A Data Scientist's Value: Are you worth more than the median Salary?

Victor Palacios

Who should care?

- DS Employers trying to evaluate their workers
- DS Job-seekers trying to figure out how best to stand out
- DS Students wondering which major will help most with a Data Science career

Problem: Can these predict salary?

1. Gender (i.e., Male or Female)
2. Age
3. Education (i.e., Bachelors, Masters, or PhD)
4. Experience (i.e., # of years in current position)
5. Major
6. Industry

Data: 2018 Kaggle ML & DS Survey

Original Dataset: 23,859 rows and 395 columns

This Capstone: 751 rows and 7 columns

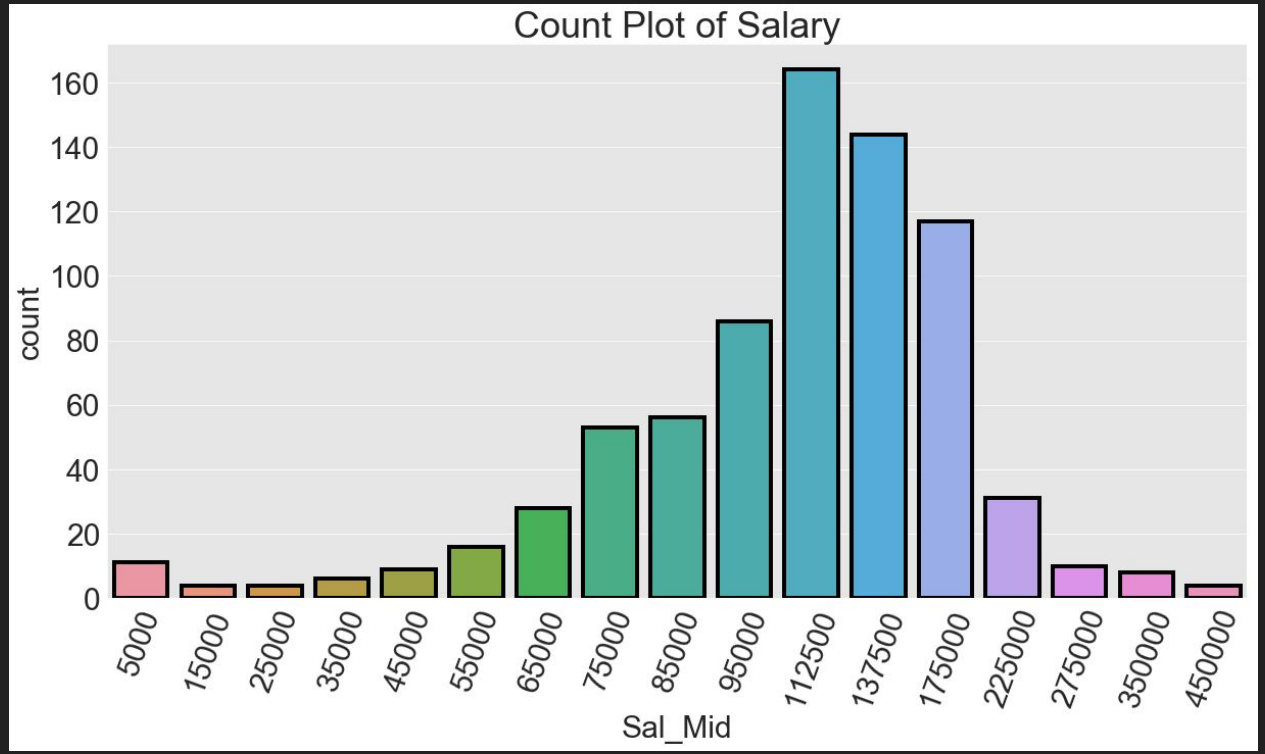
Why? Limited to:

(1) Residents of the USA,

(2) Share their salary details, and

(3) Are employed as Data Scientists

Salary Distribution for Kaggle Data Scientists



Hypothesis'

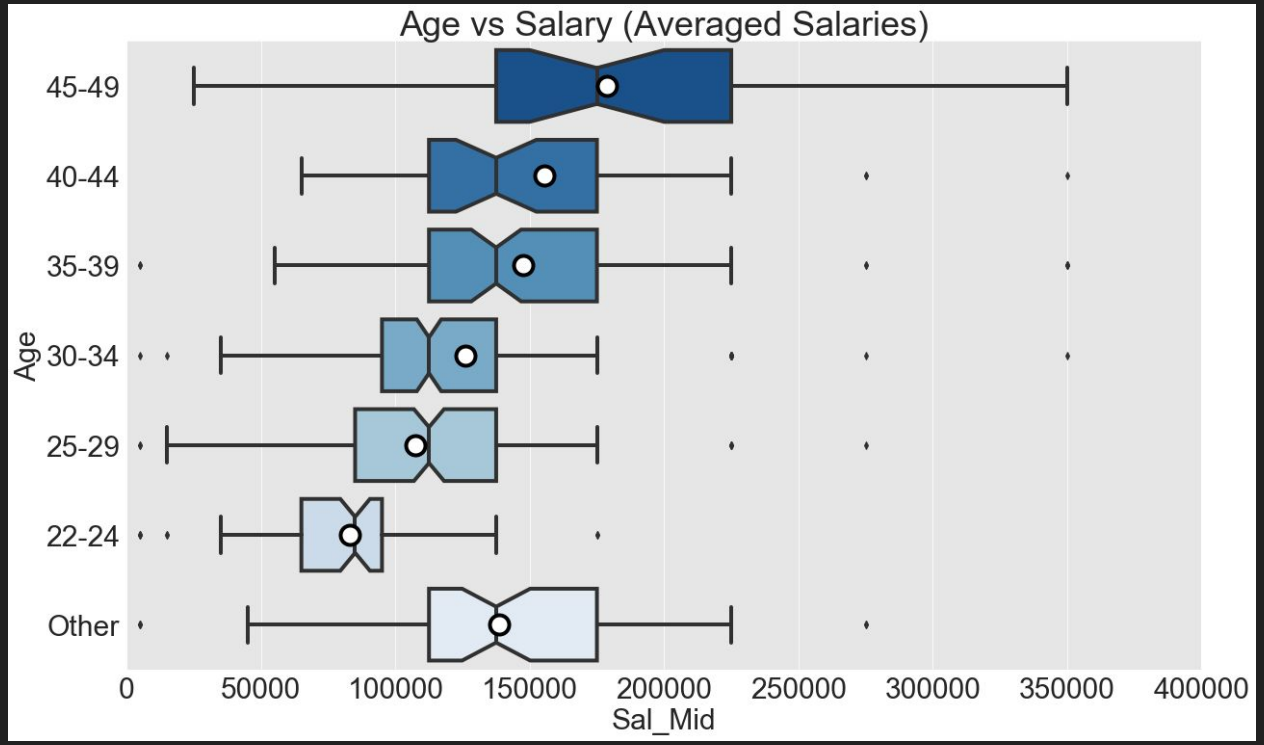
Positively Effect

1) Older Age
2) More Experience
3) STEM Majors
4) PhD

Negatively Effect

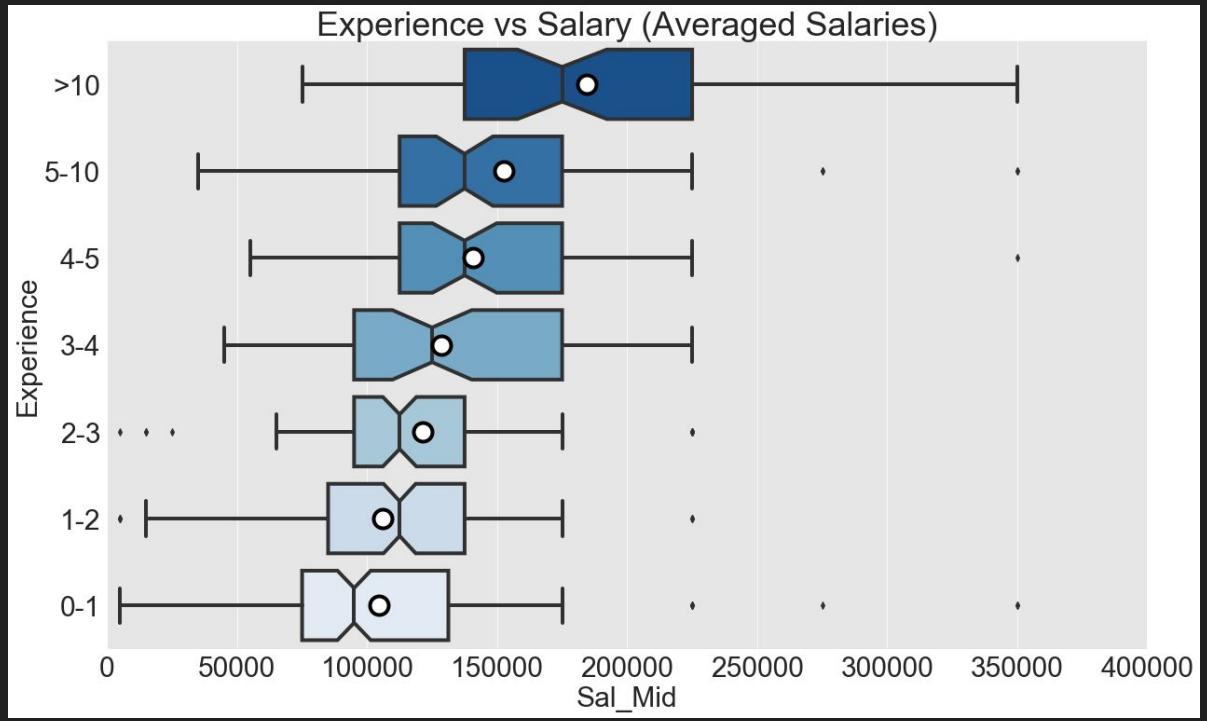
1) Younger Age
2) Less Experience
3) non-STEM Majors
4) Bachelor's

