What Influences a Developer's Starting Salary? - Predicting a Developers Salary Based on their Background

By Kenneth Heisler and Eli Nottonson

Date: 5/4/2022

Introduction

Data scientists, computer programmers, and web developers all congregate to a single platform in search of help or advice. This platform is known as Stack Overflow, a website populated by dedicated developers who assist other upcoming or even long standing developers. Recently, Stack Overflow has released a data set containing a myriad of information about their current users and developers. Our team intends to apply regression models and other statistical techniques to analyze this large data set in an attempt to find any strong correlations.

Domain/topic and Main Question/Goal

Our goal is to determine which factors most influence a developer's yearly salary? We intend to perform a feature analysis on our data set to create several models with the goal of predicting the salary for a developer. There are forty seven possible features that can impact this prediction, we hope to create models based on a subset of these with a consistent accuracy in predicting the starting salary.

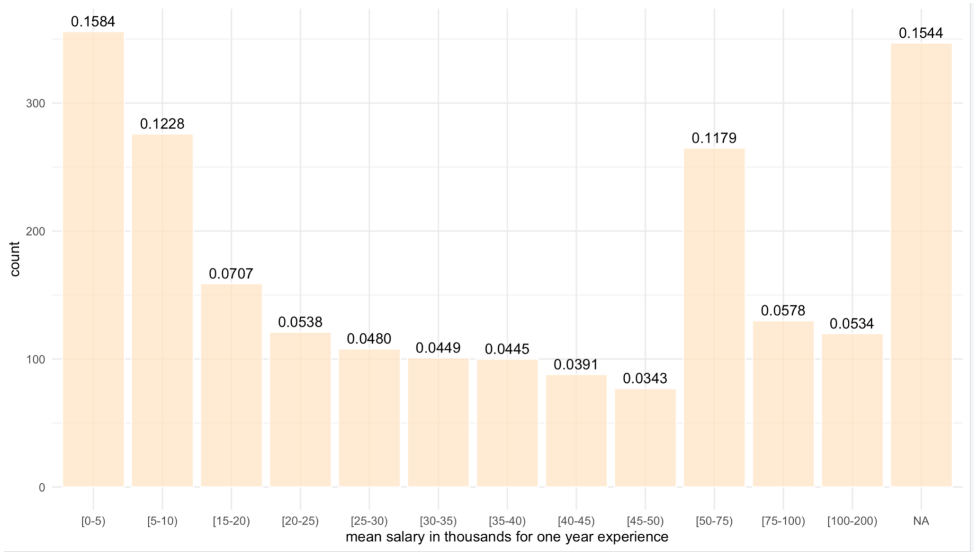
Data Source(s)

The data was collected from users by having them complete a semi-closed question survey. Our team realizes the pitfalls associated with this technique of data collection and have tried to mitigate any superfluous replies i.e eliminating extreme outliers and empty values. We are using the 2021 survey data for our project. There are 83439 entries with 48 columns each: https://insights.stackoverflow.com/survey

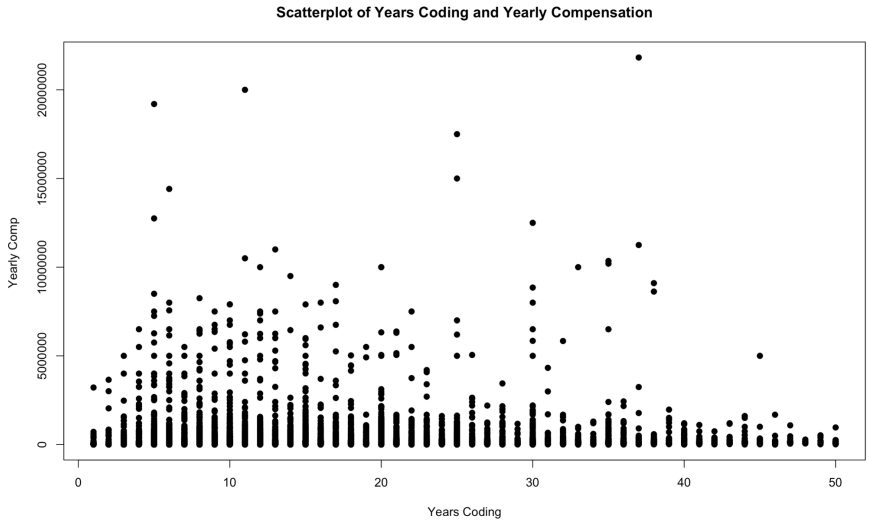
Data Exploration

Starting off our data analysis we chose to seek a correlation between numerous features and their impact on the starting salary of a developer. We have analyzed our data set, during which, we iteratively eliminated outliers and other data we deemed beyond the scope of our project. This was not an easy task because our data was not influenced the way we expected it to be. We had many difficulties trying to properly plot and constrain our data to a digestible format.

One large hurdle was removing outliers as the data we are using is survey data and is subjected to human error. The graphs shown to the right depict our data analysis as we decided to focus on starting salary versus salary for all developers. We can easily see that the influence of human error is prevalent among the data. The



NA category includes salary means that are comically large. When analyzing the highest salary means of the dataset we saw inconsistencies with other survey answers leading us to believe this data is inaccurate. One new technique we learned in this process is how to utilize bins to effectively display our data in the factored version we will be using as our variable to predict.



While manipulating the data and predictors in search of our initial goal we found that other predictors interestingly and oddly impact others within the dataset. These anomalies peaked our curiosity and we soon changed our goal as well as broadened the scope of our project. Again, we chose predictors that seemed important to us at first glance, factored the necessary predictors, and altered our predicted variable to be the yearly salary for all developers or, more precisely, the variable ConvertedCompYearly. We constructed models for our initial set of chosen predictors and found a correlation between them and our predicted variable. However, we then implemented stepwise regression in order to determine the best fit linear regression model. The stepwise regression returned one significant predictor, Age. We see this as holding true to generally held assumptions about developer salary. Stepwise regression concluding age was the most significant makes sense when taking a step back from the data and analyzing it logically. For example, a developer may start coding with minimal experience, probably when they are young, and thus only be able to obtain a starting position with a low salary. As the

developer gains experience and continues to age they are presented with more opportunities to receive a larger income. In conclusion, it seems true that an older developer would have an increased salary compared to a younger developer. However, there are always caveats, one of them being an unrealistically small error when using the predicted model. Our team came to the conclusion that we might be overfitting the data. Below are the respective summary statistics:




```
Residual standard error: 448200 on 44916 degrees of freedom
Multiple R-squared:  0.008121,  Adjusted R-squared:  0.007768
F-statistic: 22.98 on 16 and 44916 DF,  p-value: < 0.00000000000000022
```

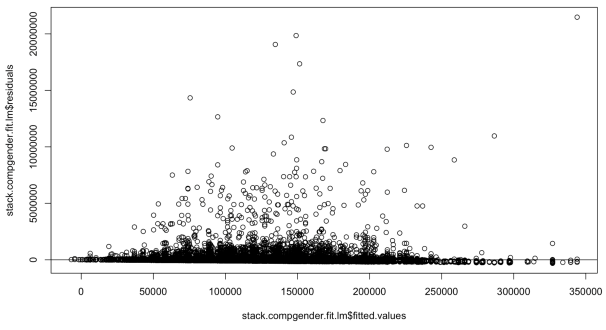
With our newfound knowledge of how the stepwise regression function manipulates our predictors, we used our own chosen predictors to conduct further research. Our next hurdle was the gender predictor. The gender predictor was not binary like most other datasets, this data set was fully inclusive of all gender preferences so much so that the developer taking the survey can input their own unique option. This enormous range of genders heavily complicated the predicting process and respective factor levels.

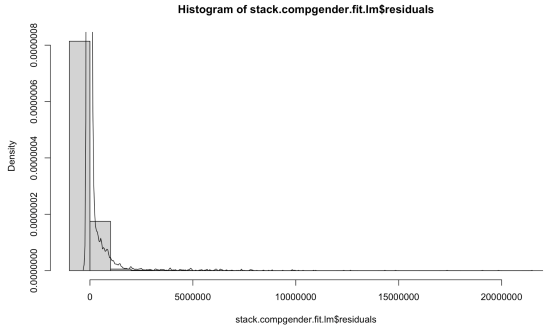
The team came together and made a decision to conform to precedent gender data and set the gender as binary between “Male” and “Female”. When closely analyzing the new data we saw the enormous disproportionality between the amount of men and women developers. Men dominated the number of responses and heavily skewed any plots we depicted when relating to

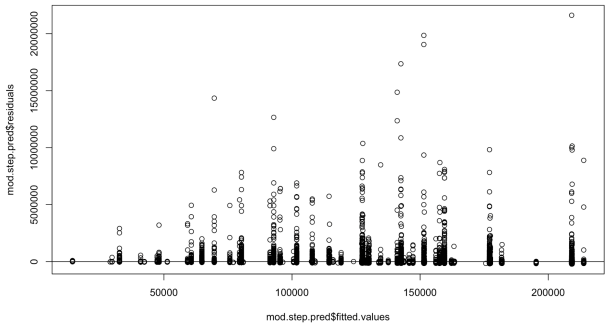
gender. In an attempt to combat the disparity between population sizes we constrained the number of male participants in our future models to be equal to that of the female population. Using this newly constructed data we wanted to see if we could predict gender based on a couple predictors. Unfortunately, we hit a wall when we were unable to create a linear regression model that is predicting a factor. The team realized that the only consistent way to conduct a meaningful prediction was to predict the converted yearly compensation of all developers based on other significant predictors.

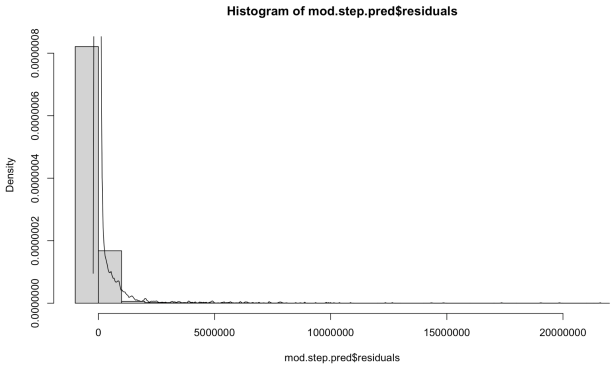
Summary

We found that the biggest factor that impacted a person's compensation was age. This is not really surprising as it makes sense to be paid more given more experience in that given field, which is what a higher age allows. What did surprise us however is how little other predictors impacted salary on the models. Years coding, organization size, and gender all had significantly less impact on our linear model created through all variables and through stepwise regression. Once we had the linear models created for all variables and those selected through stepwise regression, we fit our models to training data that made up 80% of our dataset. From there we created predictions for the test set and compared them to the actual labels. Below are the histogram and residual plots for each. We found that our models had very high mean squared









errors of 374457.6 for the model on all

predictors, and 374556.2 for the model with predictors chosen through stepwise regression. In the end, we found our results were not significant enough through our model selection to show a difference between the model chosen by step and the model made by all predictors. We found this project very challenging especially due to the dataset. The fact that we were working through survey results made data cleansing and factor creation very difficult. This made it a lot more work than we thought to just create a simple model. Through lots of trial and error we were able to overcome this and finish our project with some actual results.