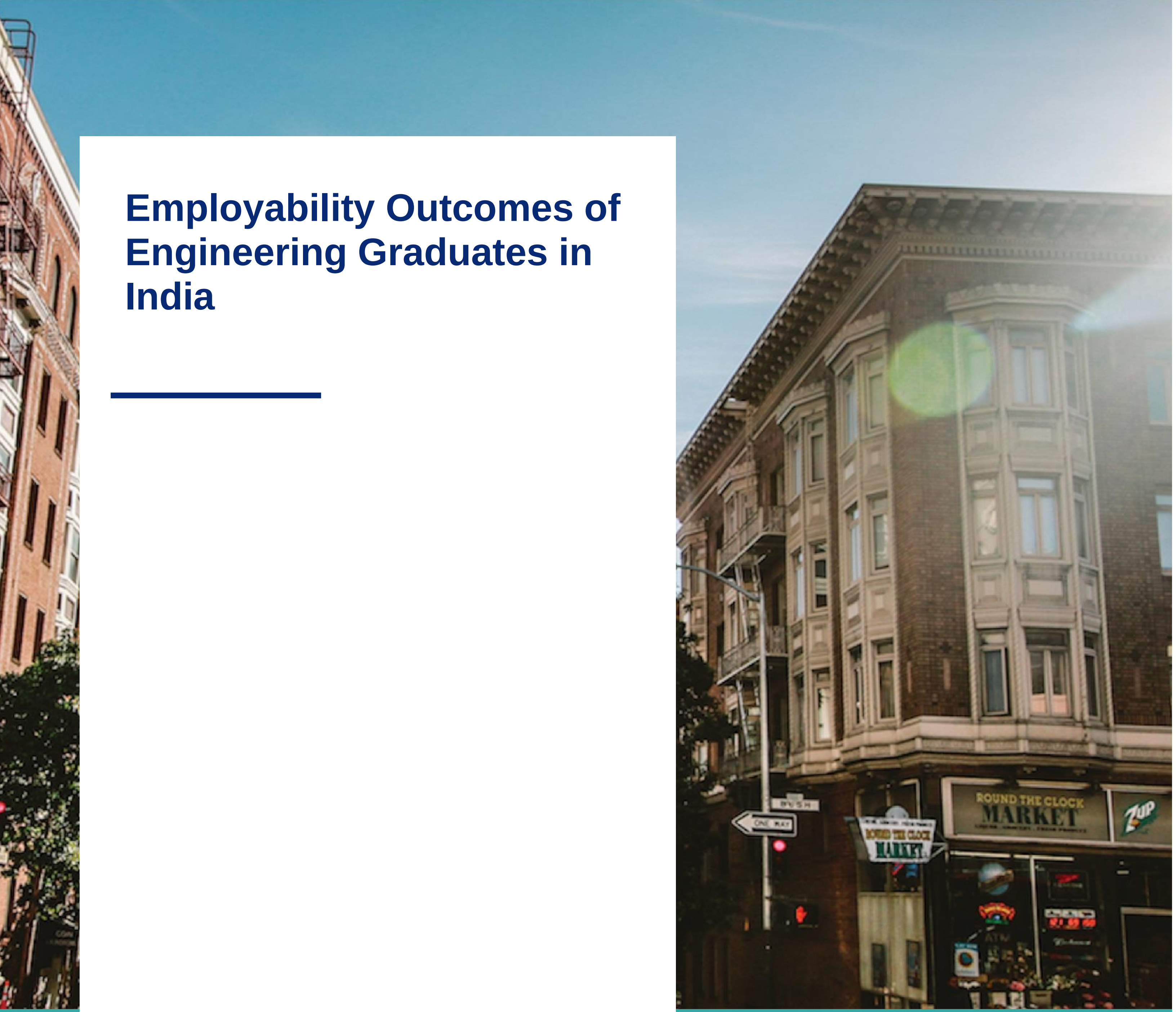
Employability Outcomes of Engineering Graduates in India



AUGUST 22

Coursera Applied Data Science Capstone Project
Authored by: Kandasamy Ramanujam

Problem Statement

According to All India Council for Technical Education, more than 1.8 Million students are enrolled in engineering courses across 10,000+ engineering institutions in India. Of these, only around 42% of them are placed in jobs at the time of completion of their engineering education. Additional information can be found at the AICTE website here.

