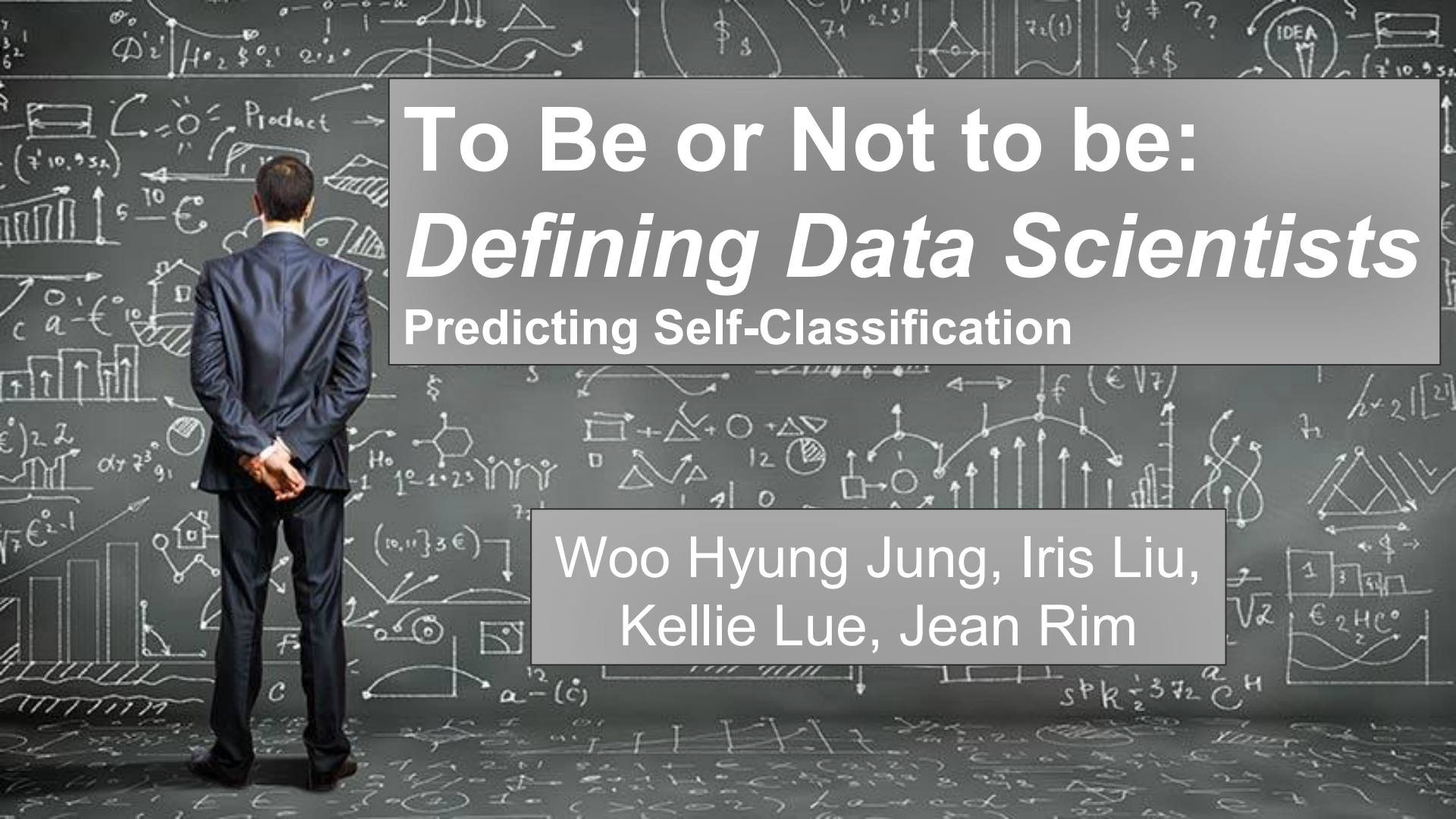


To Be or Not to be: *Defining Data Scientists*

Predicting Self-Classification

Woo Hyung Jung, Iris Liu, Kellie Lue, Jean Rim

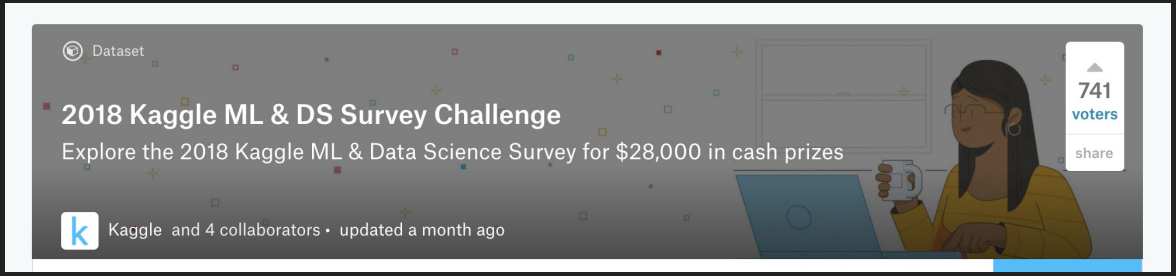
Research Question and Data

Research Question

We hope to determine what variables are significant when predicting whether an individual considers themselves a data scientist, and then use those variables to build a predictive model.

