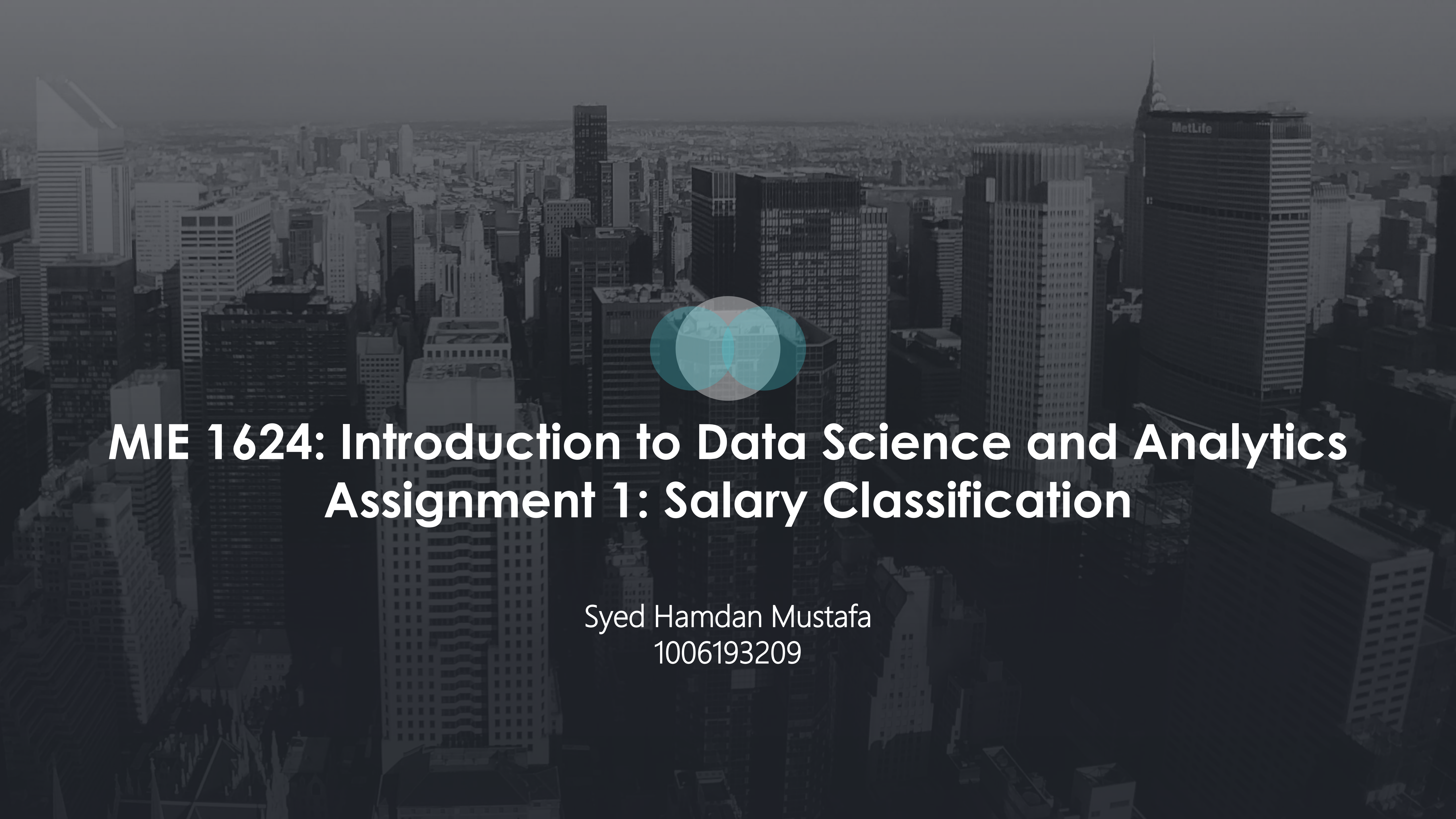
