Data Science Salaries

Data Analysis Project

Corey Feld, Emiliana Geronimo, David Scheff

MATH 664 - 102

Professor Loh

April 27th, 2023

I. Introduction

The salaries of data science professionals is a popular topic of discussion in the technology space. For the last few years, data scientists have been featured as a top career by GlassDoor and named “one of the sexiest job titles” by Harvard University (1). The demand for data scientists has increased exponentially over the past few years due to the massive influx of accessible data across the internet. According to TC Global, 2.5 quintillion bytes of data are extracted every day (1). Companies have hired data scientists in hopes that they can extract meaningful insights to generate additional profits. Data scientists are responsible for utilizing algorithms and technology to answer complex prediction and inference questions. With this big undertaking comes a bigger price.

We will be analyzing the salaries of those who work in the data science industry and other features of these types of jobs. Our goal is to develop models and statistical analyses which we can use to identify how specific variables are contributing to the outcome, salary. The main goal is not to create a model which can generate the most accurate predictions, but rather to develop simple models that will allow us to effectively interpret and measure the impact the explanatory variables have on the response variable. Section 1 describes the exploratory data analysis and data preprocessing steps. Section 2 describes our research methodology in which we develop a linear regression model and random forest model. We conduct an ANOVA test and calculate feature importance to better interpret the models, respectively. Lastly, Section 3 details the results and limitations of our work.

II. Research Methodologies

I. Exploratory Data Analysis and Data Preprocessing

The Data Science Job Salaries dataset contains 3,755 entries (rows) with 11 variables (columns) including Work Year, Experience Level, Employment Type, Job Title, Salary, Salary Currency, Salary in USD, Employee Residence, Remote Ratio, Company Location, and Company Size. The variables “Salary” and “Salary in USD” are the only numeric variables with the rest being categorical variables. There are no null values in this data set.

We develop bar charts for the categorical variables to analyze the frequency distributions of their values. We find that many of these variables have many possible values but only a few dominant values. For example, the variable “Job Title” has 93 unique values with the 20 most frequent values accounting for about 92% of the data. The variables Salary Currency, Employee

Residence, and Company Location are distributed similarly with one or a few dominant values out of many unique values. The ten most frequent values of these variables account for about 99.5%, 94%, and 95% of the data, respectfully. It is important to note that the other categorical variables have much fewer unique values, but still contain one or two dominant values. For example, the variable Work Year has four unique values with two values, 2022 and 2023, that comprise about 92% of the data. Most notably, Employment Type has the most imbalanced distribution with one out of four possible values, “Full Time”, accounting for 99% of the data (Figure 1). The rest of the categorical variables behave similarly.

We analyzed the continuous variables, Salary and Salary in USD, by obtaining summary statistics and creating histograms (Figure 2). The Salary variable is rather irrelevant because there is a pool of salary values that are all on different scales that depend on their respective currencies. This variable would be more useful in a setting where we are analyzing one or more countries that use the same currency. The Salary in USD variable is approximately normal but slightly skewed right, with a mean of about $137,570 and a standard deviation of about $63,056. It appears that most people are making between about $50,000 and $250,000 with a small population of people making more. There is a minimum of $5,132 which seems like it could be a potential outlier. However, given that different currencies translate to the US dollar differently, we cannot say whether this is an outlier or not so we treat it as a reliable data point.

