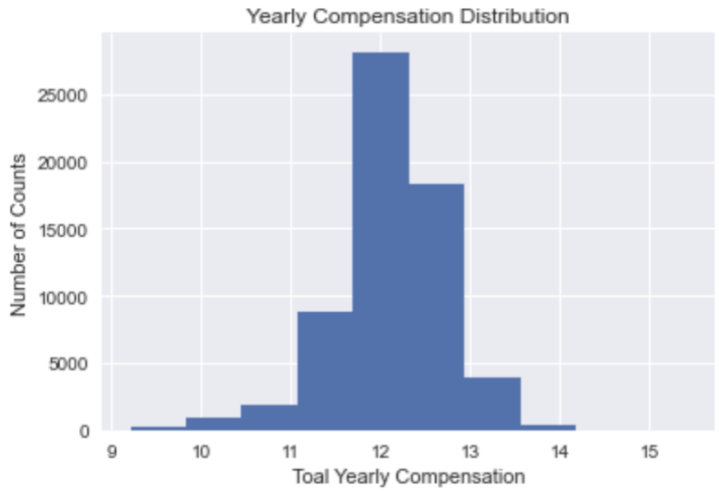

Figure 2.1 Comparison among seven companies

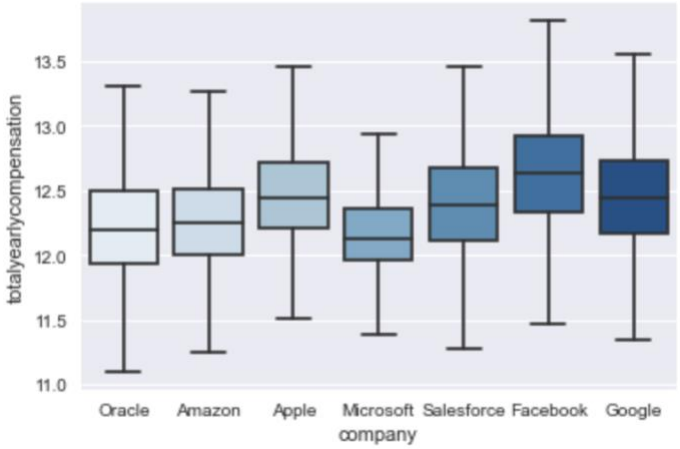

Figure 2.2 Comparison between Facebook and Microsoft

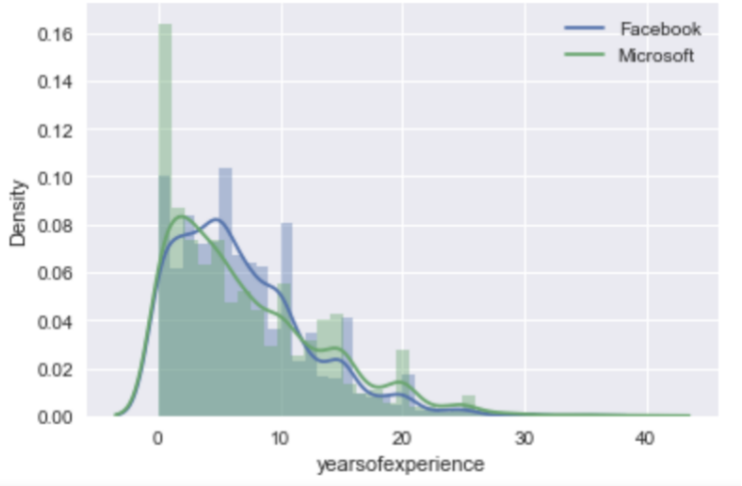

Figure 2.3 Salary comparison between SDE and PM among companies