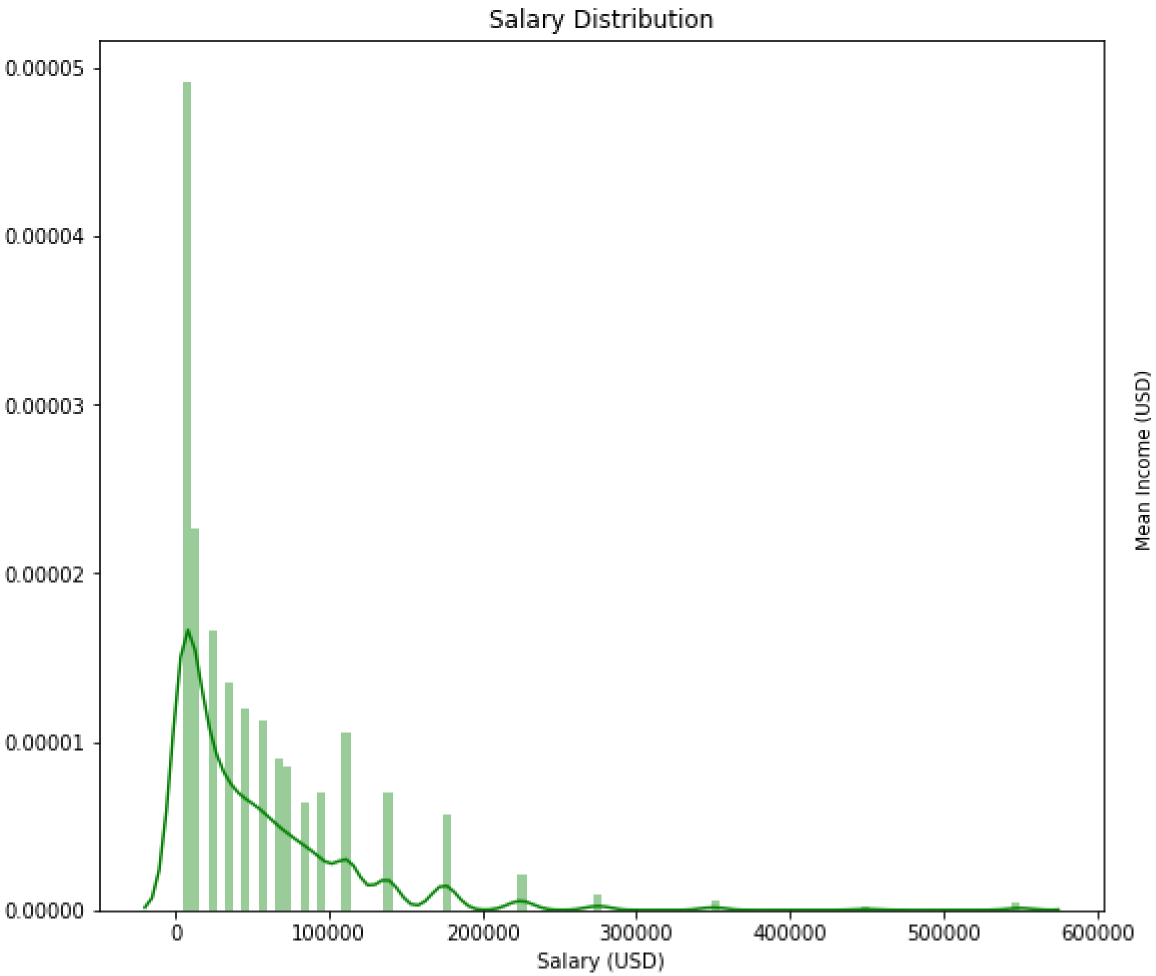
