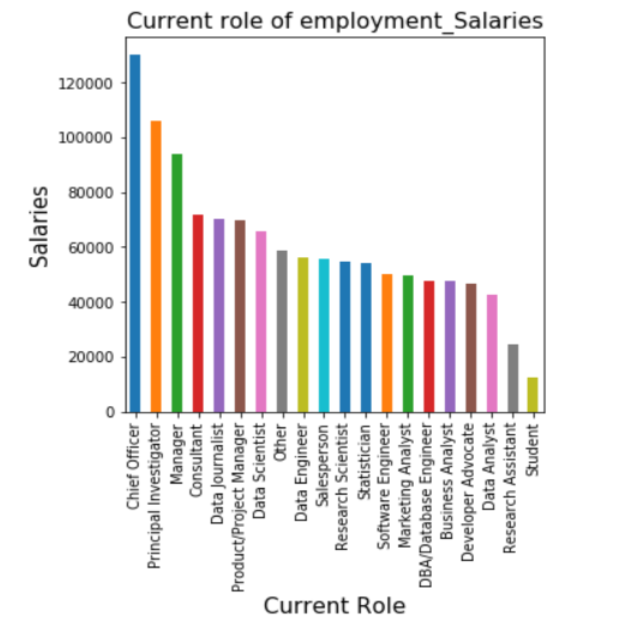
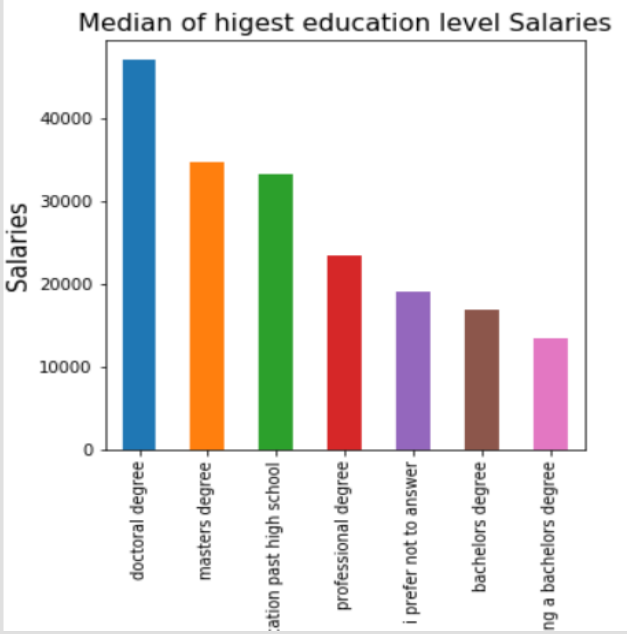
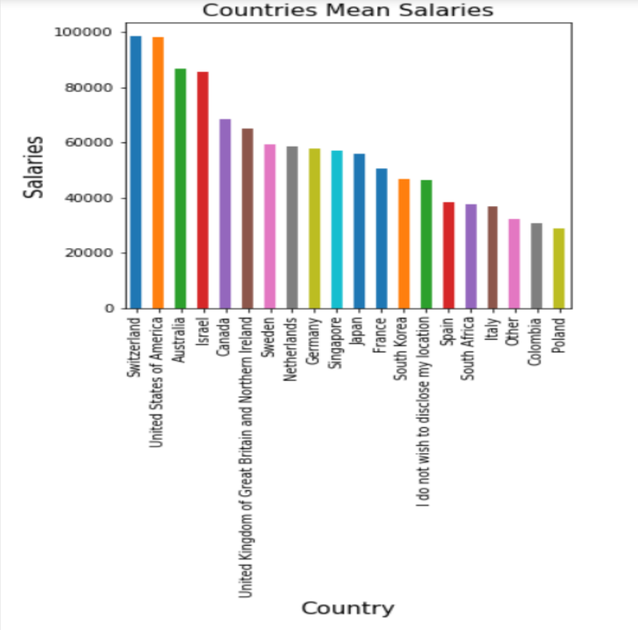
Exploratory Analysis:

In the exploratory analysis I have plotted the features like highest level of education, Current role and country against salaries and I have found that:

1. PhD students and Master students are getting high salaries.
2. Chief Officer gets high salary and student gets the lower as salary as students doesn’t work full time.
3. Switzerland and United States of America share the privilege of highest paid salaries