The Older, The Richer



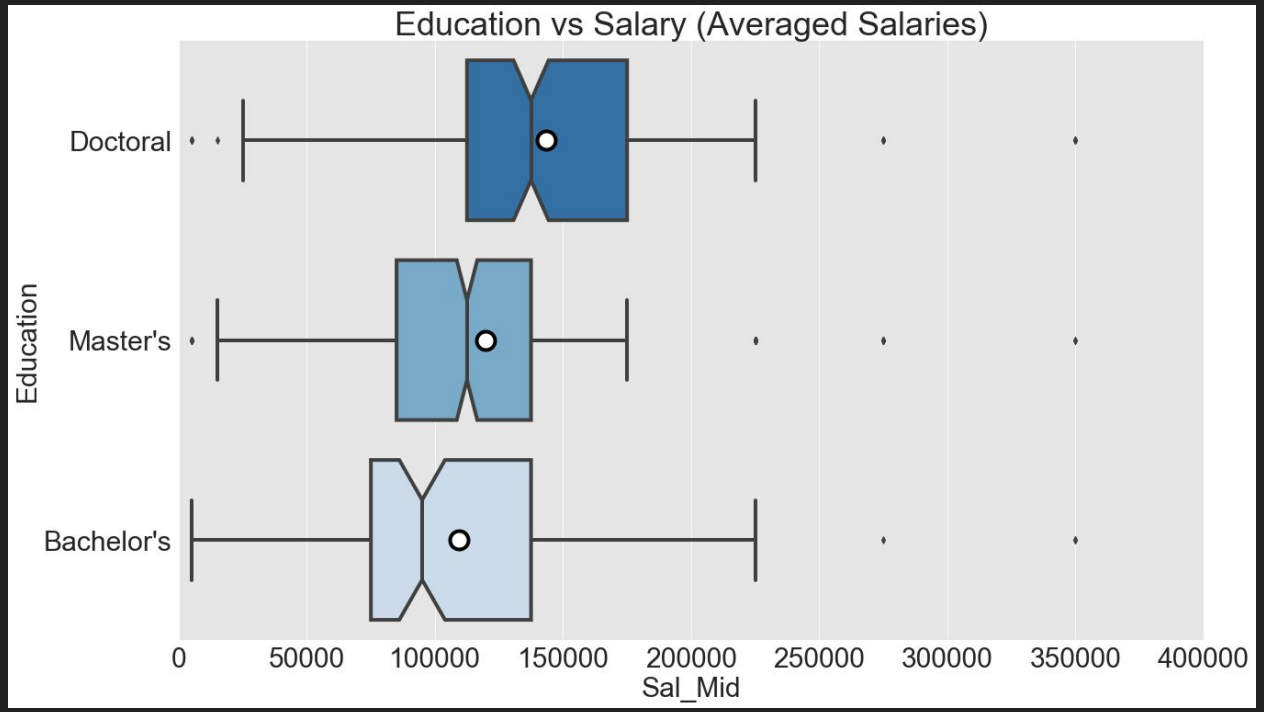
Notch = Median, Dot = Mean

More Experience, More Salary



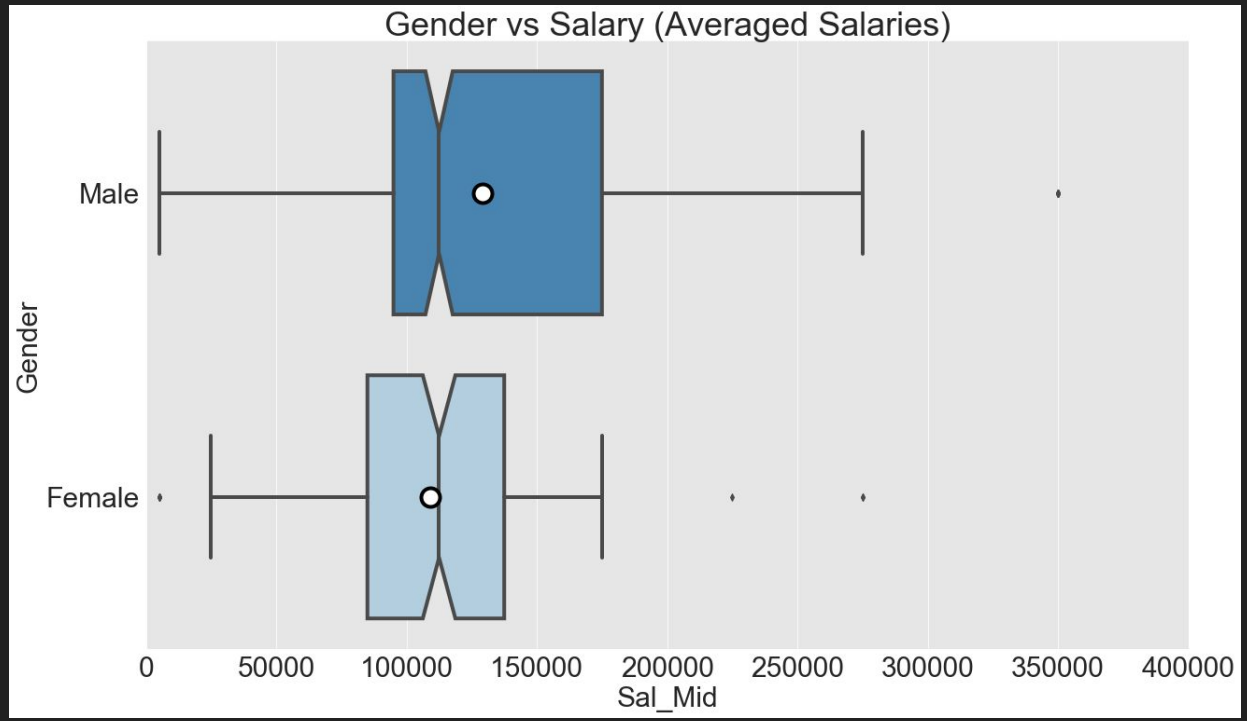
Notch = Median, Dot = Mean

