Leveraging Data Analytics and Machine Learning to Predict Data Scientists Salaries

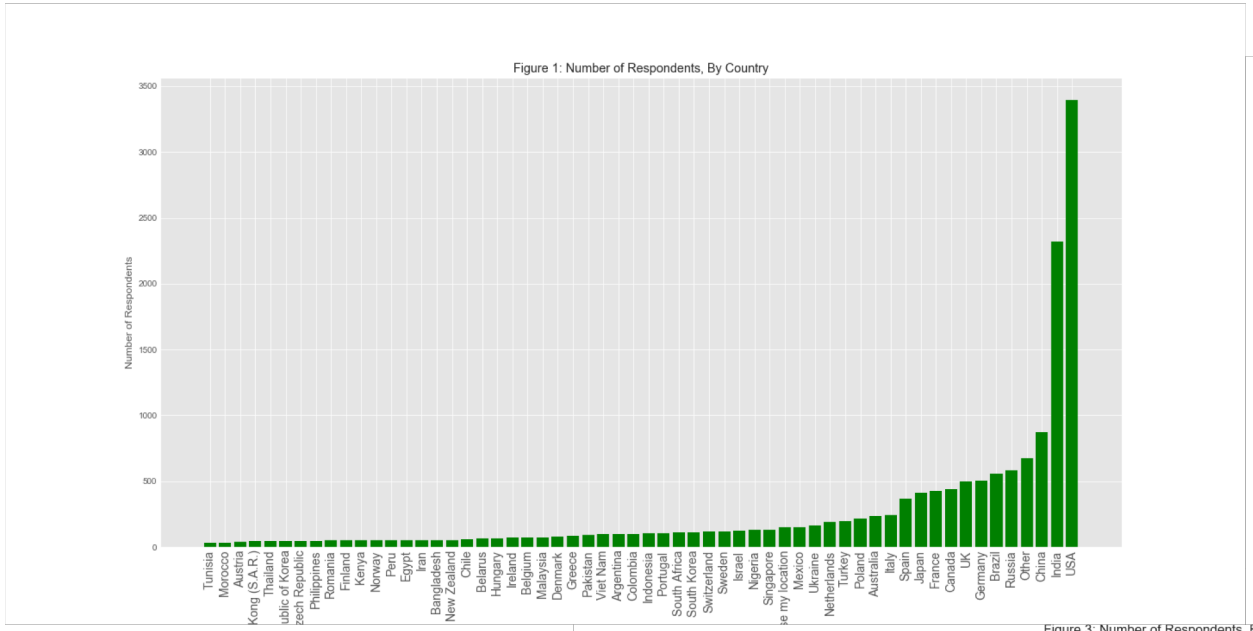
Tammi Hawa

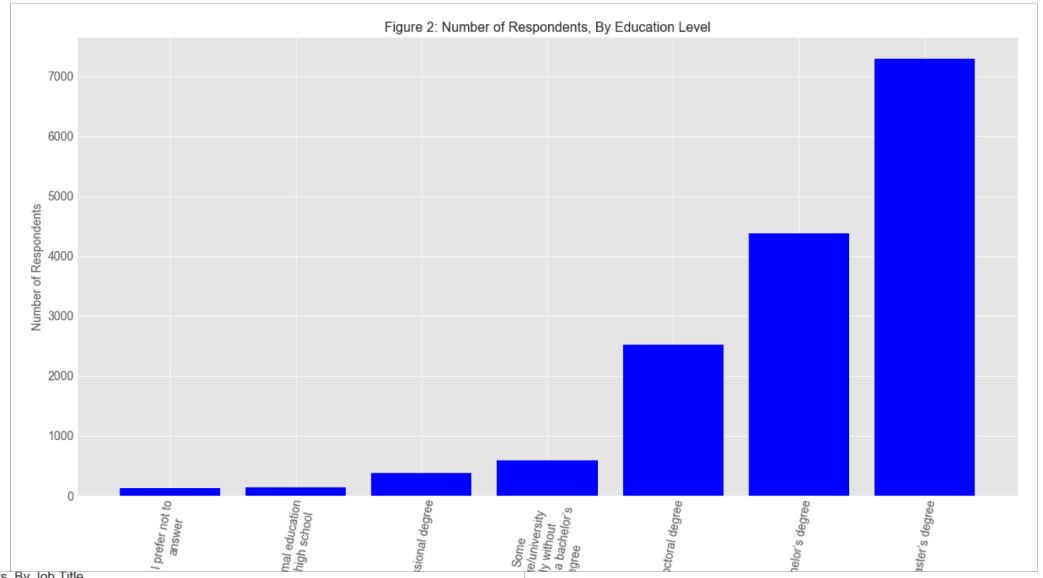
1000311655

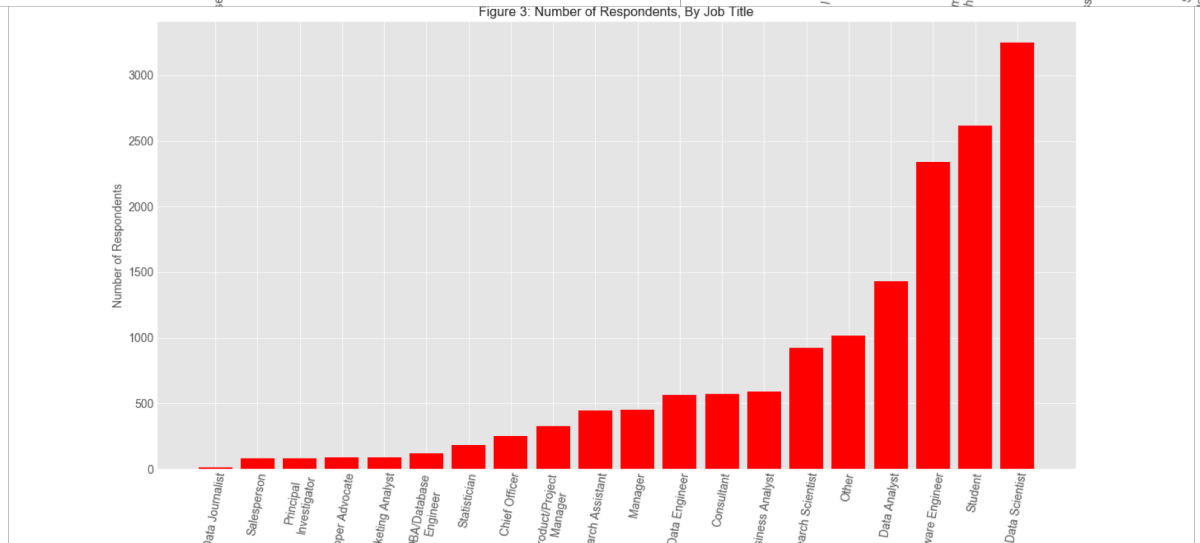
Exploratory Analysis

- The location of respondents, level of education, as well as job title were all investigated because these seemed like the obvious reasons why salaries would vary.
  - FOUND: Most of the respondents were from the United State or India, had attained either a master's or bachelors degree, and worked under the title of data scientist, software engineer, or data analyst (with a large number of respondents also being students).
- Additionally, a correlation matrix was constructed to determine which features had a strong correlation to yearly compensation.
  - FOUND: age, years of experience, years writing code to analyze data, years use machine learning methods, and living in the USA, all had a noticeable positive correlation with yearly compensation (>=0.2)

Exploratory Analysis - Graphs







Model Feature Importance

- One-hot encoding was used to handle categorical features
- Lasso Regression was used to determine which features in the models were significant. Coefficients with a value of 0 were discarded from the model.
  - Highly positive coefficients (represent a positive correlation with yearly income):
    - country_United States of America (58,351)
    - country_Switzerland (55,551)
  - Highly negative coefficients (represent a negative correlation with yearly income)
    - Data_vis_library_Altair (-30,921)
    - language_Julia (-21,893)
  - Coefficients with value 0 (does not affect yearly income, removed)
    - country_Italy, data_type_Numerical Data, education_Master's degree, industry_Marketing/CRM, language_R, major_A business discipline, Data_scientist?_Probably_yes

