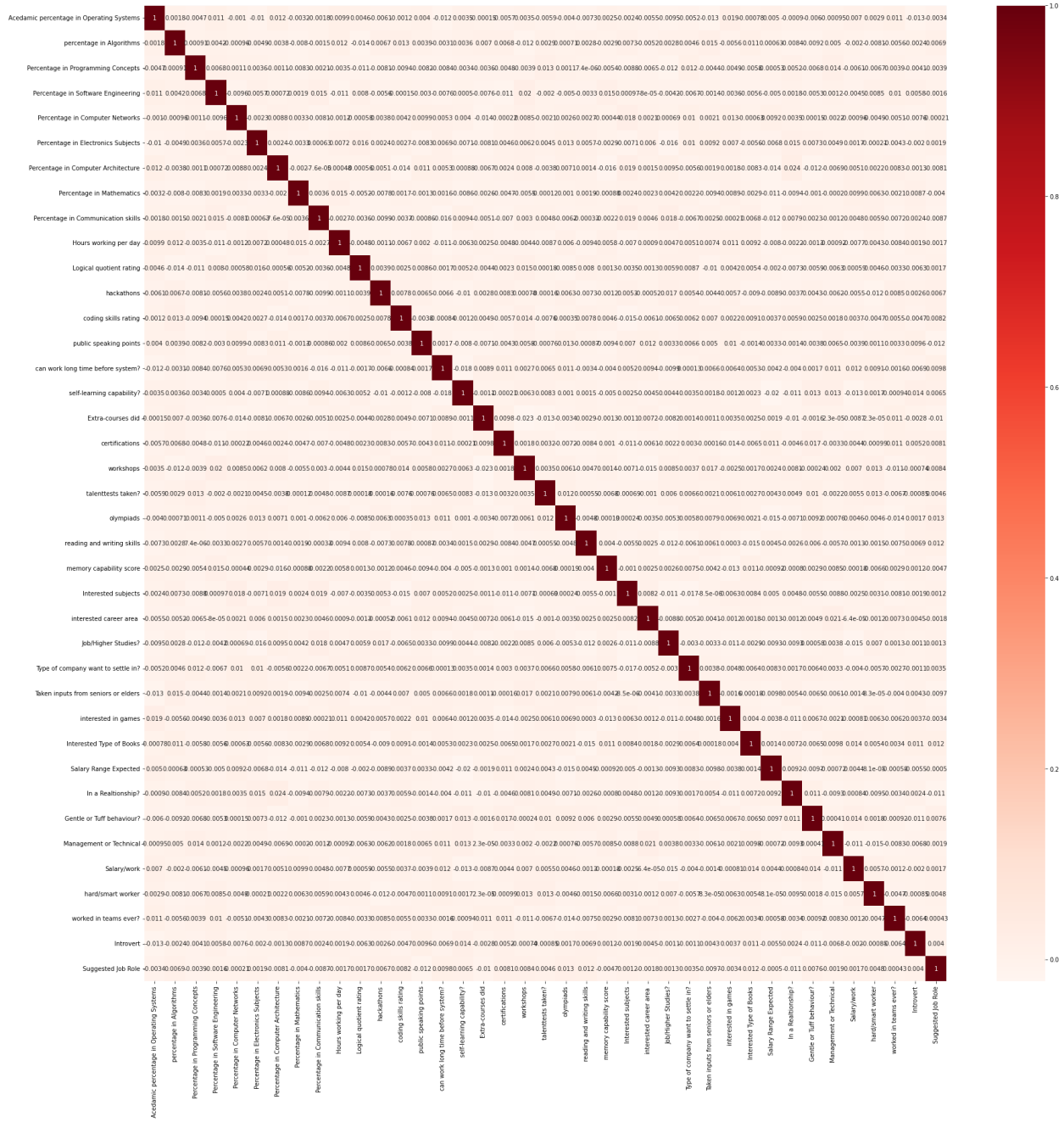
AI Assignment - 4 (Abhinav Saurabh MT20127)

