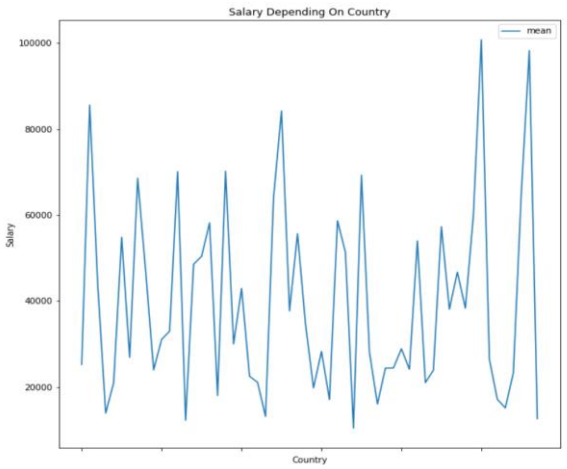
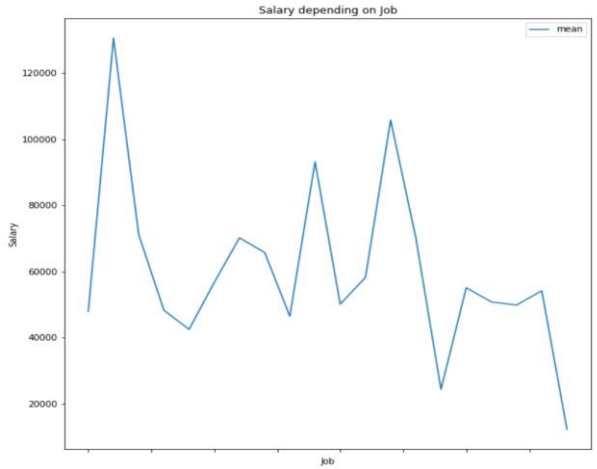
MIE1624 Assignment 2