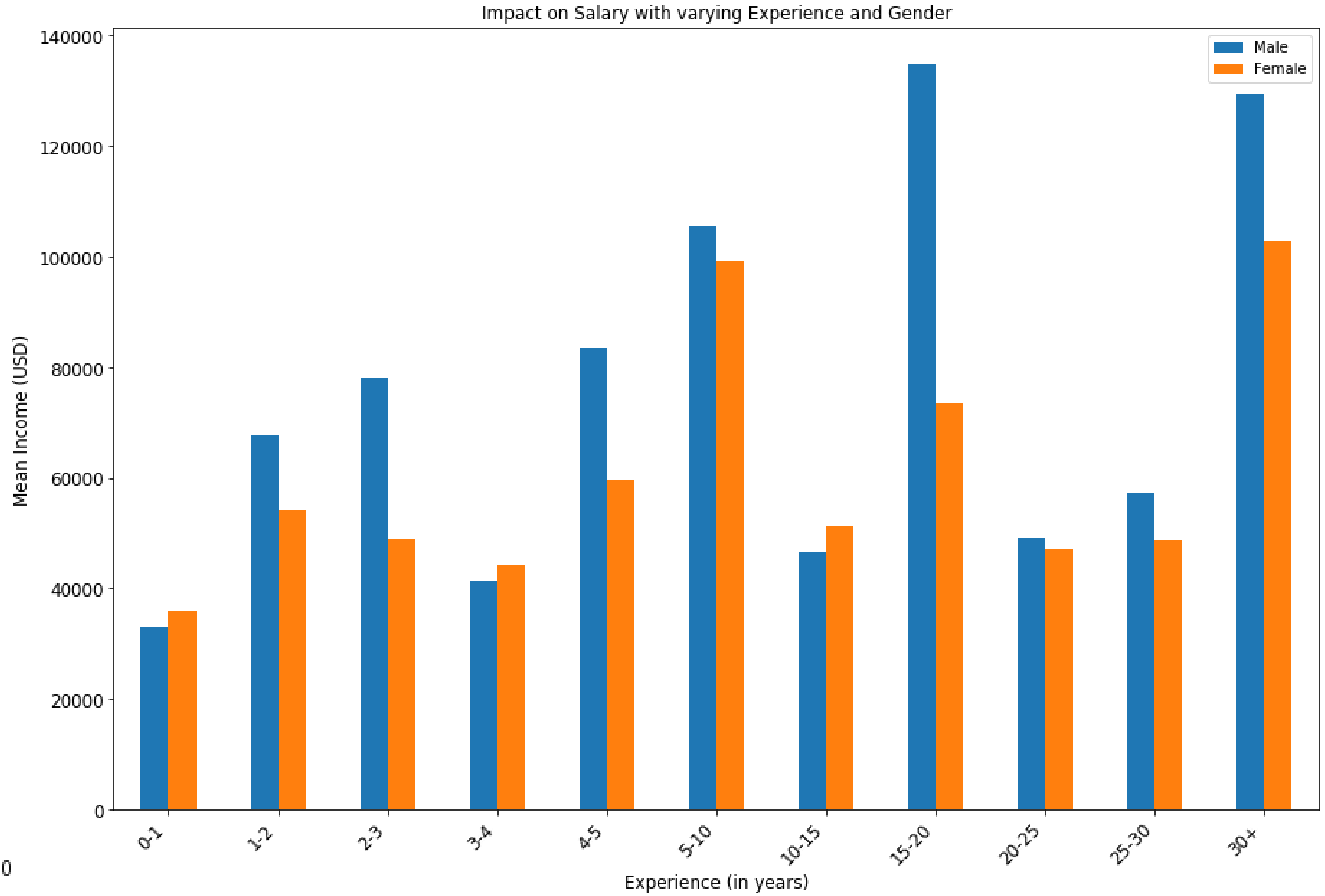


MIE 1624: Introduction to Data Science and Analytics
Assignment 1: Salary Classification

Syed Hamdan Mustafa
1006193209

Exploratory Analysis

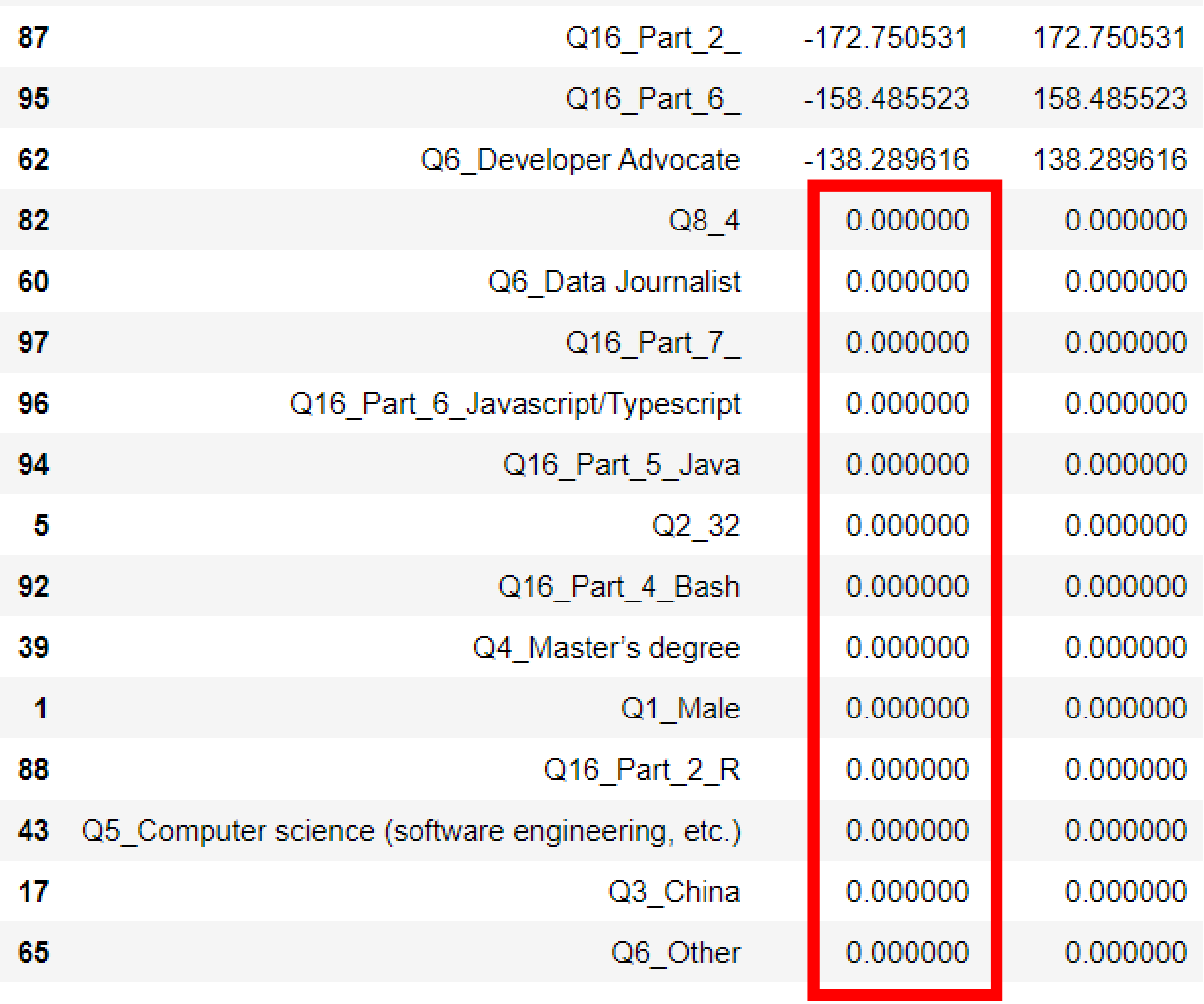




Salary was heavily skewed to the right. Hence it was difficult to predict the higher classes. Salaries were divided into 4 brackets.

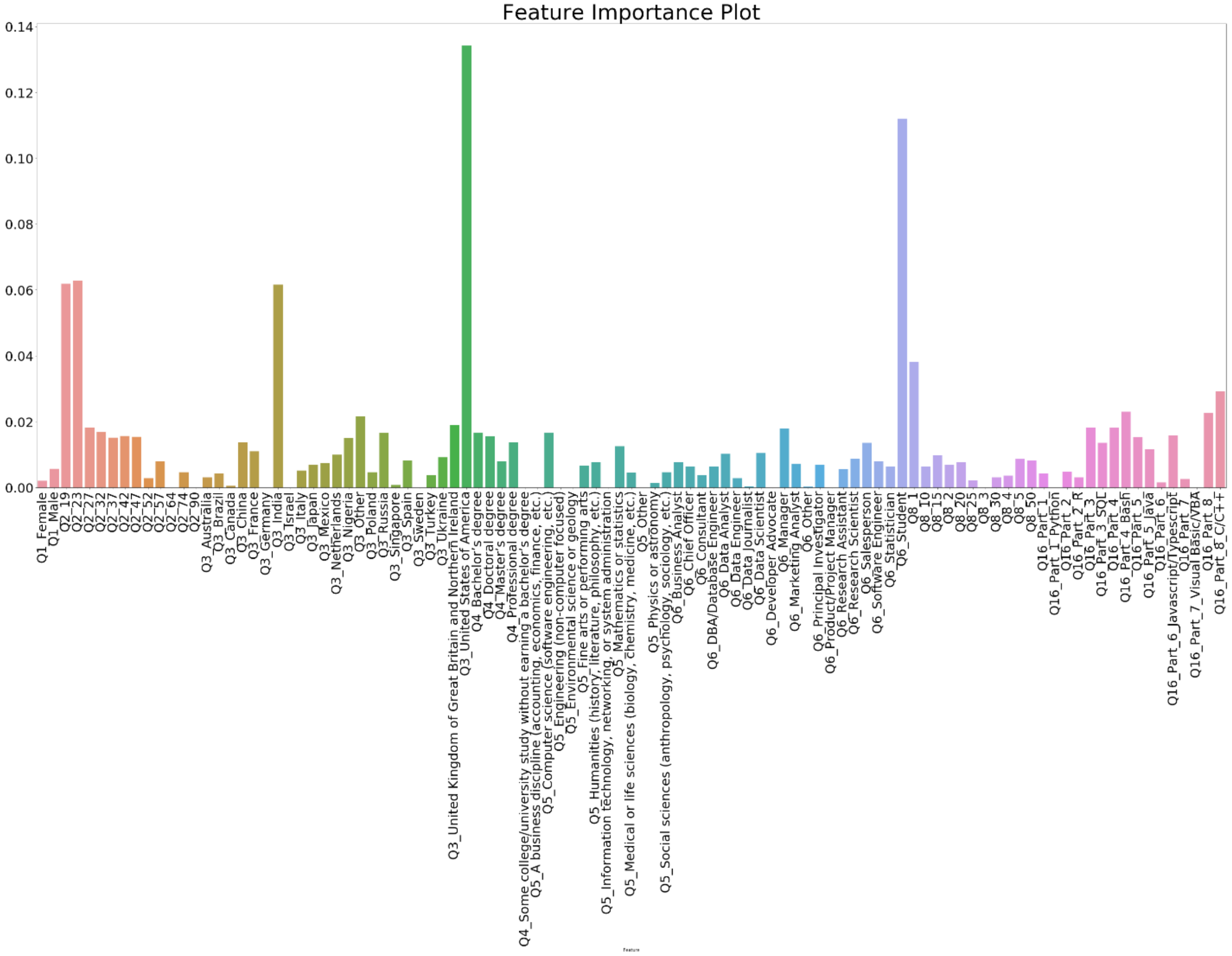
Salary varied for the different features selected. They were also separately displayed for males and females as seen.

Feature Selection and Importance



87	Q16_Part_2_	-172.750531	172.750531
95	Q16_Part_6_	-158.485523	158.485523
62	Q6_Developer Advocate	-138.289616	138.289616
82	Q8_4	0.000000	0.000000
60	Q6_Data Journalist	0.000000	0.000000
97	Q16_Part_7_	0.000000	0.000000
96	Q16_Part_6_Javascript/Typescript	0.000000	0.000000
94	Q16_Part_5_Java	0.000000	0.000000
5	Q2_32	0.000000	0.000000
92	Q16_Part_4_Bash	0.000000	0.000000
39	Q4_Master's degree	0.000000	0.000000
1	Q1_Male	0.000000	0.000000
88	Q16_Part_2_R	0.000000	0.000000
43	Q5_Computer science (software engineering, etc.)	0.000000	0.000000
17	Q3_China	0.000000	0.000000
65	Q6_Other	0.000000	0.000000

