|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group # | Following Instructions (11 pts) | Writing Quality (11 pts) | Abstract (12 pts) | Data Collection / Cleaning / Exploration (11 pts) | Data Exploration Insights (11 pts) | Methodology (11 pts) | Predictions (11 pts) | Inference (11 pts) | Conclusion (11 pts) | Additional Discretionary  Points | Numeric Grade | Letter Grade | Notes |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 4 | 11 | 11 | 9 | 11 | 11 | 11 | 11 | 11 | 11 | 0 | 97 | A |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**TERM PROJECT REPORT**

**STEM Salaries Prediction**

Under the Supervision of

Prof. Willard Williamson

Logo

Description automatically generated

Syracuse University, Syracuse,

New York – 13210

Submitted by Group 4:

Trishla Jain

Eric R. Pfeiffer

Kyle McGee

David A. Garcia

**Table of Contents**

[**1 - Abstract**](#h.yhevzrfpt16o)[**3**](#h.yhevzrfpt16o)

[**2 - Data Collection/Cleaning**](#h.79h0kcct68r8)[**3**](#h.79h0kcct68r8)

[**3 - Data Exploration Insights**](#h.1mt04ytwyjkg)[**4**](#h.1mt04ytwyjkg)

[**4 - Methodology**](#h.dfuf5ypc38d6)[**8**](#h.dfuf5ypc38d6)

[4.1 - RF/GBT Regression Methodology](#h.r47xplaxou7i) [9](#h.r47xplaxou7i)

[4.2 - RF/GBT Classification Methodology](#h.dz9ipemvjb0d) [10](#h.dz9ipemvjb0d)

[4.3 - Linear Regression Methodology](#h.abnuguhsw4mx) [11](#h.abnuguhsw4mx)

[4.4 - Logistic Regression Methodology](#h.hoxyzvn7vkt7) [12](#h.hoxyzvn7vkt7)

[**5 - Model Prediction**](#h.gl38vdldtd7x)[**13**](#h.gl38vdldtd7x)

[5.1 - RF/GBT Regression Prediction of Base Salary](#h.xosb4f5dazb7) [13](#h.xosb4f5dazb7)

[5.2 - RF/GBT Classification Prediction of Base Salary Threshold](#h.b6d9i8vjs2ou) [13](#h.b6d9i8vjs2ou)

[5.3 - Linear Regression Prediction of Base Salary](#h.4lm2o5oy1bt4) [14](#h.4lm2o5oy1bt4)

[5.4 - Logistic Regression Prediction of Base Salary Threshold](#h.quafn9ha2hi7) [15](#h.quafn9ha2hi7)

[**6 - Model Inference**](#h.d0kk4ic9ws8p)[**15**](#h.d0kk4ic9ws8p)

[6.1 - RF/GBT Regression Inference of Base Salary](#h.qz016zb5eb4y) [15](#h.qz016zb5eb4y)

[6.2 - RF/GBT Classification Inference of Base Salary Threshold](#h.uaee8fbmcv6q) [16](#h.uaee8fbmcv6q)

[6.3 - Linear Regression Inference of Base Salary](#h.sr2ag79sdeov) [17](#h.sr2ag79sdeov)

[**7 - Conclusion**](#h.pmmx25es7b8o)[**17**](#h.pmmx25es7b8o)

# **1 - Abstract**

As STEM students, we are constantly monitoring the job market. The goal of this project is to understand how different metrics and demographic variables affect STEM employees' salaries. By utilizing random forest, GBT, linear, and logistic regression models, we examine the importance of multiple characteristics in determining an employee’s base salary, such as company, location, years of work experience, and job title. The data used comes from Levels.fyi, a website where employees can post information about their jobs and themselves. Out of the 15 job titles in the dataset, software engineering managers had the highest average base salary as well as years of work experience. The end goal was to find the best model at accurately predicting base salaries. Additional information was gathered, providing insight into what features each model found to be most important. The predictions and inferences that were made on this data provide a wealth of information to current/future STEM employees.