Models Used

- The following models were trained against the training data and accuracy score of R^2 was used to measure the fit of the trained model to the data:
  - LASSO Regression – leveraged because it is normalized using the L_1 norm, is supposed to decrease error due to variance.
  - Ridge Regression – leveraged because it is normalized using the L_2 norm, is supposed to decrease error due to variance (and wanted to compare its performance to the Lasso regression)
  - KNN Regression – leveraged because it works based on feature similarity
  - Random Forest Regression – leveraged because its an ensemble method and can handle diversity and a large number of features

Model Results

- I first ran 10-fold cross validation on each model, then conducted hyperparameter tuning to identify the optimal parameters for each model, and then again ran 10-fold cross validation
- The selected model was: Random Forest Regression
  - Ideal parameters: samples with bootstrapping, has 30 decision trees, considers the square root of the number of features when making a split, and splits on the mean absolute error.
  - In the hyperparameter tuning stage, this model had a R^2 value of 0.89, significantly higher than the other models (all hovering around an accuracy score of 0.49)
  - However, after the ideal parameters were identified and cross validation was performed, the R^2 value of dropped to 0.45, possibly because less data was available in each training stage
  - When I trained the model identified in the hyperparameter tuning and tested it on the test data, the R^2 value comparing the test set to the predictions was 0.41
    - The model was overfitted to the training data, possibly because not enough trees in the forest were built and the high complexity of each decision tree