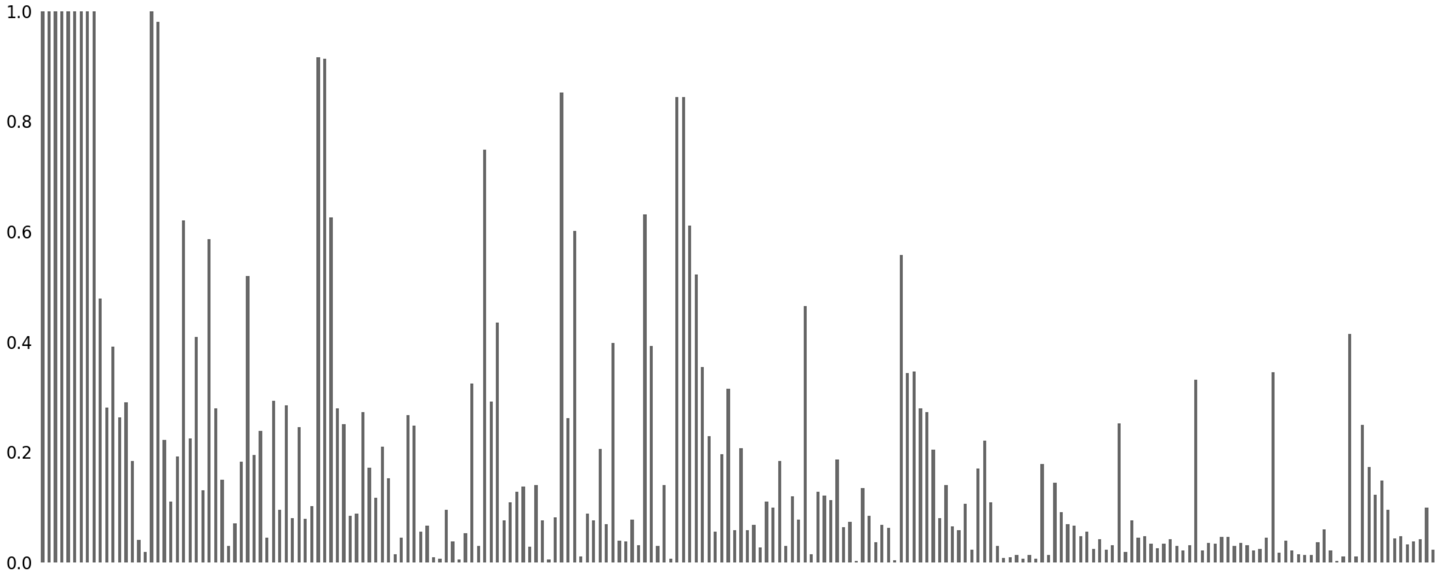
DATA CLEANING AND ENCODING

Prepared by : Yuchao Wu (1000651984)



msno bar graph (visualize the nulls)

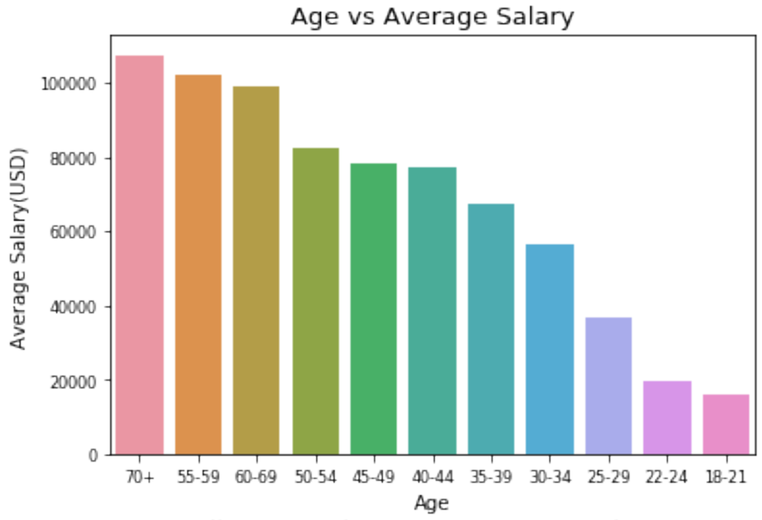
Data Cleaning Strategies:

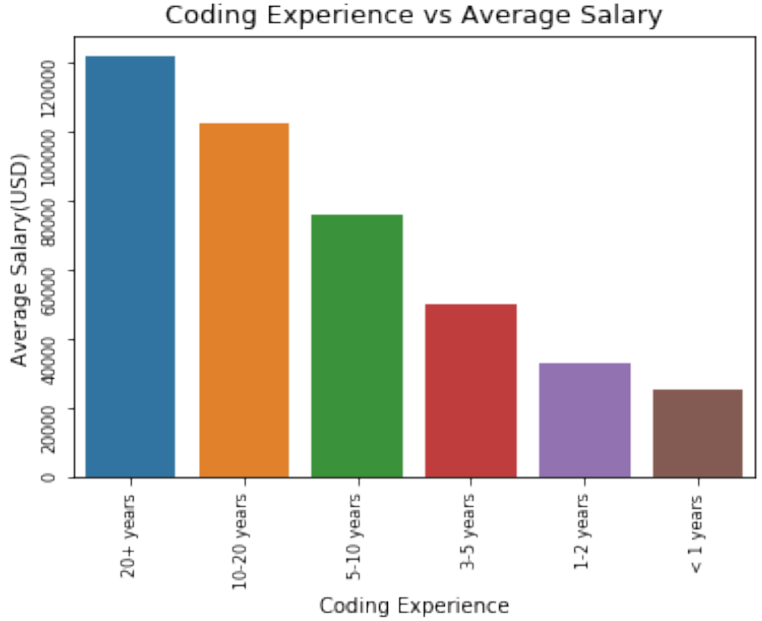
- Features with less than 10% missing data should have the missing values filled with mode.
- Features with more than 80% missing data should be dropped.
- Features with missing data between 10% to 80% should have the missing values filled with 0.

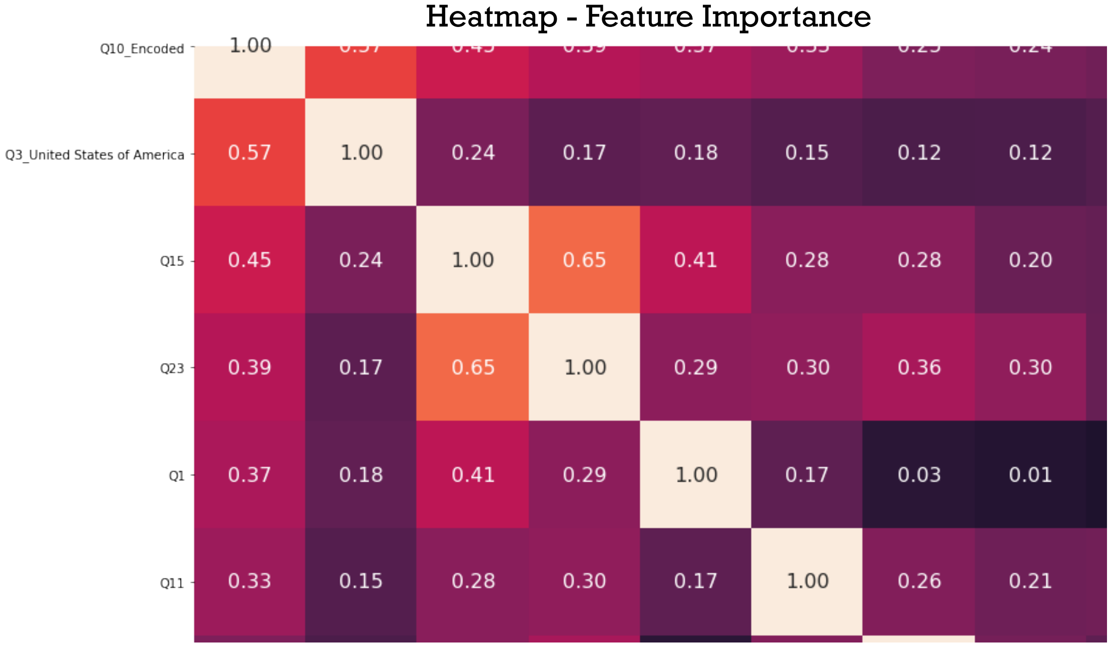
Encoding Strategies:

- One Hot Encoding is applied to the data without order (e.g., Gender).
- Ordinal Encoding is applied to the data with order (e.g., Salary level)

EXPLORATORY ANALYSIS

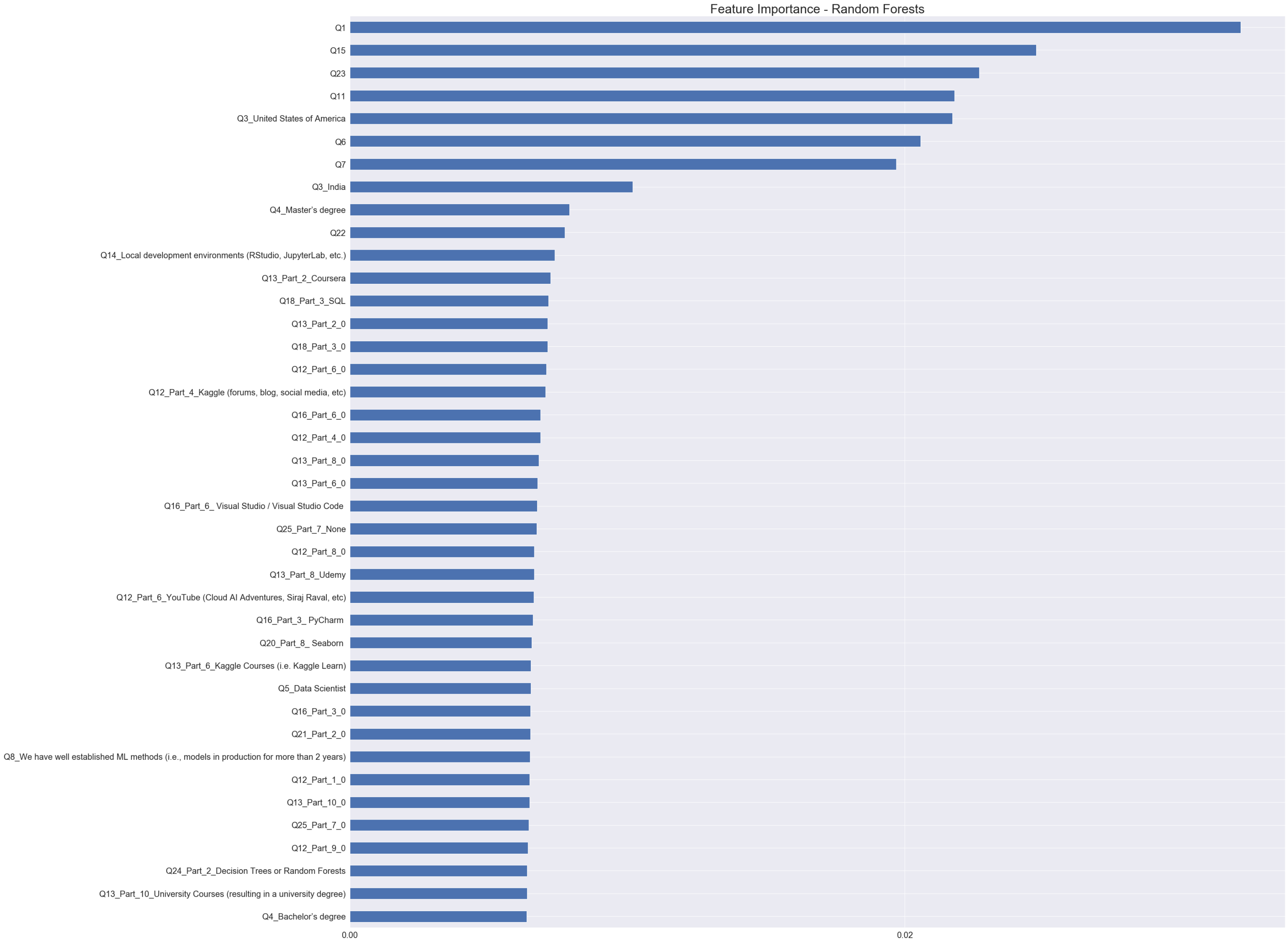

Age vs Average Salary


Coding Experience vs Average Salary


Heatmap - Feature Importance

- Two bar graphs show that coding experience and age have a direct and positive relationship with the salary. People with longer coding experience and older age tend to make more money.
- The heat-map also shows Q15 (coding experience) and Q1 (age) are very important to the salary prediction.

FEATURE SELECTION

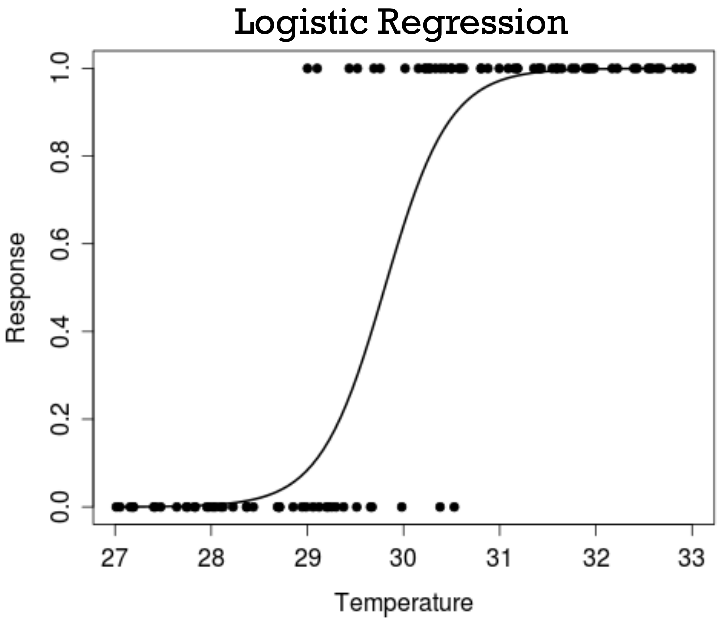


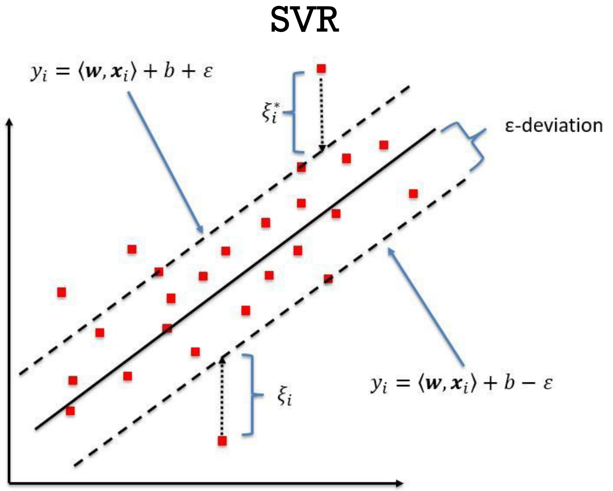
- Random Forest is implemented to select top features.
- The graph on the left shows the top 40 important features.
- The dataset dimension is reduced from 70 to 40 features.
- Q1 (age) has the most importance which is then followed by Q15 (coding experience)and Q23 (machine learning experience)