The objective of this project is to understand the factors that influence the employability of engineers – specifically indicated by the initial salary offered to the engineering graduates. Some of the factors that influence this could be academic performance – both in engineering institution and prior to that in the schools, demographic factors such as gender, location of the college, proficiency in English, other factors such as aptitude for quantitative skills.

Description of Data

Aspiring Minds Employability Outcomes 2015 (AMEO 2015) is a unique data set that contains engineering graduates' employability outcomes (salary, job title, city of employment) along with data on assessment scores and other demographic data. This includes the following:

- Scores from school final exams – 10th and 12th standard
- Scores from Engineering course
- Engineering branch
- Tier of college and the city in which the college is located
- Demographic data such as gender, state
- Scores on English, logical ability, quantitative aptitude, and Computer Programming from standardized assessment test conducted by Aspiring Minds

The training data set has 3998 entries with 39 columns. There is a test data set with 1500 entries. However, the salary information is not provided here. So, this cannot be effectively used to compare the predictions from the models built using training data. The test data set is not used in this project.

The data can be downloaded from the Aspiring Minds website. Location given below.
http://research.aspiringminds.com/resources/#ameo

Methodology for Analysis

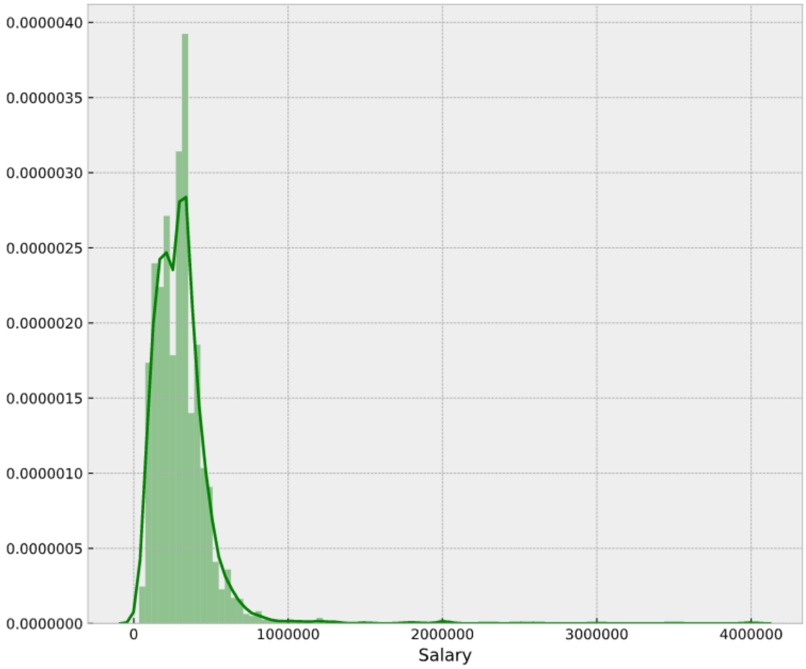
The proposed approach is as below.

- Exploratory Data Analysis to understand the data: Based on the initial analysis done, the data has non-null values on all rows and columns.
- Identify types of modeling to be used: As the outcome (salary) is a continuous variable, regression analysis is an obvious candidate. However, multi-class classification can be considered if the salary could be grouped into a small number of buckets and the problems is reframed as predicting the salary range.
- Build and test models: A subset of training data to be used for testing the models.