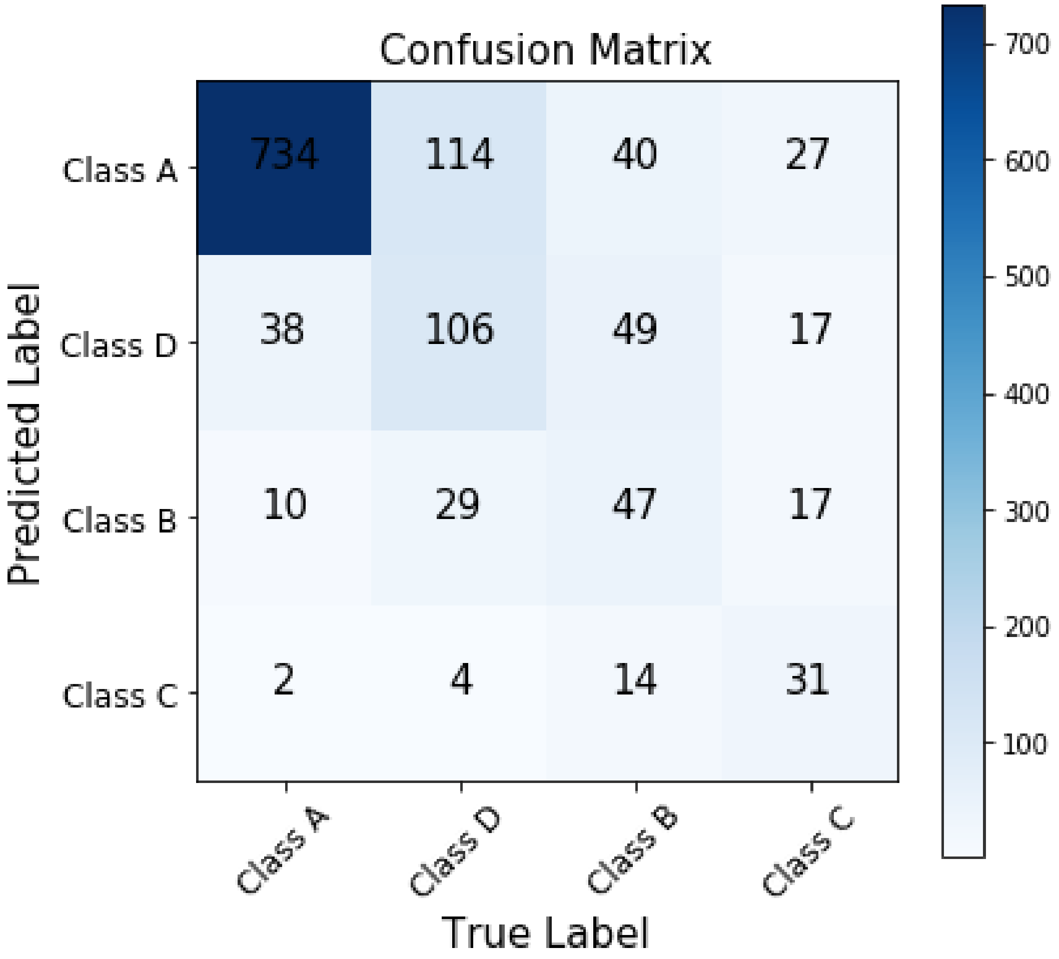
Lasso (l1) Regression was used to remove all the features with zero coefficients. Note that the features were one-hot encoded.

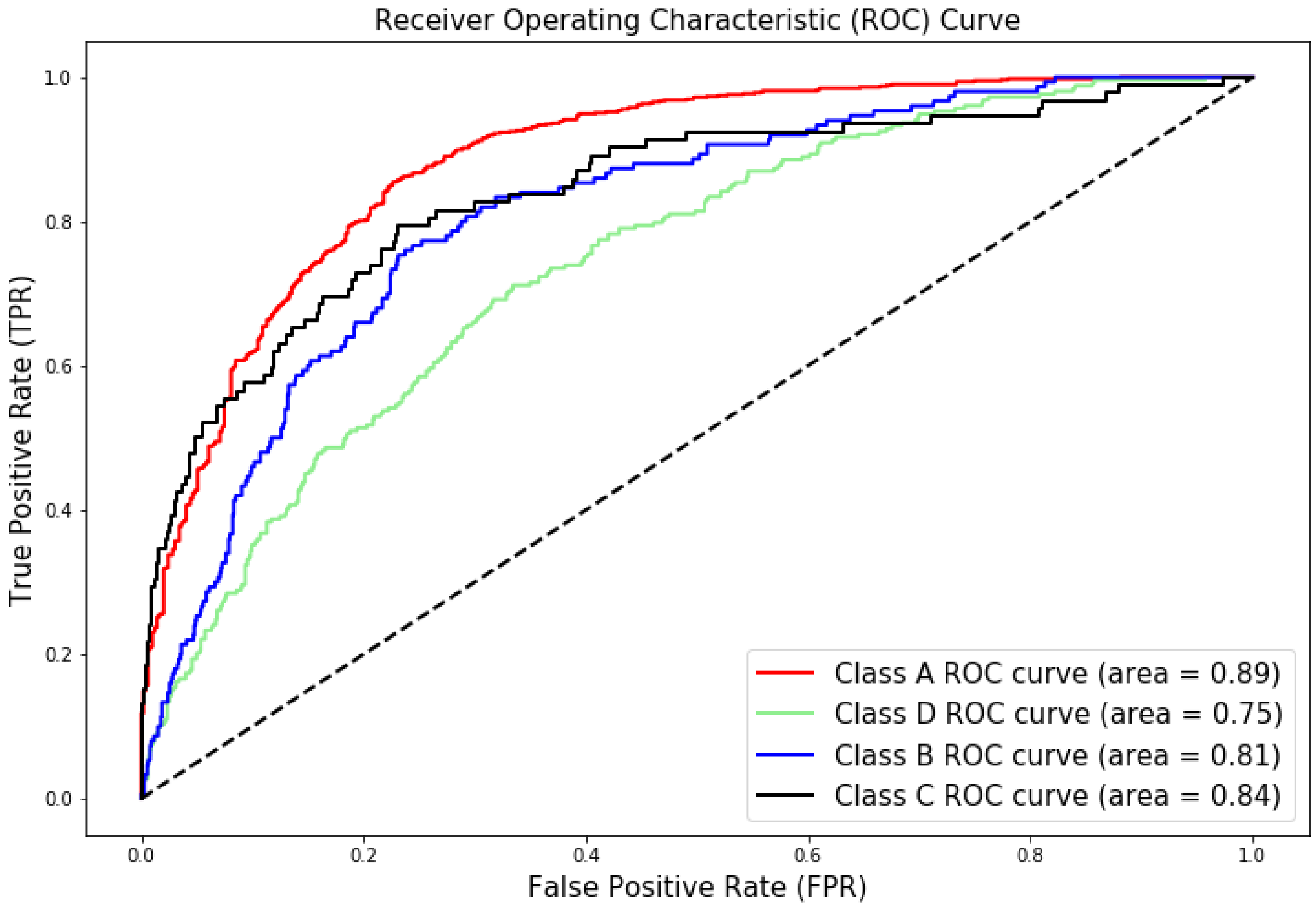


Feature Importance Plot using Mutual Information (MI).

Student and USA had the biggest impact on Salary.