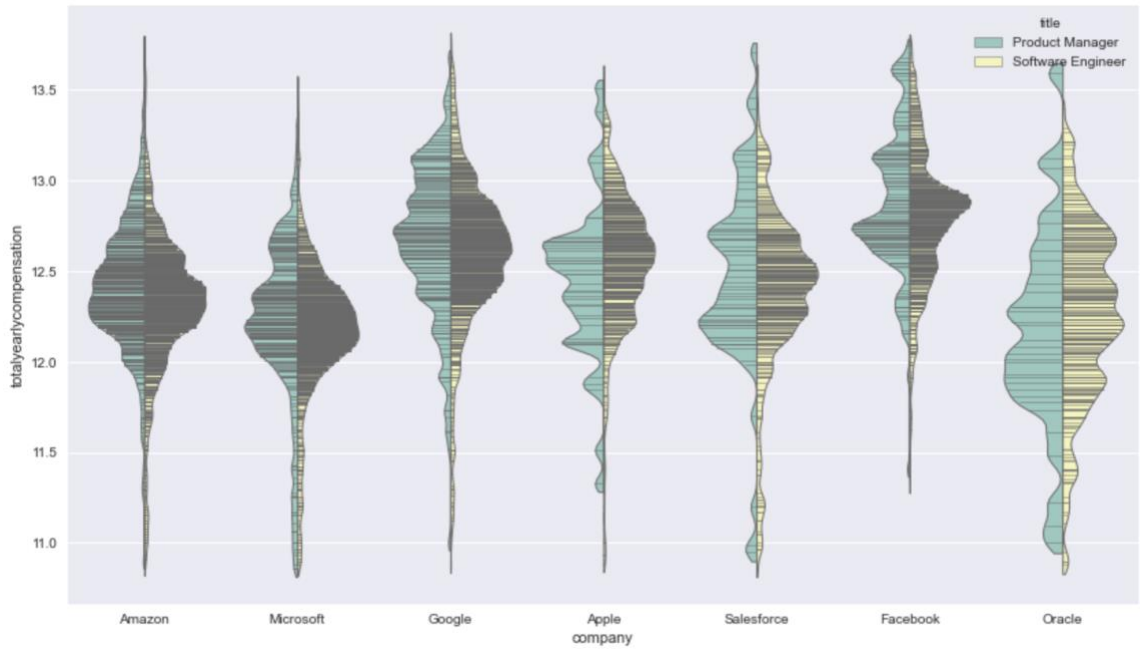

Figure 2.4 Population and salary comparison by groups of gender

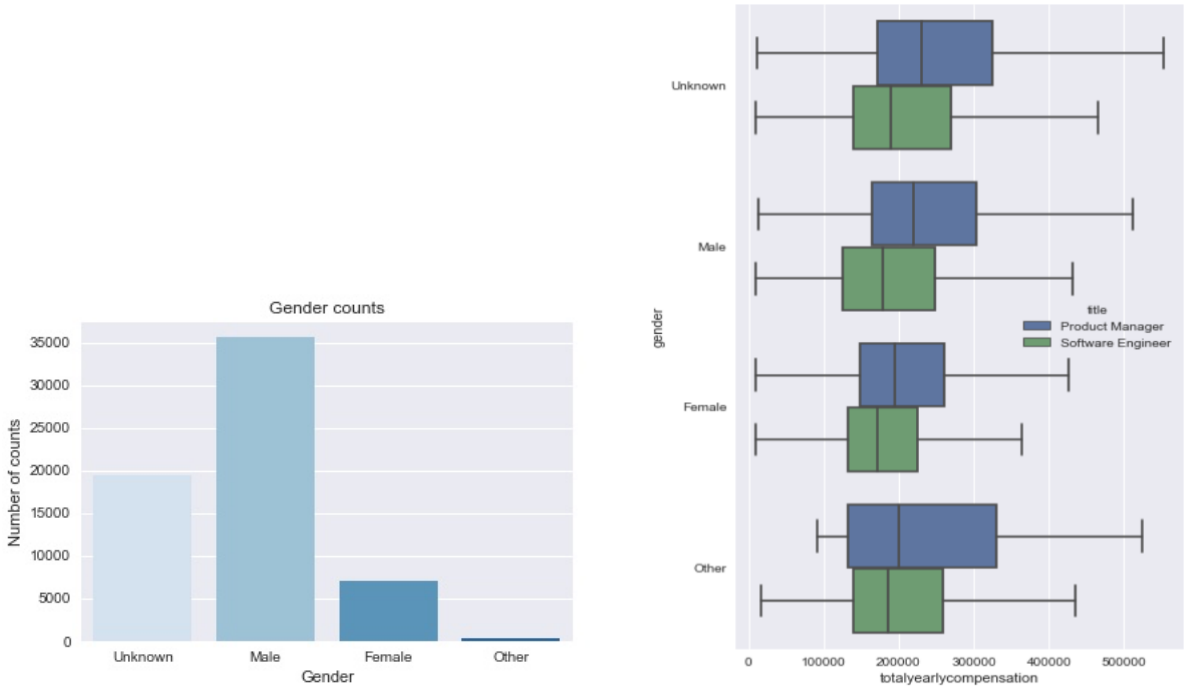

Figure 2.5 Population and salary comparison by groups of race

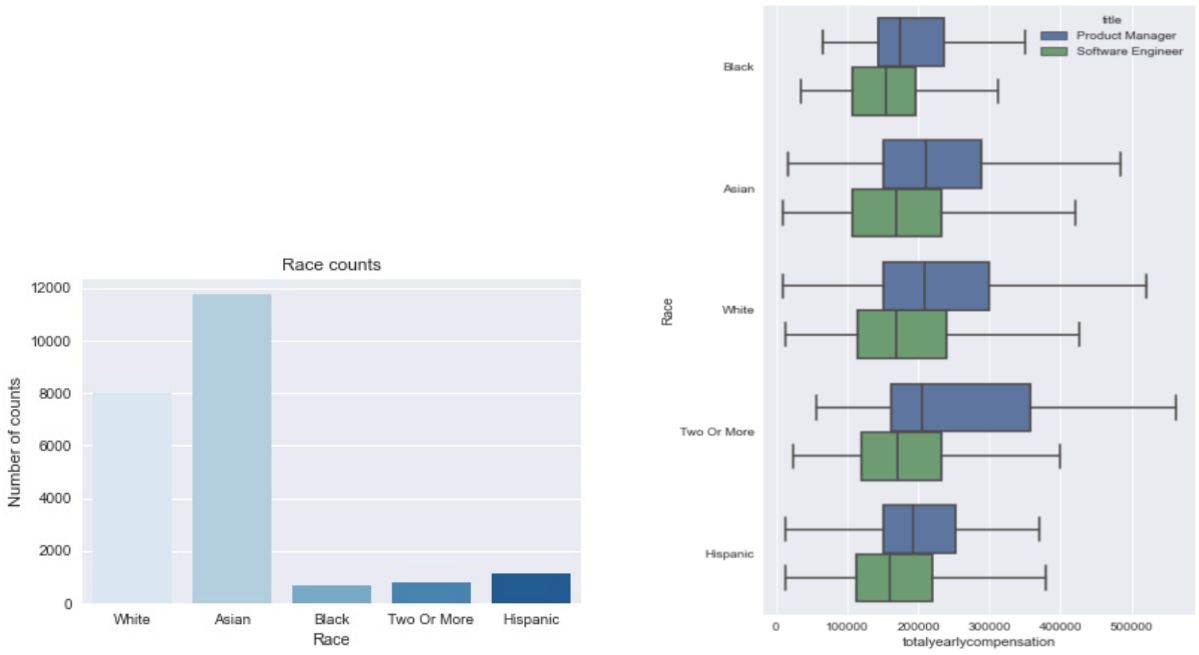