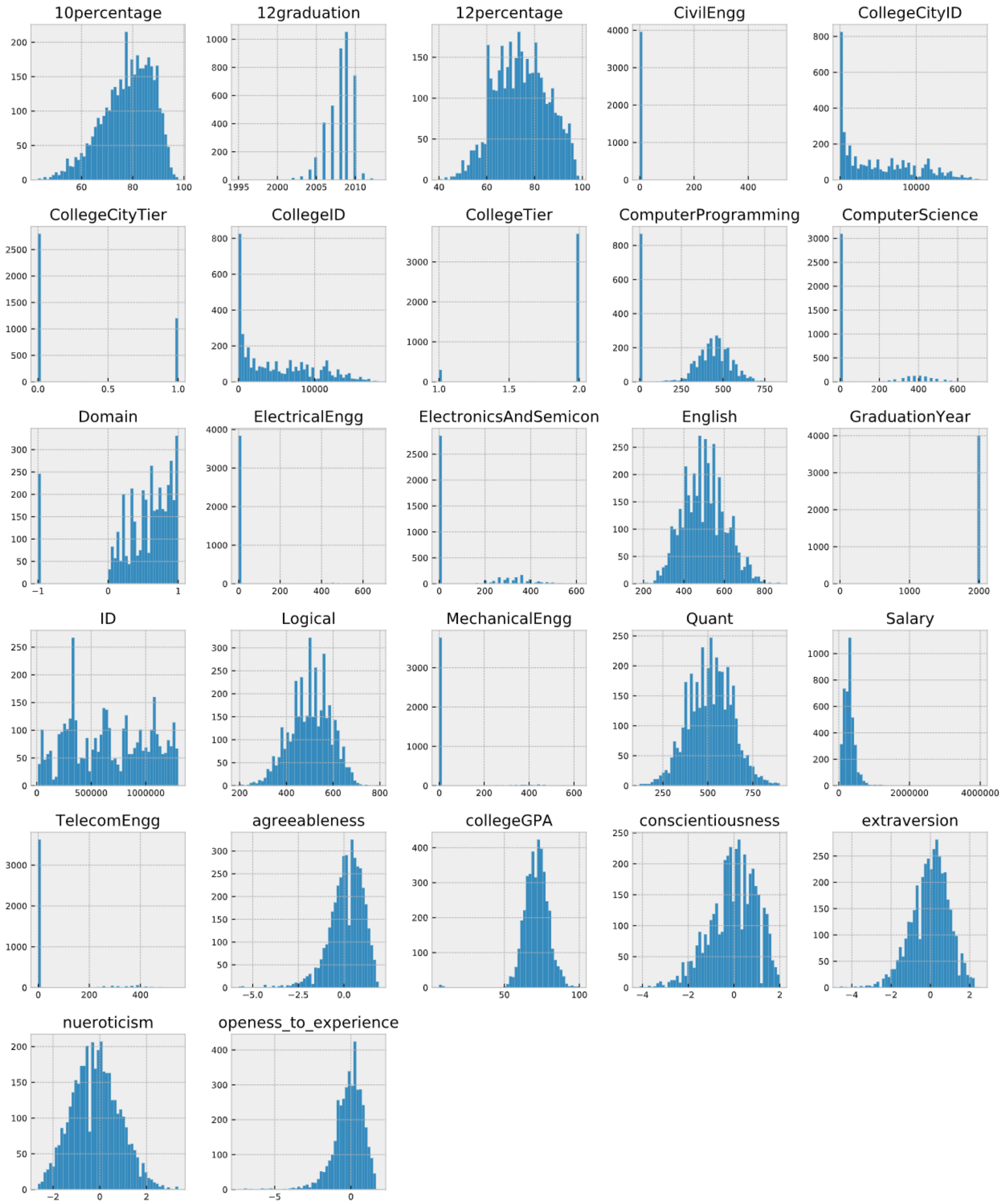
Exploratory Data Analysis

Highlights of exploratory data analysis is given below:

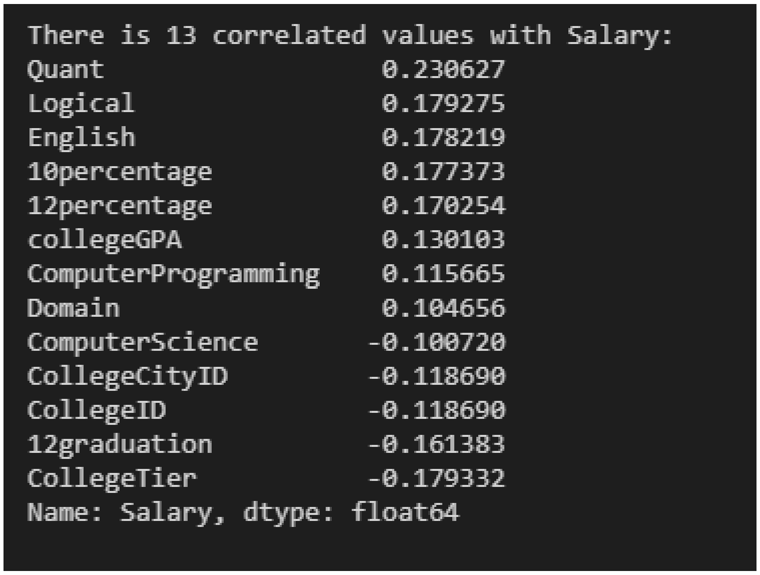
- Shape of data: 3998 rows and 39 columns – all non-null values
- Salary ranges from 0 to 4 million (Indian Rupees). However, most values are less than 1 million.



- Though all the data is non-null, data distribution charts of numeric values indicates that several columns have zeros in many rows. Based on this only a subset of columns (10percentage, 12percentage, Domain, English, Logical, Quant) are chosen for regression analysis. Some of the other columns with complete data (conscientiousness, extraversion, neuroticism, openness_to_experience) are not chosen because of lack of detailed description of the meaning of this data.

