**DATA MINING**

**PREDICTING BEST TECH COMPANY BASED ON EMPLOYEE’S RATING**

**Munahil MurrieumPooja Bhansali**

**Table Of Contents**

|  |  |  |
| --- | --- | --- |
| **No.** | **Report Section** | **Page Number** |
|  | Summary | 2 |
|  | Introduction | 3 |
|  | Data Overview | 4-6 |
|  | Data Analysis | 7-13 |
|  | Model Application: | 14-35 |
|  | Model 1: Multiple Linear Regression Model | 13-17 |
|  | Model 2: Decision Tree | 18-21 |
|  | Model 3: Bootstrap Forest | 22-24 |
|  | Model 4: Boosted Tree | 25-27 |
|  | Model 5: Neural Network | 28-32 |
|  | Model Comparison | 33-35 |
|  | Conclusion | 36-38 |
|  | Thoughts | 39 |
|  | Future Goal | 40 |
|  | Bibliography | 40 |

# **SUMMARY**

Being from the tech background, we decided to analyze the dataset of 6 top most tech companies: Google, Amazon, Facebook, Apple, Microsoft and Netflix. We compared which one is the best company to work in from an employee’s perspective, based on the review of current and former employees of these companies. We used this dataset so that we can help HR consultant and recruiting companies to help people choose companies based on different review criteria’s and also what factors employers can incorporate to make their employees happy.

We choose our dataset from [www.kaagle.com](http://www.kaagle.com). This dataset contains 65534 records and 17 predictors. We ran different model approaches and then compared and contrasted them together. Our data types are numeric-continuous and some of the variables are characters. We partitioned our data set to 75% training data and 25% validation data. We used Multivariate Normal Imputation for least square prediction for the non-missing values in each row. Appendix 1 lists each of the variable names and its respective translation.

In our analysis, we used 5 different models and approaches, including Multiple Linear Regression Model, Regression Tree Model, Regression Tree with Bootstrap Forest Model, Regression Tree with Boosted Tree Model and Neural Network Model. Each approach provided useful insights and we mainly focused and compared few determining factors in our analysis, which are RSquare and RMSE/RASE. After comparing all five models mentioned, we learned that Neural Network model can best predict overall rating because it has the lowest RASE/RMSE value and highest RSquare value for the validation data.

We also found out that Facebook has the overall highest rating compared to other companies, which indicates that the employees are happiest there.

# **INTRODUCTION**

Work and Workplace are a big part of our lives, and often they became a significant factor in our identities. We will see how different things that matter to employees relate to each other. We use reviews from all over the world to have a reasonable similarity yet diverse enough to consider as many people as possible.

We compared 6 corporations: Amazon, Apple, Facebook, Google, Microsoft and Netflix. These are companies which are leading the tech industry and many people dream of working in them. We focused on which of them is the best from an employee’s perspective. This data will help the HR/Recruiting companies in providing insights to the perspective employees when choosing from these companies. This will also help employers see what employees most are interested in when looking for the job.

Differential participation rates tell us what issues matter most to our people. Predictive Analysis can enable a customized employment value proposition that maximize mutual benefits for our organizations and their talent.

We will see how different things that matter to employees relate to each other.

We decide to Predict the Reviews from all over the world to have as many Employees as possible. Let’s find out more about some of the largest employers using the review of current and former employees of all these companies. This analysis will focus on all the different aspects of the ratings.

We will see how different things that matter to employees relate to each other. We use reviews all over the world to have a reasonably similarity yet diverse enough to consider as many people as possible.

**DATA OVERVIEW**

These datasets compare 6 top most corporations: Google, Amazon, Facebook, Apple, Microsoft and Netflix. This data is comparing the 6 largest companies, but we are comparing which one is the best from an employee perspective. Let's find out more about some of the largest employers using the review of current employees of these companies. This analysis will focus on all the different aspects of ratings.

Datasets is for 17 predictors and numbers of records are 65534.

Modelling type of “Predicting Best Tech Company based on Overall Rating”.

Data Type are “Numeric-Continuous” and some of variables are “Character”.

Data are partitioned:75% Training Data and 25% Validation Data.

We mainly used numeric variables for our dataset. We used Culture Values, Career Opportunities, Senior Management, Work Balance, Compensation Benefits, Job Title, Company, Location and Helpful Count to predict Overall Rating. In general, we found out that biggest correlation is between Culture Values and Overall Rating. We also found out that employers gave higher rating to Compensation Benefits. Therefore, we can say that employers mostly are happy when they get paid leaves. Overall Rating comparison showed us that Facebook has the highest rating followed by Google and Apple.

For our project we will predict Overall Rating based on the variables mentioned above. We will only be using Numeric variables for now. We will use different models to see which one provides us with the best results. We will mainly compare these models based on RSquare and RMSE/RASE value.

The rating is based on a widespread scale from 1 the worst to 5 the best. Some rating values in rating columns are in between levels, i.e. 1.5, 2.5, 3.5, or 4.5. We correct these rating by lowering them one level. E.g., 1.5 to 1, 2.5 to 2. Note There are no 0.5 nor 5.5 ratings to be corrected. Our approach is conservative here, and all mid-stars ratings are truncating to lower star rating. The g

|  |  |
| --- | --- |
| Company | Company Name (Numeric and Continuous Modeling type) (0 = Amazon, 1=Apple, 2=Facebook, 3=Google, 4=Microsoft, 5=Netflix) |
| Location | This Datasets is global, that’s why we include only the Country’s name. ( 1 = Africa, 2 = Algeria, 3 = Arabia, 4 = Argentina, 5 = Australia, 6 = Austria, 7 = Bahrain, 8 = Bangladesh, 9 = Belgium, 10 = Brazil, 11 = Bulgaria, 12 = Canada, 13 = Chile, 14 = China, 15 = Colombia, 16 = Denmark, 17 = Egypt, 18 = Emirates, 19 = Estonia, 20 = Ethiopia, 21 = Faso, 22 = Finland, 23 = France, 24 = Germany, 25 = Ghana, 26 = Greece, 27 = Guatemala, 28 = Guyana, 29 = Herzegovina, 30 = Hungary, 31 = India, 32 = Indonesia, 33 = Iran, 34 = Iraq, 35 = Ireland, 36 = Islands, 37 = Israel, 38 = Italy, 39 = Jamaica, 40 = Japan, 41 = Jordan, 42 = Kazakhstan, 43 = Kenya, 44 = Kong, 45 = Korea, 46 = Lanka, 47 = Latvia, 48 = Lebanon, 49 = Libya, 50 = Luxembourg, 51 = Malaysia, 52 = Maldives, 53 = Mauritius, 54 = Mexico, 55 = Morocco, 56 = Netherlands, 57 = Nigeria, 58 = Norway, 59 = Pakistan, 60 = Peru , 61 = Philippines, 62 = Poland, 63 = Portugal, 64 = Qatar, 65 = Republic, 66 = Rica, 67 = Romania, 68 = Russia, 69 = Rwanda, 70 = Salvador, 71 = Serbia, 72 = Singapore, 73 = Slovakia, 74 = Spain , 75 = Sweden, 76 = Switzerland, 77 = Taiwan, 78 = Tanzania, 79 = Thailand, 80 = Tunisia, 81 = Turkey, 82 = Uganda, 83 = UK, 84 = Ukraine, 85 = USA, 86 = Venezuela, 87 = Vietnam, 88 = Zealand) (Numeric and Continuous Modelling type) |
| Year Posted | Year is mention instead of Dated when the reviews is posted. |
| Job-Title | This string is including whether the reviewer is a ‘Current Employee’ or ‘Former Employee’ at the time of review. |
| Summary Sentiment | Short summary of Employee review. (Character Data Type) (Not Included in Variable) |
| Pros | Employee Reviews Pros (Character Data type) (Not Included in Variable) |
| Cons | Employee Reviews Cons (Character Data type) (Not Included in Variable) |
| Advice to Management | Advice given by reviewer (Character Data type) (Not Included in Variable) |
| Overall Rating | 1-5 |
| Work/Life Balance stars | 1-5 |
| Cultures and Values stars | 1-5 |
| Career Opportunities stars | 1-5 |
| Comp and Benefits stars | 1-5 |
| Senior Management stars | 1-5 |
| Helpful Review Count | A count of how many people found the review to be helpful |
| Link to Review | This will be provided you with a direct link to the page that contains the review. However, it is likely that this link will be outdated. |

**DATA ANALYSIS**

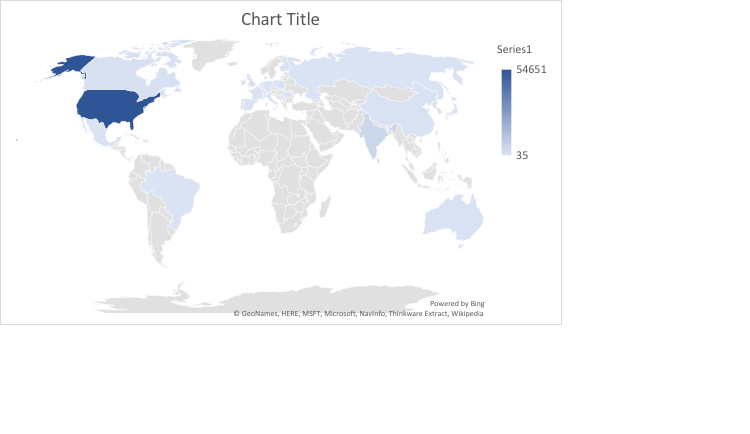
**NUMBER OF REVIEWS:**

|  |  |  |
| --- | --- | --- |
| No. | Company Name | Number of Reviews |
| 1 | Amazon | 22080 |
| 2 | Microsoft | 12134 |
| 3 | Apple | 10248 |
| 4 | Google | 5604 |
| 5 | Facebook | 1438 |
| 6 | Netflix | 494 |



The Most reviews in these datasets are from Amazon, Microsoft and Apple. Facebook and Netflix don’t have so many reviews, but around thousand is still enough to making good analysis.

**LOCATION**



This Datasets is global, that’s why we include only the Country’s name and this dataset are from 89 different countries.

The Majority Employees in these datasets are from

|  |  |
| --- | --- |
| USA | 54651 |
| INDIA | 4788 |
| UK | 1493 |
| Ireland | 1014 |
| Canada | 673 |

**OVERALL-RATING**

Summarizing the column ‘Overall Rating’: Let’s start with the column overall rating, is the column that gives us a general idea of the impression of those five companies.

|  |  |
| --- | --- |
|  | Average rate |
| Amazon | 3.6057 |
| Apple | 3.9589 |
| Facebook | 4.5180 |
| Google | 4.3784 |
| Microsoft | 3.8650 |
| Netflix | 3.4979 |

We can see that overall Facebook has the highest overall rating. People are happy working with Facebook

**Comparing all the Rating variable by COMPANY:**

Amazon have a good Compensation Benefits and Career Opportunities for their employees.

Apple have a good Cultures Values and Compensation Benefits for their employees.

Facebook have a good Compensation Benefits and Career Opportunities for their employees.

Google have a good Cultures Values and Compensation Benefits for their employees.

Microsoft have a good Compensation Benefits for their employees.

Netflix have a good Compensation Benefits and Career Opportunities for their employees.

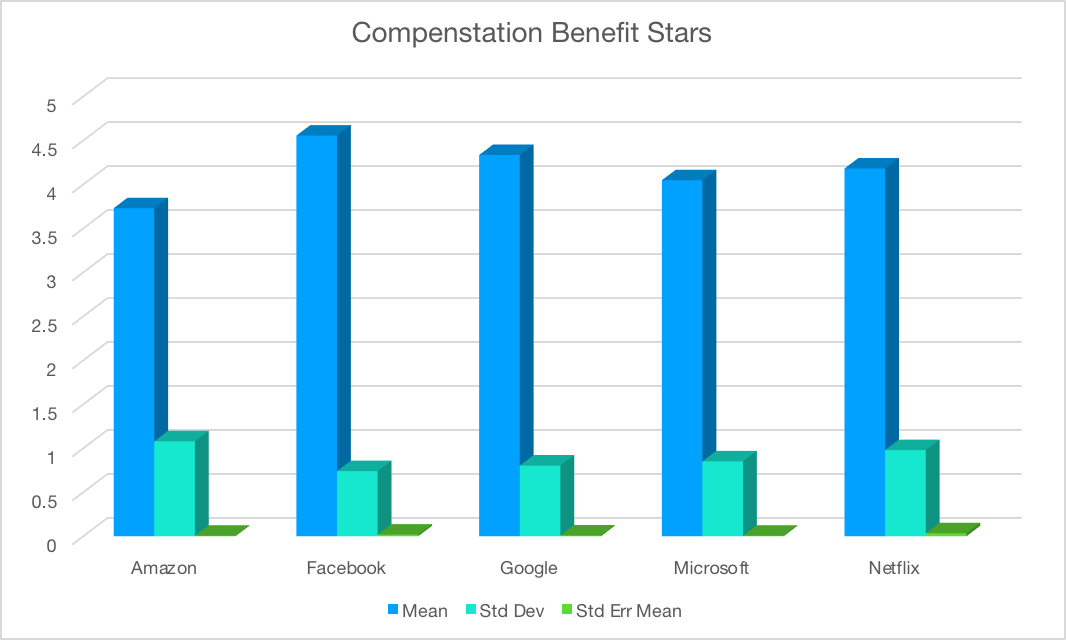
**Compensation Benefit Stars:**

Majority of the company’s employee higher review on Compensation Benefits stars except Apple.

Each employee believes that the existing company have a good compensation Benefits than the others benefits provided by companies.

Now we compare which company provide better compensation benefit than the others company.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Std Dev | Std Err Mean |
| Amazon | 3.73 | 1.080 | 0.0066 |
| Facebook | 4.555 | 0.741 | 0.0186 |
| Google | 4.335 | 0.804 | 0.009 |
| Microsoft | 4.048 | 0.852 | 0.0067 |
| Netflix | 4.181 | 0.980 | 0.0344 |
|  |  |  |  |

****

Compensation company benefits is what employers can get in this job not counting on salary. Facebook and Google offer best compensation benefits to the employees.

In this aspects Amazon needs to improve in Company compensation benefits to employee. Employees mostly are happy about their pay, or maybe this means tech employees only work where they feel happy about their pay.

This mean very good situation in this regard for all companies.

**DATA MODELLING**

## **Model 1: MULTIPLE LINEAR REGRESSION**

We created a regression model with “Overall-Ratings” as the dependent variable (response) and Company, Location, Job-Title, Work-Balance-Stars, Culture-Values-Stars, Carrer-Oppurtunities-Stars, Comp-Benefit-Stars, Senior-Managemet-Stars, Helpful-Count as independent variables.

We used the Effect Summary table in the Response "Overall-Ratings" output to remove predictors with p-value greater than 0.01. The resulting 9 predictors (all their p-values < 0.01) are listed in the table below.

**Effect Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **LogWorth** |  | **PValue** |
| Culture-Values-Stars | 1287.941 | https://lh6.googleusercontent.com/nx0itSIcT0uHacDAMaFbsNm-uyg1wT3U2YDa21An_6P4YVlCxoPfGAHO16i7KATaX61YiPjH_Trrtyn4P30PNhca8yUudg2R_lTLuD9Mxve7Y_WmBDPvICxd9Yv4I4oLi8P_gTXCihPuROUR4w | 0.00000 |
| Carrer-Opportunities-Stars | 892.054 | https://lh6.googleusercontent.com/GUYPVQiLWvKFYNm6hn1SB21TBh6rgLhwp3C8GlFUEoRl9H4sB28a9NpAVceP6XbSX6kvt6qVAxsY0W_aa60wrwMyaAjle93hN9OAO98av9r_ci2KyMlXs63dMhqcDoS8VZebhVs1Nd91MmfPpw | 0.00000 |
| Senior-Mangemnet-Stars | 364.346 | https://lh6.googleusercontent.com/GUYPVQiLWvKFYNm6hn1SB21TBh6rgLhwp3C8GlFUEoRl9H4sB28a9NpAVceP6XbSX6kvt6qVAxsY0W_aa60wrwMyaAjle93hN9OAO98av9r_ci2KyMlXs63dMhqcDoS8VZebhVs1Nd91MmfPpw | 0.00000 |
| Work-Balance-Stars | 297.302 | https://lh6.googleusercontent.com/GUYPVQiLWvKFYNm6hn1SB21TBh6rgLhwp3C8GlFUEoRl9H4sB28a9NpAVceP6XbSX6kvt6qVAxsY0W_aa60wrwMyaAjle93hN9OAO98av9r_ci2KyMlXs63dMhqcDoS8VZebhVs1Nd91MmfPpw | 0.00000 |
| Comp-Benefit-Stars | 153.866 | https://lh6.googleusercontent.com/GUYPVQiLWvKFYNm6hn1SB21TBh6rgLhwp3C8GlFUEoRl9H4sB28a9NpAVceP6XbSX6kvt6qVAxsY0W_aa60wrwMyaAjle93hN9OAO98av9r_ci2KyMlXs63dMhqcDoS8VZebhVs1Nd91MmfPpw | 0.00000 |
| Job-Title | 36.061 | https://lh5.googleusercontent.com/nJuTxhGH_H1Byv0jsRW86vpcq--Z0P1dXsVbbTVlIEdAYvQiLZDcdPSJdHf2jPMaXpAW-oWUJmZX2NJZKDsrKgQ0kUQ-Tqo561pg0AZTrCHHpYWEkUS7BPT1pzmRxFh1WPNFNtiH3nflI30qyQ | 0.00000 |
| Company | 33.443 | https://lh4.googleusercontent.com/py6c5uCI3UYjJcDUV8PkGqwCfnw7aPuOUFn7gZnxoqVJvwN5qLM7P7jkpsn7Hv8USYfvtG2DbuDXgfE3ADBT-NVq9U7VtLaF0dBYOEkbcTHFXOu32bQaUSFgIqgPg0mszn0n-Tlqfw0LaPGPwA | 0.00000 |
| Location | 7.627 | https://lh4.googleusercontent.com/OvLncnCN2CYBz-Jl8Ij_Mm0RV_yvRZnrBuvqJMJlhwA-mBKRG1zJ9FPkY0w8WmrGCIX0odTpFtXrl_Tyd93-oLfHH5csNqx4DyS8n88oHVkUrEKPbhxgjVspBTHGwyazdNSKfXEqrOtCiOdo6Q | 0.00000 |
| Helpful-Count | 5.209 | https://lh5.googleusercontent.com/A7B-qs1UO22CLgI3HVaQ1ooi6DR5aJ1HVS1I8mIXhVUND0BZyRLwaeuvNI3HpsARpacmlVyB0cFu8x7R7TIFzhz7LOAcAlNSj4S1lgm76L10VzuJjqMIFPURXfn58Oy739qAEzS9mdHd_CWfTw | 0.00001 |

**Summary of Fit**

|  |  |
| --- | --- |
| RSquare | 0.650192 |
| RSquare Adj | 0.650128 |
| Root Mean Square Error | 0.686657 |
| Mean of Response | 3.825456 |
| Observations (or Sum Wgts) | 49237 |

**Analysis of Variance**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Ratio** |
| Model | 9 | 43141.522 | 4793.50 | 10166.53 |
| Error | 49227 | 23210.451 | 0.47 | **Prob > F** |
| C. Total | 49236 | 66351.973 |  | <.0001\* |

The model, has a relatively good RSquare = 0.0650 which means that 65% of variation in *Overall-Ratings* can be explained by variations of the following predictors: *Company, Location, Job-Title, Work-Balance-Stars, Culture-Values-Stars, Carrer-Oppurtunities-Stars, Comp-Benefit-Stars, Senior-Managemet-Stars, Helpful-Count* Ratio has a probability of <0.0001 . P-values for the predictors’ regression coefficients are less than 0.01, and thus these coefficients are statistically significant.

The regression equation is statistically significant with RSquare = 0.650, F Ratio probability less than 0.01, and all p-values of the regression coefficients less than 0.01. The regression equation is the following:

**Parameter Estimates**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Estimate** | | **Std Error** | **t Ratio** | **Prob>|t|** |
| Intercept | | 0.4271579 | 0.019556 | 21.84 | <.0001\* |
| Company | | 0.0230756 | 0.001892 | 12.20 | <.0001\* |
| Location | | -0.00094 | 0.000168 | -5.58 | <.0001\* |
| Job-Title | | -0.082568 | 0.006512 | -12.6 | <.0001\* |
| Work-Balance-Stars | | 0.1225737 | 0.003299 | 37.15 | <.0001\* |
| Culture-Values-Stars | | 0.3178946 | 0.004007 | 79.33 | <.0001\* |
| Carrer-Opportunities-Stars | | 0.239449 | 0.003662 | 65.38 | <.0001\* |
| Comp-Benefit-Stars | | 0.1023865 | 0.003852 | 26.58 | <.0001\* |
| Senior-Mangemnet-Stars | | 0.1635559 | 0.003968 | 41.21 | <.0001\* |
| Helpful-Count | | -0.000819 | 0.000181 | -4.52 | <.0001\* |

**PROFILER**



Overall-Ratings = 0.4271579 - 0.023\*Company - 0.00094\*Location - 0.0825\*Job-Title + 0.12257\*Work-Balance-Stars + 0.3178946\*Culture-Values-Stars + 0.239449\*Carrer-Oppurtunities-Stars + 0.1023\*Comp-Benefit-Stars + 0.163555\*Senior-Managemet-Stars - 0.000819\*Helpful-Count

**Cross-Validation**

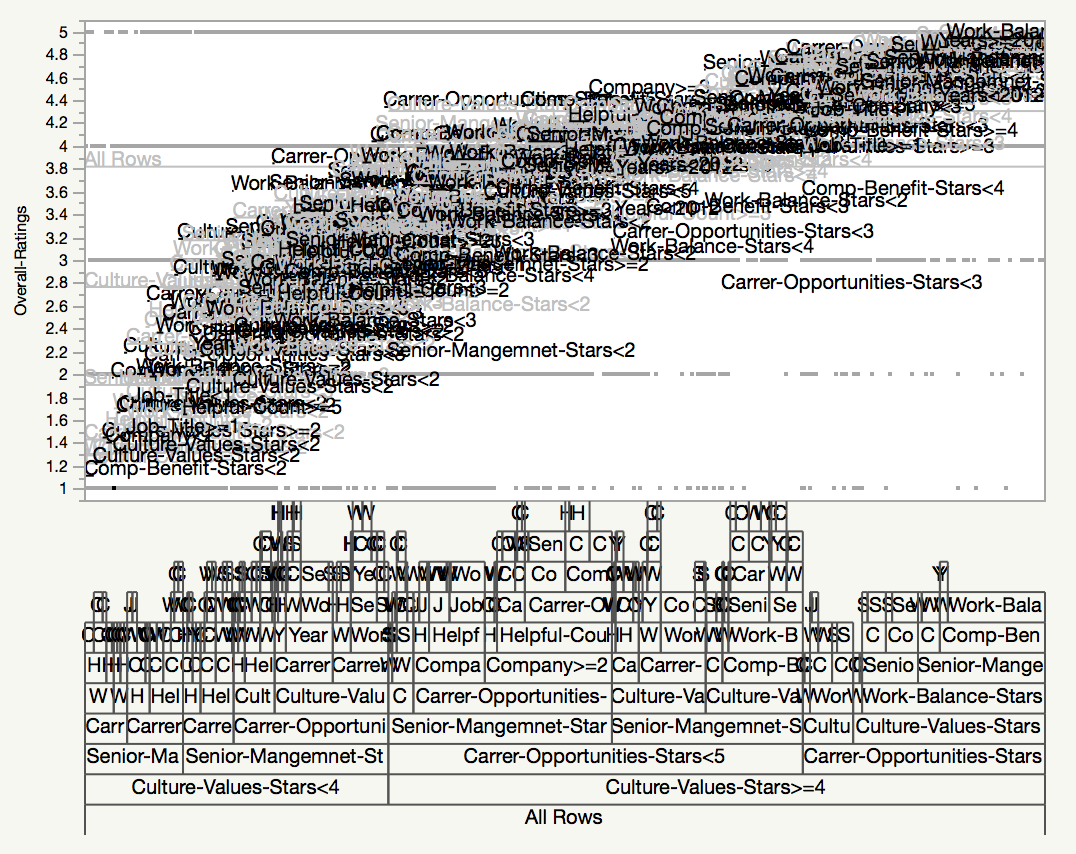
|  |  |  |  |
| --- | --- | --- | --- |
| **Source** | **RSquare** | **RASE** | **Freq** |
| Training Set | 0.6502 | 0.68659 | 49237 |
| Validation Set | 0.6532 | 0.67769 | 16297 |

The Cross-validation table below shows that the RSquare for the validation set is higher (0.6532) than that of the training set (0.6502), and the increase is only 0.461% ((0.6532-0.6502)/0.6502). Decrease in RASE for validation set vs. training set is also small, 1.29% ((0.67769-0.68659)/0.68659). Therefore, the multiple regression model fits well into the validation set and will not be an overfitting issue.

## **Model 2: DECISION TREE (Regression Tree)**

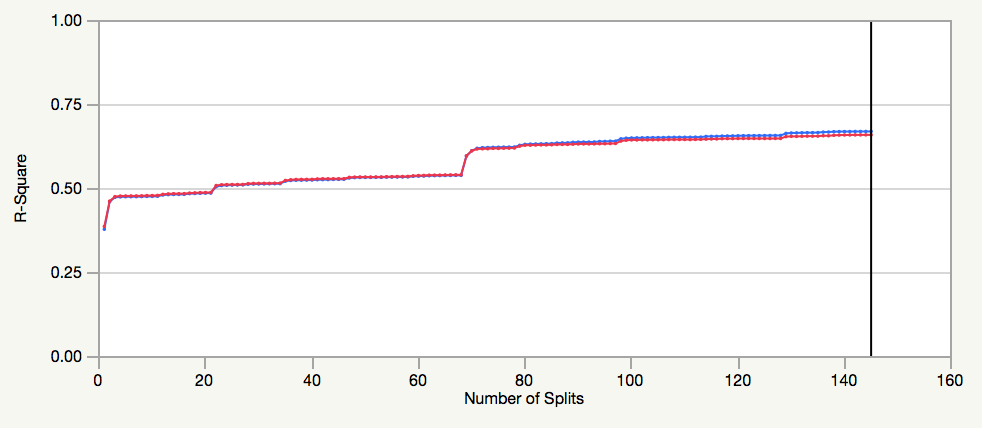
A decision tree was also trained on the dataset and the fit details of the model are seen below. As seen in the results, we observe a RSquare value for training and validation data at 0.671 and 0.661 respectively, which tell us that this model does not suffer from overfitting. Moreover, we see that there were a total 145 splits used for the final regression tree. Moreover, we observe that Training and Validation data RMSE values are at 0.6655 and 0.6696; which will be used for model performance comparisons.

**Partition for Overall-Ratings**



|  | **RSquare** | **RMSE** | **N** | **Number of Splits** | **AICc** |
| --- | --- | --- | --- | --- | --- |
| Training | 0.671 | 0.6655108 | 49237 | 145 | 99924.8 |
| Validation | 0.661 | 0.6699677 | 16297 |  |  |

**Split History**



Validation Data in Red

From the split history, we can see that the validation data (red line) is split with n = 145 as this represents the most optimal split value which was chosen due to its highest R Square value.

**Prediction Profiler**



A screenshot of a cell phone

Description automatically generated

From the leaf report we can Culture Value stars < 4 & Senior Management Stars < 2 & Career Opportunities Stars < 2 & Work Balance Stars < 2 & Helpul Count >=1 & Compenstation Benefits Stars < 2 than MEAN will be 1.12 and Count will be 453 records.

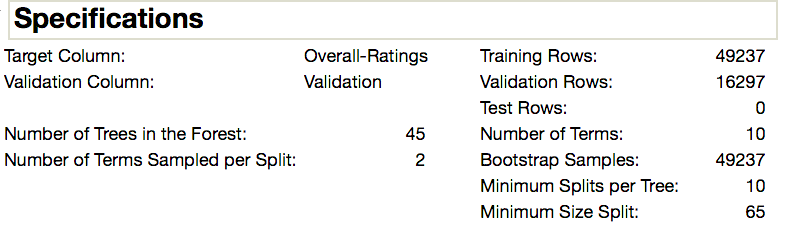
**Column Contributions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Term** | **Number of Splits** | **SS** |  | **Portion** |
| Culture-Values-Stars | 20 | 27194.7513 | https://lh6.googleusercontent.com/AM9SWEG9dirVhPxS-8rOSTkK3StkVCDHGaR4m-v22s0oGYzlq6hUxa8TgvxTNv49ZgADbzUzoawLzVUaL0SZ5z4ChsQ0wa6iRmPFp2Yd4Na0oLM9a5SRjstIRr3lv_k3ioGyhdv21uNoyrb7vw | 0.6105 |
| Carrer-Opportunities-Stars | 19 | 7114.4857 | https://lh4.googleusercontent.com/YNBECQR1N6Y4CkQYxVTr2R5SkoV__IQBhFsP3agED-KzH_ueuY7N5Zl8IXsfum45QO3CDUdYRrJ3KK-da4fR046ZO75MpYysXKXMnv3sXB6jaYJ6lcVtbznVzgr9sgRHhzvMBWIKY00vLA0hJA | 0.1597 |
| Senior-Mangemnet-Stars | 19 | 6859.28101 | https://lh6.googleusercontent.com/PjE-udLx3MWr7jtw7ClHZ-eSOz2gd819E9VWcDCrvq8zljnpknNM977QTCKdZr7aKZECdH9I7JbaqMlKWB29PUgxd089_8p-HtAukacm2hC7HEy-4oKzDsZlSmeNs5KUEA3wPjQC6HaE5z_wKA | 0.1540 |
| Helpful-Count | 15 | 1160.53246 | https://lh6.googleusercontent.com/pu8sln7dUNgxhDnotdMJEP9dhVGQ1Mgj7as4DlS34B5UPFRaWezMKwiSX0rhxpiWeyuwEQBM7DeVCVd6T4In9g5jxjBqyJ58BDqp83YKegAY0HS9Y8g7tTwzcVCkJC_U-DJfRKAHIALcd5Y3Fw | 0.0261 |
| Work-Balance-Stars | 37 | 1119.8963 | https://lh3.googleusercontent.com/lDDnmI0FLARI6SNB_y7sD1gdRE5HCo1WO6ATDWO5yzRsvlVOW-xNE4fE9fGeAq3wpncSkqhDM_tklZKtx_PvnvWGKo8WyHDcl-NPydUsxBYNxEB4TlYDon68Uy6NPHy4ZmF8Sl5jLPzG0tK13g | 0.0251 |
| Company | 8 | 530.28314 | https://lh6.googleusercontent.com/rRDfTEP1NFpIYD-ZIL4pyIwbXRTaaRg5uTK8vn3I8OZhkqULj0dEz-AqH9STfTlgOSaZi44TucNUMW-8Q0Lq6Tddzinhn-ehAlMm1COw40lsHo_xXR6Bisgq8I7i6jBvTv4UkuKWr1oH_Oyipg | 0.0119 |
| Comp-Benefit-Stars | 16 | 319.151511 | https://lh4.googleusercontent.com/KcnYy--BJrXk-9sHIjMDxGMm_Azfl8a22F1h5gSKp-ZjcfuPlk0zFuZ-laxWLFm32YJdUzyoCiZ6yCt_OR7-Y3b4BTNGvehdYTPZagI6yj1ium9F8qU1VA8TawNoRVWVZD10T8_R4Ut_G6n2-Q | 0.0072 |
| Years | 7 | 154.699939 | https://lh4.googleusercontent.com/YUp42xAH8k_LL24aNPZfsimwwA5FHUyLMlrwUPOby-13CE3ghWRUB3K0af-rq9g3-2vXF0jw46nnFexx549ppszqnUHlhN41T-PbPhkAFx40nIGIW6UPD434pgdbtIyxv8vv9xkFSfdGQB62yg | 0.0035 |
| Job-Title | 4 | 91.5977467 | https://lh4.googleusercontent.com/YUp42xAH8k_LL24aNPZfsimwwA5FHUyLMlrwUPOby-13CE3ghWRUB3K0af-rq9g3-2vXF0jw46nnFexx549ppszqnUHlhN41T-PbPhkAFx40nIGIW6UPD434pgdbtIyxv8vv9xkFSfdGQB62yg | 0.0021 |
| Location | 0 | 0 | https://lh6.googleusercontent.com/rLOiAluDeDqM8fjB4kgGaeBCKkyRA-TQ0ZaXSMg2ITRdoY6n5ydDjq-d3elvJAFNHZuPazqNUq3FYXYYev0I71s34-uDg9_IRSZEVSwmOX4_ioHdG_f6go2Uvna1vJh0sMA3YZZHPuJHNTzHiw | 0.0000 |

We observe from the column contributions the predictors that have been used in the decision tree. We can see that the feature with the highest contribution to the *Overall-Ratings* is Culture-Values-Stars, Career-Opportunities-Stars and Senior-Management Stars.

**Model 3 : Bootstrap Forest**

To further improve the performance of the decision tree, we implemented a bootstrap forest. A bootstrap forest is a variation of a random forest in which random predictors and samples are drawn from the dataset which is then repeated until the average of all trees is produced. From the snippet below we can see that there was a total of 45 trees used in the forest with 10 terms sampled per split.

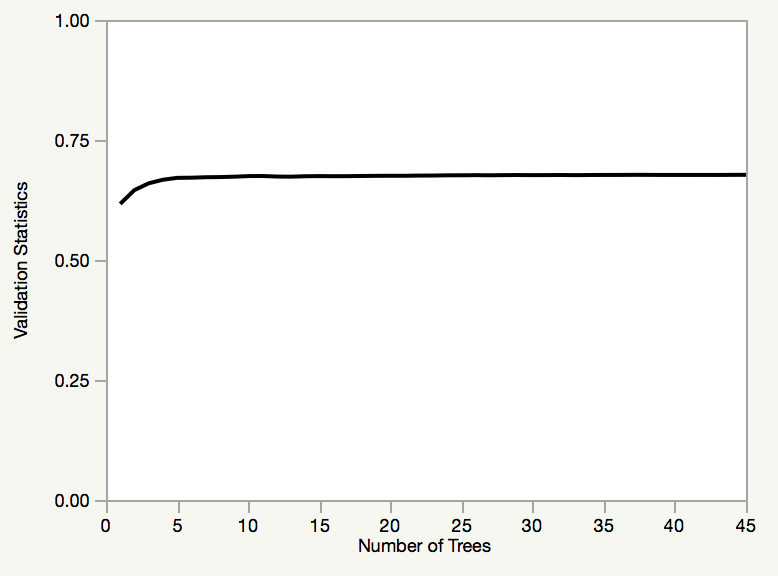
****

From the fit summary, we can see that the RSquare and RMSE values for both training and validation data have made improvements. We can see that the RSquare has improved using the bootstrap as compared to the regression tree as our model shows an increase to 68.4% of variation in *Overall-Ratings* is explained by the predictors for the training dataset and 67.8% for the validation dataset. Moreover, we observe that RMSE is lower at 0.6524 and 0.6525 for training and validation data respectively

|  |  |
| --- | --- |
| **Individual Trees** | **RMSE** |
| In Bag | 0.5617385 |
| Out of Bag | 0.7171377 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RSquare** | **RMSE** | **N** |
| Training | 0.684 | 0.652403 | 49237 |
| Validation | 0.678 | 0.6525462 | 16297 |

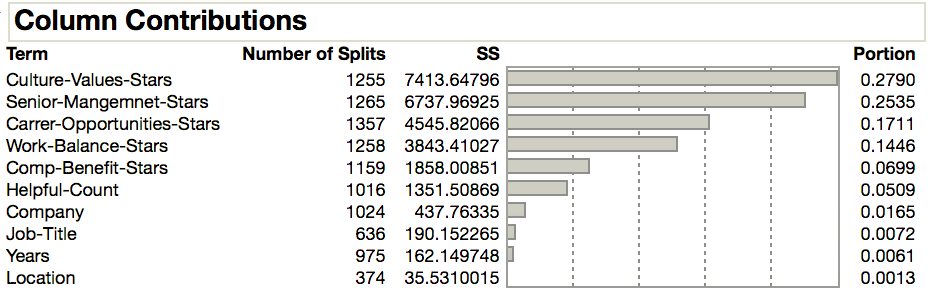
**Cumulative Validation**



**Prediction Profiler**



From the column contributions, we can see that more predictors were used as compared to the decision tree, which is because bootstrapping method has been applied. The involvement of more predictors in the model to predict *Overall-Ratings* influences the increase in RSquare and lowering of RMSE.



## **Model 4: BOOSTED TREE**

Another improved model of a decision tree is the boosted tree. Like a bootstrap forest, a boosted tree is a small tree that is repeatedly fit to the data using a random sample of predictors. The difference is that boosted trees make use of residuals based on misclassifications or errors which are reduced until an optimal tree is generated. From the snippet below we can see that a small tree was used with 3 splits and 50 layers were used for the model.

**Specifications**

|  |  |
| --- | --- |
| Target Column: | Overall-Ratings |
| Validation Column: | Validation |
| Number of Layers: | 50 |
| Splits per Tree: | 3 |
| Learning Rate: | 0.1 |

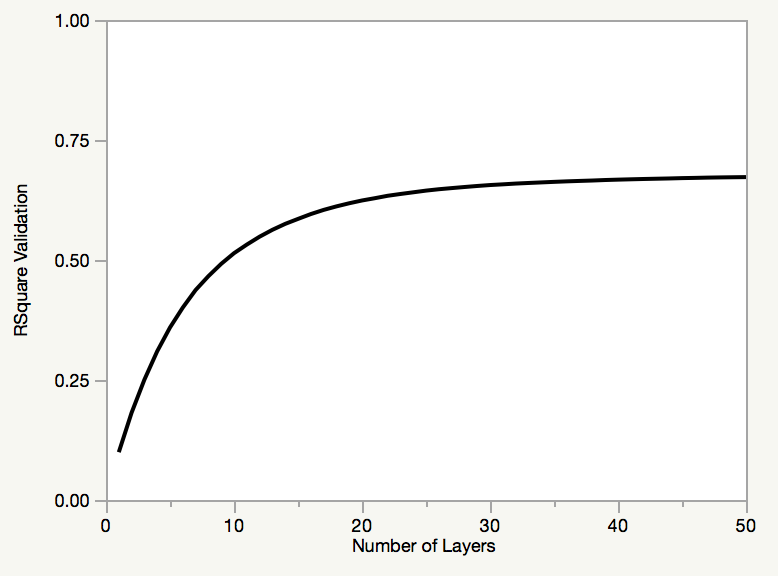
|  |  |
| --- | --- |
| Number of training rows: | 49237 |
| Number of validation rows: | 16297 |

From the fit details, we can see that the boosted tree has similar performance to the bootstrap forest which have higher RSquare and lower RMSE values as compared to the regression tree.

**Overall Statistics**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **RSquare** | **RMSE** | **N** |
| Training | 0.672 | 0.6648492 | 49237 |
| Validation | 0.673 | 0.6583306 | 16297 |

**Cumulative Validation**

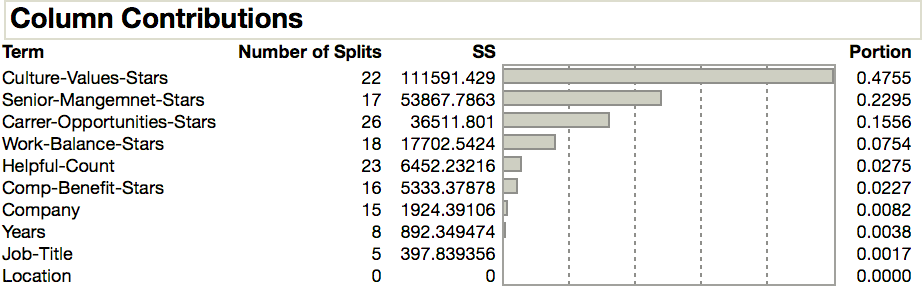


**Prediction Profiler**



As seen in the snippets from the profiler generated by Boosted Tree model, we can observe the relation between the predictors and the response variable (Overall Rating). By examining, the profiler we can observe which predictor has the highest contribution to the model. We observe that Location, Years anf Job Title is represented by a horizontal line which represents a “no relationship”.

As seen in the column contributions, we can observe the predictors that were used the most in the sampling. We can see that Culture-Values-Stars plays a significant role in our model; which is consistent with the findings from the regression tree, bootstrap forest and boosted tree models



## 

## **Model 5: NEURAL NETWORKS**

The model output including the performance results of training and validation

**Model NTanH(3)**

**Training**

**Overall-Ratings**

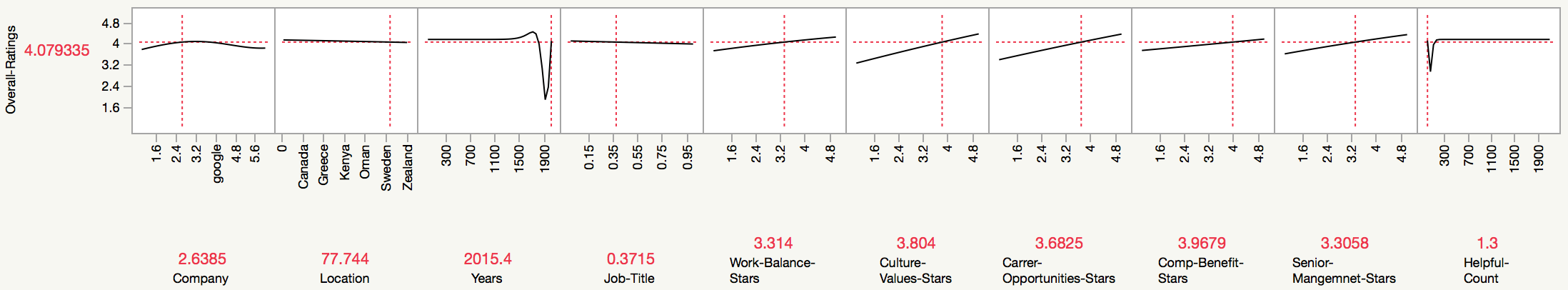
|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.6692405 |
| RMSE | 0.6676321 |
| Mean Abs Dev | 0.4915643 |
| -LogLikelihood | 49970.628 |
| SSE | 21946.092 |
| Sum Freq | 49236 |

**Validation**

**Overall-Ratings**

|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.6665443 |
| RMSE | 0.6645214 |
| Mean Abs Dev | 0.4899092 |
| -LogLikelihood | 16464.05 |
| SSE | 7196.5712 |
| Sum Freq | 16297 |
|  |  |

**Prediction Profiler**



The validation RSquare is only slightly lower than the RSquare for the training dataset ((0.6692405-0.6665443)/0.6692405 = 0.402%). The RMSE for the validation dataset is only 0.46% lower ((0.6676321-0.6645214)/0.6676321) than that of the training data. Overall, the validation dataset fits well with the training neural network model, the chance of overfitting is really low.

**Model NTanH(10)NBoost(2)**

**Training**

**Overall-Ratings**

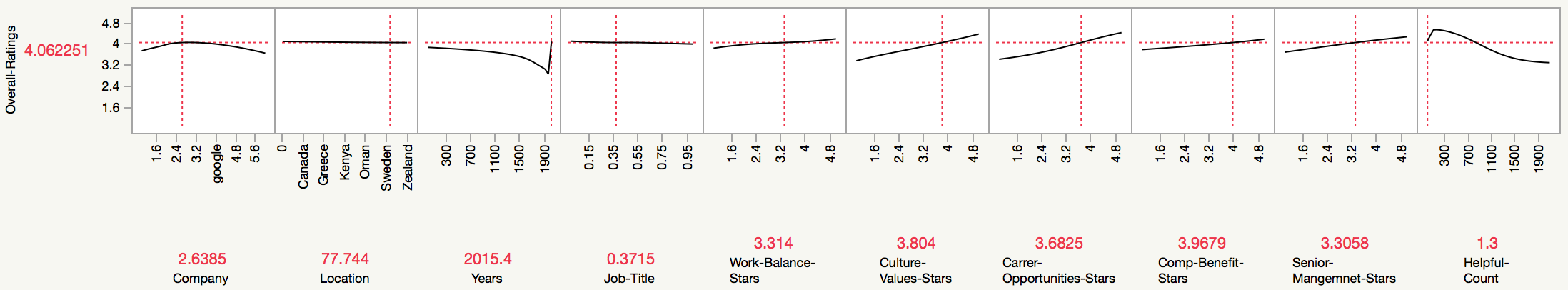
|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.6839155 |
| RMSE | 0.6526535 |
| Mean Abs Dev | 0.4787416 |
| -LogLikelihood | 48853.414 |
| SSE | 20972.396 |
| Sum Freq | 49236 |
|  |  |

**Validation**

**Overall-Ratings**

|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.681222 |
| RMSE | 0.6497317 |
| Mean Abs Dev | 0.4781778 |
| -LogLikelihood | 16097.245 |
| SSE | 6879.8002 |
| Sum Freq | 16297 |

**Prediction Profiler**



**Model NTanH(3)NLinear(3)NGaussian(3)NTanH2(3)NLinear2(3)NGaussian2(3)**

**Training**

**Overall-Ratings**

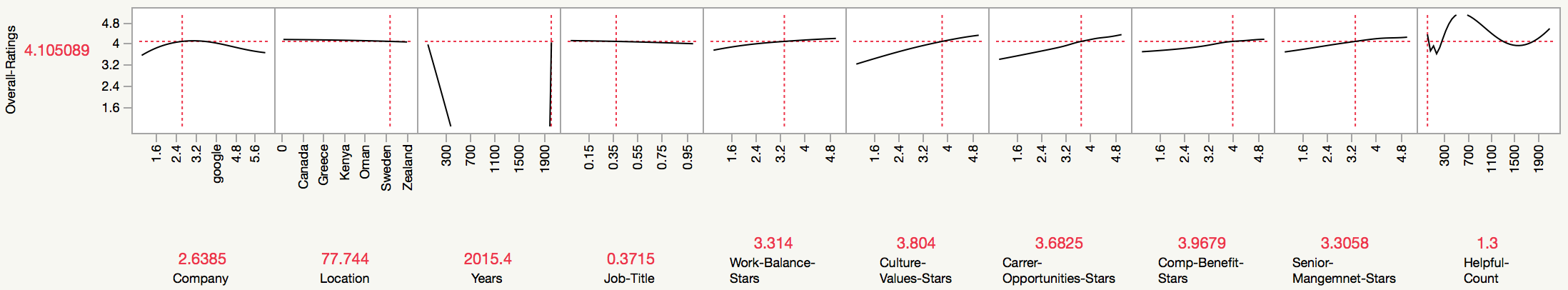
|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.6830981 |
| RMSE | 0.6534968 |
| Mean Abs Dev | 0.4819167 |
| -LogLikelihood | 48916.991 |
| SSE | 21026.628 |
| Sum Freq | 49236 |

**Validation**

**Overall-Ratings**

|  |  |
| --- | --- |
| **Measures** | **Value** |
| RSquare | 0.6810616 |
| RMSE | 0.6498952 |
| Mean Abs Dev | 0.4796517 |
| -LogLikelihood | 16101.344 |
| SSE | 6883.262 |
| Sum Freq | 16297 |

**Prediction Profiler**



The Prediction Profiler of the input predictors show that some of them have approximately linear relationships with the response (Overall Rating), for example, Culture Value Benefits, Career Opportunities and Compensations Benefits. Some input predictors have nonlinear relationships with Overall Rating, e.g., Helpful count. Several variables like Job title and Locations have minimum to no relationships with Overall Rating (horizontal or almost horizontal lines in the Prediction Profiler), and thus may be potentially excluded as predictors in the neural network modeling.

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures**  **(Validations)** | **Model**  **(3 Nodes)** | **Model**  **(10 Nodes & Boosting)** | **Model**  **(2 layers , 3 nodes each)** |
| **RSquare** | 0.6665443 | 0.681222 | 0.6810616 |
| **RMSE** | 0.6645214 | 0.6497317 | 0.6498952 |

Based on comparing the validation results, the best neural network model is the one with 10 nodes and boosting. It has the highest validation *RSquare* and lowest *RASE/RMSE* as compared with other two models.  The comparison of the neural network models results demonstrate that increasing the number of nodes, adding boosting capabilities, and adding the second layer with the nodes may increase (at least marginally) the model performance results, specifically in terms of the higher validation *RSquare,* and lower *RASE/RMSE*.

# **OVERALL COMPARISONS:**

**Measures of Fit for Overall-Ratings**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Creator** |  | **RSquare** | **RASE** | **AAE** | **Freq** |
| Regression Model Pred Formula | Fit Least Squares | https://lh3.googleusercontent.com/sVsjSU8dj3PIyD-IWEHhIcsO4ge0rtT50WOo0wlNT9w0MQJahkEMbMfES3LFdiBVeD8JHTrghM7h-y5ZLwDaX0mFHB2xem9_wtyDBCqZ3ur48dY02WhaOeU6yFMDE3lEdw00bVcG | 0.6532 | 0.6777 | 0.4999 | 16297 |
| Decision Tree Pred Formula | Partition | https://lh4.googleusercontent.com/wPmNlV1vT50mi5pToclaPtm5-Ux4iRh5LVLW6vUligxyVnHu9h9HVVQg53Hrok-HyfRFh61QAJE-xx5OSPrTL7iyb8ibMPGB5vEpjbQnn7MXrRFgKiKuy5KS-LNliJImX_SobPla | 0.6611 | 0.6700 | 0.4954 | 16297 |
| Bootstrap Forest Pred Formula | Bootstrap Forest | https://lh4.googleusercontent.com/Jztfi1OFcjTvoOIQcB3ptRIqpiccD4gtCQYjTargYzCB04qneHRPgclqXkPkeKOeTJ5DhHXCJcOg7z1QEqDUwP7v8r--aJqrmNd17mwb9htIjawtli3JogPjrTX-XIXLRBrMtMga | 0.6785 | 0.6525 | 0.4923 | 16297 |
| Boosted Tree Pred formula | Boosted Tree | https://lh3.googleusercontent.com/rS00cys16OrDBcFLMwOM85zR5BPPipt1OrFcjON-qoJ7083Cw-NgwKsFktjHEHfDh6ZcQp2ogjqsAOgtkXQT9QoAC8EQ4wgdYmEkM6D7IHo-EKVOGZQzTSrCNWzU5f2kvKsOgus3 | 0.6727 | 0.6583 | 0.4860 | 16297 |
| Neural Model NtanH(3) Pred formula | Neural | https://lh3.googleusercontent.com/rS00cys16OrDBcFLMwOM85zR5BPPipt1OrFcjON-qoJ7083Cw-NgwKsFktjHEHfDh6ZcQp2ogjqsAOgtkXQT9QoAC8EQ4wgdYmEkM6D7IHo-EKVOGZQzTSrCNWzU5f2kvKsOgus3 | 0.6665 | 0.6645 | 0.4833 | 16297 |
| Neural Model NtanH(10) Pred formula | Neural | https://lh4.googleusercontent.com/Jztfi1OFcjTvoOIQcB3ptRIqpiccD4gtCQYjTargYzCB04qneHRPgclqXkPkeKOeTJ5DhHXCJcOg7z1QEqDUwP7v8r--aJqrmNd17mwb9htIjawtli3JogPjrTX-XIXLRBrMtMga | 0.6812 | 0.6497 | 0.4782 | 16297 |
| Neural Model All 3 Pred formula | Neural | https://lh3.googleusercontent.com/3oN1TnGsDsXBMuxAdfEeulo7quuKAgUX69aBeugJ6Y9tIA8nOxR5x3mPV3mrQNTOzDDmtyUBw6aqlckwwBXpQMHdTxB2MUqawrceBZoDYR8nVQ65NMhf9pYZYNGCmKPbbxFhYk2I | 0.6811 | 0.6499 | 0.4797 | 16297 |

From the model comparison, we can see a summarized table which contains RSquare and RASE/RMSE values for each model applied. We can see that from all models that were trained on the data, the Neural Network with 10 nodes and boosting has the lowest RASE value and highest RSquare value for the validation data. The second-best model is the Neural Networks with 2 layers, and three nodes each has the second highest RSquare value and second lowest RASE value for validation data.

**Actual by Predicted Plot**

|  |  |
| --- | --- |
| Training |  |
| Validation |  |

**Residual by Predicted Plot**

|  |  |
| --- | --- |
| Training |  |
| Validation |  |

As seen in the snippets from the profiler generated by neural network model, we can observe the relation between the predictors and the response variable (Overall Rating). By examining, the profiler we can observe which predictor has the highest contribution to the model. We observe that Location, Job Title and Years is represented by a horizontal line which represents a “no relationship”. However, we observe that Cultures Vales and Compensation Benefits play significant roles in the neural network’s layers.

# **CONCLUSIONS**

After understanding our data and exploring the various predictors involved to predict *Overall-Ratings*, we have arrived at the conclusion that the Neural Network Model has the best performance upon training on the dataset. Moreover, we recommend that this model be used because it has the lowest RMSE value and the highest RSquare value for the validation data as compared to all other models. Both statistical measures, which are useful for comparison of performance of machine learning models, both indicate that the neural network has the best performance.

From our findings, we find that the regression model and the regression tree perform the worst as represented by having the highest Validation RMSE values and competitively high Validation RSquare values. Also, we have applied the Bootstrap Forest and the Boosted Tree. These machine learning models have all shown an improved performance as compared to the Linear Regression model and Regression Tree Model, but has a slightly lower performance on validation data as compared to the Neural Network

# **ANALYSIS CONCLUSIONS**

Google and Facebook have received the best overall ratings and when dug deep to know what other factors apart from, work life balance, culture& values, compensation benefits, senior management and career opportunities have been spoken about by the employees, it was found that ‘food’ is also one of the most popular words used by the employees of these two companies. So, food can also have an impact on the ratings.

Based on the all the analysis above, we can see that Facebook will be a ideal company for tech fellows to work. It appears on top on different aspects of rating, either with other companies, or alone.

**Overall Rating**: Facebook appears to be the clear winner, followed by Google, Apple, Netflix, Amazon, and Microsoft

**Work Balance**: Google appears to provide better work balance followed by Facebook, Microsoft, Apple, Netflix, and finally Amazon

**Culture Values**: Facebook comes out on top again followed by Google, Apple, Netflix, Amazon, Microsoft

**Career Opportunities**: Facebook appears to provide better career opportunities followed by Google, Amazon, Microsoft, Apple, Netflix

**Compensation Benefits**: Facebook pay/benefits is valued more over the other companies followed by Google, Netflix, Apple, Microsoft, Amazon

**Senior Management**: The senior management appears to be better with facebook followed by Google, Apple, Netflix, Amazon, and Microsoft

**THOUGHTS**

Facebook and Google rank among the top consistently Amazon is ranked as not so good when it comes to Work Balance, Culture values, Compensation benefits, and Senior management Microsoft doesn't get any better when it comes to Culture Values, Career Opportunities, Compensation benefits, and Senior management

Now these are just peer ratings and each person is entitled to their opinion. Whether or not this holds true is to be experienced by the employee themselves. Will this influence my opinion if I was applying to one of these organizations? YES.

I'd like to think that it comes down to how compatible an employee's ideals/goals/ideas align with the management that they're working for. Or rather how flexible an employee can be with their ideals/goals/ideas when it is not aligned with those of the company

## **FUTURE GOALS**

We see that our highest RSquare is 0.6812. This is because we used the numeric variables from our dataset. However, this value can be improved, if we apply different data analytics and machine learning techniques, like creating more variables from the categorical dataset. This might improve our RSquare value, and further lower our RASE/RMSE value.

# **BIBLIOGRAPHY**

Kaggle : <https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews>