Supervised Learning Capstone

Are you a Data Scientist?

Logistic Regression Model

Kim Nowell-Berry

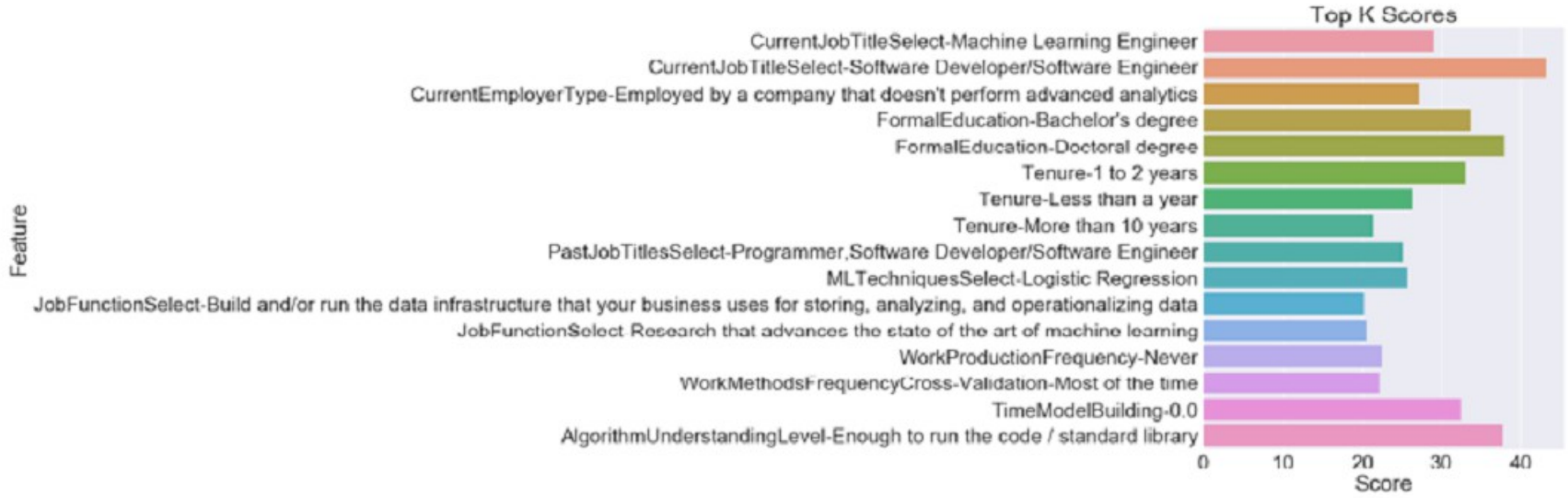
Data

- Kaggle Survey Data
- 16,000 Professionals working in the field of Data Science
- What features ultimately define the people that call themselves Data Scientists?