Figure 2.6 Working base of STEM employees Heat Map

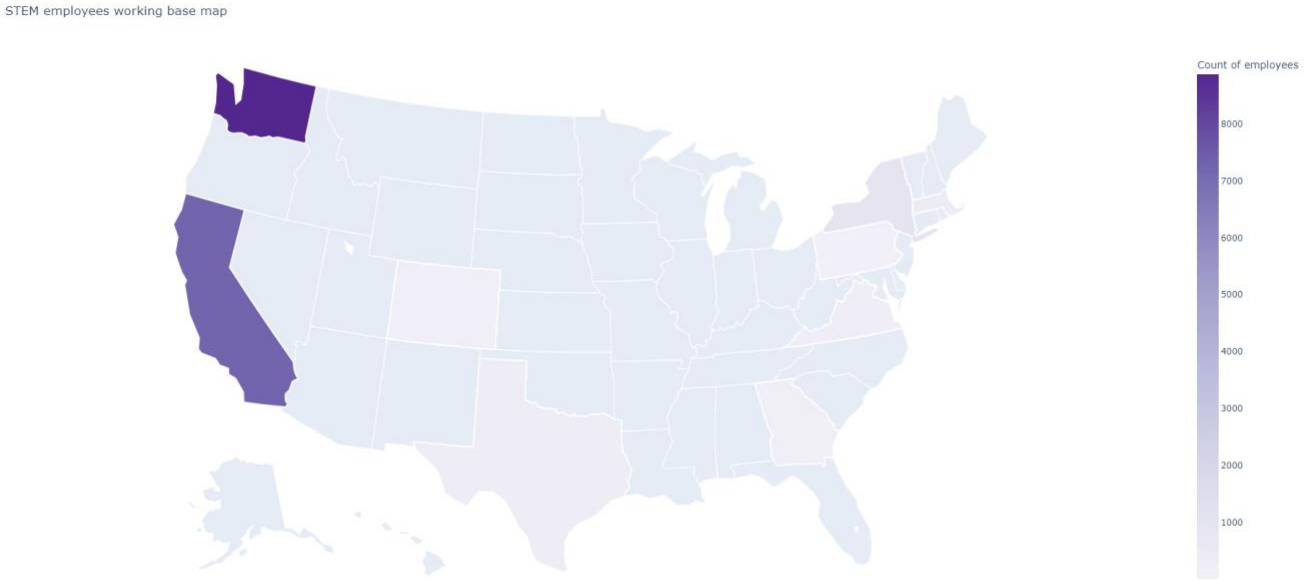

Figure 3.1 Feature Importances - Linear Regression

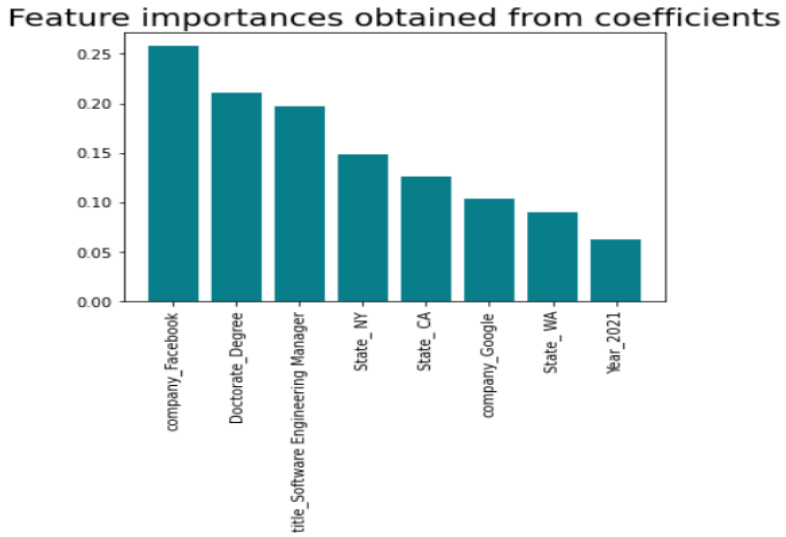

Figure 3.2 - Number of features vs. R^2 and MSE

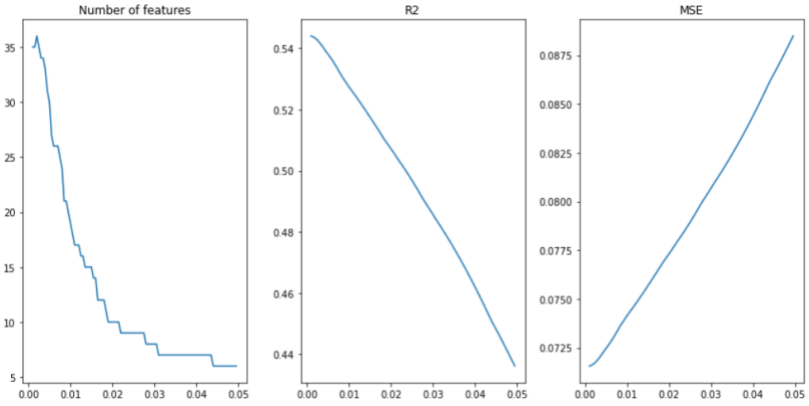