# **2 - Data Collection/Cleaning**

Started in 2017, Levels.fyi is a website that promotes pay transparency in the tech industry by providing employees a platform to anonymously post information about their base salary, job title, years of work experience, and more. This project uses a dataset hosted on Kaggle.com of posts web scraped from Levels.fyi. The original shape of this dataset was 62,646 rows by 29 columns. There were 15 different job titles represented in this dataset. Some of the key columns used in this project are listed below in Table 1, and a data dictionary describing all of the original columns is provided in the appendix.

The initial cleaning process involved correcting the dataset’s schema as well as removing incorrect and/or unneeded values across many rows and columns. For example, the “basesalary”, “yearsofexperience”, and “yearsatcompany” columns were set to integer-type. Entries like “99”, “SE1”, and null values were removed from the “Title” column. Hardcoded “NA” and “null” values were removed from the “Education”, “Race”, and “Gender” columns, and values such as “0” and “1” were removed from the “Education” and “Gender” columns.

One major issue with this dataset was the inconsistency of company names across rows. Companies like JP Morgan & Chase, for example, were originally listed as “JP Morgan”, “JPMorgan Chase”, “JPMORGAN”, and more. With 1,636 unique values originally in the “Company” column, it was important to find a way to programmatically standardize company names. Our solution to this problem was a multi-step function that relied on Python dictionaries to reduce processing time. First, all values in the “Company” column were converted to be fully lowercase. Then, regular expressions were used to find “sub-patterns” across rows. If the string “JP'' was found in the strings “JP Morgan”, “JPMorgan Chase”, or “JPMORGAN”, then our function would either make or update the corresponding key-value pair. After processing the entire dataset, the key-value pairs were sorted alphabetically, and we manually went through each pair to find any company names that needed rectification. Once an incorrect name was found, we used the regexp\_replace pyspark function to correct the name. 61 companies had their names standardized, and ~1800 values were updated. After cleaning, the shape of the dataset was 21,581 rows by 19 columns.

# **3 - Data Exploration Insights**

The average base salary across all job titles was $133,852. Figure 1 shows the average base salary for each of the 15 jobs in this dataset. Managerial positions, such as software engineering manager, technical program manager, and product manager, grouped around high average base salaries. Of all jobs, only software engineering managers had an average base salary ($183,898) above the 3rd quartile ($165,000). Technical program and product managers fall above the median ($135,000) at $157,205, and $154,620, respectively. Positions like sales, recruiter, and business analyst fall below the median at $114,784, $114,341, and $104,381, respectively. None of the job titles in this dataset had an average base salary below the 25th percentile ($100,000).

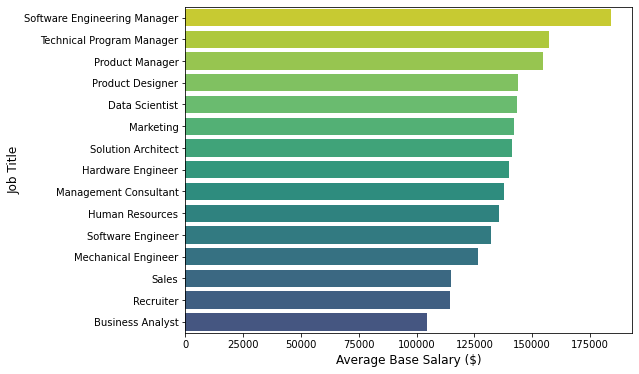


Figure 1. Average Base Salary in U.S. Dollars for all 15 Job Titles

Figure 2 shows the average number of years of work experience for each job. Like Figure 1, some of the managerial positions group around a high average number of years of work experience. Interestingly, while software engineering and technical program managers retained their 1st and 2nd place positions from Figure 1, program managers dropped to number 6. This could be partially explained by the high correlation between base salary and years of experience for product managers (0.451). However, technical program managers have a similarly high correlation (0.441) but did not change positions between Figures 1 and 2 (Table 1).

