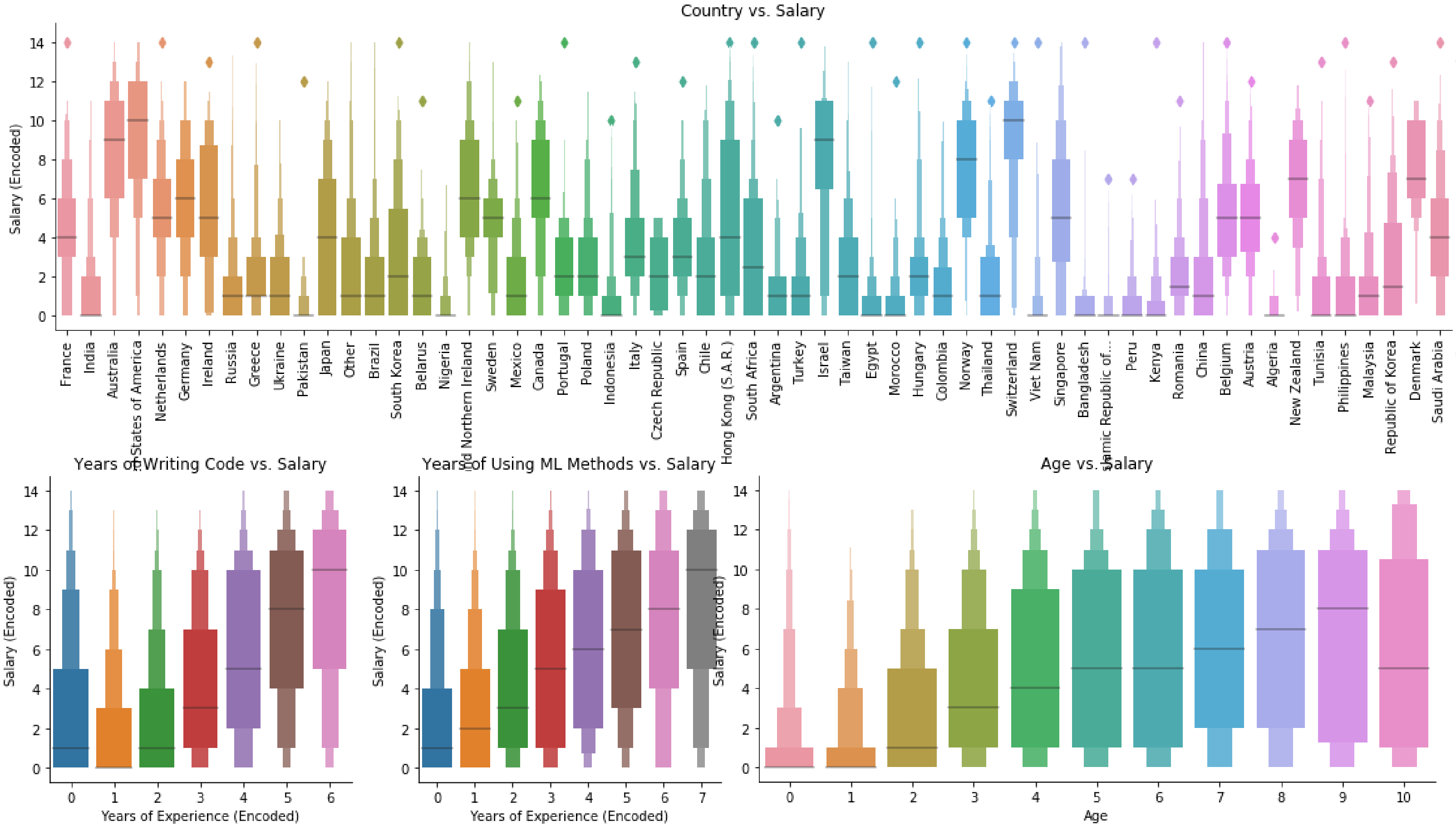
MIE1624 : Intro to Data Science and Analytics

Sai Anirudh Basamsetty : 1006042747

Assignment - 2

Exploratory Data Analysis


Country vs. Salary

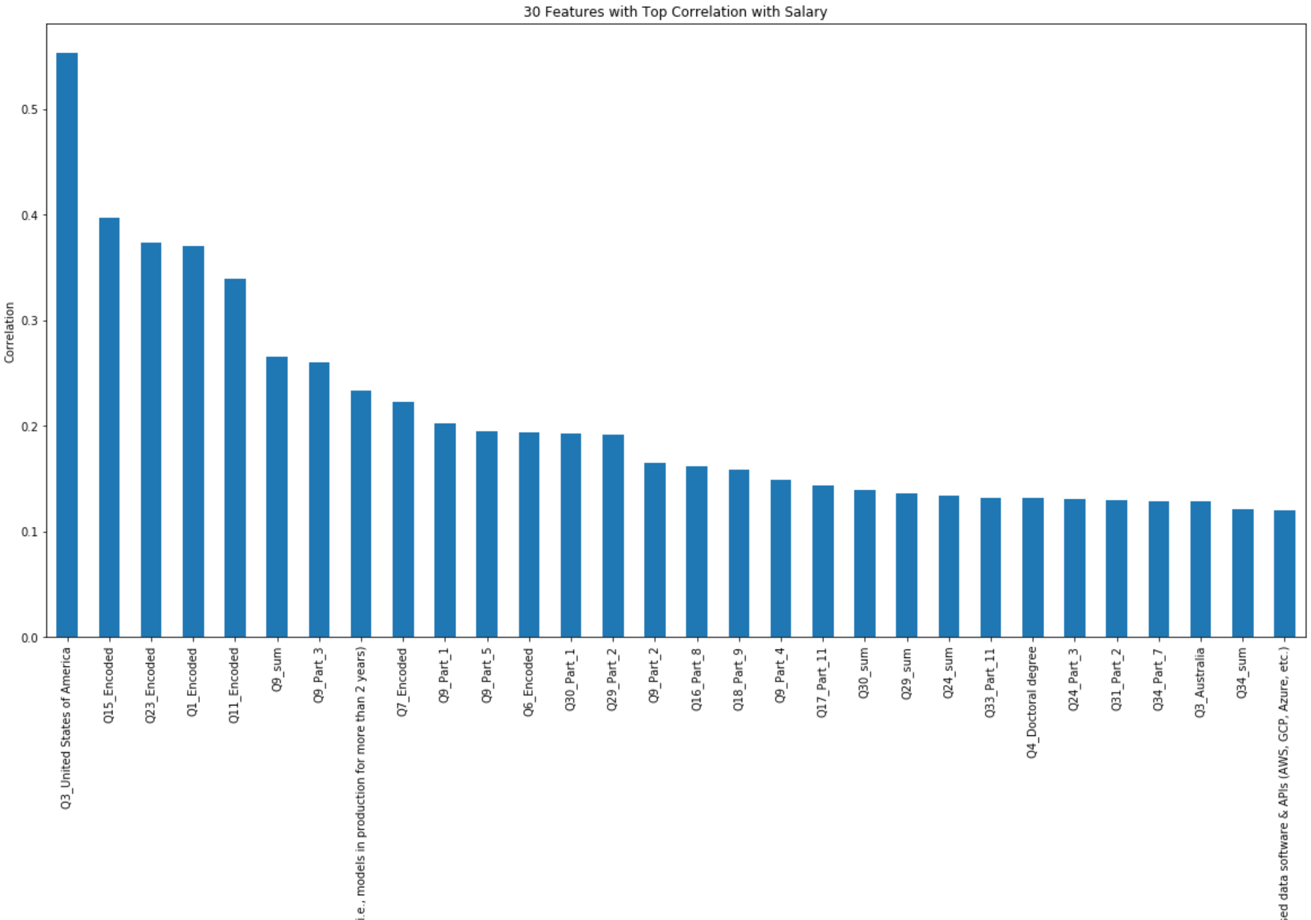
Years of Writing Code vs. Salary

Years of Using ML Methods vs. Salary

Age vs. Salary

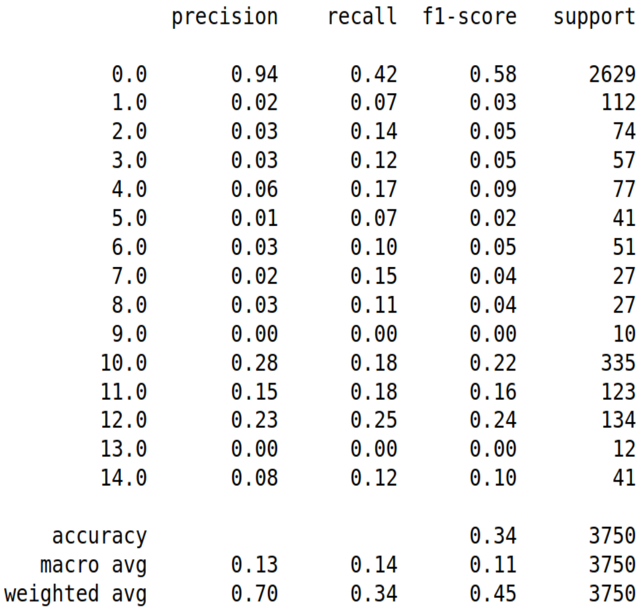
- Employees in top countries USA, Hong Kong and Singapore have highest salaries
- Linear relation can be observed with age, years of experience and salaries

Feature Importance



The graph shows the top 30 features that have the strongest correlation with employee's annual salary.

Model Results



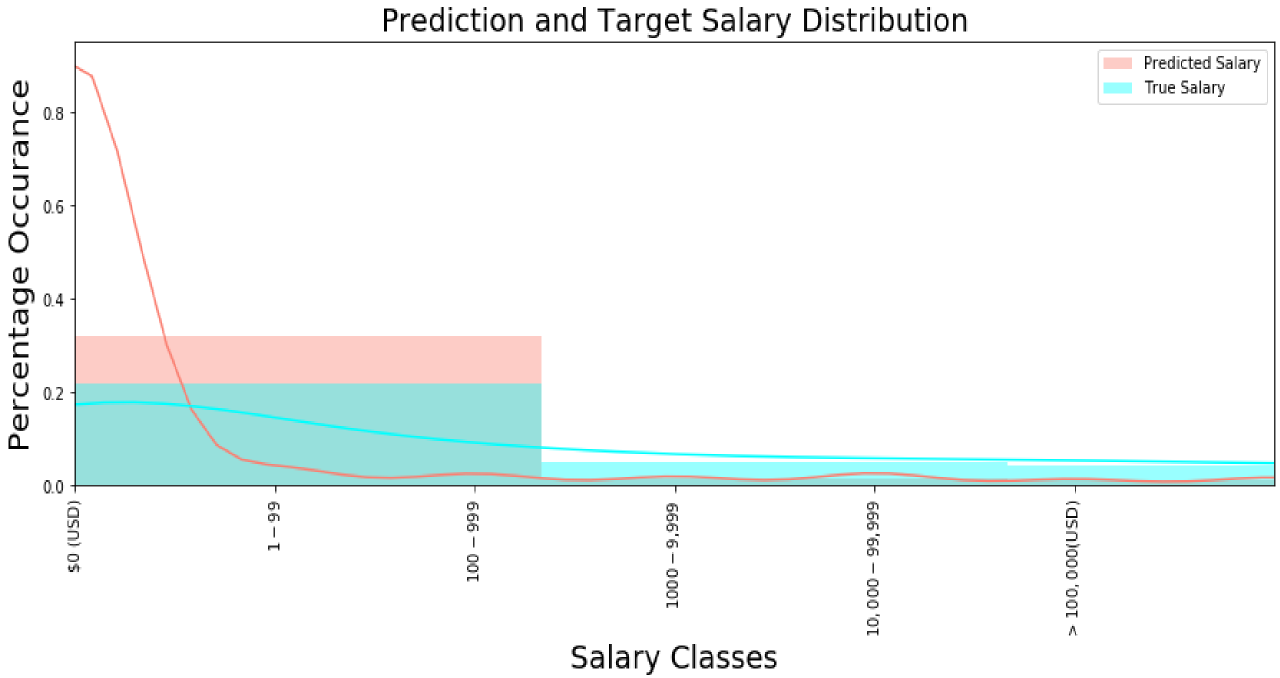
```
              precision    recall  f1-score   support

         0.0       0.94      0.42      0.58      2629
         1.0       0.02      0.07      0.03       112
         2.0       0.03      0.14      0.05        74
         3.0       0.03      0.12      0.05        57
         4.0       0.06      0.17      0.09        77
         5.0       0.01      0.07      0.02        41
         6.0       0.03      0.10      0.05        51
         7.0       0.02      0.15      0.04        27
         8.0       0.03      0.11      0.04        27
         9.0       0.00      0.00      0.00        10
        10.0       0.28      0.18      0.22       335
        11.0       0.15      0.18      0.16       123
        12.0       0.23      0.25      0.24       134
        13.0       0.00      0.00      0.00        12
        14.0       0.08      0.12      0.10        41

    accuracy                           0.34      3750
   macro avg       0.13      0.14      0.11      3750
weighted avg       0.70      0.34      0.45      3750
```

- The best parameters for Logistic regression are C=1, L1 regularization and solver ‘liblinear’.
- The precision, recall and f1-score are listed in the table. 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.

Results



- According to the distribution plot, the model over predicts the lowest salary bucket and cannot accurately predict the higher ones. This may be due to the uneven distribution of samples. Most respondents fall into the first salary categories while the higher salaries received fewer samples. Generally, the performance of the model is not very good. More features can be developed by using feature engineering that might contribute to the accuracy of the model.
- To improve the accuracy of training set, one way is to be more careful while encoding the data-set 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 other option. For training and test sets generally, improved accuracy can be achieved by obtaining more samples, especially those that fall into buckets with higher salaries. Additionally, the classes can be combined so the samples are more aggregated and better for prediction.