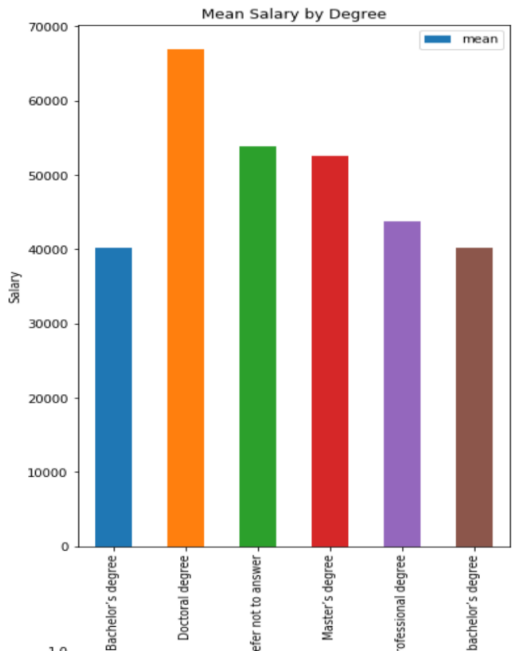
Exploratory Analysis

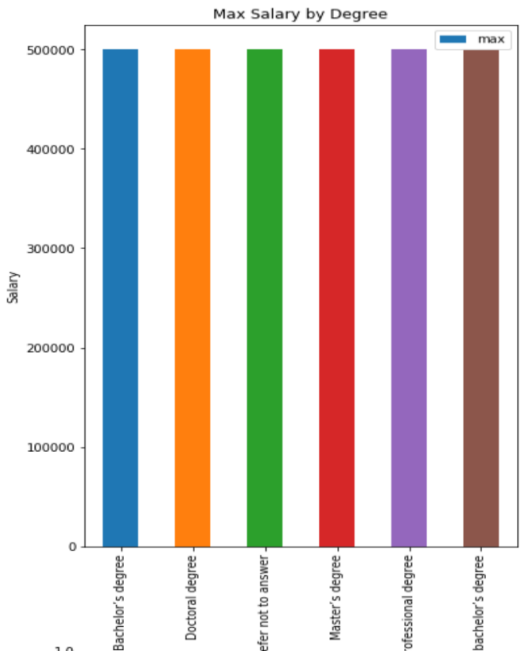


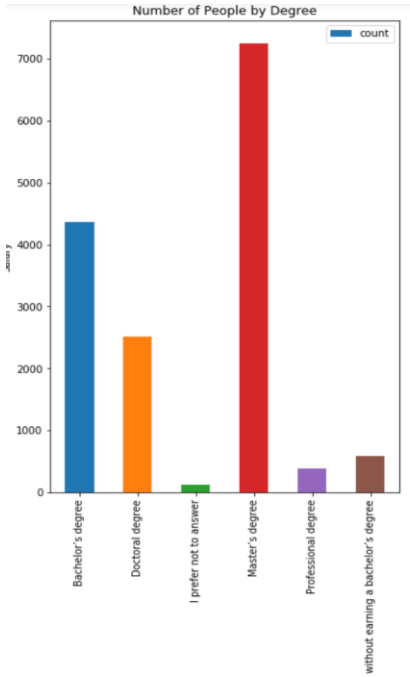


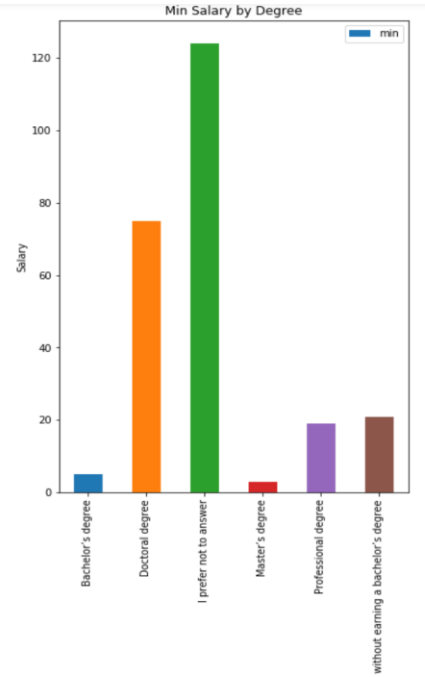
The left figure show the relationships between Job and yearly salary, Country and yearly salary. The average yearly compensation vary with these two variables.

The right four figure shows the mean yearly salary of people depending on their education degrees. It proves that people with higher degree earn more salary. The Doctoral degree has the highest average. It should be also noted that people with different degrees all have chance to earn salary as high as 500,000. Most people have master degree. From the last figure that shows minimum compensation by degree, it shows that that some people earn as less as 20. This may be fake because some respondents may have given false response.