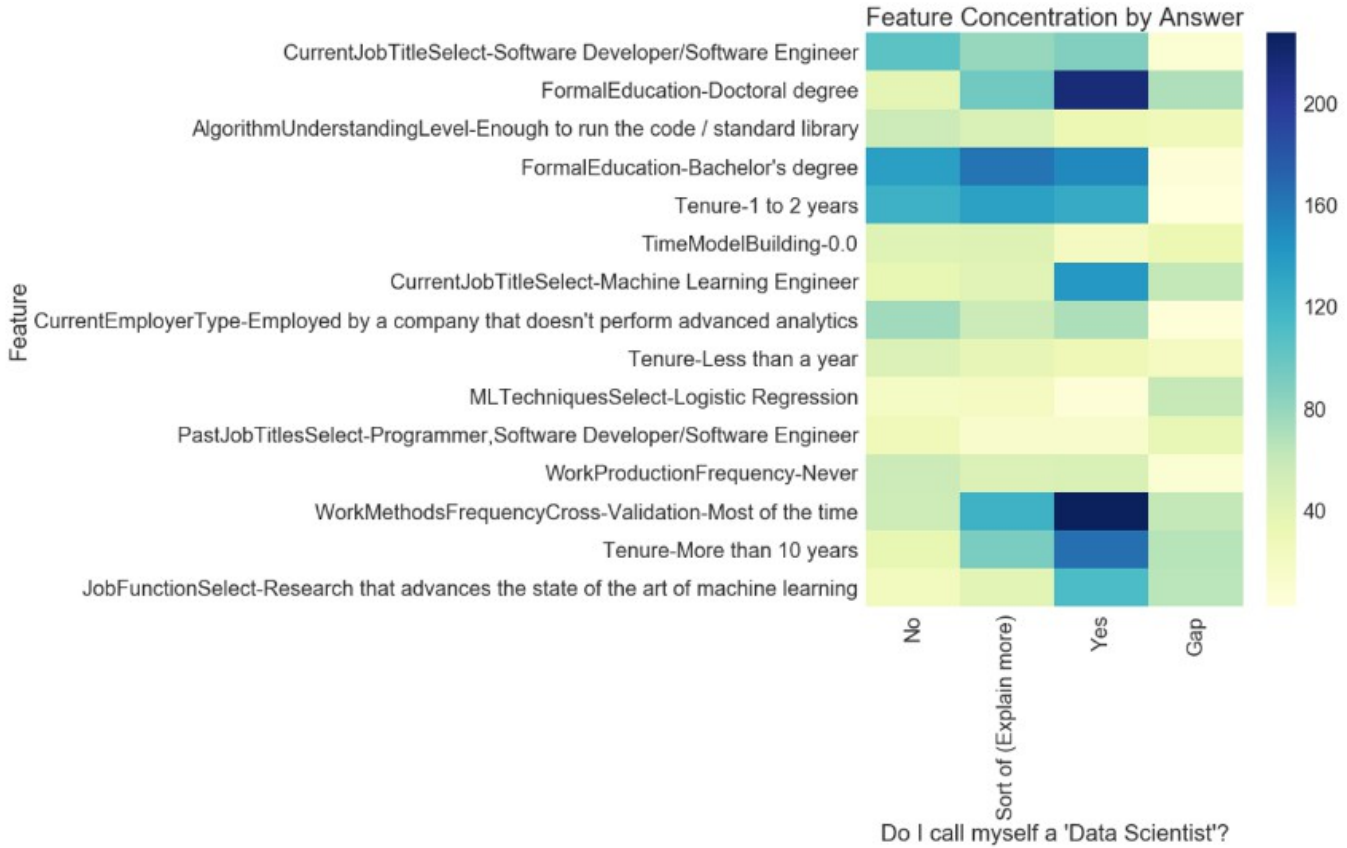
Modeling Approach

- The features of the model were the questions and responses to several survey questions.
- Feature engineering to transform some of the features
- Feature selection and reduction
- K-Best with Logistic Regression