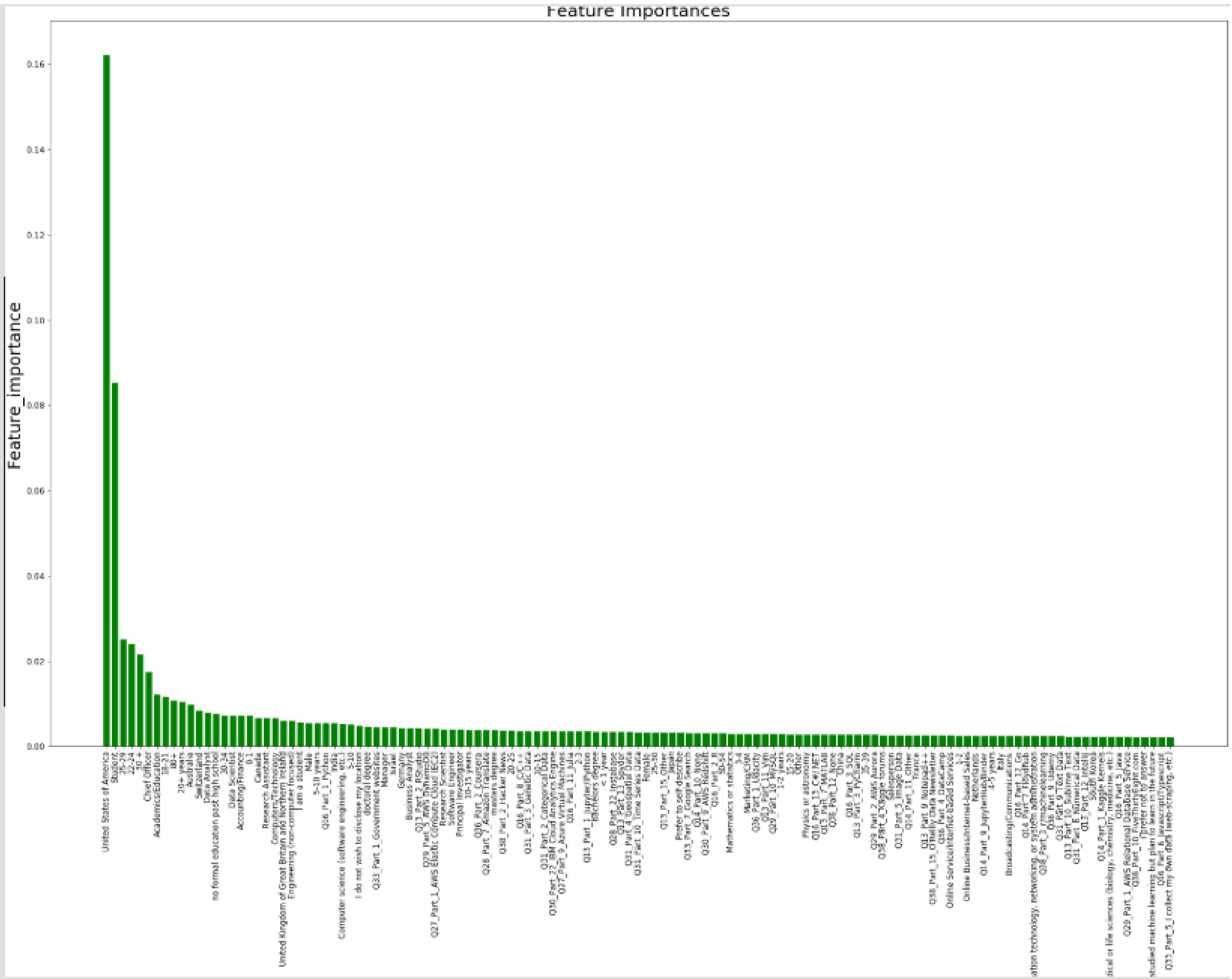






Feature Importance:

In the exploratory analysis I have plotted the order of feature importance and it seems that features like location (USA), age, current role have high importance and are shown in the figure. I have used Random Forest Regressor feature selection technique .



