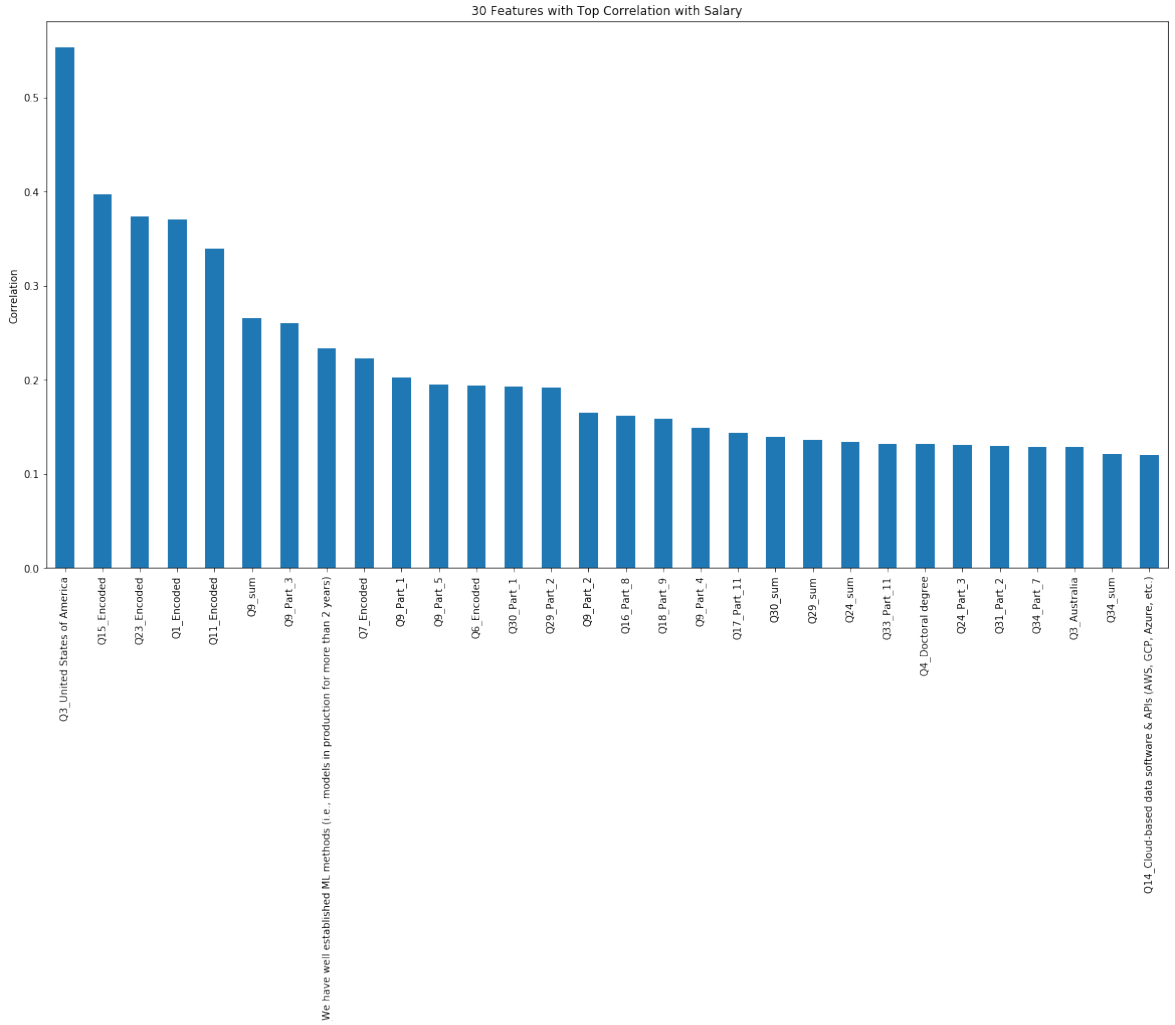
MIE 1624 Assignment 1

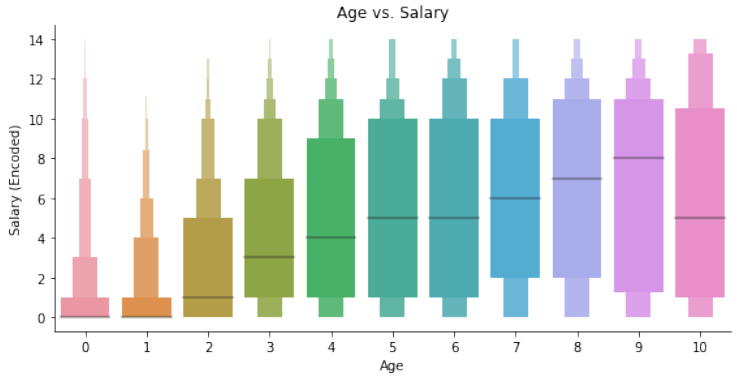
Tianyi yu

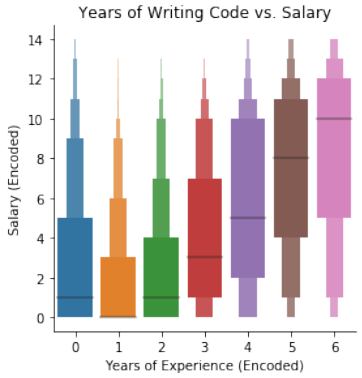
1005898502

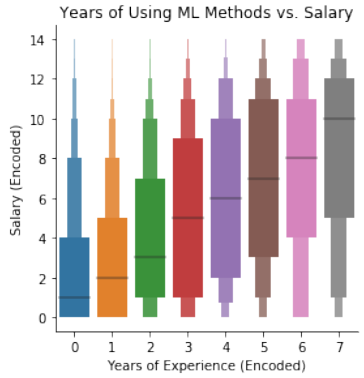


Feature Importance

Exploratory Analysis







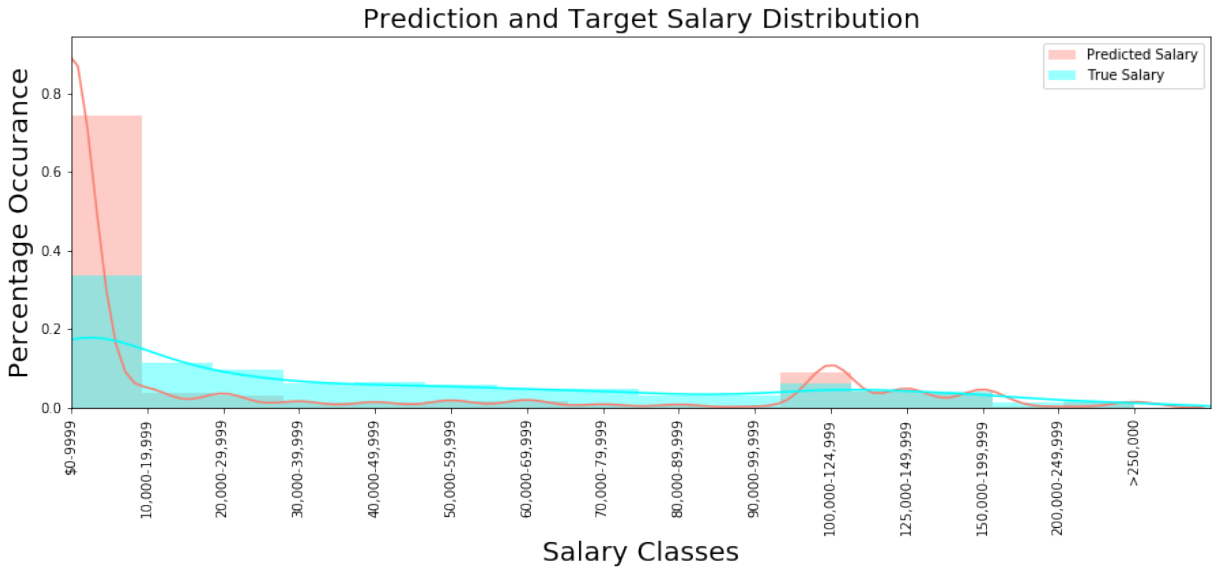
- The first four are listed below and th ey each will have a plot to show the t rend in the next section.
- Whether the respondent works in the United States of America
- Number of years the respondent has been writing code to analyze data
- Number of years the respondent has been using machine learning methods
- The age of the respondent

Model Results

	precision	recall	f1-score	support
0	0.424	0.938	0.584	1177
1	0.096	0.033	0.049	395
2	0.124	0.038	0.058	340
3	0.125	0.028	0.046	215
4	0.140	0.027	0.045	226
5	0.109	0.029	0.045	209
6	0.138	0.047	0.070	170
7	0.192	0.029	0.050	175
8	0.083	0.018	0.030	111
9	0.000	0.000	0.000	109
10	0.171	0.249	0.203	217
11	0.221	0.207	0.214	150
12	0.276	0.255	0.265	145
13	0.000	0.000	0.000	47
14	0.159	0.109	0.130	64
accuracy			0.345	3750
macro avg	0.151	0.134	0.119	3750
weighted avg	0.226	0.345	0.241	3750

- The precision, recall and f1 score are listed in the table above. The overall accuracy is 0.34 which generally aligns with that of the training set, only slightly lower. Classes with more support tend to have higher precision, recall and f1 score.
- The best model obtained from Grid Search is the logistic regression model with C = 1 and L1 regularization. The solver used is ‘liblinear’.

Results Discussion



- According to the distribution plot, the model overpredicts the lowest salary bucket and underpredict the higher ones. This may be attributed to the unevenly distributed samples. Most respondents fall into the first several buckets while buckets with higher salaries receive fewer samples. Generally, the performance of the model is not very good.
- To improve the accuracy of training set, one way is to be more careful when encoding the dataset. Some of the options in the multiple choice questions such as 'Other' may require more appropriate handling as we do not know what 'Other' actually is. Performing cross validation with more folds is also an option. For training and test sets generally, improving accuracy can be achieved by obtaining more samples, especially those fall into buckets with higher salaries. Additionally, the classes can be combined so the samples are more aggregated and better for prediction.