Data Explanation

- Multiple Choice Response data from “2018 Kaggle ML & DS Survey Challenge”
- 23860 rows and 395 columns
- Each row is a response to the survey
- 50 questions were surveyed → Answers are distributed over 395 columns

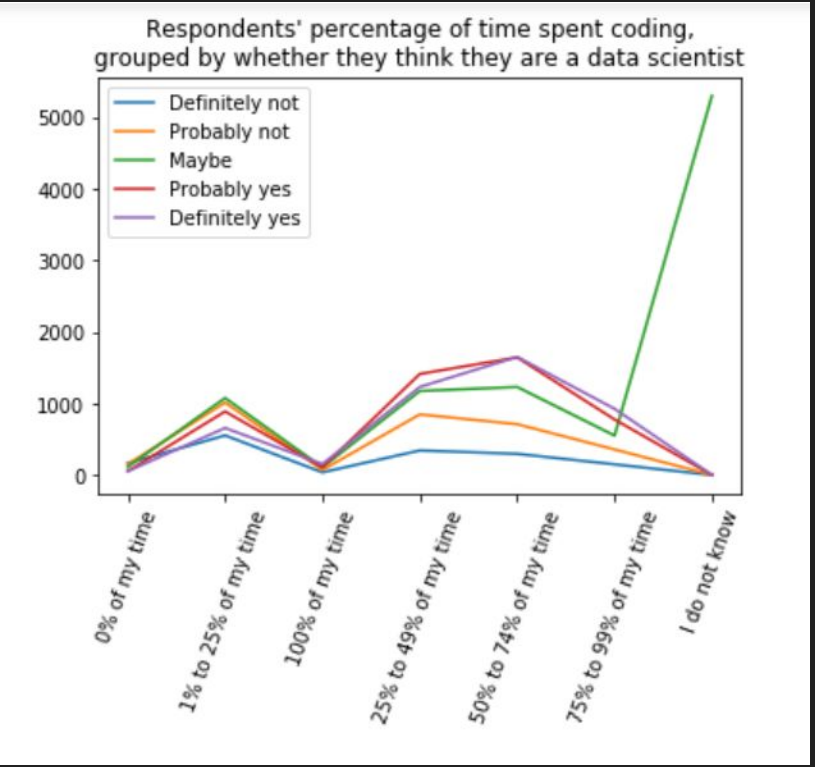


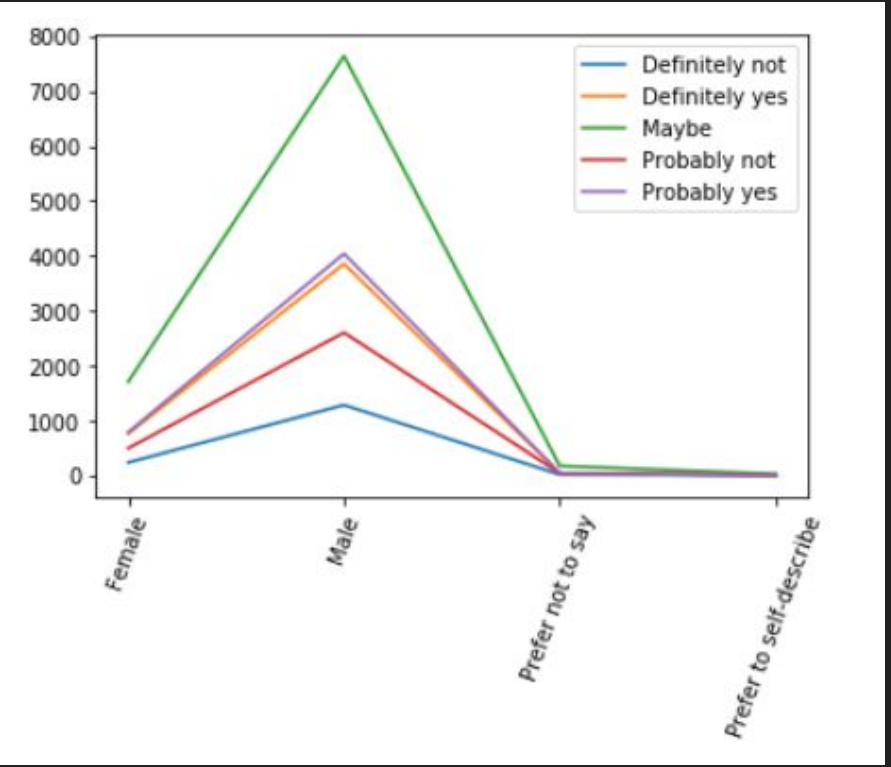
Data Cleaning

Began with data difficult to work with

- Subsetted data set based on interesting and intuitively important variables for our model
- Dealt With NAs column by column
- Converted categorical and string variables to int in order to be able to model more easily
    - Dummy variables
    - Binary classification
    - Medians

Exploratory Analysis





Model

- Logistic Regression Model
- Predict whether or not an individual considers themselves a data scientist

Data Scientist ~ Gender + Undergrad Major + Current Title + Years of Experience + Yearly Compensation + Primary Tool + % Time Coding + Coding Experience

- 60.74% accuracy
- Cross validation confirms relative accuracy

Conclusion

Determination of a data scientist

The most important factors in predicting whether or not an individual considers themselves a data scientist include those related to gender, experience coding, and educational background.

In order to create a better predictive model

- Further analyze the groups we created within our columns to turn our categorical and string data numerical
- Break up each column into more groups
- Work on a way to better deal with non numerical data for predictive methods