Dataset

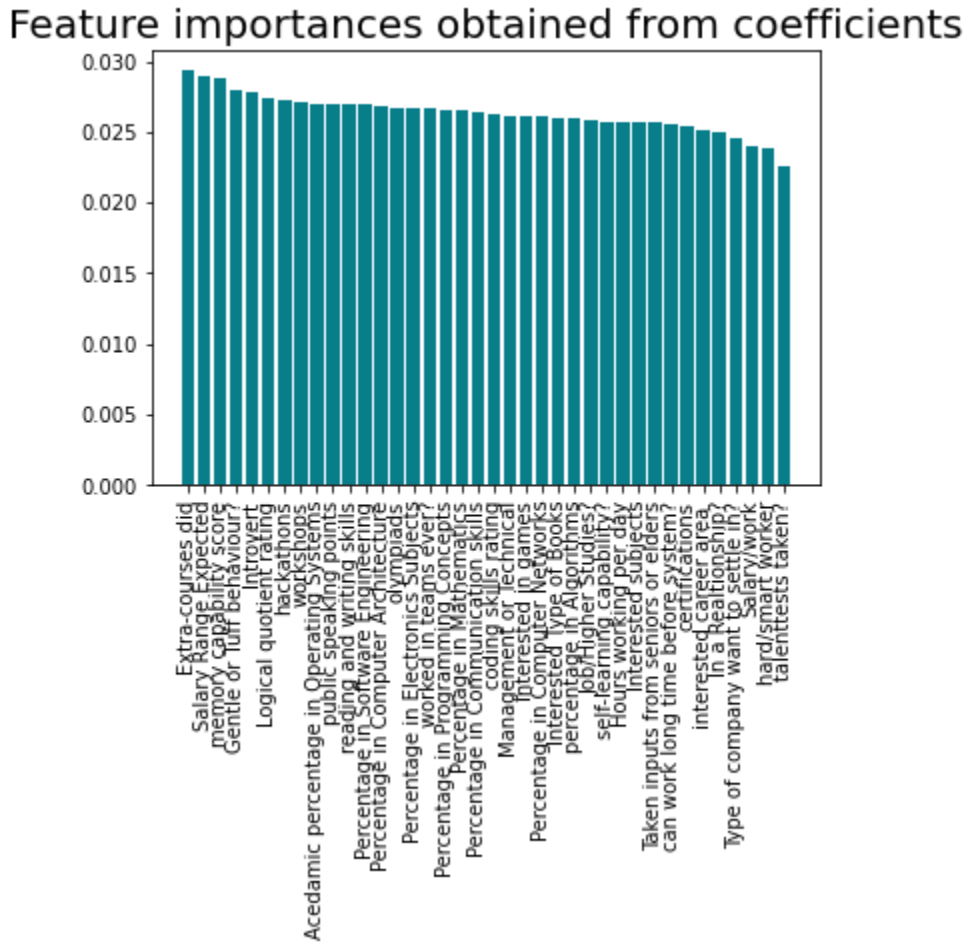
- The dataset contains 20000 rows × 39 columns.
- That means it has 38 features and 1 target variable.
- And dataset contains 20000 instances.

Further Analysis

- Further analysing the dataset we found very less correlation among the features.
- Some of the graphs are below. This is Heatmap.



- Another graph for the correlation.



- I can see the features have a very low correlation with the target variable.

Reclassifying Target variables and other modifications

- I reclassified the target variable among 7-8 classes to improve the prediction accuracy.
- I have used classes such as Admin, Analyst, Developer, Manager, Architect, Support, Engineer, Associate Roles.
- After Reclassification the accuracy score improved from 5% to 20%.
- This reclassification worked well for the model.
- Further bucketing grades didn't improve any accuracy scores for the model. So I undid it.

Label Encoding

- I performed label encoding to further make it suitable for machine learning Algorithms.
- I used label encoder fit transform on all the columns to convert them into numbers and classes.

ANN Model:

Using MLP Classifier I have implemented a neural network with 5 Layers.(128,64,32 & relu)

With 60:40 Split we obtain accuracy : 18.9%
With 70:30 Split we obtain accuracy : 18.51%
With 80:20 Split we obtain accuracy : 18.6%
With 90:10 Split we obtain accuracy : 19.3%

Other Models:

- 60:40 Split
    - Logistic Regression: 18.225%
    - Random Forest: 16.98%
    - SVM: 18.05%
    - XGB: 17.5%
- 70:30 Split
    - Logistic Regression: 18.65%
    - Random Forest: 17.5%
    - SVM: 18.2%
    - XGB: 18.93%
- 80:20 Split
    - Logistic Regression : 18.65%
    - Random Forest : 17.15%
    - SVM : 18.075%
    - XGB : 17.9%
- 90:10 Split
    - Logistic Regression :19.15%
    - Random Forest : 17.3%
    - SVM :18.8
    - XGB :18.3