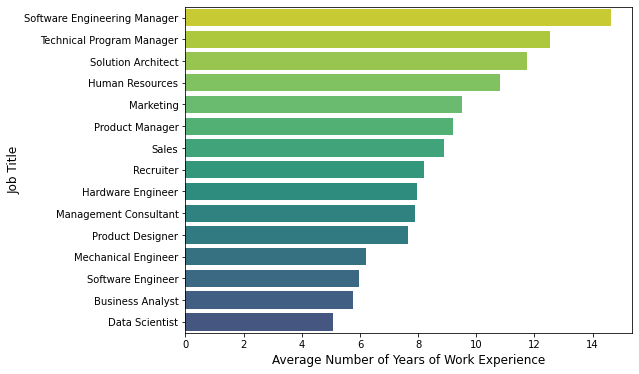


Figure 2. Average Number of Years of Work Experience for all 15 Job Titles

Table 1. Correlation between Base Salary and Years of Experience for Managerial Positions

|  |  |
| --- | --- |
| **Job Title** | **Correlation between Base Salary and Years of Experience** |
| Software Engineering Manager | 0.227 |
| Technical Program Manager | 0.441 |
| Product Manager | 0.451 |

When looking at the distribution of years of experience for each managerial position, the range is smallest for product managers (Figure 3). This helps to explain the change in position for product managers between average base salary and average years of experience. Combined with the fact that the average base salary for product managers is only $2,585 less than that of technical program managers, product management seems like a strong position for early/mid-career professionals trying to obtain a higher base salary.

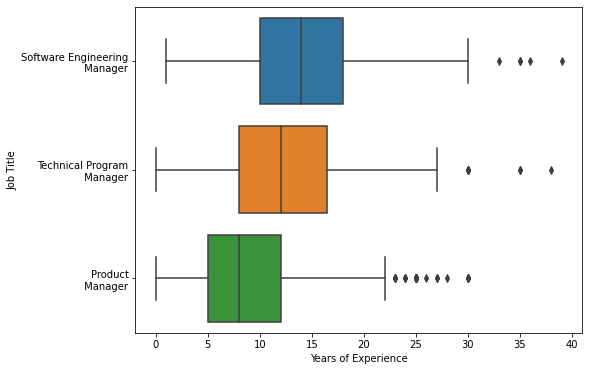


Figure 3. Distribution of Years of Experience for each Managerial Position

Figure 4 shows the average base salary by country. Countries with the highest average base salaries were Switzerland ($162,580), Denmark ($156,500), the United States ($149,706), the United Arab Emirates ($143,222), and Saudi Arabia ($140,000). With 78% of the observations in this dataset coming from the U.S., Figure 5was created to show the average base salary by U.S. state. The top five states with the highest average base salaries were California ($166,699), New York ($155,913), Washington ($150,147), Oregon ($138,530), and Massachusetts ($137,870).

