Capstone Project 1: Final Report

Introduction

As a job-seeker, salary negotiation is one of the most difficult parts of an interview process. As an employer, being sure that you are not undervaluing (or overvaluing) an employee is important for long-term business prospects. Therefore, this capstone tackles how to evaluate the salary of a Data Scientist. This capstone uses experience, age, gender, education, major, and industry to build a supervised machine learning model to predict whether or not an individual will earn more than the median salary of a Data Scientist.

It is hypothesized that age and experience will be positively correlated with salary, gender will not be correlated, and certain education levels (such as PhD), STEM majors (science, technology, engineering, and mathematics majors), and industry (such as finance) will be positively correlated with salary. In contrast, education levels (such as a bachelor's degree), non-STEM majors, and certain industries (such as government) will be negatively correlated with salary. Furthermore, it is expected that these trends are also shown in the subset of Data Scientists who make more than the median salary. The machine learning model is expected to make use of these features for robust predictions of what features determine whether a Data Scientist will earn more than their median salary.

Data Acquisition and Cleaning

Data was obtained from the "2018 Kaggle ML & DS Survey". The survey received 23,859 usable respondents from 147 countries and territories. As well, it was conducted from October 22nd, 2018 to October 29th, 2018. To protect the respondents' identity, Kaggle separated and randomized the open-ended responses. Therefore, open-ended responses will not be used in this capstone.

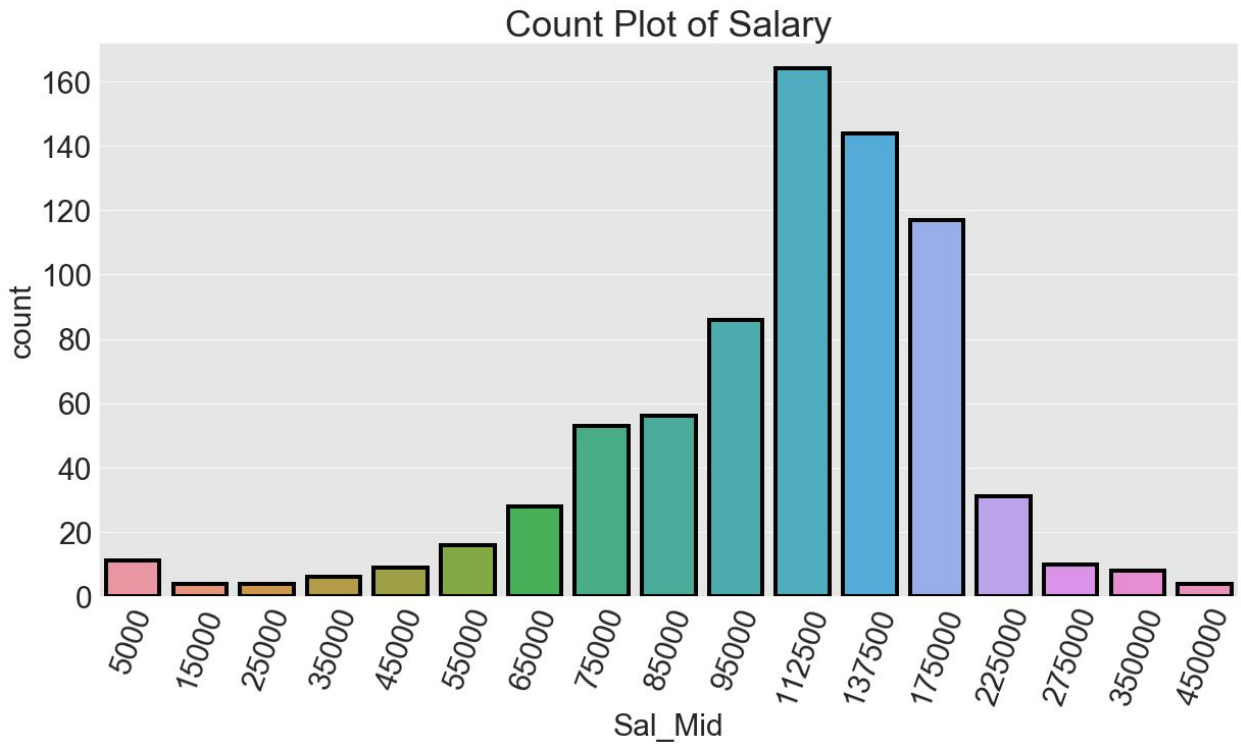
The original dataset contains 23,859 rows and 395 columns. This capstone will focus on those who: (1) live in the United States of America, (2) share their salary details, and (3) are employed as Data Scientists. As well, this capstone will limit its scope to 7 columns that inquire about: gender, age, education, major, industry, experience, and salary. After limiting to these parameters, the dataset was left with 777 rows and 7 columns. There was one missing value in the "Major" column, so the missing value was changed to "unknown". After removing subgroups that could (1) not be merged together into an "other" category and (2) contained less than 30 samples, the final dataset was reduced to 751 rows and 7 columns.

