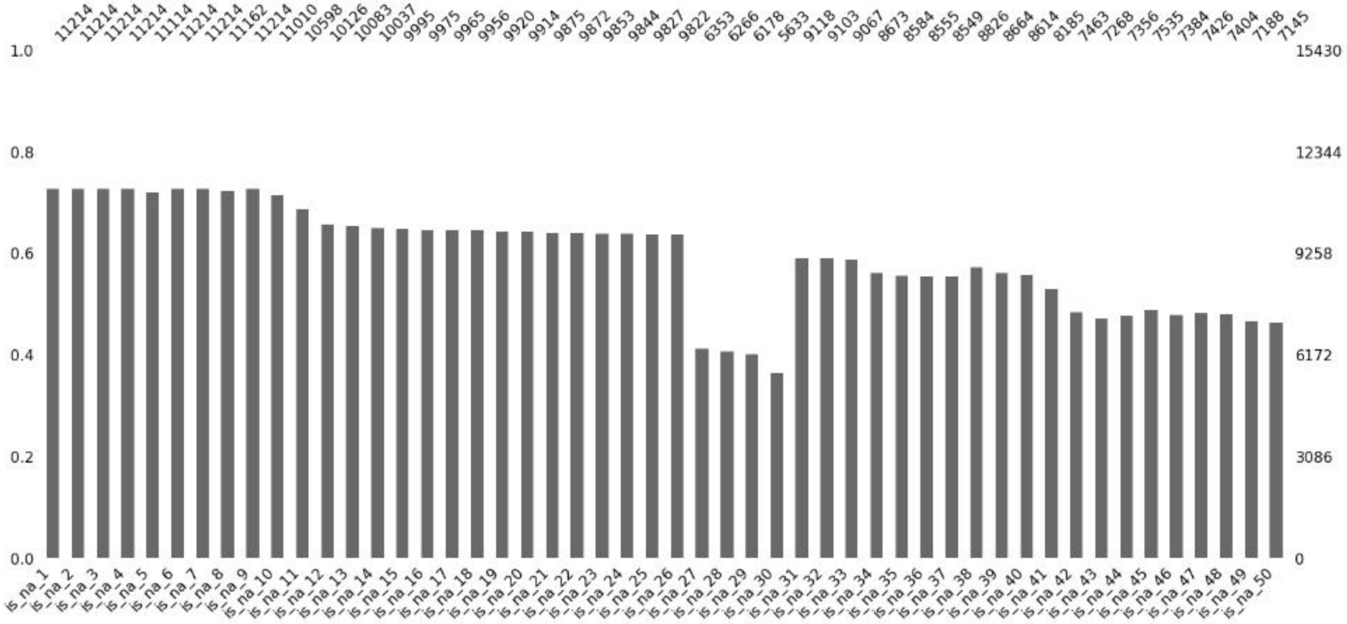
# Kaggle Survey Salary Prediction

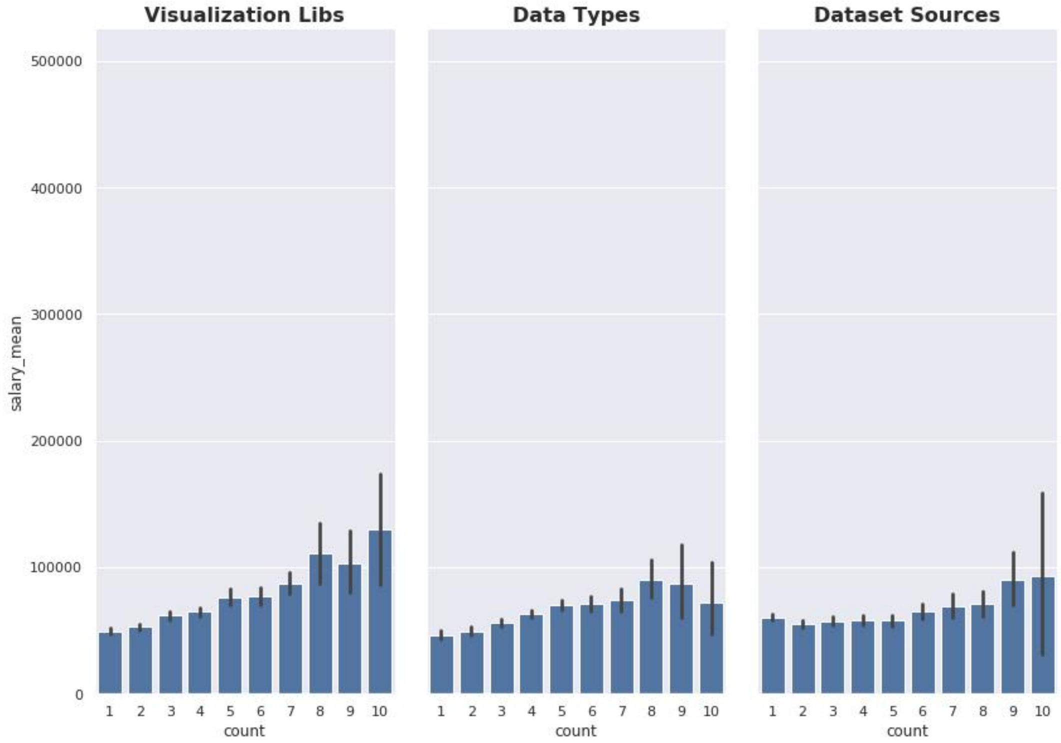
Han Hu | 1000555348 Dec 14 2018

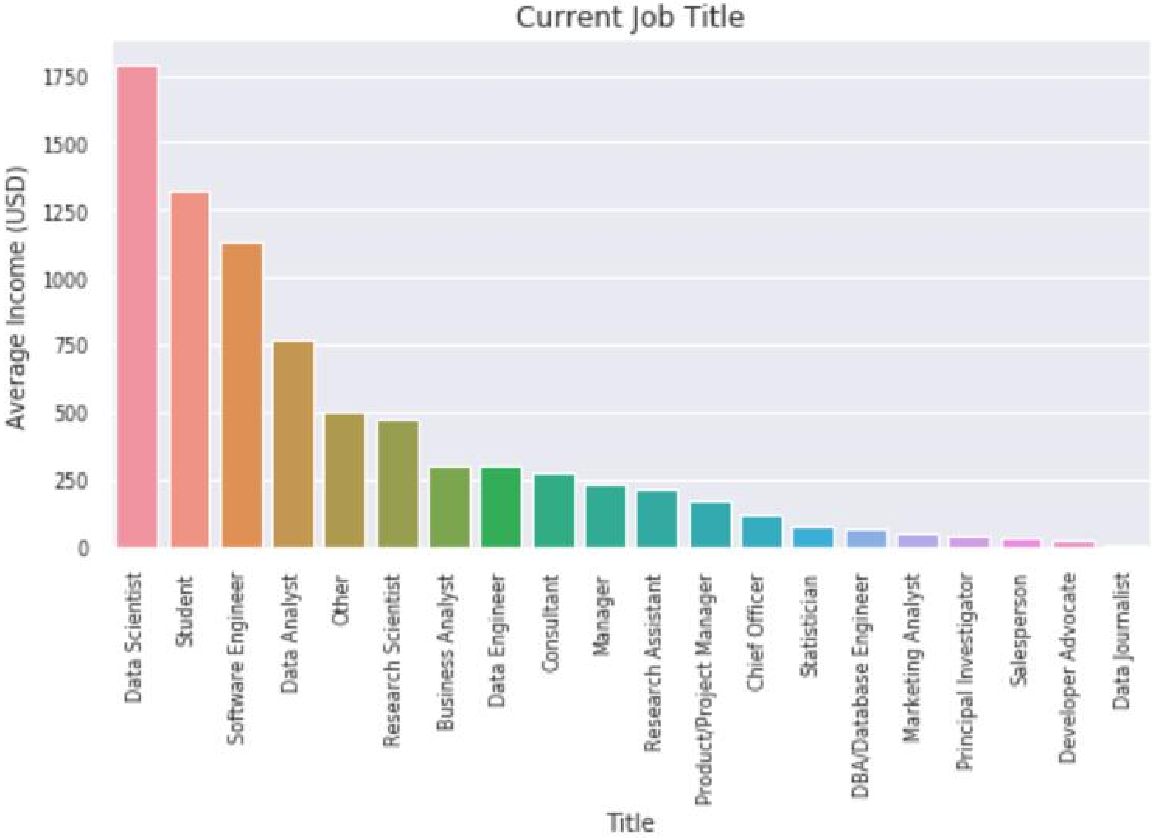
# Data Cleaning

- Questions with over 15% of its data missing are dropped, then missing data entries in the remaining questions are dropped.
- One hot encoding used to encode categorical data into numerical data so it can be inputed into the model.
- The missing questions are likely not only because of respondants losing interest over time.
- Questions regarding industry computing products have a low response rate