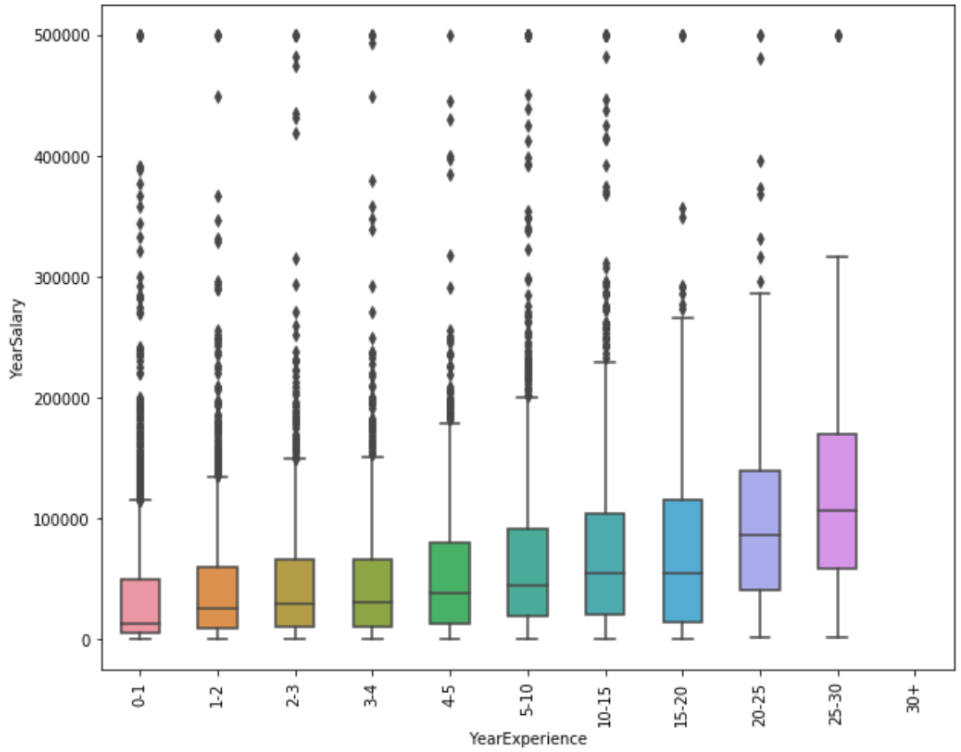




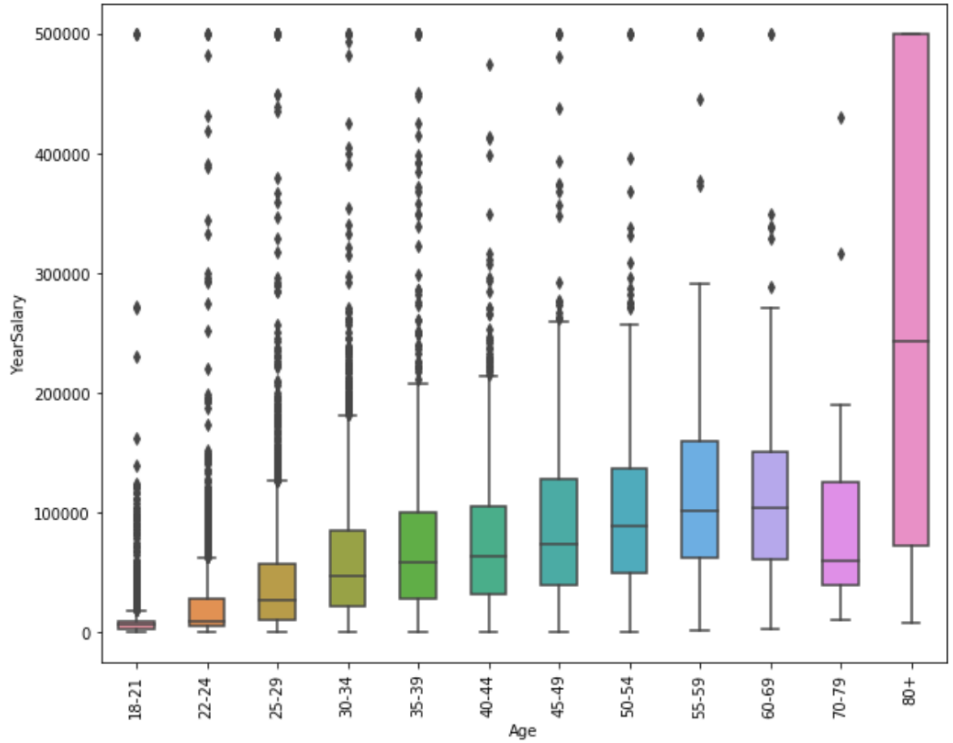




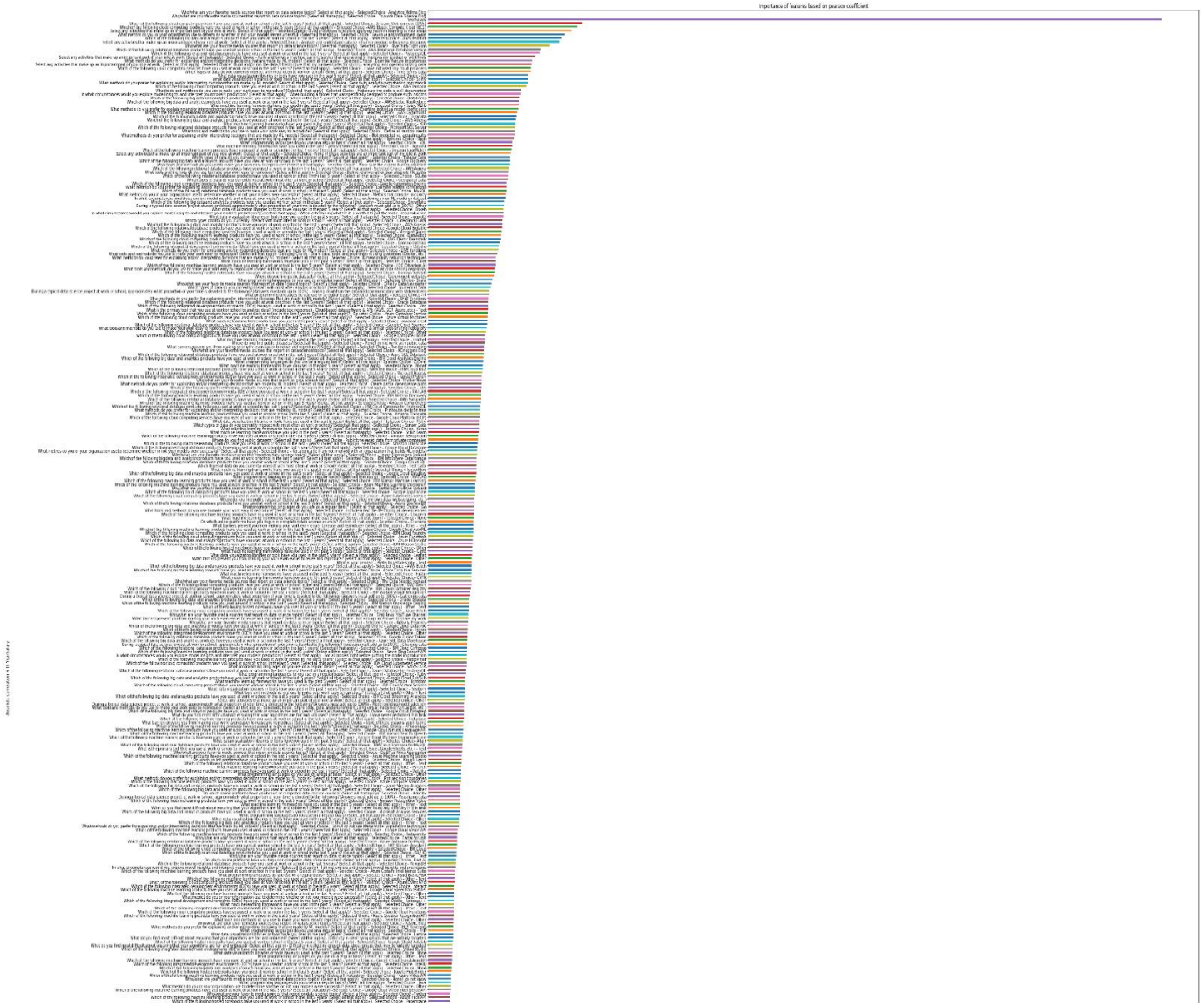
Exploratory Analysis



These figures show the relationships between Age and yearly salary, years of working experience and yearly salary. Based on the compensation distribution, it can be concluded that older people or those who have longer work experience are more likely to have a higher yearly salary.



Visualization



In the figure, the features are ranked according to their pearson correlation with yearly salary.

TOP 5 most important features according to correlation with Salary are:
1.Using AWS in the past
2. Using EC2 in the past
3. Building prototypes to explore applying machine learning to new areas in work
4. Using Revenue and/or business goals to determine wheter or not their model was successful
5. Using AWS Redshift product in the past

The top 3 less important features according to correlation with Salary are:
1. Completing Datacamp Data science course
2. Completing Udemy Data science course
3. Spending most of their other online platform
4. Towards Data Science Bog as their favorite media source
5. Analytics Vidhya Blog as their favorite media source

Feature Selection and Model Implementation

Two kinds of feature selection methods, “Select From Model-Lasso” and “Principal Component Analysis(PCA)”, were implemented. Then, four machine learning methods, which are "Linear Regression", "Lasso", "Random Forest", "Gradient Boosting“ to do regression.

In the list below, R-Square(R2) and Root Mean Square Error (RMSE) scores on testing set were documented. Note that during the training, 10-fold cross-validation were done on each model and the scores are the mean of 10 folds. The list shows that “Select From Model-Lasso” is better than “PCA” from every measurement. Therefore, the models were developed based on “SelectFromModel-Lasso” feature selection.

Linear Regression:
- Select from Lasso Model: R2: 0.586 RMSE: 33281.99
- PCA Selection: R2: 0.508192 RMSE: 36259.826012