Feature selection can be used for the following reasons:(From Wiki)

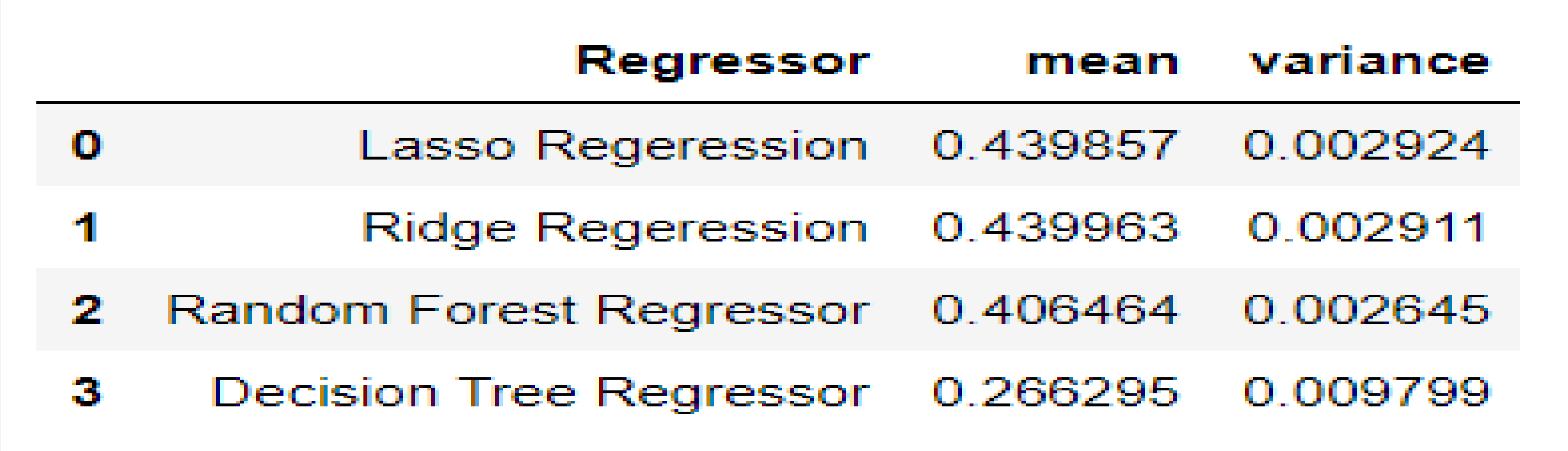
1. Simplification of models to make them easier to interpret by researchers/users
2. Shorter training times
3. To avoid the curse of dimensionality,
4. Enhanced generalization by reducing overfitting (formally, reduction of variance)

I have manually picked the features from the data set after preprocessing the data and then I created a data frame with 343 features. Then I used Random Forest Regressor Technique to generate the feature importance and then I have removed those features which aren't significant to predict the salary. Finally I have selected 130 features.

Random Forest regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size. I have used feature_importaces_ as my metric.

MODEL RESULTS

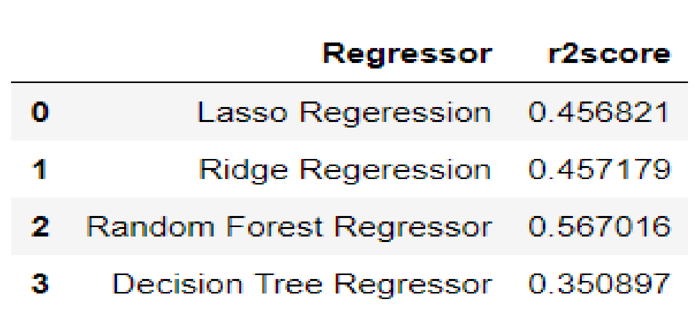
- I have taken four algorithms like Lasso regression, Ridge Regression, Decision Tree regressor and Random Forest Regressor.
- I have performed 10 fold cross validation along with hyper tuning of the parameters for the respective models.
- Below table gives the mean fold accuracy and variance of the folds .



	Regressor	mean	variance
0	Lasso Regeression	0.439857	0.002924
1	Ridge Regeression	0.439963	0.002911
2	Random Forest Regressor	0.406464	0.002645
3	Decision Tree Regressor	0.266295	0.009799

MODEL RESULTS

- After running the hyper tuning of the parameters my r2 scores are as follows:



	Regressor	r2score
0	Lasso Regeression	0.456821
1	Ridge Regeression	0.457179
2	Random Forest Regressor	0.567016
3	Decision Tree Regressor	0.350897

- After running the models with 10 fold cross validation and hyper tuning the parameters random forest regressor was performing the best for me.
- The training score was 0.56
- The testing score was 0.42
- We can say that , there is a overfitting of the data. The model is not able to find the exact pattern of the model whereas it is just remembering it which is leading to a less test score. To improve the accuracy of the model I can include more parameters for tuning like number of features used for a split and increasing number of estimators too increases the accuracy.