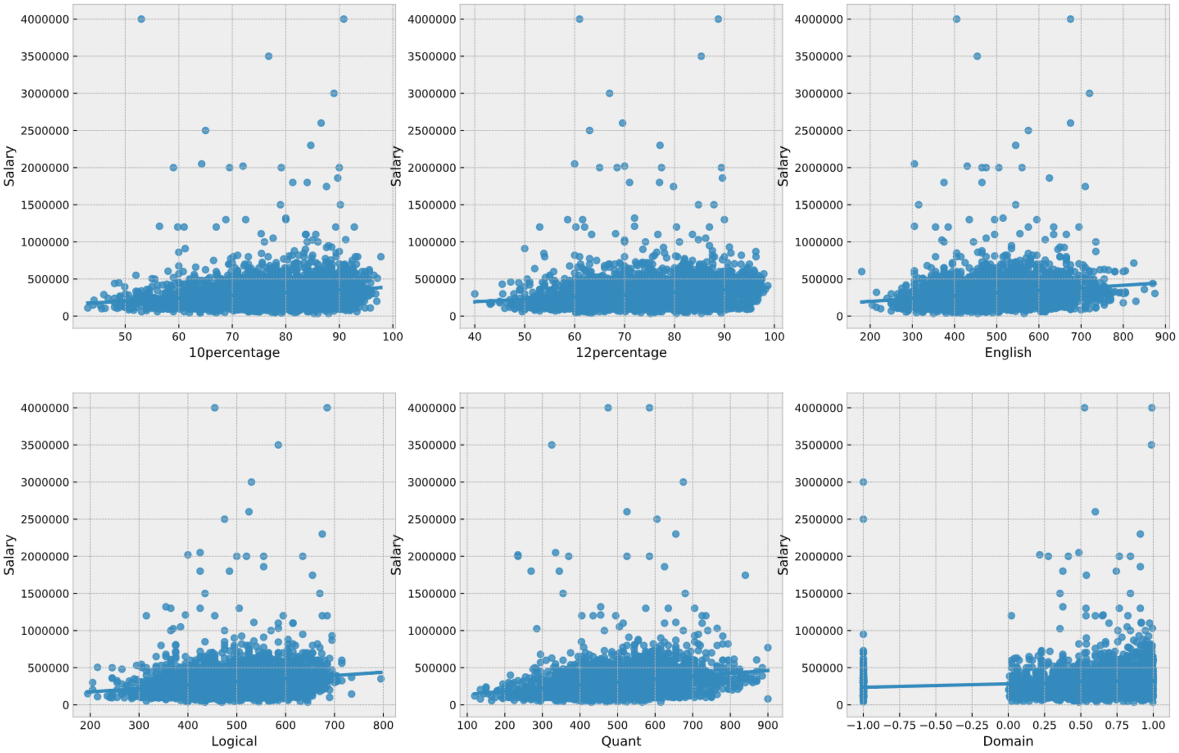


- Correlation between Salary and other parameters was evaluated. This indicated that the level of correlation is not very high.



```
There is 13 correlated values with Salary:
Quant                    0.230627
Logical                  0.179275
English                  0.178219
10percentage             0.177373
12percentage             0.170254
collegeGPA               0.130103
ComputerProgramming      0.115665
Domain                   0.104656
ComputerScience         -0.100720
CollegeCityID           -0.118690
CollegeID               -0.118690
12graduation            -0.161383
CollegeTier             -0.179332
Name: Salary, dtype: float64
```

- The correlation was plotted visually for some of the variables

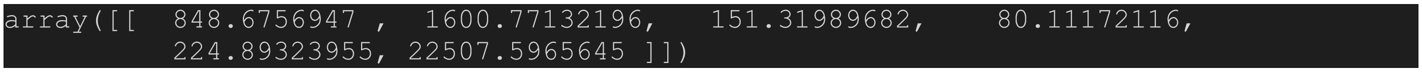


Results

Model Development – Regression

Multiple Linear Regression was done using six parameters – ‘10percentage', '12percentage', 'English', 'Logical', 'Quant', and 'Domain'. 20% of the records in the data set were reserved as test data.

The Regression coefficients discovered were as below.



```
array([[  848.6756947 ,  1600.77132196,   151.31989682,    80.11172116,
          224.89323955, 22507.5965645 ]])
```

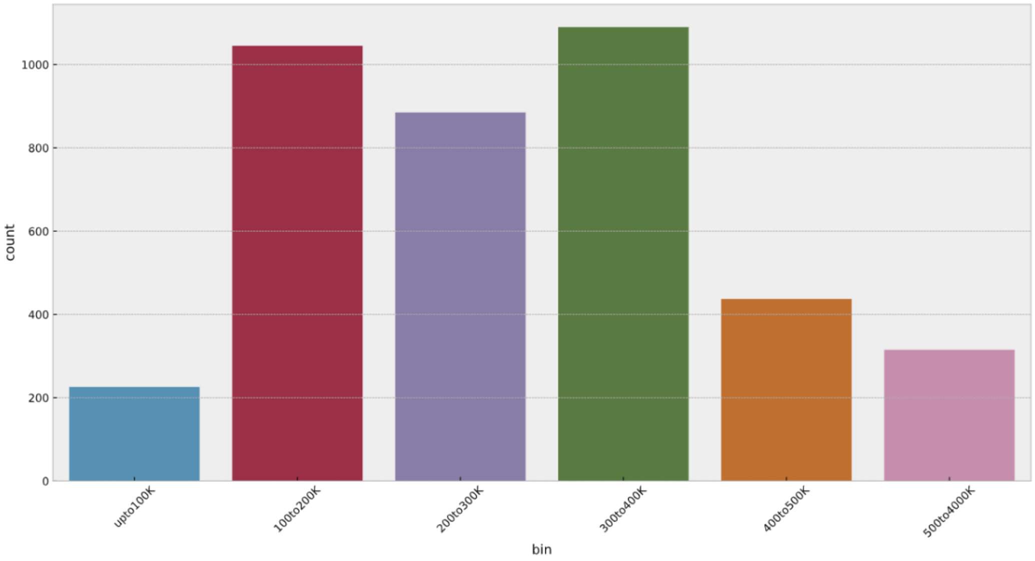
This indicates high level of influence for 10percentage and 12percentage. The coefficient for Domain is also high. The influence of Quant, English and Logical were relatively low.