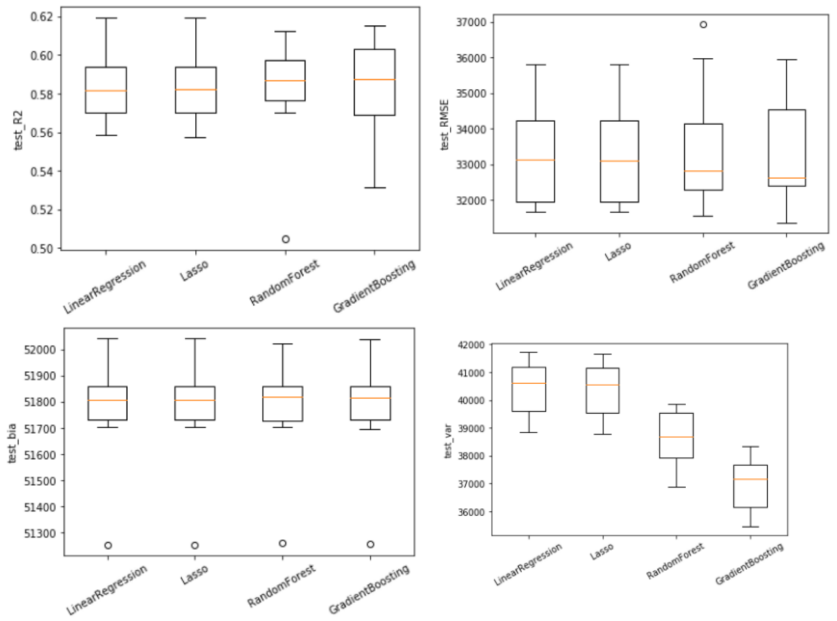
Lasso Regression:
- Select from Lasso Model: R2: 0.586 RMSE: 33280.845
- PCA Selection: R2: 0.508335 RMSE: 36254.577446

Random Forest:
- Select from Lasso Model: R2: 0.581 RMSE: 33479.51
- PCA Selection: R2: 0.335 RMSE: 42181.732

Gradient Boosting:
- Select from Lasso Model: R2: 0.583 RMSE: 33394.71
- PCA Selection: R2: 0.391928 RMSE: 40332.249468

	R2 Mean	R2 Variance	RMSE Mean	RMSE Variance
Linear Regression	0.585748	0.020902	33281.9889	1432.597005
Lasso	0.585777	0.021063	33280.84464	1441.114786
Random Forest Regression	0.580787	0.029581	33479.506408	1790.835260
Gradient Boosting Decision Tree	0.582955	0.025714	33394.714399	1665.112675

To compare the models performance on chosen feature bunch, the R2 and RMSE score were used to measure the accuracy. Lasso has the best R2 score and RMSE score.
To better view the distribution of scores of different algorithms, we visualize the score using boxplot.



The left four figures shows not much difference in R2, RMSE across the algorithms.
The bias of each algorithms are almost the same, but the variance are quite different. Linear Regression and Lasso have the larger variance, while Random Forest have lower and Gradient Boosting have even lower variance. That's because Gradient Boosting is based on weak learners (high bias, low variance).In terms of decision trees, weak learners are shallow trees, sometimes even as small as decision stumps (trees with two leaves). Boosting reduces error mainly by reducing bias. Taking into account that Gradient Boosting performs better than Random Forest with regard to variance, Gradient Boosting has been chosen as best model.

Results

GridSearch Cross-Validation was used to tune the hyperparameters.
For each algorithm, we define the hyperparameter space beforehand.
With every candidate hyperparameter, GridSearchCV will first split the data into K folds, train the model on K-1 folds and then score on test data. We choose to use R2 as performance measure score. R2 measures how well the regression line approximates the real data points, it also portrays percent of variance in the data explained by regression model. GridSearchCV will give the best hyperparameters based on the average R2 score cross K times training.

Algorithm	R2	RMSE
LinearRegression	0.609854	32337.127780
Lasso	0.609823	32338.420809
RandomForest	0.857235	19561.372166
GradientBoosting	0.841806	20591.270771

According to above table, the models that perform better are Random Forest and Gradient Boosting. Random Forest seems to be better. However, according to the analysis before, Gradient Boosting has much lower variance than Random Forest and the bias are almost the same. So we choose Gradient Boosting as our best model.

Measurement	Train	Test
R2	0.841805	0.622567
RMSE	20591.270	31562.426438
Bias	51771.136	51377.113905
Variance	43157.7047	40938.415630
Total	94928.84164	92315.529536

The results above shows the train and test result based on Gradient Boosting model. It shows the performance on training set is better than testing set.
It shows the R2 score is much higher for training than testing, which means that the model fits training set so well that it has captured the noise of the data. The variance difference between training set and testing set is much higher than bias difference, and the variance in training stage is higher. Therefore it can be concluded that the model is overfitting.

To improve the testing performance, we can fit multiple models and use validation or cross-validation to compare their predictive accuracies on test data. In the case of Gradient Boosting, the maximum tree depth also plays a huge role in determining the fit, so we can decrease the max depth to significantly reduces overfitting.