Figure 3.3 Feature Importance - LASSO

company_Facebook	0.258234
title_Software Engineering Manager	0.222238
Doctorate_Degree	0.199690
company_Google	0.107383
company_Apple	0.052885
Year_2021	0.047808

Figure 3.4 Feature Importance - Random Forest

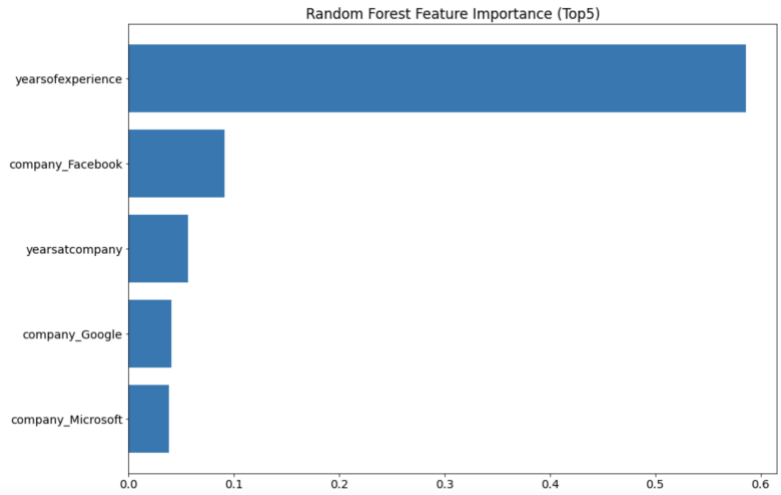