The R^2 for the model is 0.11 indicating that the data is not a great fit for the model.

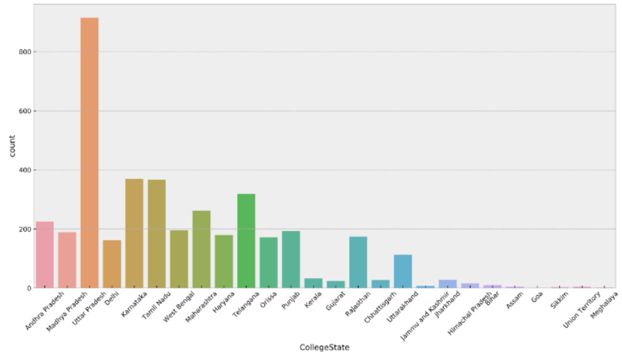
Model Development – Classification

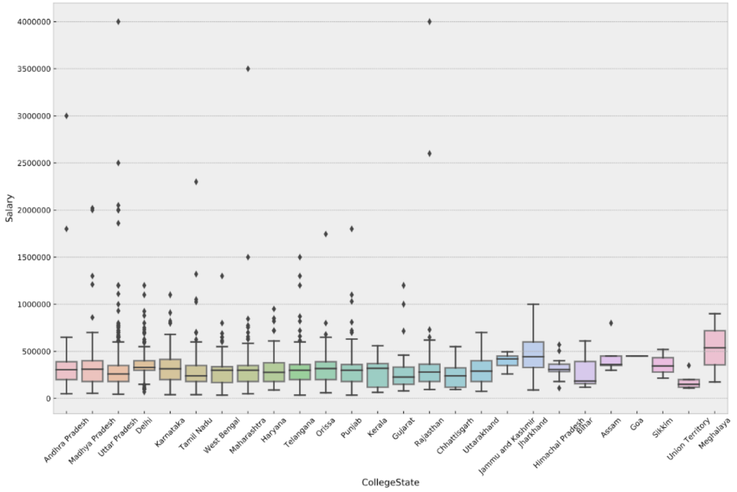
As the level of regression was less on any specific variable, supervised multi-class classification is considered as an alternate modeling option. The outcome – salary – is converted into a set of discrete ranges. Techniques such as Support Vector Machines can be used to learn to classify a set of features into one of the finite set of salary ranges.

The data set is broken into six bins of approximately similar number of values in each bin. This is shown below. The salary ranges (in Indian Rupees) are defined in five 100K bins up to 500K. All values above 500K up to 4000K are kept in a single bin.



In addition to the six features used in regression, a set of categorical variables are also identified. Count plots were done for all categorical values to identify number of distinct values. A subset of features with a limited number of values were chosen as candidates. Some of these variables are shown below visually.





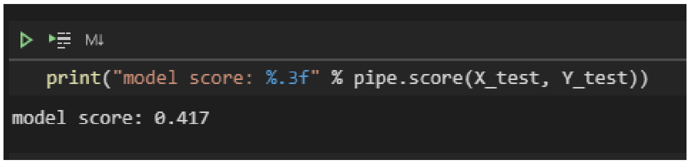
The set of variables chosen for modeling is as below.

- Numeric features: 10percentage, 12percentage, English, Logical, Quant, Domain
- Categorical features: Designation, JobCity, Gender, 10board, 12board, Degree, Specialization, CollegeState, CollegeTier, CollegeCityTier

The following process was followed for modeling:

- Pipeline segment for preprocessing

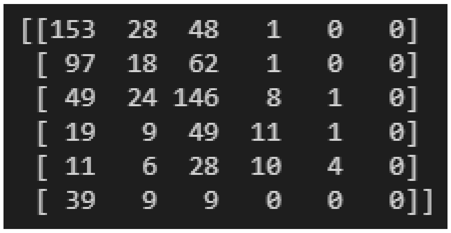
  - Categorical variables were encoded using One Hot Encoder
  - Numeric variables were scaled using Standard Scaler
- The outcome variable – bin (to denote the salary bin) was encoded using Label Encoder
- SVM using rbf kernel was used for modeling
- Model accuracy score is 0.417



```
print("model score: %.3f" % pipe.score(X_test, Y_test))
```