PhD = Higher Salary



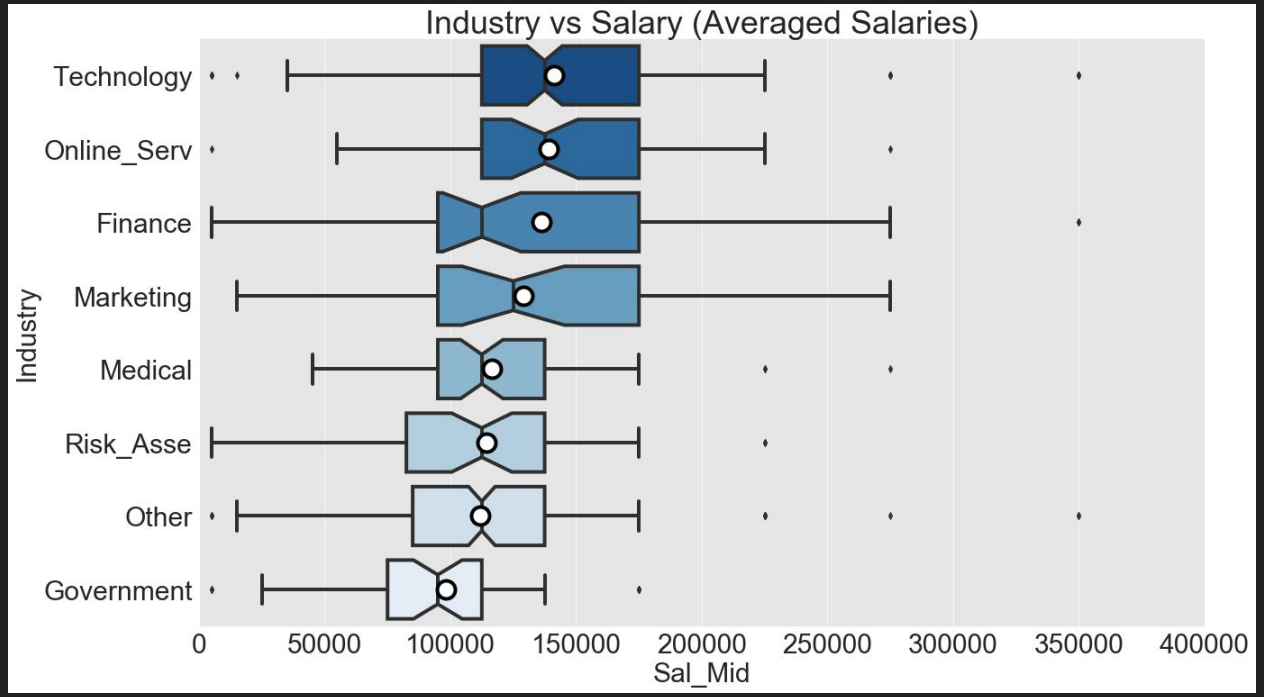
Notch = Median, Dot = Mean

Male? Higher Salary



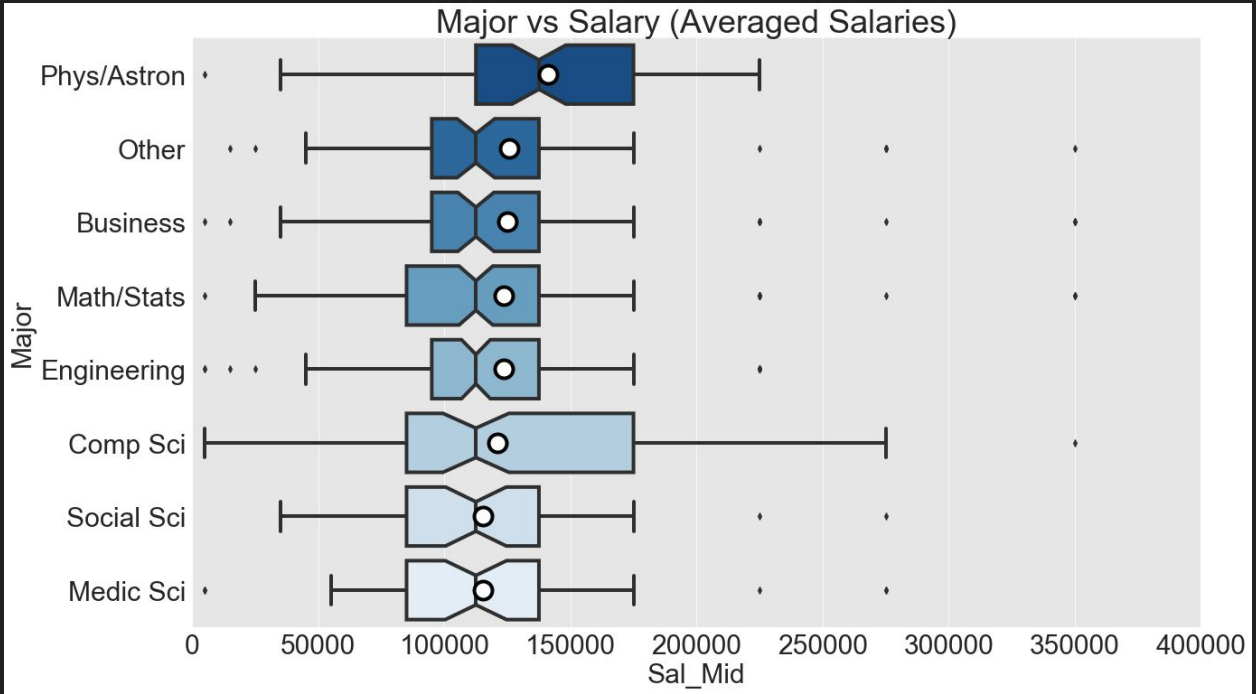
Notch = Median, Dot = Mean

A Trumped Government Does Worst