MODEL IMPLEMENTATION

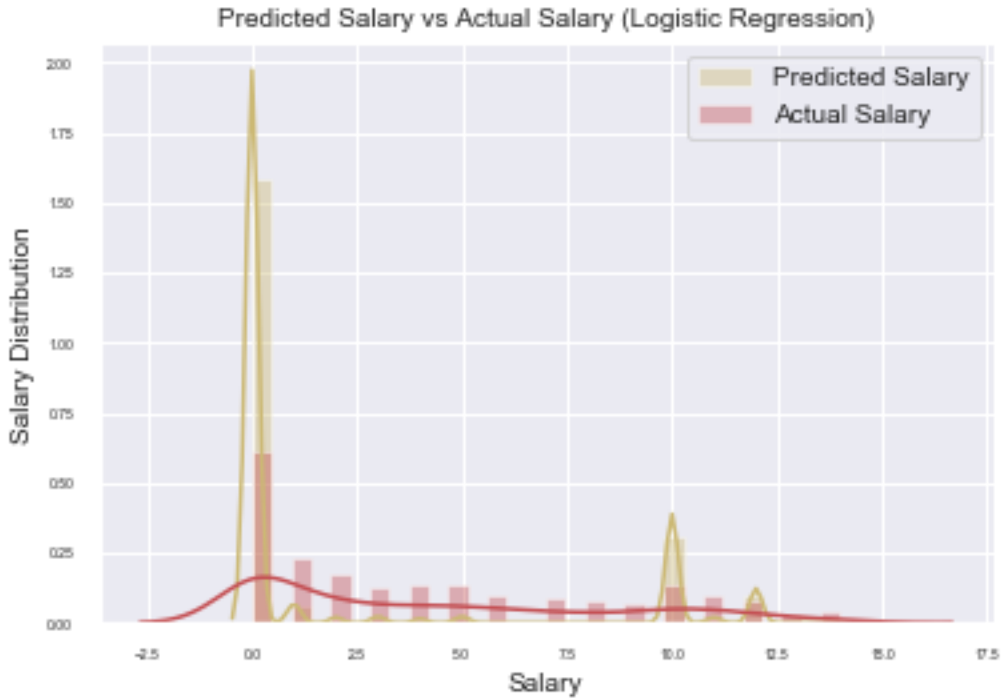
- Two algorithms:
  - Logistic Regression with multiple parameters
  - Support Vector Machine (SVM)
- Logistic Regression – after grid search tuning
  - Cross validation (10 folds) = 0.33(+/- 0.02)
- SVM – after grid search tuning
  - Cross validation (10 folds) = 0.55(+/- 0.05)

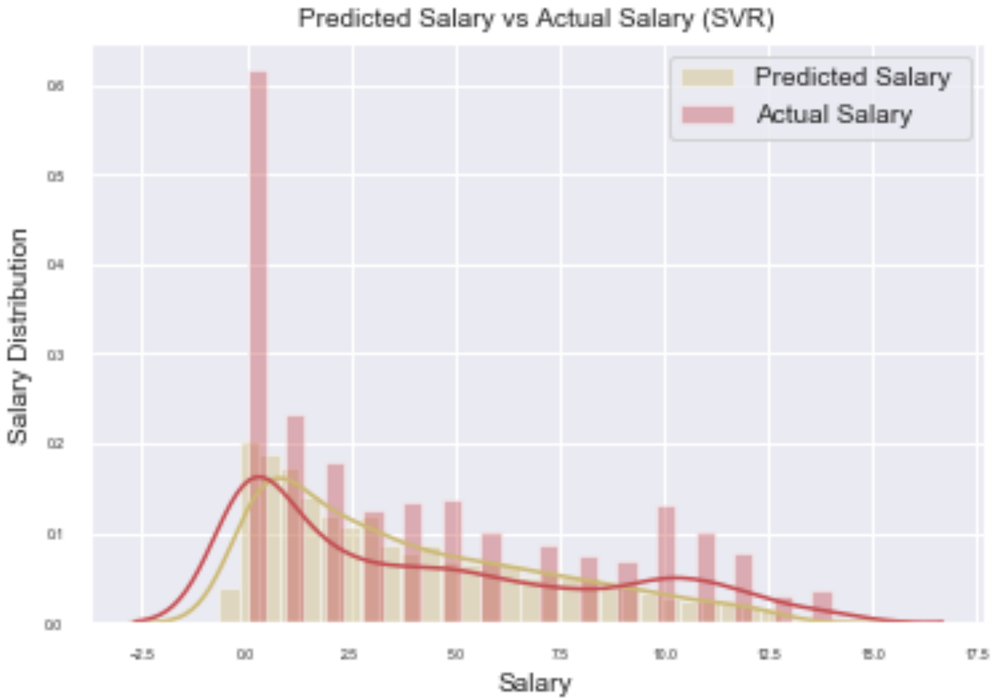
SVM performs better !


Logistic Regression


SVR
$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - \varepsilon$

MODEL RESULTS AND VISUALIZATION


Predicted Salary vs Actual Salary (Logistic Regression)


Predicted Salary vs Actual Salary (SVR)

- When evaluating the model on test datasets, both logistic and SVR tend to overfit the data.
- SVR has cross validation score of around 0.54 while logistic regression is around 0.32 on testing dataset.
- From the distribution plots, SVR model tends to have a more similar shape and trend as the actual data compared with the logistic regression model.
- Both models perform poorly on this dataset. The model could be further improved by testing more parameters and perform outlier detection on the dataset.