Data Exploration

Count Plot of Salary

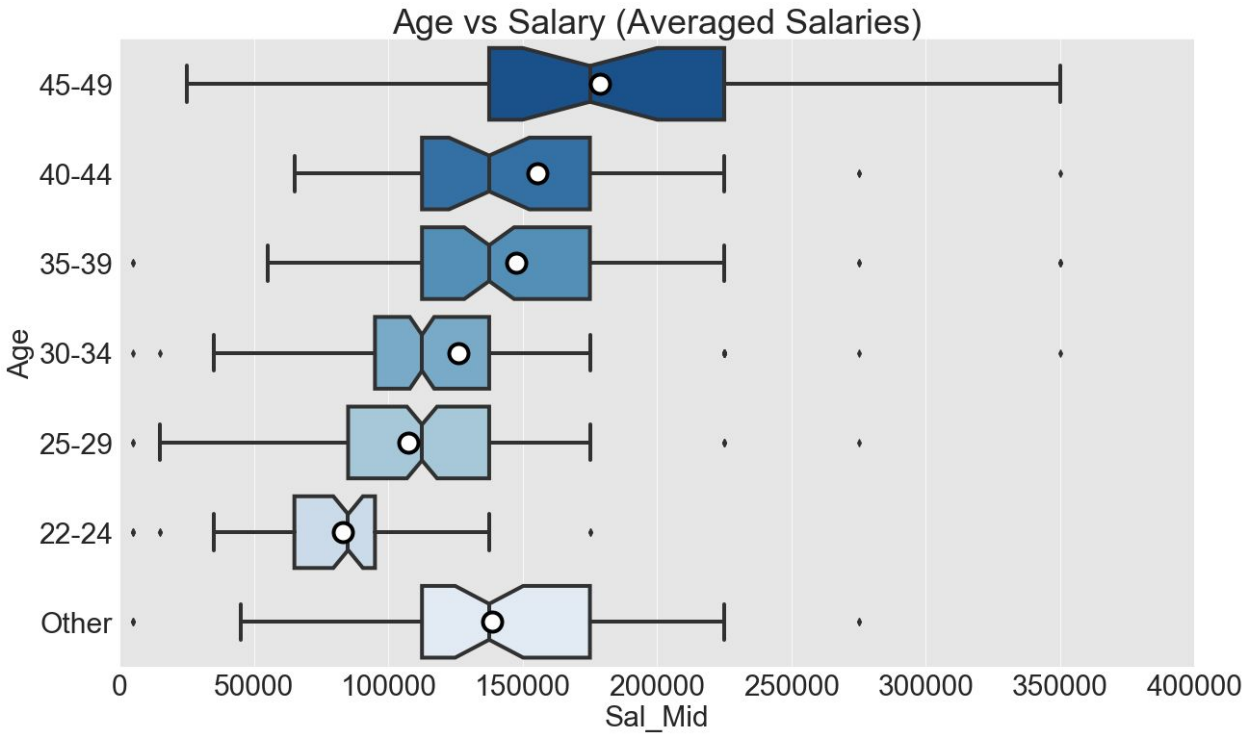
The target variable for this capstone will be salary. Therefore the data was split into two groups: Those who make more than the median salary (n=314), and those who make equal to or

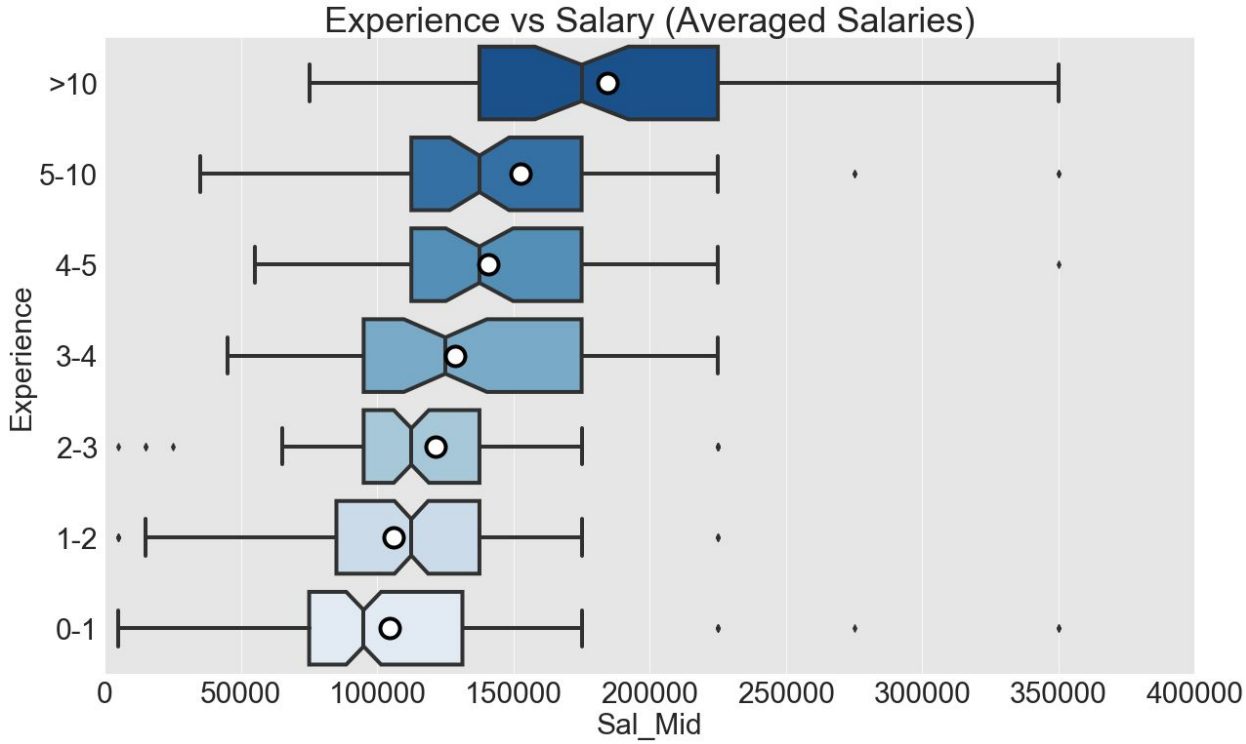
less than the median (n=437). The median salary which is the focus of this capstone was $112,500. A count plot of Salary is shown in the figure below.

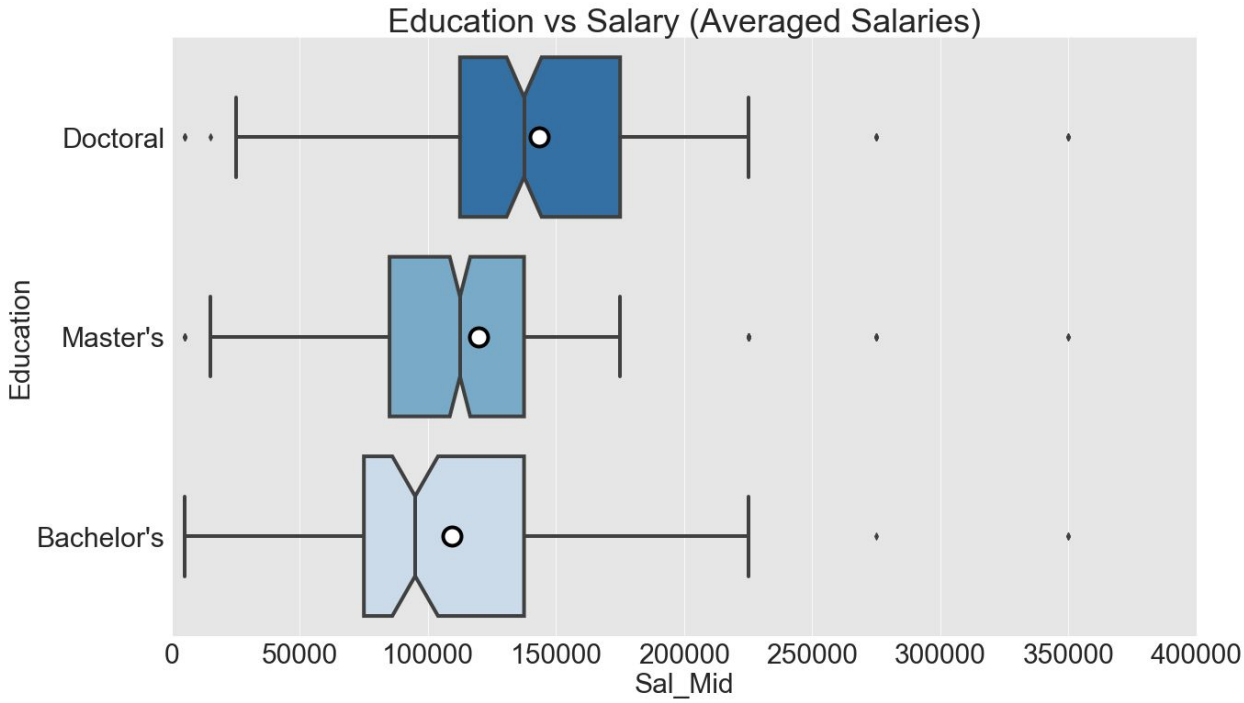


Examining salary for all Data Scientists

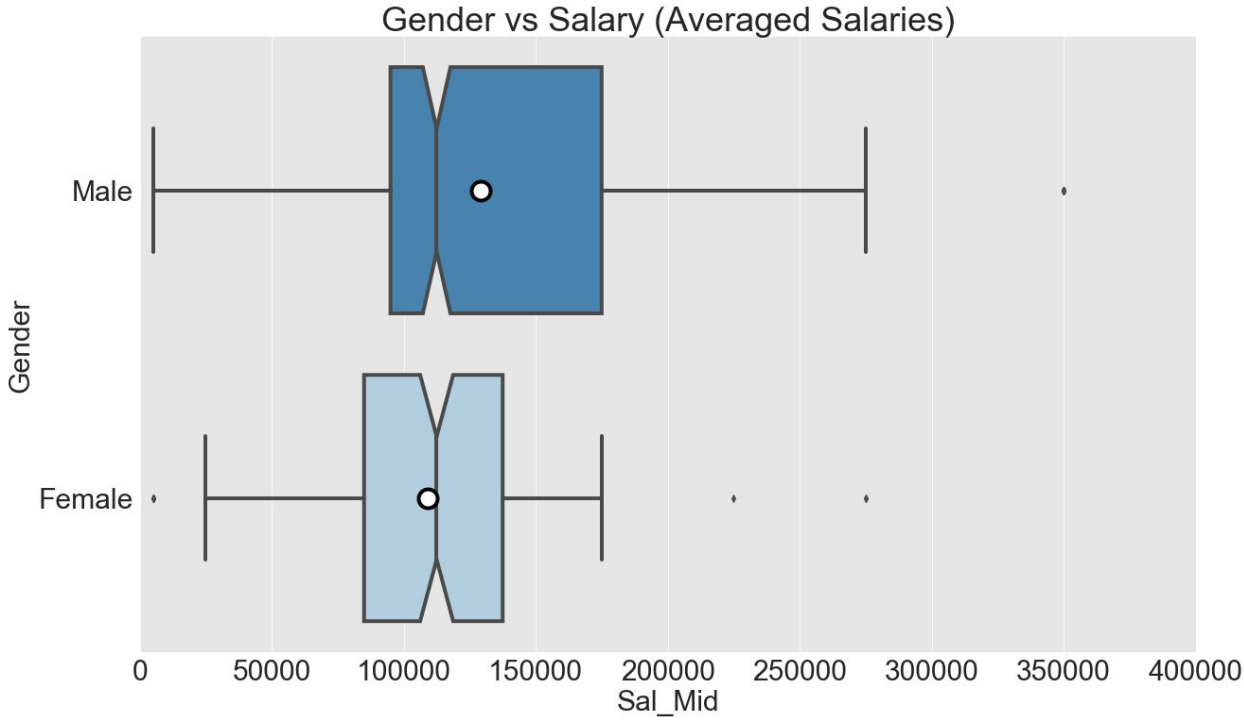
In line with the hypothesis, Salary was commensurate with Age, Experience, and Education both graphically and statistically (see figures below; notch in the box plot indicates median while the white dot indicates mean).


Age vs Salary (Averaged Salaries)
45-49
40-44
35-39
30-34
25-29
22-24
Other
Age
0
50000
100000
150000
200000
250000
300000
350000
400000
Sal_Mid

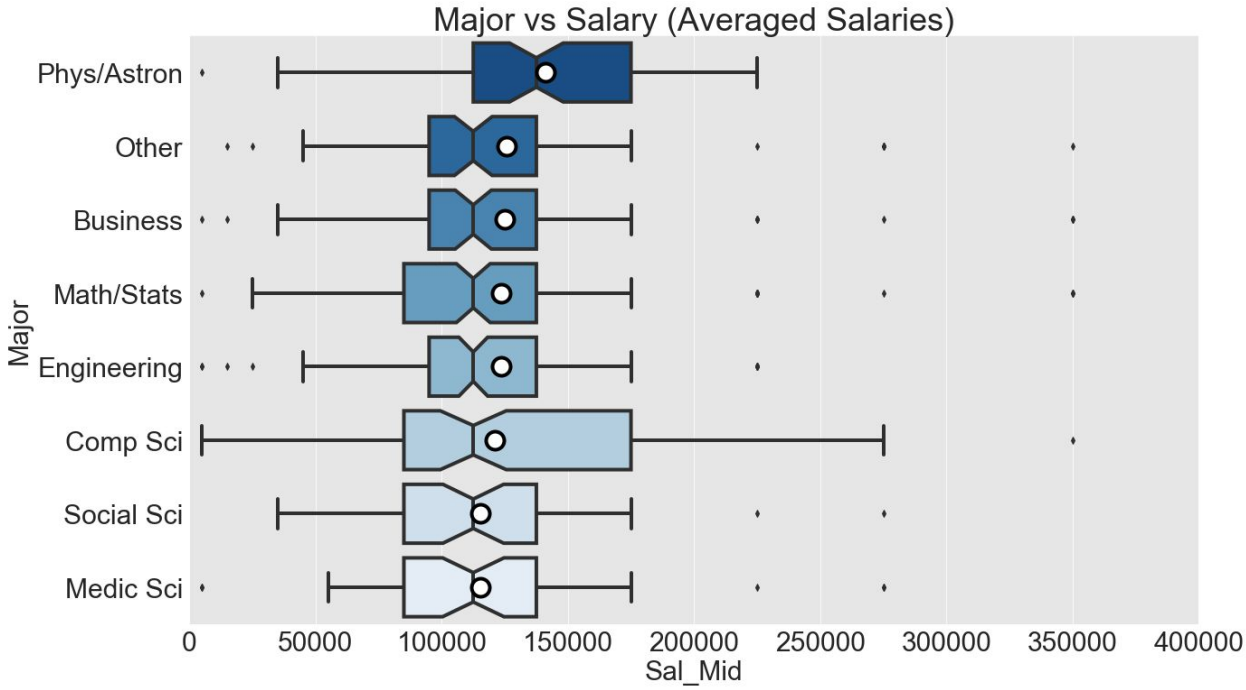

Experience vs Salary (Averaged Salaries)
>10
5-10
4-5
3-4
2-3
1-2
0-1
Experience
0
50000
100000
150000
200000
250000
300000
350000
400000
Sal_Mid



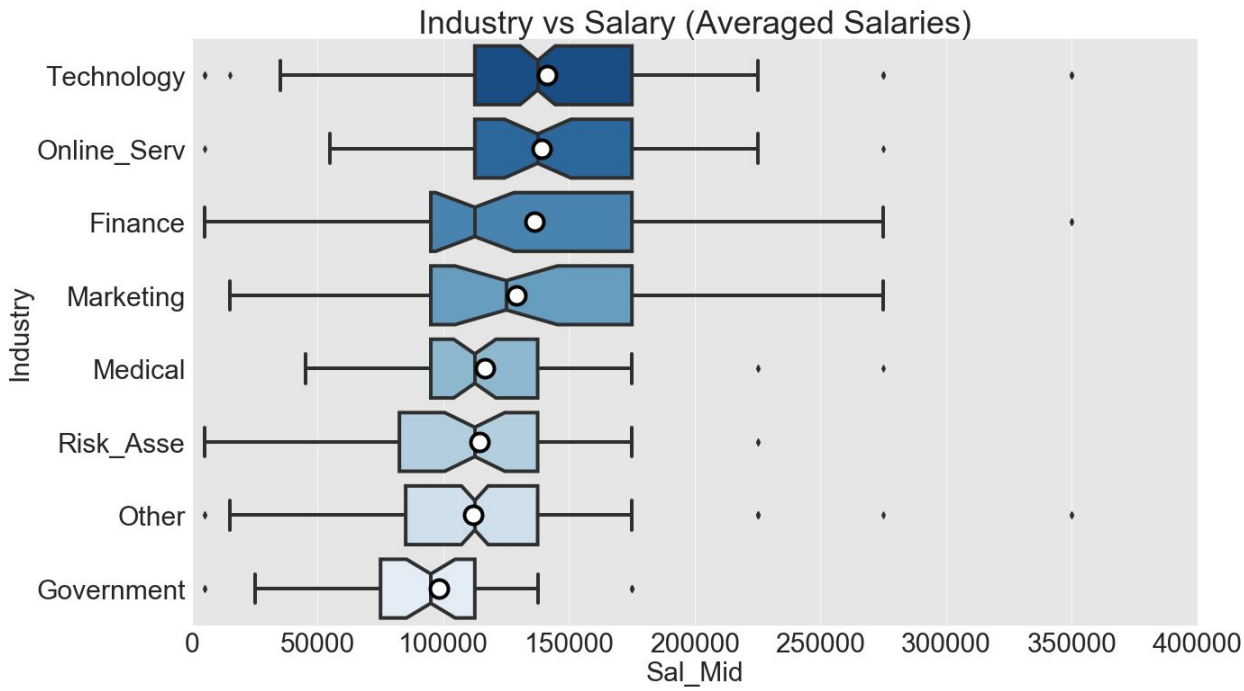
Not in line with the hypothesis was gender; Data Scientist's salaries were statistically ($p < 0.000$) biased against women (see figure below). Evaluation of Simpson's Paradox for gender bias was also conducted. It revealed that certain education levels (Master's, PhD), certain majors (Mathematics and Statistics), and certain industries (Technology, Marketing, and Risk Assessment), and other subsets are statistically biased against women.



Also not in line with the hypothesis, Major did not statistically correlate with Salary (see figure below). Even though Physics and Astronomy majors have a higher median and mean than all other groups, they were not statistically significant differences.

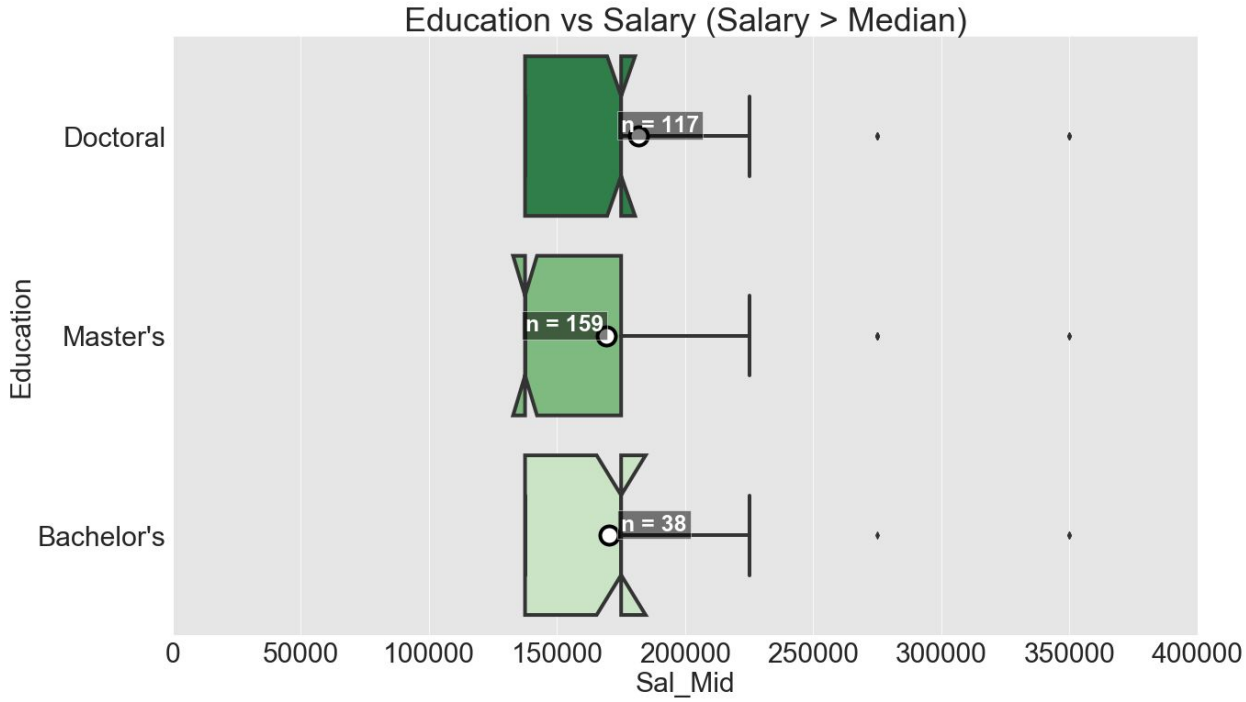


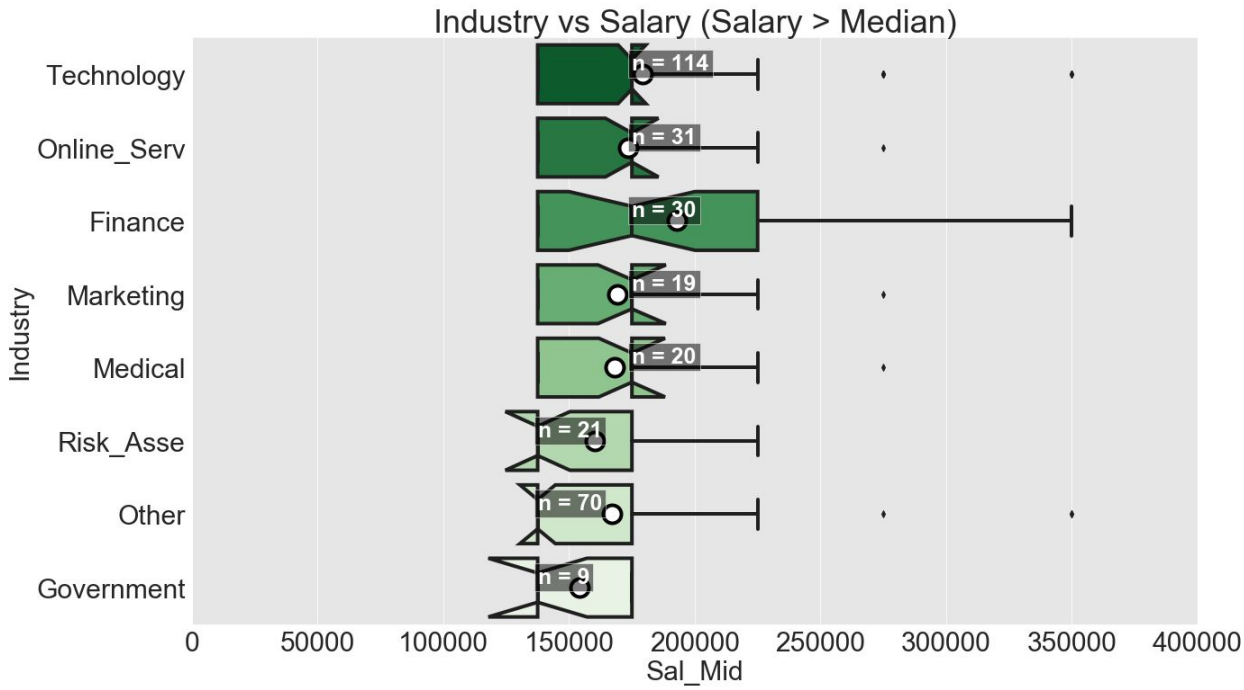
Industry had mixed results; some industries had higher salaries than others. For example, Finance was statistically more likely to yield a higher salary than Government ($p < 0.002$), as was Technology ($p < 0.000$). However, while Government had a lower average and lower median than all other industries, these differences were typically not statistically significant.



Examining the subset of Data Scientists that make more than the median salary

Exploration of the subset of Data Scientists who make more than the median salary was conducted for each feature. For that subset, Education and Industry and were no longer statistically significant factors (see figures below; sample size is shown by "n" because these groups are sometimes much smaller than the above groups). For Industry, the small sample sizes for each subgroup (some less than 30) may have caused the non-significant p values.





Machine Learning - Supervised Learning and Binary Classification

I'm interested in classification, because as a job seeker, I'm concerned with the question: how do I make at least median salary as a Data Scientist?

Preparing Data for Models

The features used in this capstone were Experience, Education, Gender, Age, Major, and Industry. They were all categorical features, and therefore standardization of any continuous features was not a necessary preprocessing step.

Models and Metric Used

The supervised machine learning algorithms used were logistic regression, k nearest neighbor, random forest, gaussian naive bayes, and extreme gradient boosting for binary classification. The metric used to choose the best model was the ROC AUC score (hereafter, simple "score") as it is independent of threshold, and therefore can indicate which model would work best across business situations.

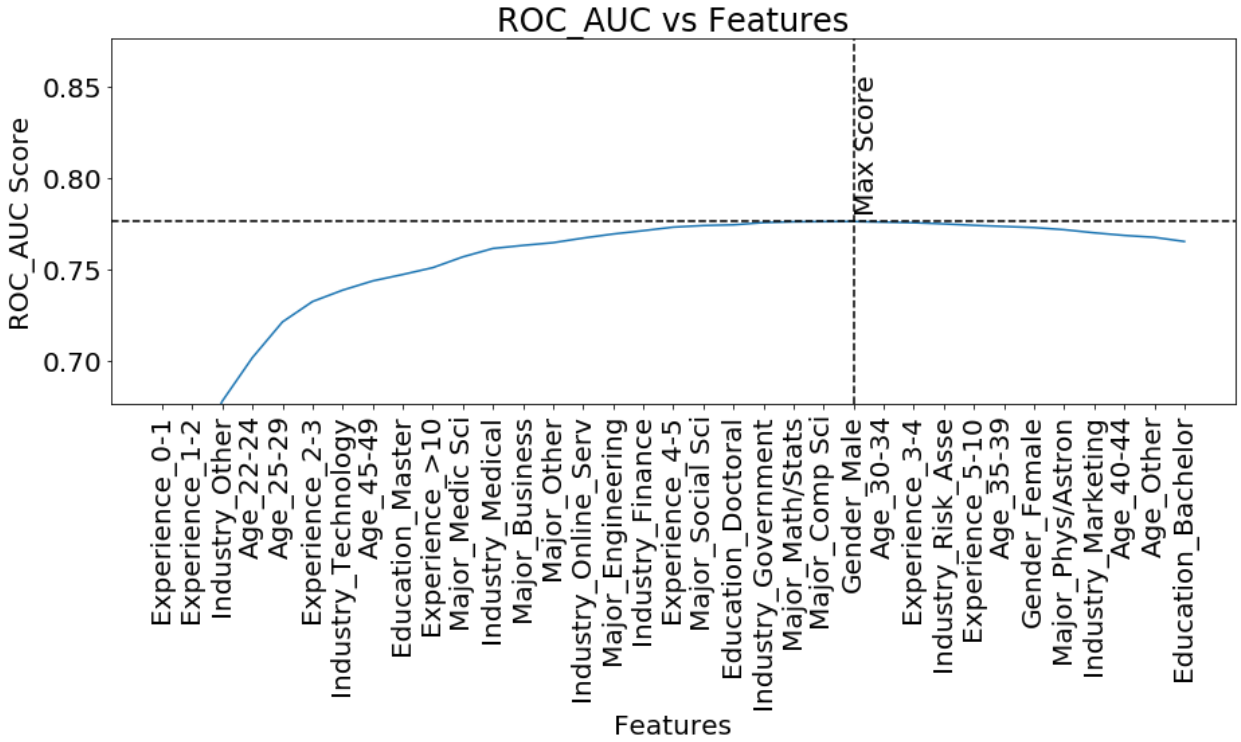
Model Selection

Each of the five algorithms were grid-searched and cross-validated (n_folds = 5). The best result was gaussian naive bayes with a score of 0.734 (see figure below for more details).

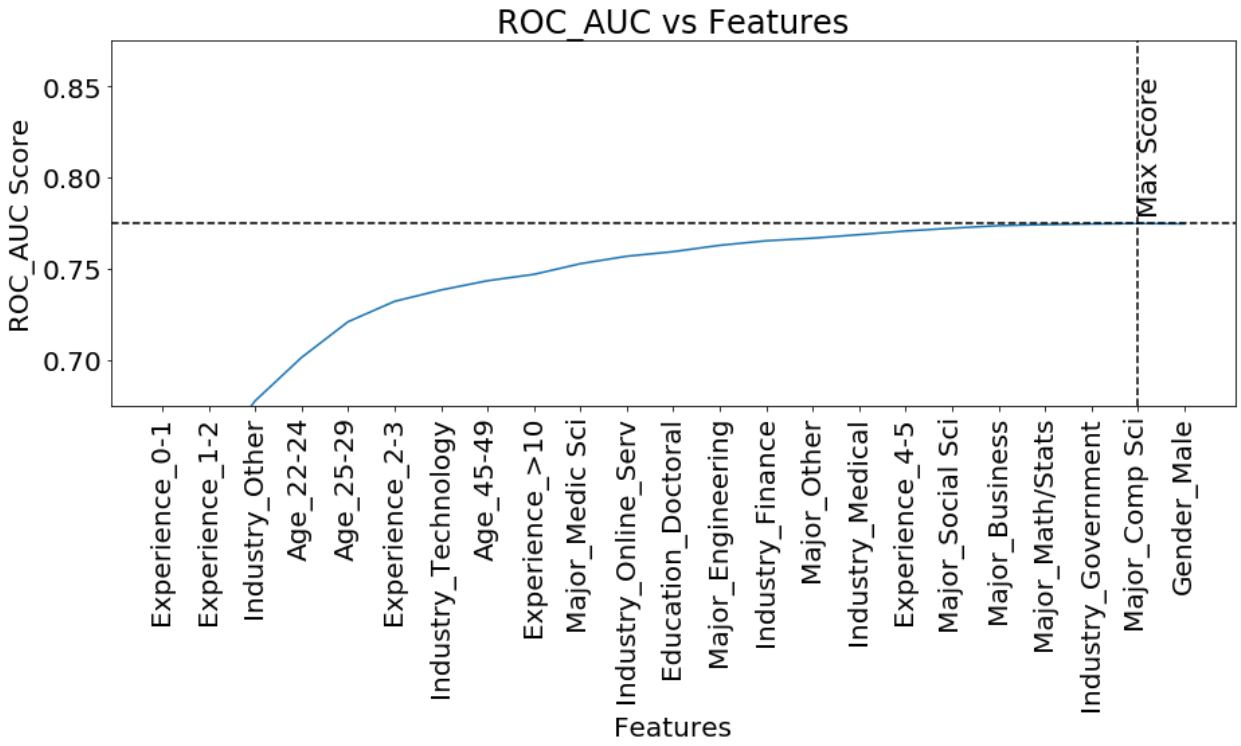
Classifier	ROC AUC Score	Best Parameters
Logistic Regression	0.729	C = 0.1
K-Nearest Neighbor	0.677	n_neighbors = 12
Random Forest	0.708	criterion = 'entropy', max_depth = 3, max_features = 'auto', n_estimators = 30
Gaussian Naive Bayes	0.734	var_smoothing = 0.1
Xtreme Gradient Boosting	0.704	alpha = 10, lambda = 10, max_depth = 2

Feature Selection

Afterwards, the features that contributed to that score were examined (see figure below). They were ordered using forward selection for which features maximized ROC_AUC score.



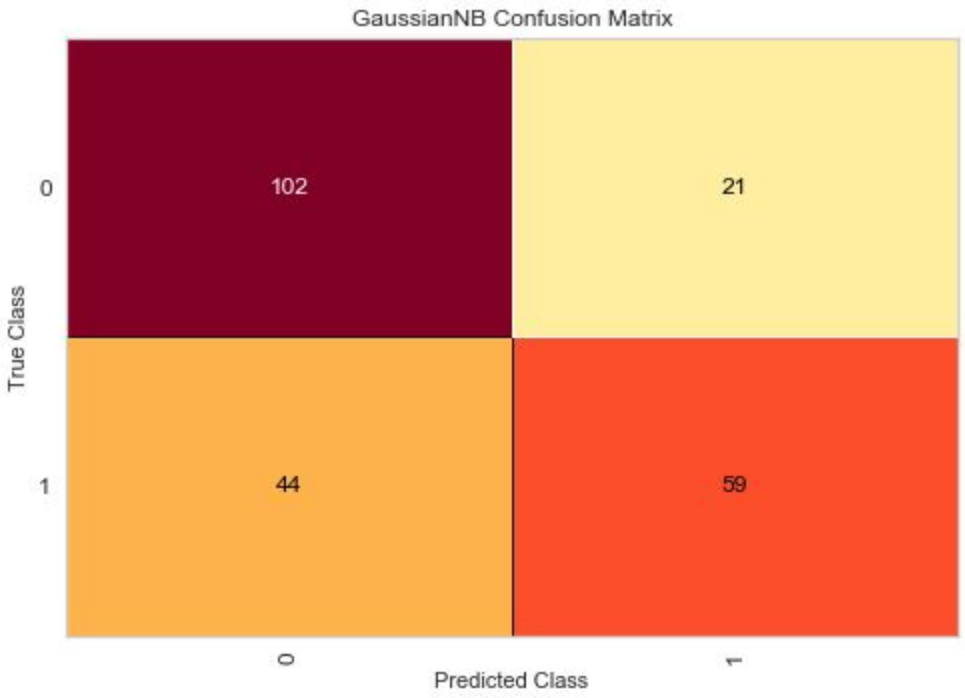
Since the maximum score was attained at Gender_Male, features after it were dropped. The score improved to 0.750. The remaining features were checked for high correlation (> 50%). Education_Doctoral and Education_Master were highly correlated (70%), and therefore Education_Master was dropped. Again feature importance was checked (see the figure below).

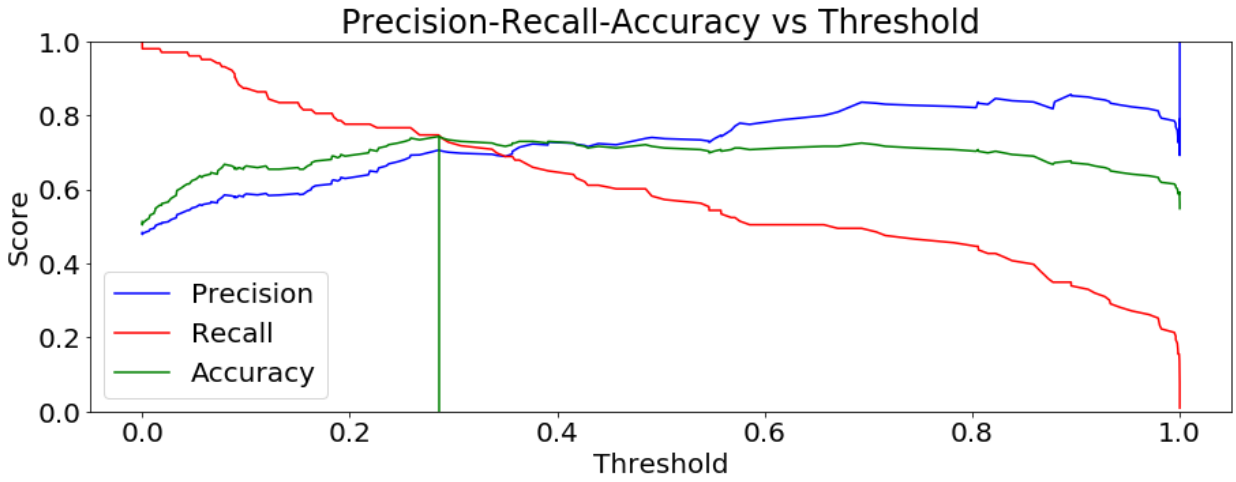


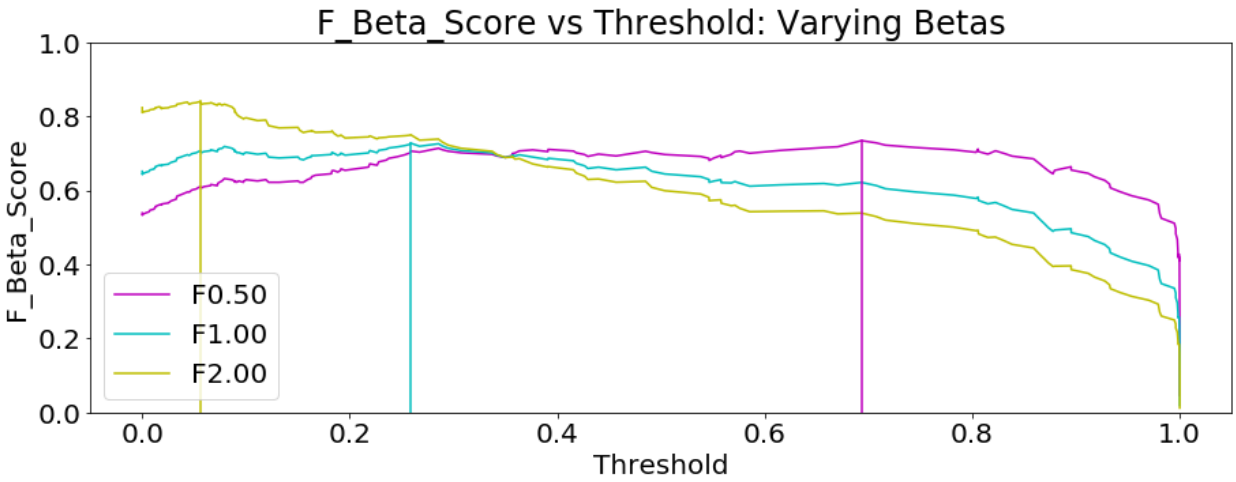
The score was 0.748, and Gender_Male became an unnecessary feature, so it was dropped. After dropping, the score returned to 0.750. Since features could no longer be dropped, highly correlated features were dropped, and a maximum score of 0.750 was obtained, the model was examined in more depth.

The confusion matrix results are shown in the figure below.

There were 102 true negatives, 59 true positives, 44 false negatives, and 21 false positives. Next, other metrics versus threshold were examined (see figures below).


GaussianNB Confusion Matrix
0
1
True Class
102
21
44
59
0
1
Predicted Class





Typically a metric such as accuracy (except in the case of imbalanced data), f1-score, R-squared, etc. are chosen for determining how well a model performs. There are three business cases considered to determine the metric and thereafter the threshold. These will be considered in the next section.

Business Case Discussion

As an employer, undervaluing an employee may cause them to leave, while overvaluing may be very costly. In this section, three business situations, their metric, and threshold will be discussed. We will look at using our model to answer the question: if a Data Scientist at a client company is asking for a raise to above average salary, should you give it to them or not?

During economic hardship it may not be feasible to pay employees their desired salary or give raises, but it is also dangerous to pay them below their market value as they may leave. In such case, giving more weight to precision (a metric sensitive to false positives) may be a valid decision so as to ensure you don't lose valuable employees who certainly are worth above average pay, but also not giving out that salary without a high degree of certainty. Typically the F1 score metric gives equal weight to precision and recall, but by varying beta, the weights can

be adjusted. A beta of less than 1 gives more weight to precision. Therefore using a beta of 0.5 may be reasonable. In such case, the F0.5 achieves its maximum score of 0.73 when the threshold is set to 0.69.

Considering the opposite economics conditions, it may be the case that business is booming and competition is especially fierce. Therefore, fear of losing a valuable employee is high. In such case, recall (a metric sensitive to false negatives) may be given more weight. A beta greater than 1 gives more weight to recall, and the maximum F2 score of 0.84 is achieved when the threshold is set to 0.06.

If neither of the above cases is applicable, another option is to use F1 - giving equal weight to recall and precision, or to use accuracy. A maximum accuracy of 0.74 can be achieved when the threshold is set to 0.29, while a maximum F1 score of 0.73 can be attained when the threshold is set to 0.26.

Assumptions and Limitations

The final dataset (limited to Data Scientists who live in the United States) was relatively small (n = 751) suggesting that conclusions drawn from such data may not be reliable. However, the median salary found in this dataset was $112,500 which is in line with H1B salary data showing the median salary as $120,000 (https://towardsdatascience.com/how-much-do-data-scientists-make-cbd7ec2b458). However, for the H1B dataset, only limited areas of the United States were covered: San Francisco Bay Area (San Francisco, San Jose, Cupertino, Mountain View, Palo Alto, etc.), Seattle (including Redmond for Microsoft), Austin, and Los Angeles (including Santa Monica).

Likewise, LinkedIn also reveals that San Francisco Bay Area Data Scientists make a $120,000 base salary ($137,000 total compensation). But for the entire United States, LinkedIn reports a base salary of $93,000 ($99,400 total compensation). This information suggests that the Kaggle dataset used here may be more in line with compensation from those who live in the west coast of the United States. For example, it may be the case that most Kagglers from the United States are also based in the west coast. It may also be the case that those from the west coast are more willing to share their salaries as the dataset used in this capstone is limited to those who shared their salary details.

Conclusion

How should companies compensate their Data Scientists? This capstone suggests that Data Scientist's median compensation is $112,500. For individual Data Scientists earning less than this, it should be noted that happiness related to financial security maxes with a $75,000 salary (https://blogs.wsj.com/wealth/2010/09/07/the-perfect-salary-for-happiness-75000-a-year/). However, businesses paying a Data Scientist less than the median may lose talent, especially talent that should be earning more than the median which was the topic of this capstone.

How important is a bachelor's, master's, or doctoral degree for a Data Scientist? What about major? Often businesses will prefer candidates with Doctoral degrees or from STEM majors (science, technology, engineering, mathematics) for Data Science roles, but this capstone was not able to show statistical evidence that those who go on to make higher salaries than the median do so because of their educational background. It may be the case that the education of a candidate may not be the best factor for including or excluding a potential Data Scientist.

Finally, it was shown that experience played an important role for the machine learning algorithm (in this case gaussian naive bayes) to predict whether a Data Scientist would make more than the median salary.