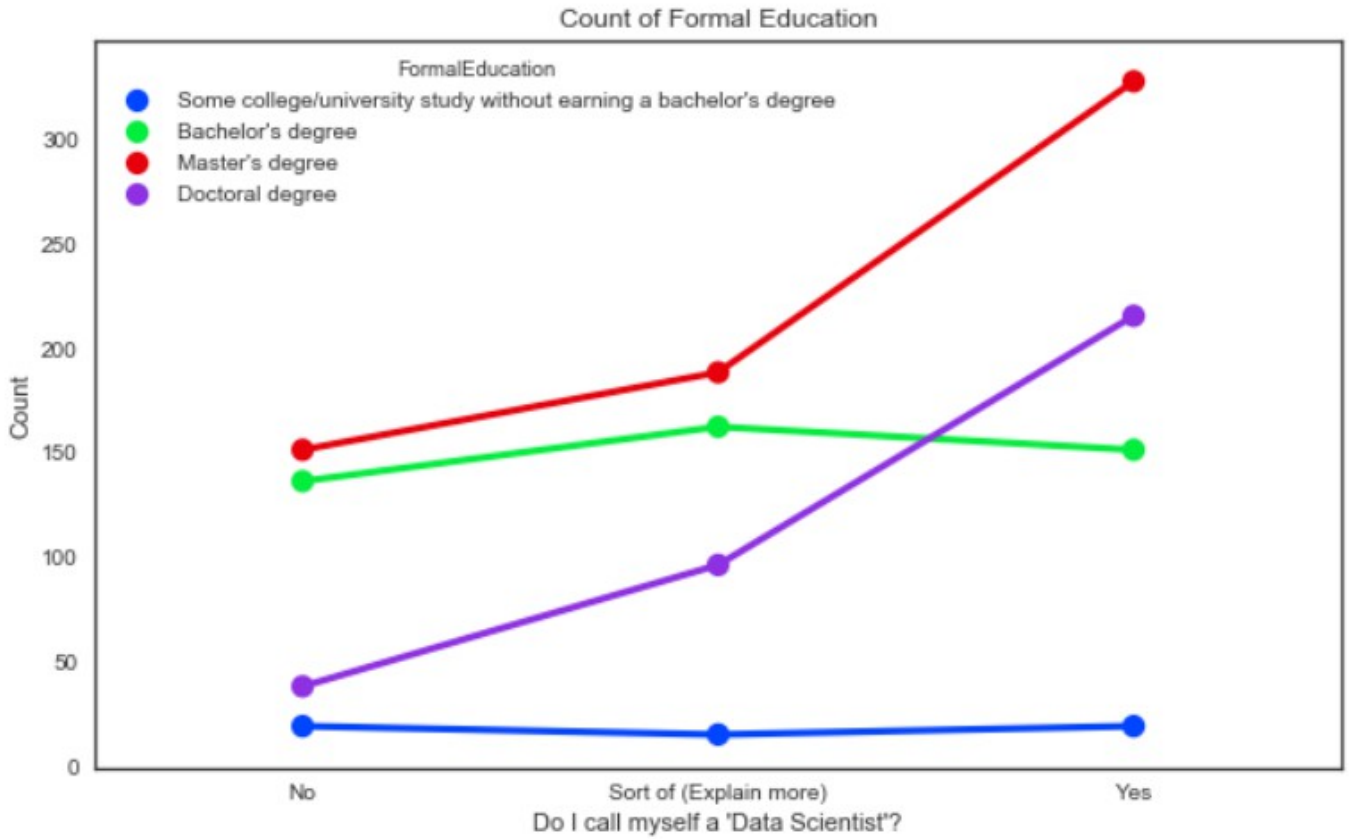
Top K Features



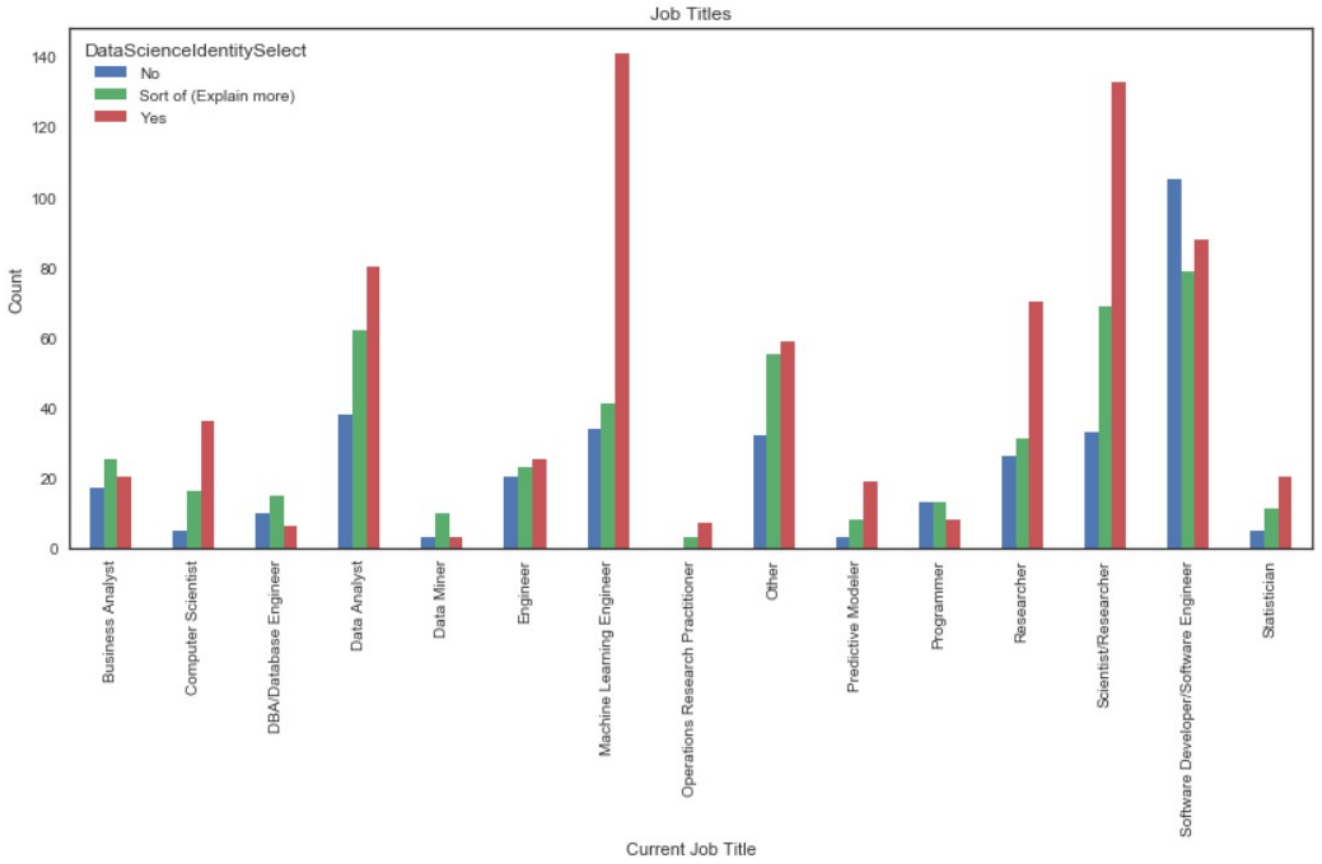
Feature Concentration



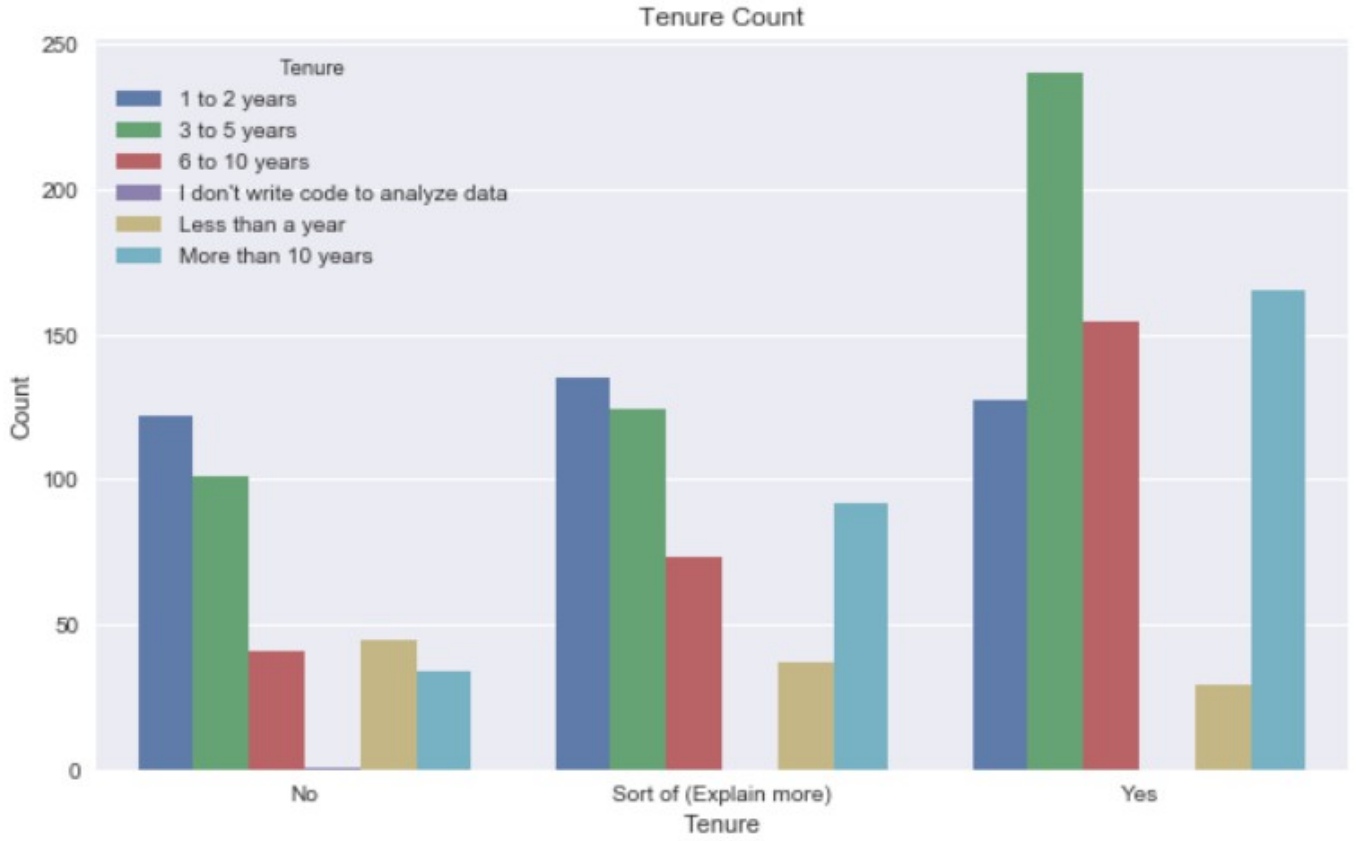
Formal Education



Current Job Titles



Tenure


Tenure Count
Tenure
1 to 2 years
3 to 5 years
6 to 10 years
I don't write code to analyze data
Less than a year
More than 10 years
Count
250
200
150
100
50
0
No
Sort of (Explain more)
Yes
Tenure

Prediction

Using myself as an example, I predict the probability that I would call myself a Data Scientist. In this case, the model accurately predicts that I would not call myself a Data Scientist.

Below are the final features used in the model:

CurrentJobTitleSelect-Machine Learning Engineer = False

CurrentJobTitleSelect-Software Developer/Software Engineer = True

CurrentEmployerType-Employed by a company that doesn't perform advanced analytics = False

FormalEducation-Bachelor's degree = False

FormalEducation-Doctoral degree = False

Tenure-1 to 2 years = False

Tenure-Less than a year = False

Tenure-More than 10 years = True

PastJobTitlesSelect-Programmer,Software Developer/Software Engineer = True

MLTechniquesSelect-Logistic Regression = False

JobFunctionSelect-Build and/or run the data infrastructure that your business uses for storing, analyzing, and operationalizing data = True

JobFunctionSelect-Research that advances the state of the art of machine learning = False

WorkProductionFrequency-Never = False

WorkMethodsFrequencyCross-Validation-Most of the time = False

TimeModelBuilding-0.0 = False

AlgorithmUnderstandingLevel-Enough to run the code / standard library = True

Results

- The model accuracy score when including all of the features was very low at 45%.
- After feature selection, the model accuracy score was still very low on the training set at 54%, but higher than the initial model.
- PCA was applied as well as K-Best for feature selection, however, K-Best worked better.

Conclusion

- People can use this model to identify what skills or features need to be enhanced to ultimately achieve their goal of becoming a Data Scientist.