# Exploratory Analysis



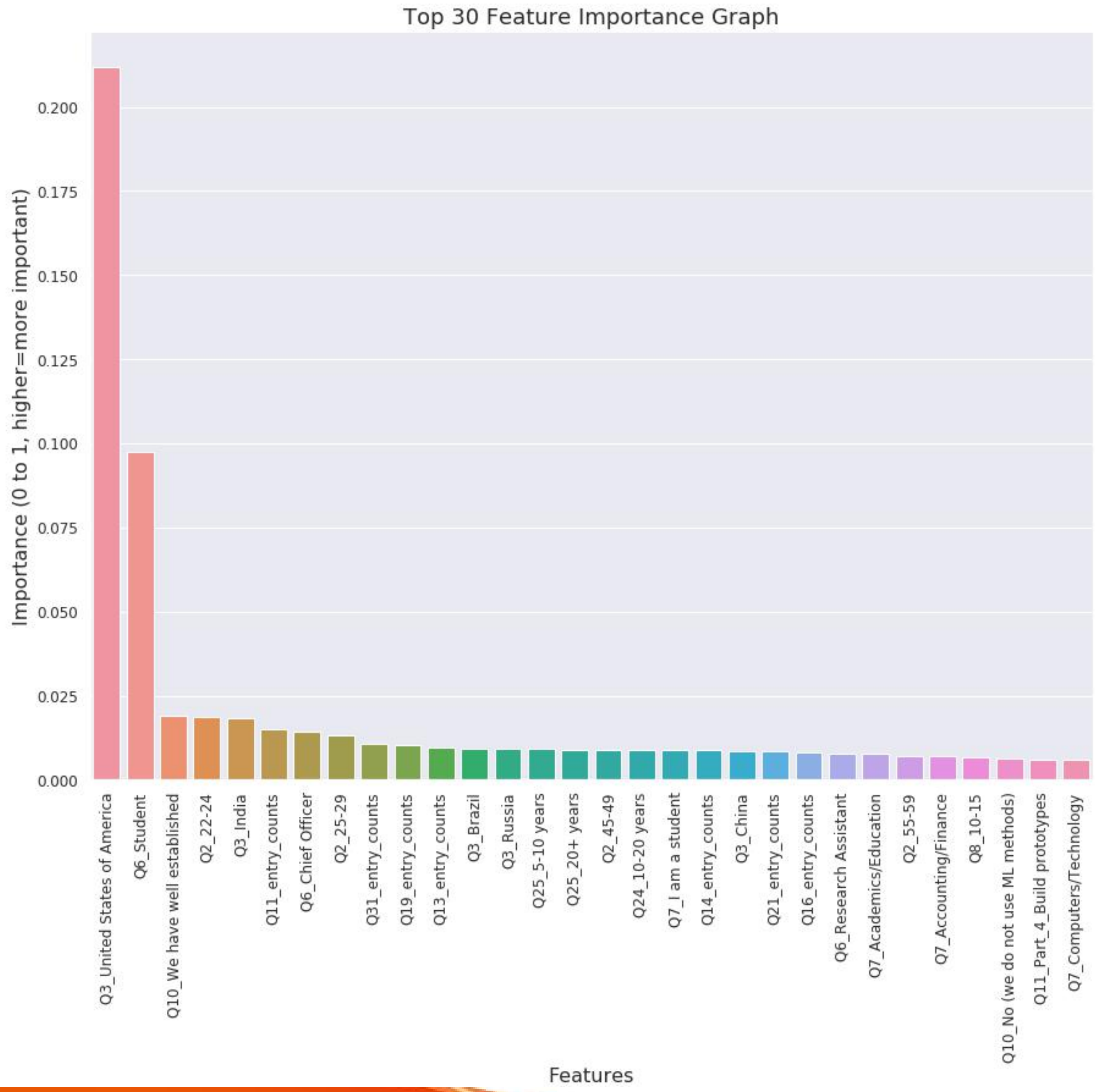


There is an obvious trend in the number of features selected

Trends can be observed in the individual questions.

# Model Feature Selection

- Used 3 different feature selection methods:
  - Boruta Library (Wrapper Method)
  - SelectFromModel_ExtraTreeRegressor (Wrapper Method)
  - Analysis of Variance (Filter Method)
- Reduced Model dimension from 354 dim to 133 dim, while gave better performance.
- Questions that had higher response rate were stronger predictors.

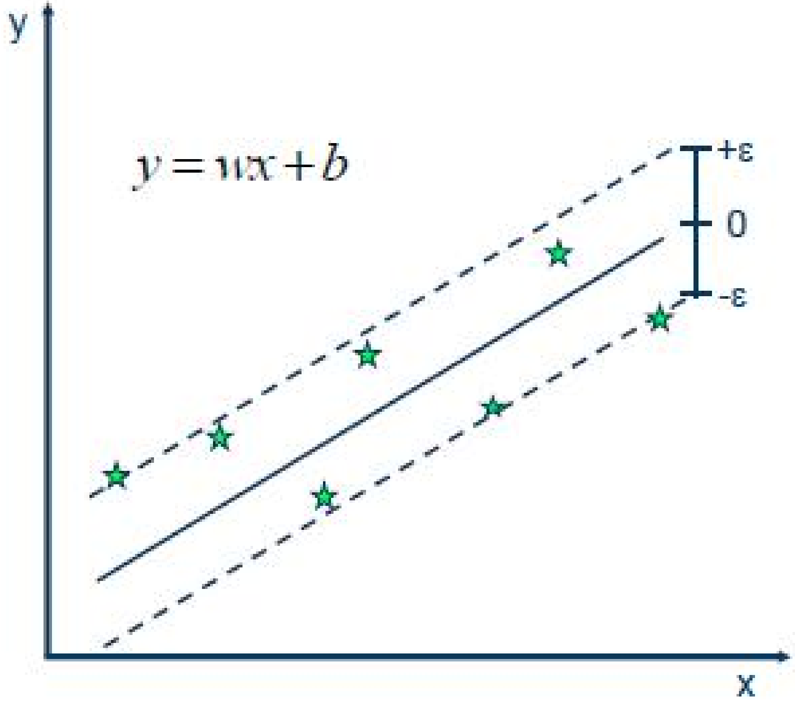


# Model Implementation

- Support Vector Machine was Implemented

- 4 algorithms were explored:
  - Ridge Linear Regression R2 score= 0.43 (+/- 0.23)
  - Stocastic Gradient Descent R2 score= 0.47 (+/- 0.17)
  - Random Forest R2 score= 0.43 (+/- 0.20)
  - Support Vector Machine: R2 score=0.50 (+/- 0.14)

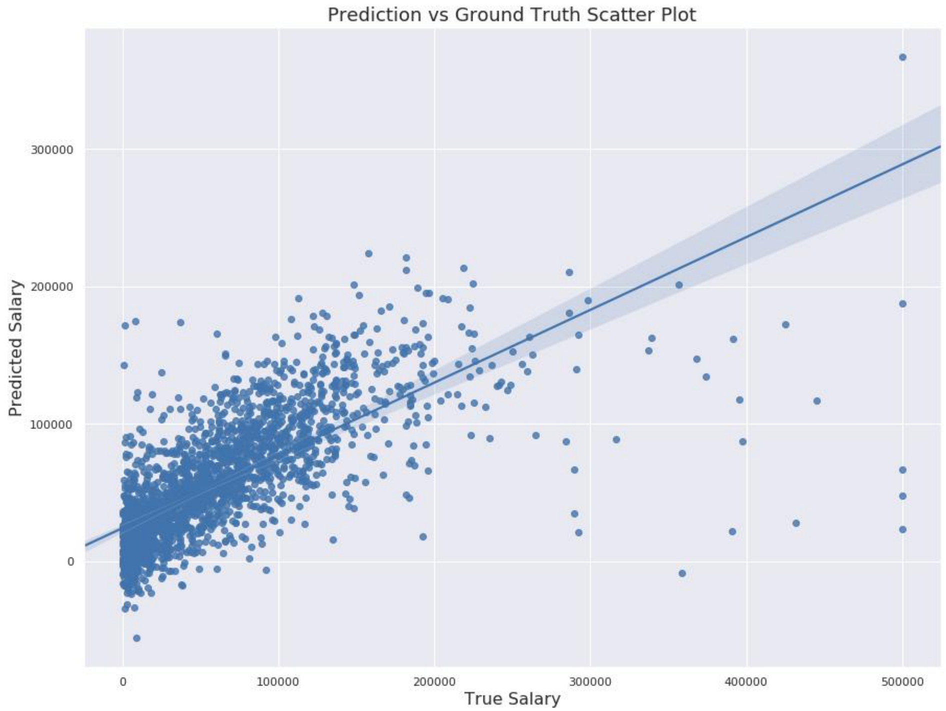
- SVM has best R2 score after grid search

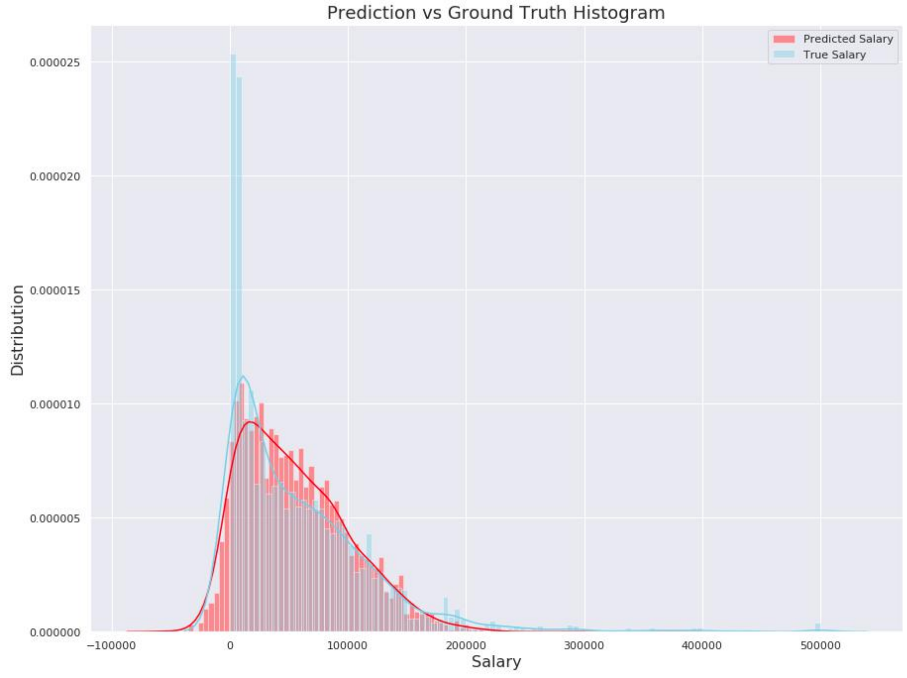


# Model Result Visualization

- Model Performed Poorly because
  1. Very high salary value at the end of the salary distribution that the model cannot predict.
  2. Not enough data points at the extreme spectrums of the salary distribution

- Potential Solutions:
  1. Perform outlier detection on the dataset
  2. More Data Points
  3. Use Ensemble Methods to fit the residual of the model


Prediction vs Ground Truth Scatter Plot
Predicted Salary
True Salary


Prediction vs Ground Truth Histogram
Predicted Salary
True Salary
Distribution
Salary