Model Implementation

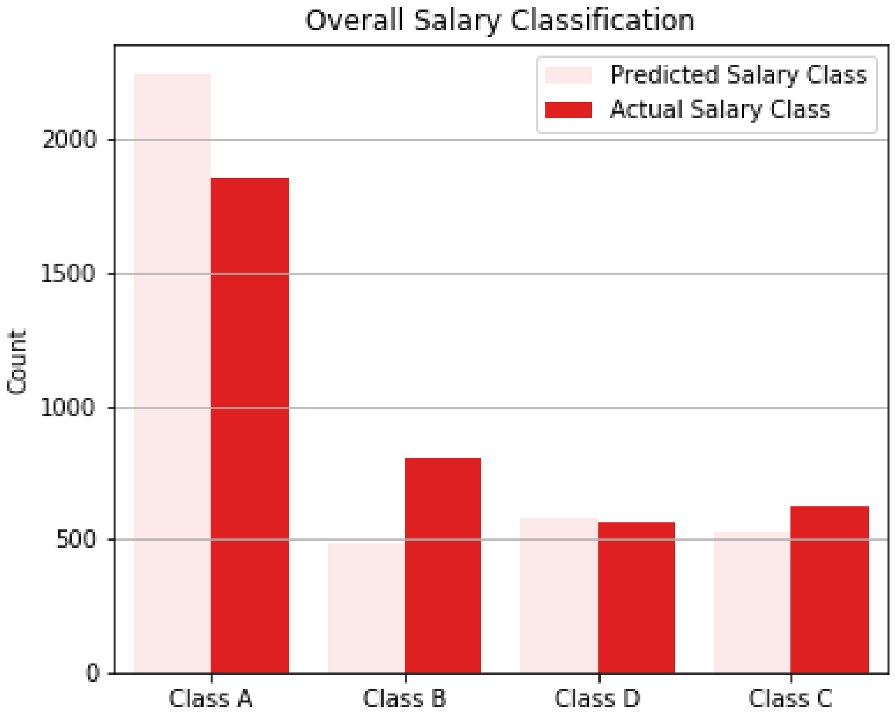


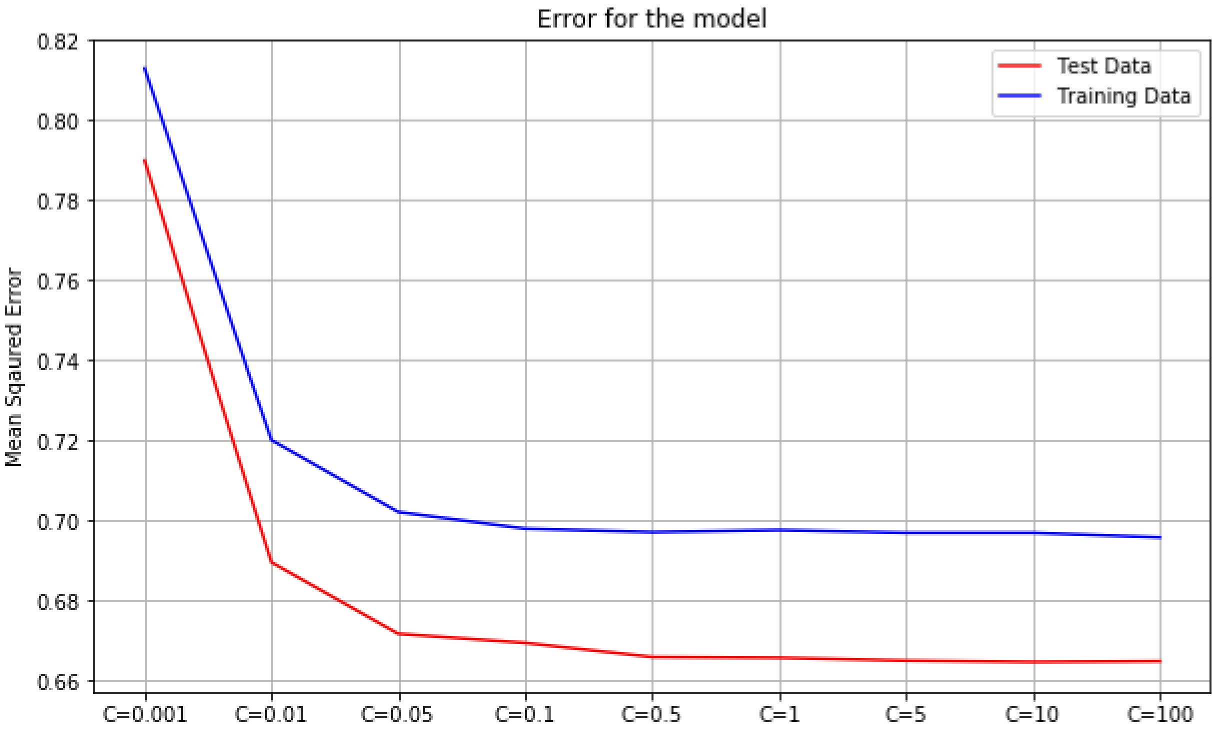


Multinomial Logisitic regression was used. The accuracy from the 10-folds was **68%**

OneVsRest Classifier was used to classify the different salary brackets. Class D showed the worst performance whereas Class A showed the best.

Model Testing





Class A- 0-30k
Class B- 30k-60k
Class C- 60k-95k
Class D- 95k+

This model got an accuracy of **69.12%** on the training set
This model got an accuracy of **66.95%** on the testing set

C=1 onwards gives low error.
In our case, the difference between the training and test error is not great, hence the variance is low. However, the bias is high since the training error itself is quite high. This is understood and expected given the bias in the data for the lower salary classes.

Summary

- 8 questions from the original survey were considered for the model.
- The features were cleaned and one-hot encoded.
- The salary target variable was differentiated into 4 classes.
  - This was done because the original 18 salary brackets were showing poor accuracy.
  - Salary distribution is heavily skewed to the left and has limited data for higher salary classes.
  - Grouping the salary brackets made the model accurate; however, the model was not able to predict exact classes (from the original count of 18).
- Multinomial logistic regression is used and the accuracy achieved is around 68% via 10-fold.
- OneVsRest Classifier is used to classify the different brackets and obtain the ROC curve.
- Model is not overfitting nor underfitting looking at the training and test error.
- Model has exhibited low variance and high bias (towards the lower salary bracket).

THANK YOU!