Figure 3.5 Feature Importance - XGBoost

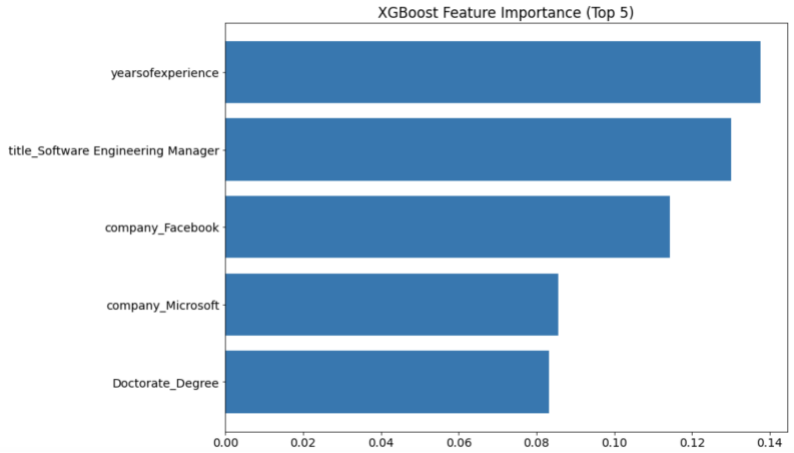

Figure 3.6 - Model Comparison

	R-Square	MSE
XGBoost	0.629	0.054
SVR	0.608	0.058
Random Forest	0.607	0.059
LR (Lasso Selected Features)	0.557	0.065
Linear Regression	0.535	0.069
KNN	0.530	0.074