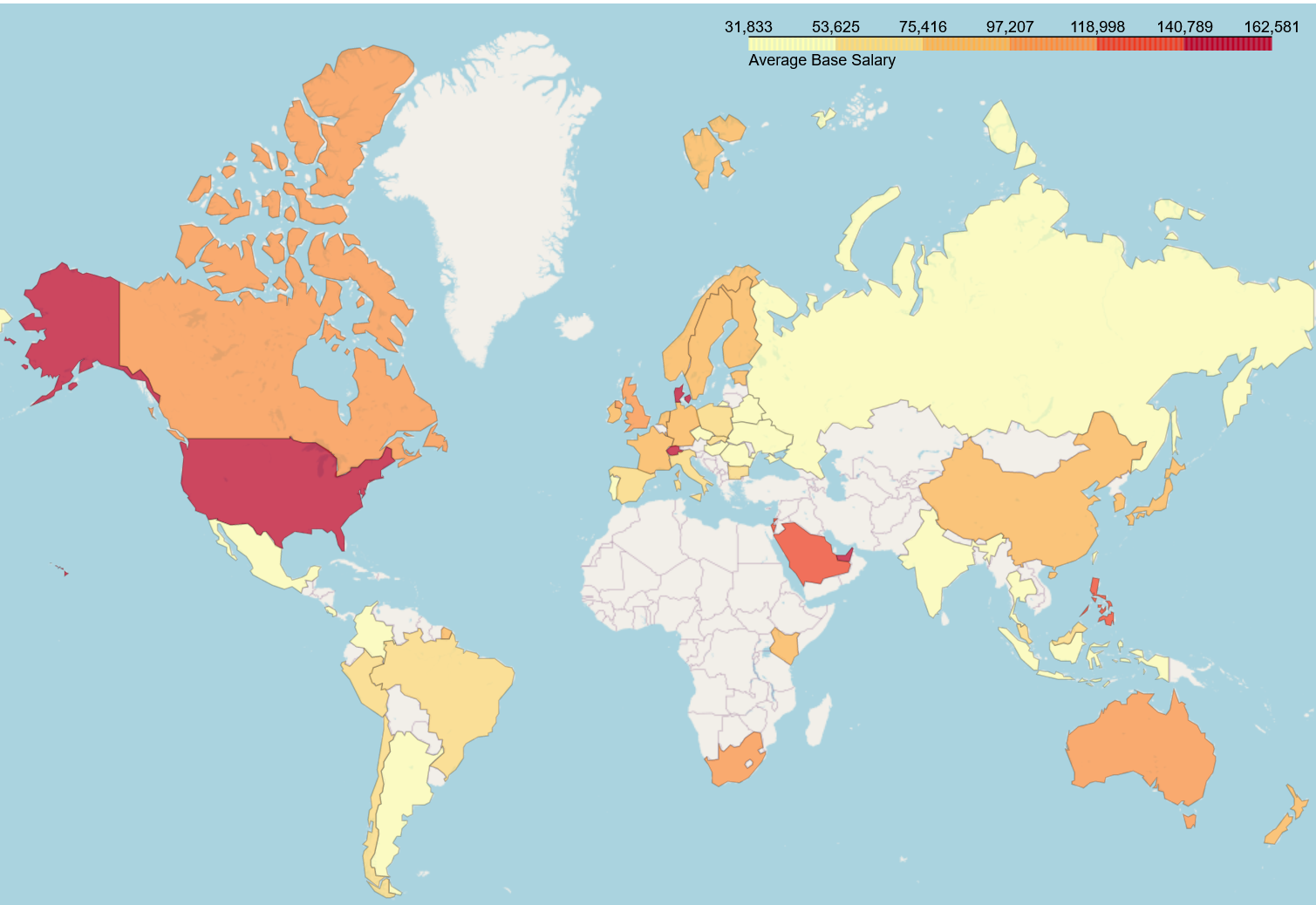


Figure 4. Average Base Salary by Country

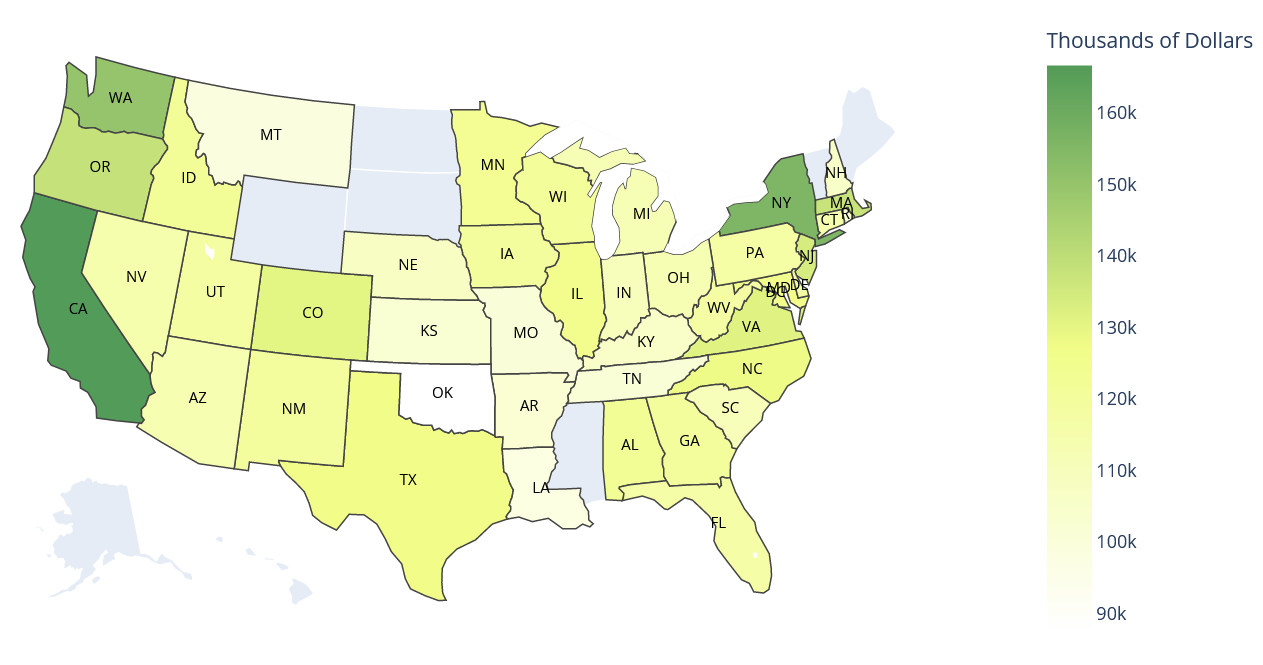


Figure 5. Average Base Salary by U.S. State

# **4 - Methodology**

The eight features used in our machine learning models are listed in Table 2. Six of these features were string-type data and needed to be indexed for use in random forest (RF), gradient boosting trees (GBT), linear regression, and logistic regression models. In addition to indexing these columns, we decided to subset the cleaned dataset by excluding companies that did not have at least 20 rows worth of data. After building an initial set of models, it was clear that the “Company” column was an important feature. We did not want our models to be influenced by high or low salaries from a company that appeared in the cleaned dataset only a handful of times. This threshold of at least 20 observations per company allowed us to keep ~16,833 rows of data and examine 145 unique companies in our models.

Each of the following subsections details the specific steps for each model as well as the scoring methodology used to evaluate them.

Table 2. Descriptions of Features Used in Models

|  |  |  |
| --- | --- | --- |
| **Column** | **Data Type** | **Description** |
| Years of Experience | Integer | Total # of Years of Work Experience |
| Years at Company | Integer | # of Years at Current Company |
| Company | String | Company Name |
| Title | String | Job Title |
| Country | String | Country of Employee’s Job |
| Gender | String | Employee’s Gender |
| Race | String | Employee’s Race |
| Education | String | Employee’s Highest Level of Education |

## **4.1 - RF/GBT Regression Methodology**

Figure 6 shows the steps used for training RF and GBT regression models. As described above, one of the first steps in training these models was to string index columns such as “Company”, “Title”, and “Country”. These string-indexed columns were then supplied to a vector assembler along with the two numerical features: years of experience and years at the company. After these two steps, RF and GBT regression models were ready to be trained.

Because there were 145 unique companies in the cleaned, subset data, the string-indexed “Company” column contained the most unique observations out of all of the features in the vector assembly. After creating train, test, and validation sets, we ensured that each of these three sets contained at least one value for each of the string-indexed companies. The RF and GBT regression models were scored using mean squared error (MSE). The best model was the one with the lowest MSE.