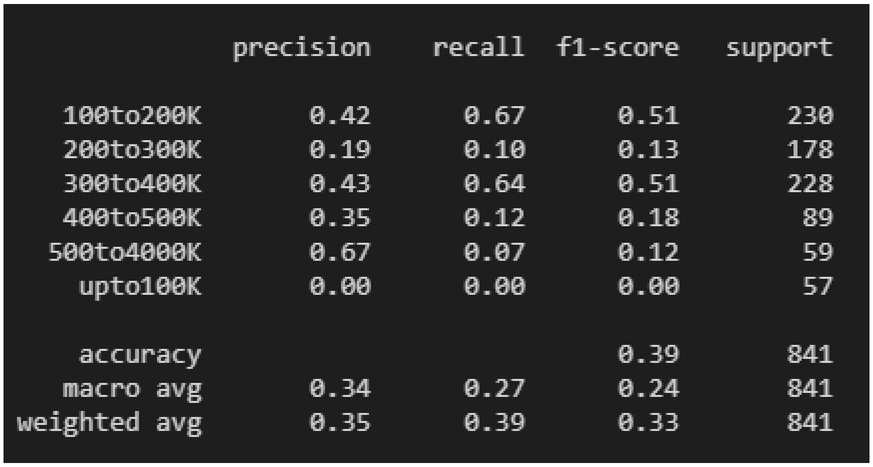
```
model score: 0.417
```

Confusion matrix based on comparing the predicted values with the test data is as below.



```
[[153  28  48   1   0   0]
 [ 97  18  62   1   0   0]
 [ 49  24 146   8   1   0]
 [ 19   9  49  11   1   0]
 [ 11   6  28  10   4   0]
 [ 39   9   9   0   0   0]]
```

The precision, recall and F1 scores are as below.

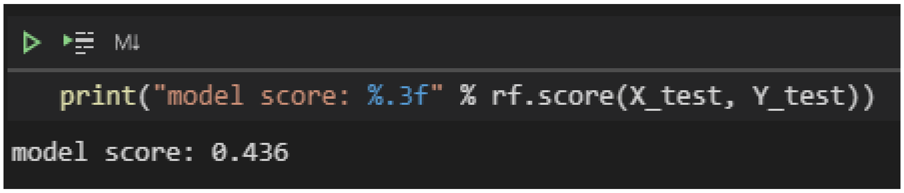


```
              precision    recall  f1-score   support

   100to200K       0.42      0.67      0.51       230
   200to300K       0.19      0.10      0.13       178
   300to400K       0.43      0.64      0.51       228
   400to500K       0.35      0.12      0.18        89
  500to4000K       0.67      0.07      0.12        59
    upto100K       0.00      0.00      0.00        57

    accuracy                           0.39       841
   macro avg       0.34      0.27      0.24       841
weighted avg       0.35      0.39      0.33       841
```

The F1 scores for the 100-200K, 300-400K were the best. The overall accuracy is 0.417. This is better than the accuracy of the regression model.
However, the ability of the model to predict other salary ranges – up to 100K, 200-300K, 400-500K and 500K+ is not good.

As classification scores were better that regression, classification models were built using other classifiers – KNN, Decision Tree, Random Forest, Ada Boost and Gradient Boost. The comparative R^2 scores indicated that the Gradient Boost classifier provided the best result. A final model is built using this classifier.

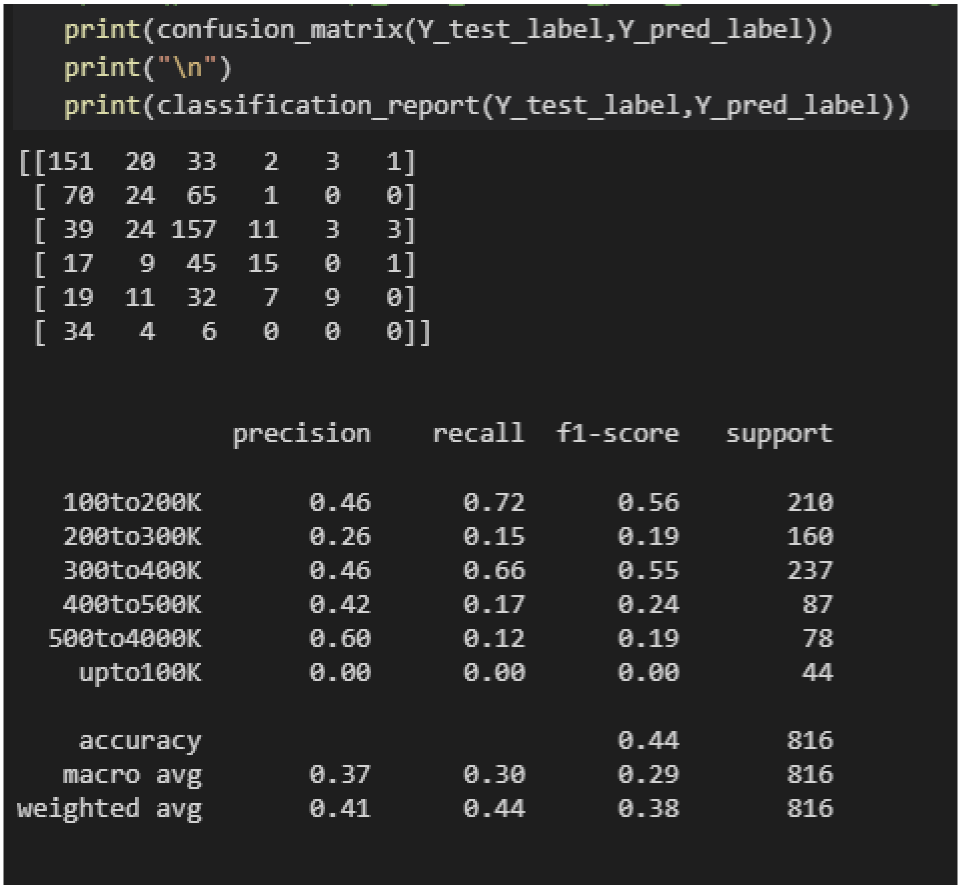
The results from this model are as below:



```
print("model score: %.3f" % rf.score(X_test, Y_test))
```

```
model score: 0.436
```

The confusion matrix is shown below:



```
print(confusion_matrix(Y_test_label,Y_pred_label))
print("\n")
print(classification_report(Y_test_label,Y_pred_label))
```

```
[[151  20  33   2   3   1]
 [ 70  24  65   1   0   0]
 [ 39  24 157  11   3   3]
 [ 17   9  45  15   0   1]
 [ 19  11  32   7   9   0]
 [ 34   4   6   0   0   0]]


              precision    recall  f1-score   support

   100to200K       0.46      0.72      0.56       210
   200to300K       0.26      0.15      0.19       160
   300to400K       0.46      0.66      0.55       237
   400to500K       0.42      0.17      0.24        87
  500to4000K       0.60      0.12      0.19        78
    upto100K       0.00      0.00      0.00        44

    accuracy                           0.44       816
   macro avg       0.37      0.30      0.29       816
weighted avg       0.41      0.44      0.38       816
```

The results are marginally better than the results from SVM.

Discussion and Conclusion

The exploratory data analysis indicated that the level of correlation between salary and any specific set of variables is not very high. A combination of features was used to prepare a regression model. The results from the regression model too indicated the same.

The data was split into several salary ranges to pursue a classification approach to predict salary ranges based on a set of features. The model was able to predict certain ranges in the salary but unable to predict certain other ranges.

Based on the equation derived using regression analysis, the features with the greatest influence are 10th standard percentage and 12th standard percentage.

Potential Next Steps

The accuracy of the models may be further improved by performing hyperparameter tuning for the selected models.
Other sources of data, with additional features with higher impact may be considered.