Figure 1: Employee Type Distribution

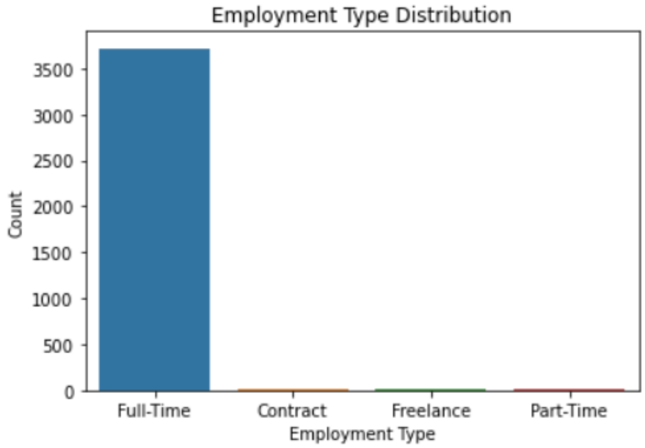
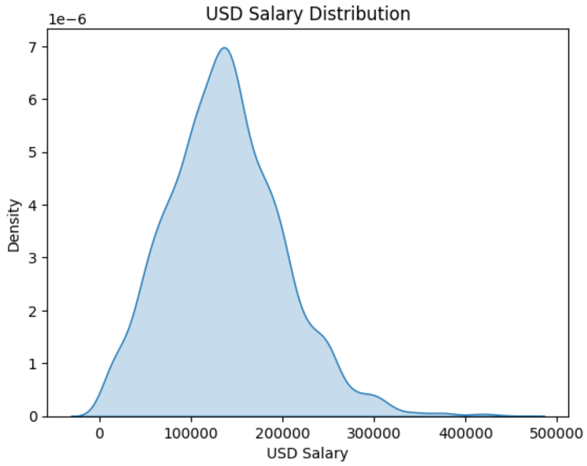


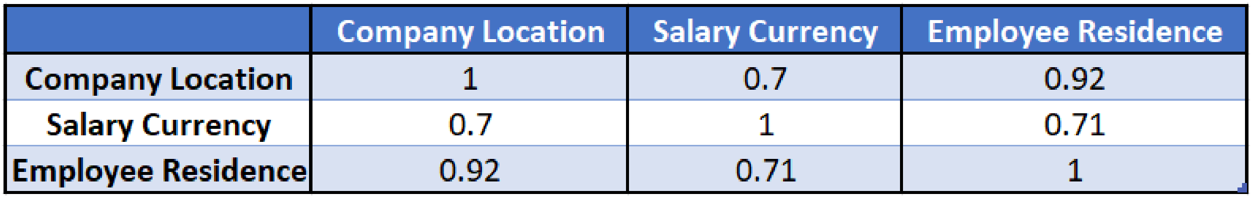
Figure 2: Salary in USD Histogram



We performed a series of “group by” operations to analyze the average salary for each value in each column. Something notable is that the value “US” (America), has the highest average salary for Salary Currency (USD), Employee Residence, and Company Location.

To address the imbalances in the variables, we assigned all values not in the top 10 most frequent values to the value "Other" (we used 20 values for Job Title). This will limit the number of unique values, making for more efficient variable encoding, and improving model performance. The model will not perform well if there are minimal data points for many of the available values of variables. We then encoded our variables using a label encoder, assigning each unique value of each variable a number zero to (n - 1). We choose label encoding because it is a simple method that allows us to easily identify the variables in model results for proper interpretation. We created a correlation matrix (Figure 3) and found that 3 pairs of variables are highly correlated (≥ 0.7). We removed two out of the three variables, Salary Currency and Company Location, and kept Employee Residence since it is the most diverse variable. The correlation matrix is pictured below:

Figure 3: Correlation Matrix



	Company Location	Salary Currency	Employee Residence
Company Location	1	0.7	0.92
Salary Currency	0.7	1	0.71
Employee Residence	0.92	0.71	1

II. Methods

a. Linear Regression

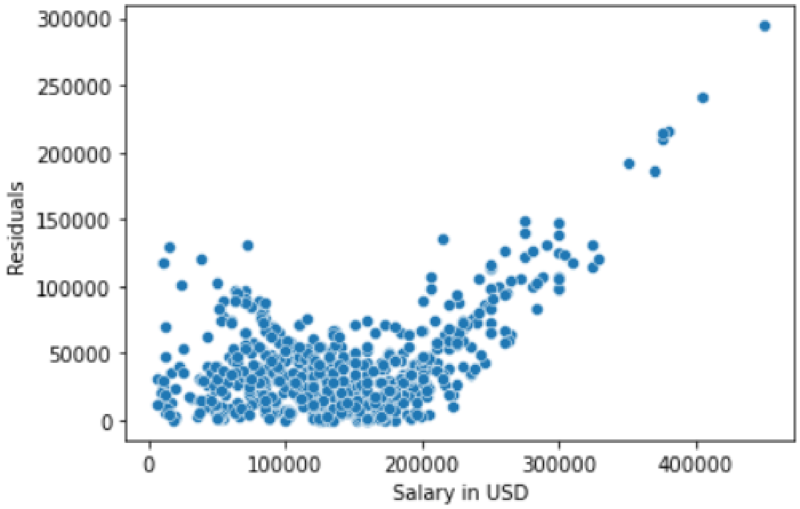
We perform a train-test-split of our data using 80% of the data as training data and 20% as test data. We fit a linear regression model to our training data using ordinary least squares with Work Year, Experience Level, Job Title, Employee Residence, Company Size, and Remote Ratio as the predictor variables and Salary in USD as the response variable. Since three of the variables are highly correlated with each other, we decided to keep Employee Residence and exclude the other variables since this variable is the most correlated with the response variable.

The results show a test R-squared value of 0.42, meaning that the model accounts for about 42% of the variation in salary. This is very similar to the R-squared of the random forest model, indicating that they may have similar predictive capabilities. The Mean Average Error (MAE) provides us with an estimate of the average error in US dollars in the predictions on unseen test data. The MAE value for this model is $37,618.29, meaning that on average, we predict the salary to be $37,618.29 away from true salary value. This model does not seem to be overfitting or underfitting considering that the training results are a little bit worse than the testing results. However, the results between the training and testing data are close enough that the model could be slightly underfitting of our data.

All predictor variables except for Remote Ratio turned out to be significant predictors. The variable Experience Level appears to be the most significant predictor. It contains three coefficients for each of the four values for this variable. These three values have coefficient values whose t-values are the three highest out of all the coefficients. So Experience Level is the most significant predictor with its t-values being the highest.

The residual plot reveals an upward quadratic trend in the residuals (Figure 4). A histogram of the residuals shows a skewed right distribution with deviations from normality. Although these results indicate that the fit is not the best and the predictive capabilities of the model are limited, we do not perform any transformations to preserve the ease and efficiency with which we can interpret the model and its variable importances. The residual plot is pictured below:

Figure 4: Residual Plot for Linear Regression Model

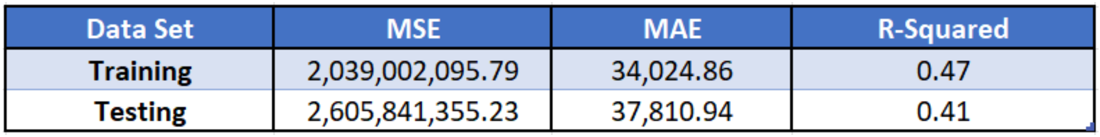


To assess feature importance, we conducted an ANOVA test which reveals the proportion of the sum of squares that each variable accounts for in the linear regression model. If a variable accounts for a large amount of the regression sum of squares, then it is deemed more important. A larger regression sum of squares means a lower p-value, and therefore more significance.

b. Random Forest Regression

We utilize the same train-test-split of our data using 80% of the data as training data and 20% as test data. We fit a random forest regressor to our training data and use grid search cross validation to tune the hyperparameters. The training and testing results (Figure 5) are given below:

Figure 5: Training and Testing MSE, MAE, R-Squared



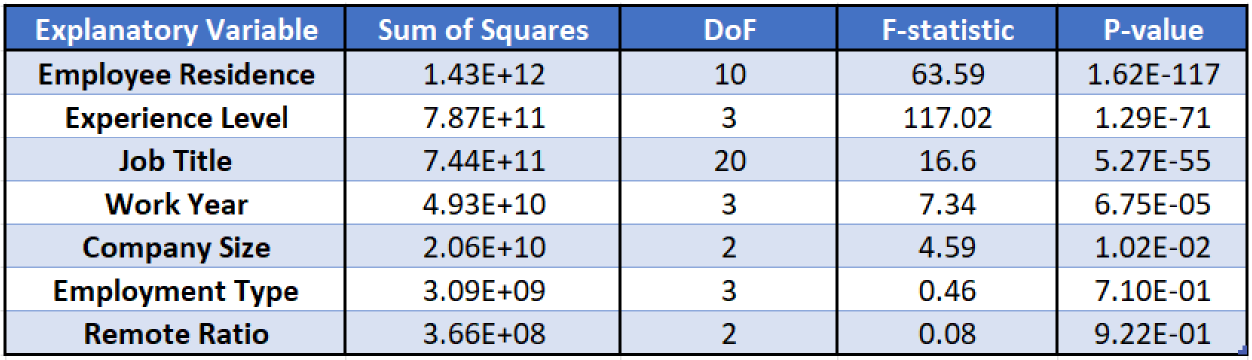
Data Set	MSE	MAE	R-Squared
Training	2,039,002,095.79	34,024.86	0.47
Testing	2,605,841,355.23	37,810.94	0.41

The results show that although the model does not fit the data very well, there seems to be very little to no underfitting or overfitting. The model performs slightly better on the training data than the test data but seems to generalize well to the unseen data. According to the results, the random forest model explains about 41% of variation in the response variable, the salary of data science employees. The MAE reveals that on average, we predict the salary to be $37,810.94 away from true salary value. While this seems like a large margin of error in the context of yearly salary, we are not so focused on the predictive capabilities, but rather the ease and efficiency with which we can interpret the model and variable impact. To assess feature importance, we calculate feature importance for the variables. This is done by calculating the total decrease in Gini impurity, or the increase in information gain. The variables that lead to the largest decreases in Gini impurity are the most important variables.

III. Results

We will be comparing the results of the ANOVA test for the linear regression model with the results of the feature importance function for the random forest model. The ANOVA results of the linear regression model are pictured below:

Figure 5: ANOVA Results for Linear Regression Model

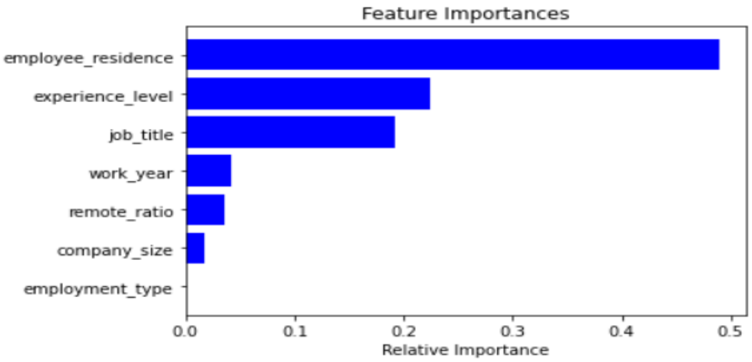


Explanatory Variable	Sum of Squares	DoF	F-statistic	P-value
Employee Residence	1.43E+12	10	63.59	1.62E-117
Experience Level	7.87E+11	3	117.02	1.29E-71
Job Title	7.44E+11	20	16.6	5.27E-55
Work Year	4.93E+10	3	7.34	6.75E-05
Company Size	2.06E+10	2	4.59	1.02E-02
Employment Type	3.09E+09	3	0.46	7.10E-01
Remote Ratio	3.66E+08	2	0.08	9.22E-01

We evaluate variable importance according to the p-values. A low p-value equates to accounting for a large amount of the regression sum of squares and vice versa. The amount of regression sum of squares that a variable accounts for is deterministic of its importance. So, the lower the p-value, the higher variable importance is. The results show that Employee Residence, Experience Level, and Job Title are by far the three most important variables. Employee Residence is more significant than any other variable, with Experience Level and Job Title coming in a bit lower. Work Year and Company Size account for a significant amount of the regression sum of squares as well, but a much smaller amount than the top three variables. Employment Type and Remote Ratio are not significant and therefore do not contribute to a significant amount of the regression sum of squares. We can see from the p-values that the three most important variables driving most of the model's predictions are Employee Residence, Experience Level, and Job Title.

Furthermore, the results of the feature importance function for the random forest is pictured below:

Figure 6: Random Forest Feature Importances



As stated before, feature importance is measured here by measuring the total decrease in Gini impurity for each variable. The results show that Employee Residence is by far the most important variable, with Experience Level and Job Title coming in as the next two most important variables. After that, there is a very sharp drop off, with Work Year, Remote Ratio, and Company Size coming in with much smaller variable importance. Lastly, Employment Type holds no variable importance. It holds no variable importance and can be removed from the model since it is not causing a decrease in Gini impurity. Overall, Employee Residence, Experience Level, and Job Title are the three most important variables with the other variables providing some to little importance.

The results seem extremely similar to those we obtained from the ANOVA test for the linear regression model. In both cases, Employee Residence, Experience Level, and Job Title are the three most important variables by far. Employee Residence is the most important variable by a lot. Work Year is way less important, but still the next most important variable. The results differ in the sense that the order of importance for the last few variables are different. However, both methods regard these three variables as the least important. In addition, both methods deem Employment Type as unimportant, being labeled as insignificant in the ANOVA test and showing to cause no decrease in the Gini impurity.

Due to the extreme commonalities between the two feature importance methods for the two models, we conclude that Employee Residence is the driving factor in determining the salary of a data science employee. Experience Level and Job Title are the next most important factors to consider, while Work Year has a small impact. We can also conclude that Employment Type is not a contributing factor given it is not significant in both cases. We are unable to make any conclusions regarding Remote Ratio and Company Size. Remote ratio is not significant according to the ANOVA test but shows to cause a small decrease in Gini impurity. Additionally, Company Size is significant according to the ANOVA test but also shows to cause a small decrease in Gini impurity.

IV. Limitations

There are some limitations to the research we conducted that must be kept in mind as they can potentially affect the results of the data analysis.

One limitation of this research is that all the explanatory variables (which are all categorical variables) are very imbalanced. This can make model computations and accurate predictions very difficult for machine learning and predictive models. Regardless of the type, machine learning models are made to “learn” trends in the training data and use this “knowledge” to make predictions about future, unseen data. If there is not enough data for the model to “learn” properly, then its predictive capabilities and accuracy will suffer. For example, it was mentioned before that Employment Type has the most severe imbalance. As we can see from the figure below/above, there is practically no data for three out of the four possible levels. It would be very difficult for a machine learning model to identify trends or “learn” from this data since there is not enough information to “learn” from. Perhaps, when more data is available, both models’ predictive capabilities will improve, making for more reliable results.

Another limitation of our research is that we used a label encoder to encode our explanatory variables. When there are many levels to a variable, there may be an ordering developed for these variables. This can cause confusion in model computation and interpretation because the variable levels may be given an implicit ordering that it should not have. Nonetheless, we used this method because it allows us to easily interpret results.

Lastly, we did not transform any of the variables for the linear regression model. The residual plot clearly revealed that transformations are necessary to improve the model fit and predictive capabilities. However, we opted not to make any transformations to preserve the interpretability of the model and its coefficients, since that is the main objective of the research.

Appendix

https://1drv.ms/u/s!AkAes9EMtUN8gw0LHF-sO4VaTRzk?e=gUI7aq

V. Works Cited

(1) Davenport, T. H., & Patil , D. J. (2022, October 19). *Data scientist: The sexiest job of the 21st Century*. Harvard Business Review. Retrieved April 26, 2023, from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century