A picture containing timeline

Description automatically generated

## **4.2 - RF/GBT Classification Methodology**

Figure 7 shows the steps used for training RF and GBT classification models. Some slight differences needed to be accounted for when training classification models compared to regression models. Most importantly, a binary variable needed to be created based on the main predictor: base salary. The predictor in RF and GBT classification models was a salary threshold. Base salaries below $125,000 were encoded as a 0, and base salaries greater than or equal to $125,000 were encoded as a 1. $125,000 was chosen as the threshold for this stratification for several reasons. First, it was close to the mean of $135,000 without going over the mean. Second, ~39% of the observations in the subset data had base salaries equal to or less than $125,000, giving the overall dataset a slight preference towards salaries above this amount. Because one of our goals was to predict/identify features that contribute to very high base salaries, we wanted to pick a threshold were the majority of observations were higher than the threshold rather than below it. This stratification is shown in step 2 of Figure 7.

After stratifying the base salary column, the necessary columns were string indexed, put into a vector assembler along with years of experience and years at the company, and supplied to the RF and GBT classification models for training. As mentioned in section 4.1, after creating train, test, and validation sets, we ensured that each set contained at least one value for each of the companies in the string-indexed “Company” column.

The RF and GBT classification models were scored using a binary classification evaluator (accuracy). The best model was the one with the highest accuracy.

Timeline

Description automatically generated

## **4.3 - Linear Regression Methodology**

Timeline

Description automatically generated

The steps for training the linear regression models are shown above in Figure 8. Like with RF and GBT regression/classification models, one of the first steps was to remove companies that did not have at least 20 observations across the entire cleaned dataset. Next, columns containing string-type data were indexed. These string-indexed columns were supplied to a vector assembler along with the two numerical features: years of experience and years at the company. After these two steps, the linear regression model was ready to be trained. The linear regression model was scored using mean squared error (MSE).

## **4.4 - Logistic Regression Methodology**

The steps for training the logistic regression models are shown in Figure 9 below. For this model, the binary threshold of base salaries above or below $125,000 described in section 4.2 was used. This is also shown in step 2 of Figure 9 below. Additionally, a grid search was used to improve the performance of this model by tuning regularization hyperparameters. This model was scored using a binary classification evaluator (area under ROC). The best model was the one with the highest AUC.

Timeline

Description automatically generated

# **5 - Model Prediction**

## **5.1 - RF/GBT Regression Prediction of Base Salary**

The MSE results of RF and GBT regression models are shown in Table 3. Prior to tuning any hyperparameters, a base, out-of-package model was trained for both RF and GBT regression. These models were initialized with the train set and MSE was calculated after transforming on the validation set. The MSE of the base GBT model was ~281 million less than the MSE of the base RF model, showing a much stronger performance with GBT.

|  |  |
| --- | --- |
| Table 3: RF and GBT Regression MSE Scores | |
| **Predictor: Base Salary** | |
| **Model** | **Validation MSE** |
| GBT (Grid Search) | 916,326,206 |
| GBT (Base) | 988,787,916 |
| RF (Grid Search) | 1,011,243,792 |
| RF(Base) | 1,197,407,661 |

After building this initial set of models, a grid search was performed to see if hyperparameter tuning could be used to improve the MSE even further. During this process, each type of model was initialized with different parameters, trained on the train set, and scored using the test set. The parameters tuned for RF models were maximum depth, number of trees, and minimum instances per node. The parameters tuned for GBT models were maximum depth, maximum iterations, minimum instances per node, and step size.

After completing each grid search, both models were initialized with the best hyperparameters identified. The models were then trained using the train set and scored using the validation set. The best performing model out of the four shown in Table 3 was the GBT model with the following hyperparameters: Max Depth = 5, Max Iterations = 20, Max Instances = 15, and Step Size = 0.1.

## **5.2 - RF/GBT Classification Prediction of Base Salary Threshold**

The accuracy scores for our RF and GBT classification models are shown in Table 4 below. Like with our regression models, a base, out-of-package model was built first. After transforming on the validation set, the accuracy of each model's predictions was calculated. The base GBT model performed about 2% better than the base RF model at 92.79% vs. 90.49%.

|  |  |
| --- | --- |
| Table 4: RF and GBT Classification Accuracy | |
| **Predictor: Base Salary Threshold (0 or 1)** | |
| **Model** | **Validation Accuracy** |
| RF (Grid Search) | 0.934242 |
| GBT (Grid Search) | 0.931375 |
| GBT (Base) | 0.927936 |
| RF (Base) | 0.904979 |

To see if the accuracy could be improved, a grid search was performed for both models. During this process, each model was initialized with different parameters, trained on the train set, and scored using the test set. The parameters tested for RF models were maximum depth, number of trees, and minimum instances per node. The parameters tested for GBT models were maximum depth, maximum iterations, minimum instances per node, and step size. The best performing model out of the four shown in Table 4 was the RF model with the following hyperparameters: Max Depth = 20, Max Iterations = 25, and Min Instances per Node = 5.

|  |  |
| --- | --- |
| Table 5: Linear Regression MSE Scores | |
| **Predictor: Base Salary** | |
| **Dataset** | **MSE** |
| Train | 731,182,660 |
| Validation | 841,455,762 |
| Test | 742,129,538 |

## **5.3 - Linear Regression Prediction of Base Salary**

We created a linear regression model with the goal of predicting a given base salary according to the set of features listed in Table 2. Table 5 shows the MSE scores of training and testing this model. Figure 10 shows the results of this model. On the y-axis is the predicted salary, and on the x-axis is the observed base salary. The red line represents the predictions that were exactly the same as the base salary.

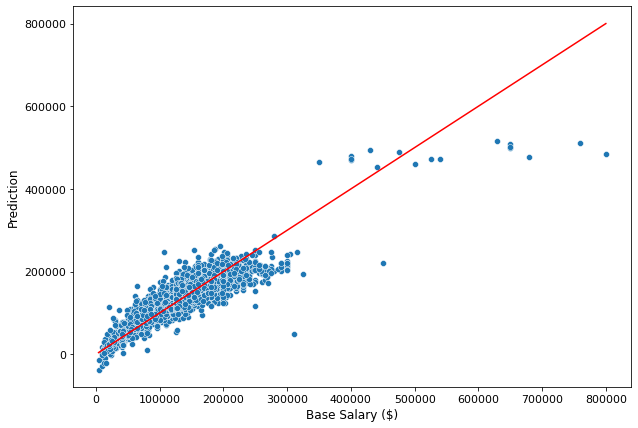


Figure 10. Deviation of Linear Regression Predictions from Original Base Salaries

## **5.4 - Logistic Regression Prediction of Base Salary Threshold**

The results of training and testing the logistic regression model are shown in table 6. The hyperparameters used for the best logistic regression model were alpha = 0.2 and lambda = 0.01.

Table 6: Logistic Regression MSE Scores

|  |  |
| --- | --- |
| **Predictor: Base Salary** | |
| **Dataset** | **Area Under ROC Score** |
| Train | 0.946232 |
| Test | 0.933573 |

# **6 - Model Inference**

## **6.1 - RF/GBT Regression Inference of Base Salary**

One of the main advantages of both RF and GBT models is that they provide a way to examine the most important features in each model. Table 7 shows the feature importance of the tuned GBT model from section 4.1.

|  |  |
| --- | --- |
| Table 7: Feature Importance of GBT Regression Model | |
| **Feature** | **Importance** |
| Company | 0.404130 |
| Country | 0.222952 |
| Years of Experience | 0.193428 |
| Title | 0.121646 |
| Education | 0.035666 |
| Years at Company | 0.012581 |
| Gender | 0.006025 |
| Race | 0.003572 |

Company, country, and years of experience have the highest feature importance. Interestingly, the GBT model assigned almost 20 times more importance to total years of experience than the number of years at a given company. Also, it is reassuring that this model found that the gender and race of an employee has comparatively very low importance when predicting base salary.

The tuned GBT model had an MSE of 916,326,206. This is a difficult metric to interpret without additional information, especially when this MSE is almost 300 million less than the MSE for a base RF model. Figure 11 shows the difference between predicted and actual salary from the validation set and offers an easier way to interpret this model. The x-axis in Figure 11 is zoomed in on a difference between $0 and $100,000 to show that over 75% of the predictions from this model are within $25,000 of the actual base salary.

## Chart, histogram Description automatically generated

## **6.2 - RF/GBT Classification Inference of Base Salary Threshold**

|  |  |
| --- | --- |
| Table 8: Feature Importance of RF Classification Model | |
| **Feature** | **Importance** |
| Country | 0.327419 |
| Company | 0.308475 |
| Years of Experience | 0.205219 |
| Title | 0.068283 |
| Years at Company | 0.038320 |
| Education | 0.026295 |
| Race | 0.019901 |
| Gender | 0.006087 |

Table 8 shows the feature importance of the tuned RF model from section 4.2. Country, company, and years of experience have the highest importance when determining if an employee’s salary is above or below $125,000. These features are the same as the ones in Table 7, although the “Country” and “Company” columns have changed positions. The repeated importance of these features emphasizes that an employee’s country, company, and years of experience are three of the key factors in determining their salary. In tables 7 and 8, these three features account for more than 80% of the feature importance.

## **6.3 - Linear Regression Inference of Base Salary**

|  |  |
| --- | --- |
| Table 9: Coefficients of Numeric Features in Dollars | |
| **Predictor** | **Coefficient ($)** |
| Years of Experience | 3,637 |
| Years at Company | 376 |

Which specific features create the highest base salary? The coefficients in Table 9 suggest that for each additional year of work experience, an employee’s base salary will increase by ~3,637 dollars. However, for each additional year of work experience at the same company, an employee’s base salary will increase by ~376 dollars. This seems to indicate that it is better to accumulate experience and then change companies rather than stay with the same company indefinitely.

# **7 - Conclusion**

The goal of this project was to understand the factors that determine salary for STEM employees. Below are our final observations:

1. Software engineering manager is the best job in terms of average base salary
2. A data scientist needs ~5 years of work experience on average for an average base salary of $150,000
3. Working in developed countries with very large tech industries such as the U.S., Denmark, and Switzerland can contribute to a higher base salary
4. Company, country, and total number of years of work experience have the most significant impact on determining an employee’s base salary
5. Switching companies after gaining some experience is a better way for employees to improve their salary than staying with the same company

**Appendix**

Table 10: Description of All Columns in the Original Dataset

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Timestamp | Datetime of When Post was Made on Levels.fyi |
| Company | Company Name |
| Level | Level within Company |
| Title | Role Title |
| Totalyearlycompensation | Total Yearly Compensation |
| Location | Job Location |
| Yearsofexperience | Years of Experience |
| Yearsatcompany | Years of Experience at Current Company |
| Tag | Tags (e.g., Full Stack, Software Engineer) |
| Basesalary | Base Salary |
| stockgrantvalue | Stock Grant Value |
| Bonus | Dollar Value of Bonus Given |
| Otherdetails | Free Form Text Field |
| Cityid | City ID |
| Dmaid | DMA ID |
| rownumber | Row Number |
| Masters\_Degree | 1 if yes, 0 if not |
| Bachelors\_Degree | 1 if yes, 0 if not |
| Doctorate\_Degree | 1 if yes, 0 if not |
| Highschool | 1 if yes, 0 if not |
| Some\_College | 1 if yes, 0 if not |
| Race\_Asian | 1 if yes, 0 if not |
| Race\_White | 1 if yes, 0 if not |
| Race\_Two\_Or\_More | 1 if yes, 0 if not |
| Race\_Black | 1 if yes, 0 if not |
| Race\_Hispanic | 1 if yes, 0 if not |
| Race | Race as a factor column |
| Education | Education as a factor column |

**Link to Dataset**

[**https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries**](https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries)