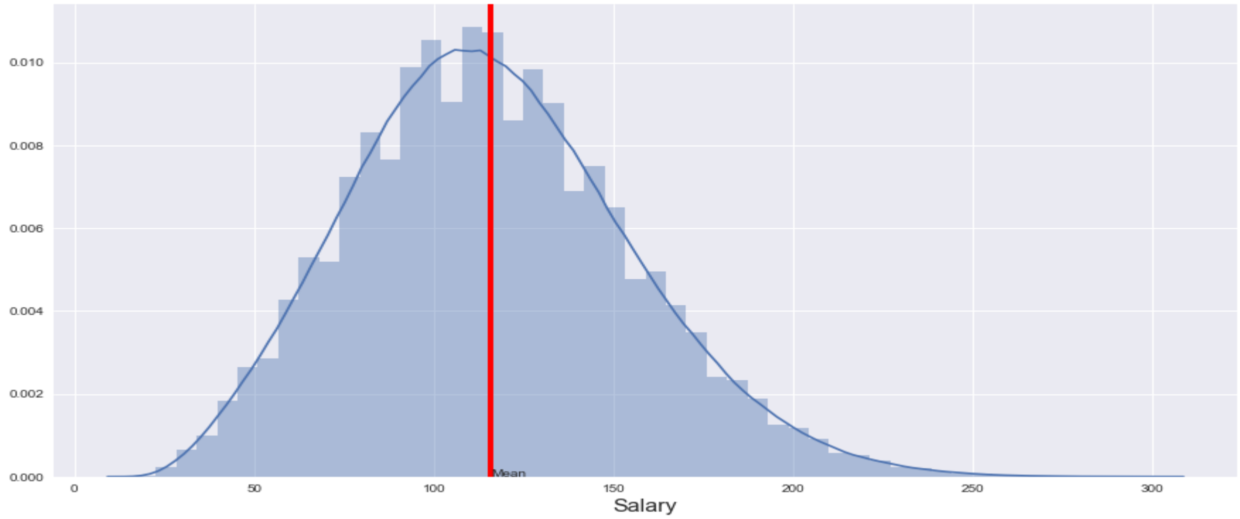
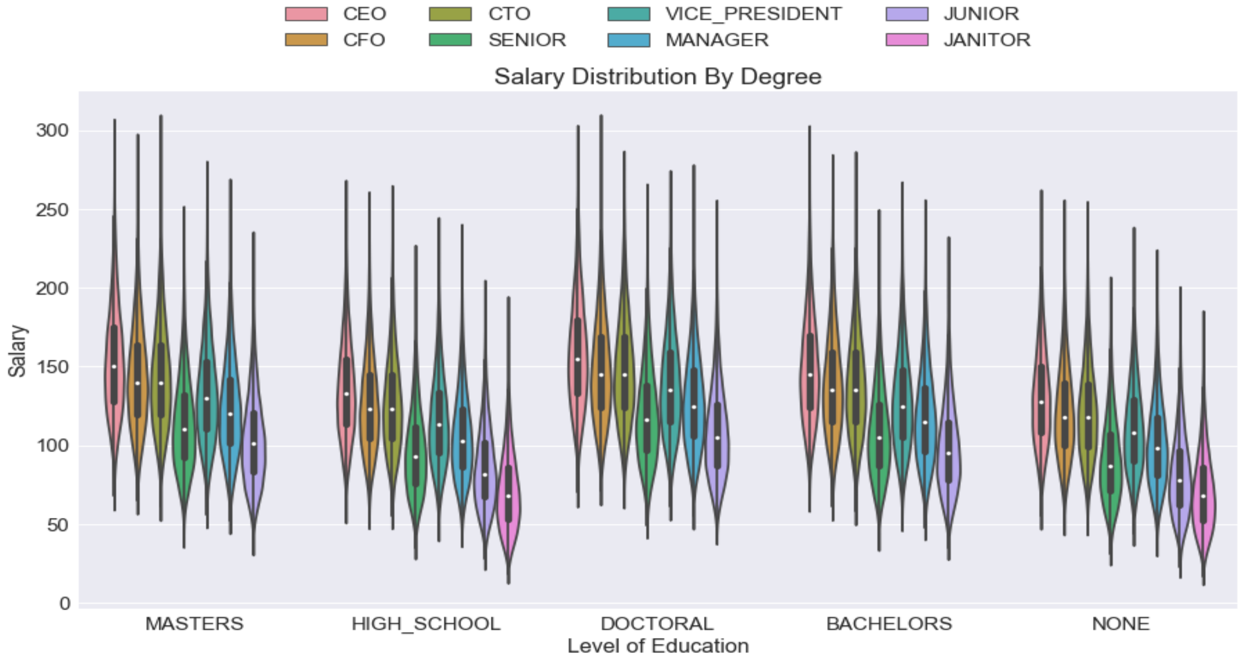
Salary Prediction Summary Report

The objective of this project is to build a model that will be able to reliably predict a salary when given a set of inputs and to find a model that produces a mean squared error between 320 and 340. The input data consisted of three files. Two training files, features and salary and one testing file.

Once the data is loaded, we looked at it in a variety of different ways to ensure the validity and integrity of the data. Should any missing data, outliers or nonsensical data be encountered, it will be dealt with programmatically and fully documented with the approval of the projects' sponsors. Understanding that this could be a very time-consuming step depending on the cleanliness and integrity of our original data.

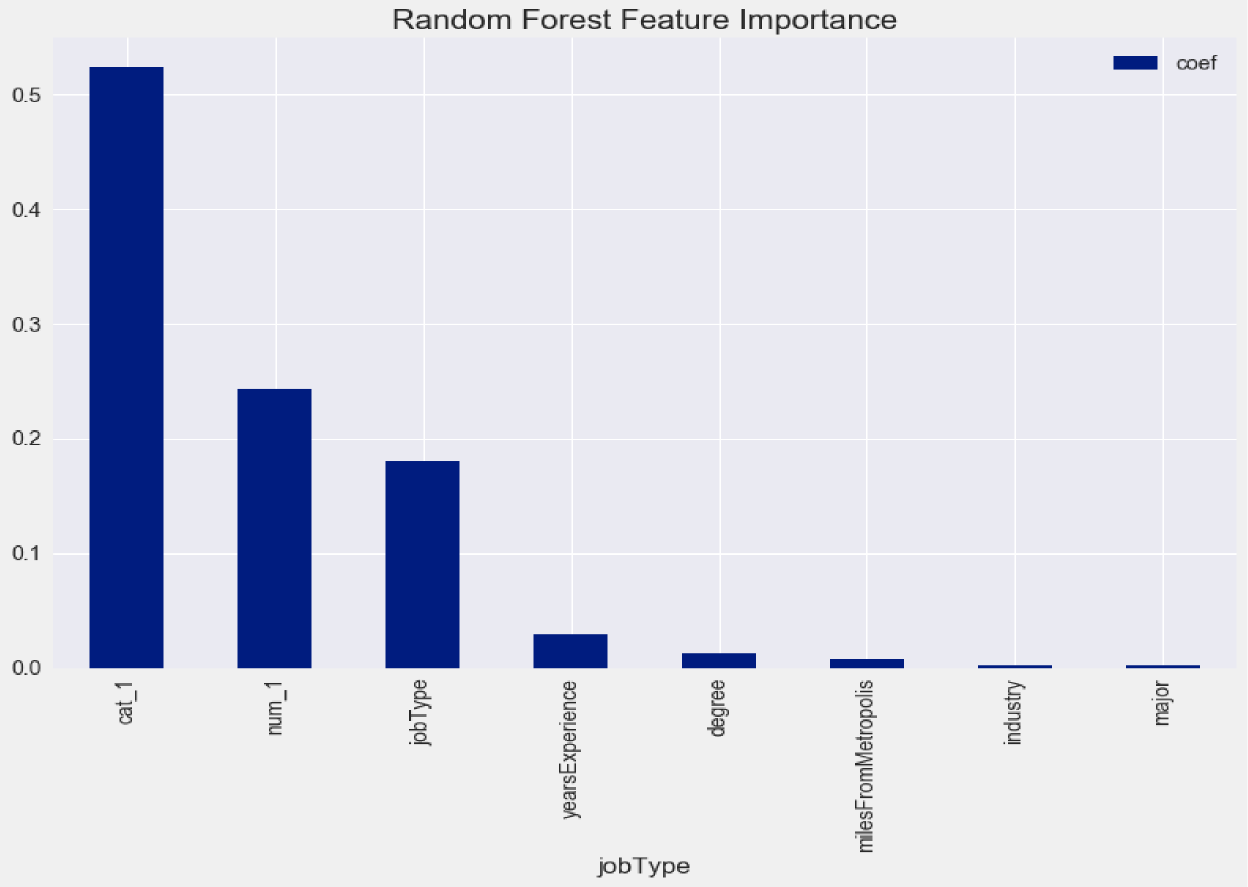


From the above salary distribution plot, we can see that the salaries are fairly normally distributed with a fat tail on the right side signifying higher salaries.



Violin plot give us quite a bit of information we see the median value, the maximum and minimum value, the interquartile range and the distribution of the values. The trend is obvious, the higher the level of education the higher the salary. Additional violin plots indicate that the oil, finance and web industries produces the highest salaries. Looking at which major produced the highest salaries we see that engineering, computer science and math lead the way.

We ran a baseline test using linear regression with limited feature engineering just to get a measure of the mean squared error. Then we proceeded to dive into some feature engineering and to improve our mean squared error. Once we've completed our features, we decided on three estimators to test. Since this is a regression problem we decided on linear regression, ridge regression and random forest regressor. Of these three the random forest was the winner producing a mean squared error of 341. Below is a graph of the features it used.



We see that the first three features were the most important to the random forest model. This leads us to needing to do more feature engineering and feature selection to see if we can get a better distribution of importance and lowering the mean squared error.

Items to consider for future revisions include: expanding on the exploratory data analysis, expanding on the feature engineering/selection process, researching additional estimators to use and tuning the hyperparameters of those estimators.