Notch = Median, Dot = Mean

Major did not correlate with Salary



Notch = Median, Dot = Mean

Results

Positively Effect

1) Older Age
2) More Experience
3) ~~STEM Majors~~
4) PhD
5) Male*

Negatively Effect

1) Younger Age
2) Less Experience
3) ~~non-STEM Majors~~
4) Bachelor's
5) Female*

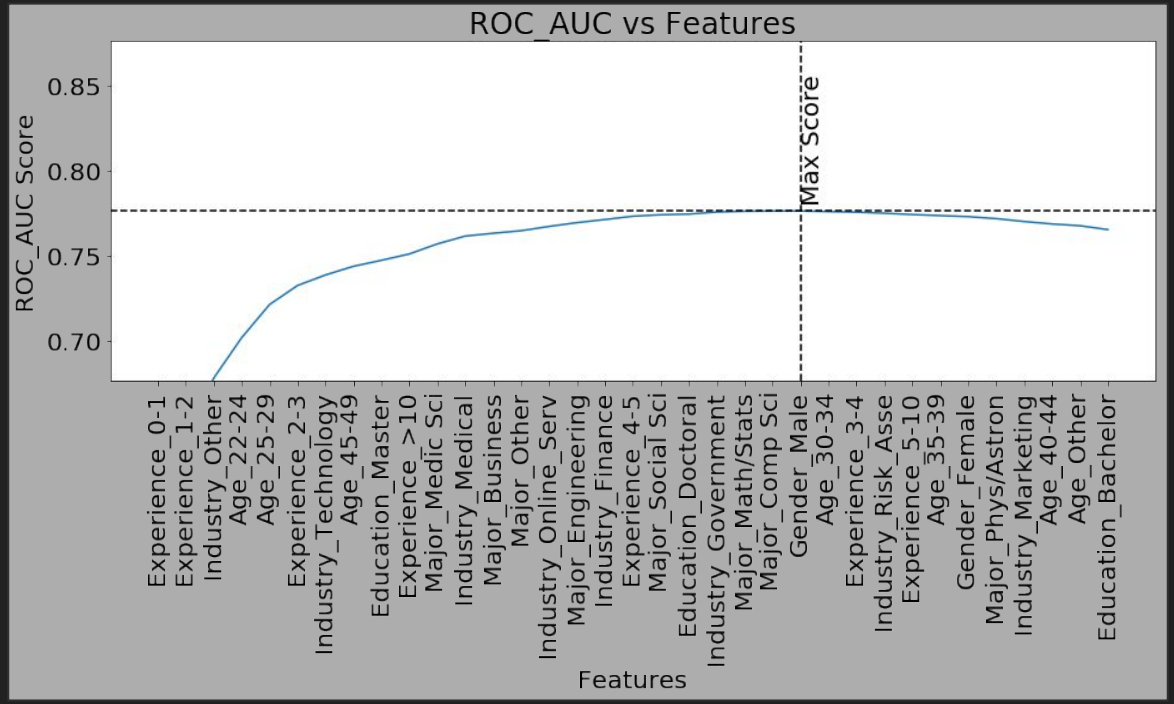
* Simpson's paradox did not appear to be at play.

Machine Learning

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

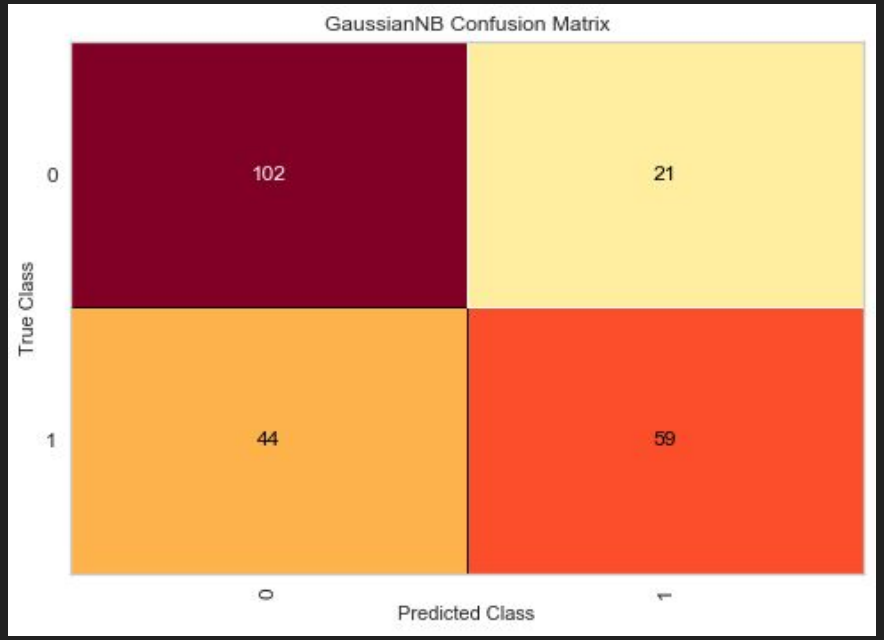
Best model was Gaussian Naive Bayes

Forward Selection Stepwise Regression



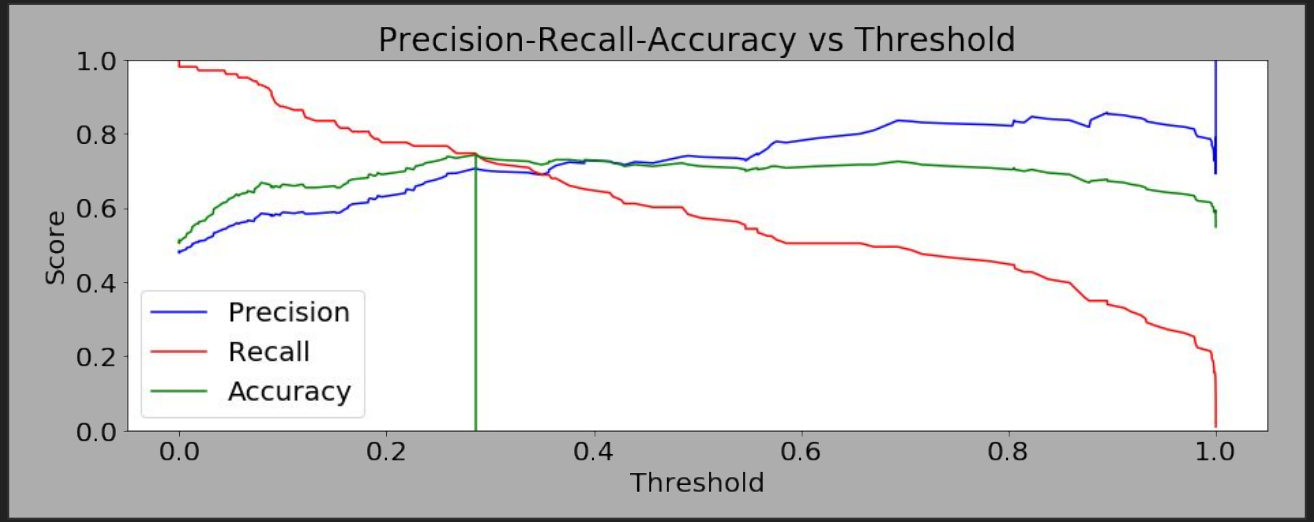
Forward selection used for feature reduction

Confusion Matrix



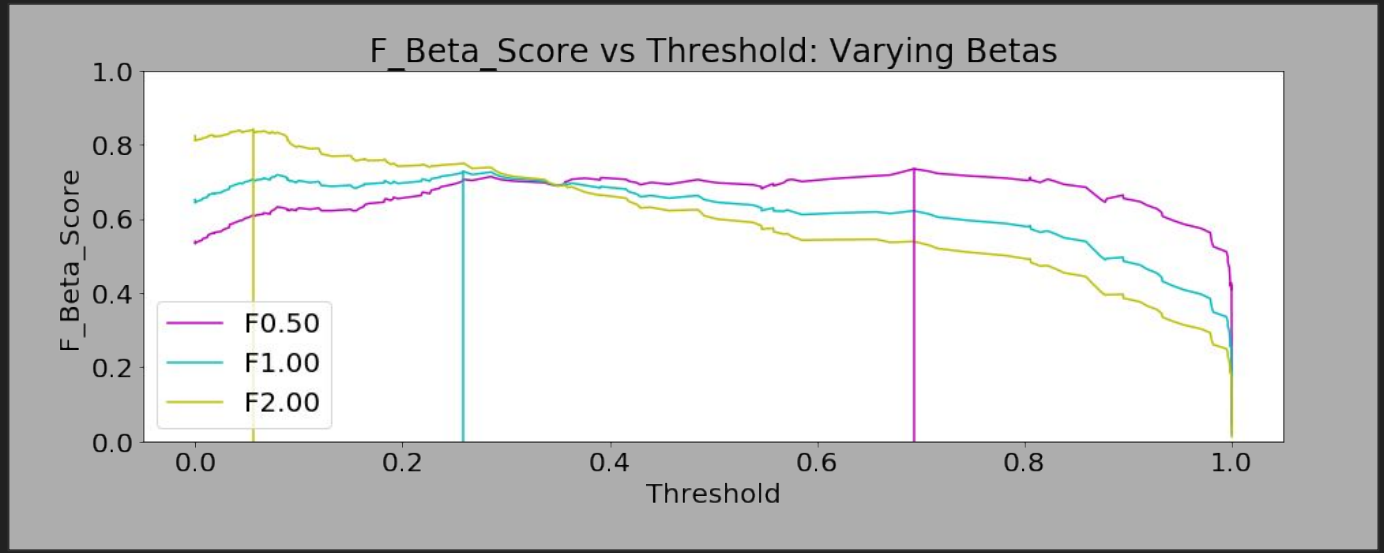
Large number of false negatives ~= true positives

P, R & ACC, vs Threshold



Benchmarks used for measuring model usefulness

F-Beta Scores vs Threshold



Varying Beta gives more weight to precision or recall

Business Case Example

Background: Economic hardship

Problem: Not feasible to pay employee desired salary or give raises, but it is also dangerous to pay them below their market value as they may leave.

Solution: More weight to precision; F 0.5 achieves its maximum score of 0.73 when the threshold is set to 0.69.

Actionable Recommendations

- DS Employers: Value women equally; Stop asking for Quantitative Majors.
- DS Job-seekers: Government typically pays lower than other sectors; avoid if a high salary is desired.
- DS Students: If in doubt, go with Physics.