MACHINE LEARNING SALARY PREDICTION

Elif Özker
Oğuz Kaan Şanlı
Elif Dilara Akkuş

DEUCENG
Dokuz Eylül University
Dept of Computer Engineering

Outline

Introduction and Dataset

- 6684 Rows
- 9 Columns
- Education level represents:

  0 : High School

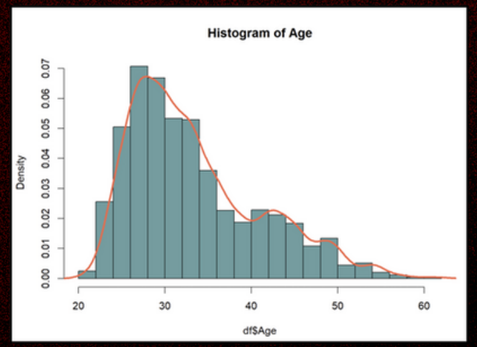
  1: Bachelor Degree

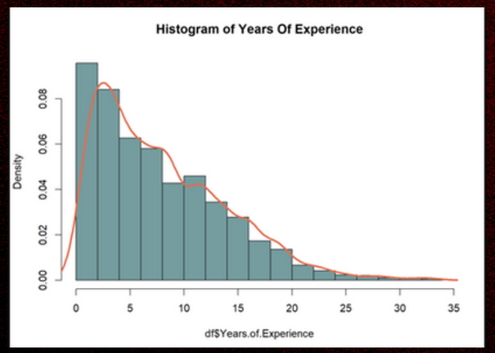
  2: Master Degree

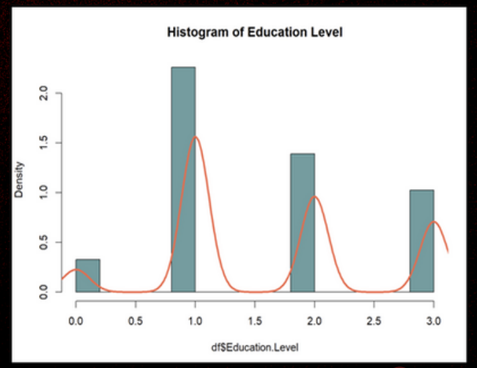
  3 : Phd
- Target Feature is Salary

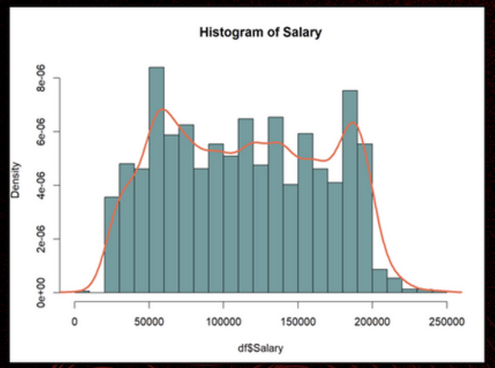
```
      Age            Gender          Education.Level  Job.Title          Years.of.Experience
 Min.   :21.00   Length:6684        Min.   :0.000   Length:6684        Min.   : 0.000
 1st Qu.:28.00   Class :character   1st Qu.:1.000   Class :character   1st Qu.: 3.000
 Median :32.00   Mode  :character   Median :1.000   Mode  :character   Median : 7.000
 Mean   :33.61                      Mean   :1.622                      Mean   : 8.078
 3rd Qu.:38.00                      3rd Qu.:2.000                      3rd Qu.:12.000
 Max.   :62.00                      Max.   :3.000                      Max.   :34.000
     Salary          Country              Race               Senior
 Min.   :   350   Length:6684        Length:6684        Min.   :0.0000
 1st Qu.: 70000   Class :character   Class :character   1st Qu.:0.0000
 Median :115000   Mode  :character   Mode  :character   Median :0.0000
 Mean   :115307                                         Mean   :0.1435
 3rd Qu.:160000                                         3rd Qu.:0.0000
 Max.   :250000                                         Max.   :1.0000
```

Preprocessing and Cleaning









Check for missing values

Remove duplicated data

Preprocessing and Cleaning

Encoding

Gender_Female	Gender_Male	Education.Level_0	Education.Level_1
0	1	0	1

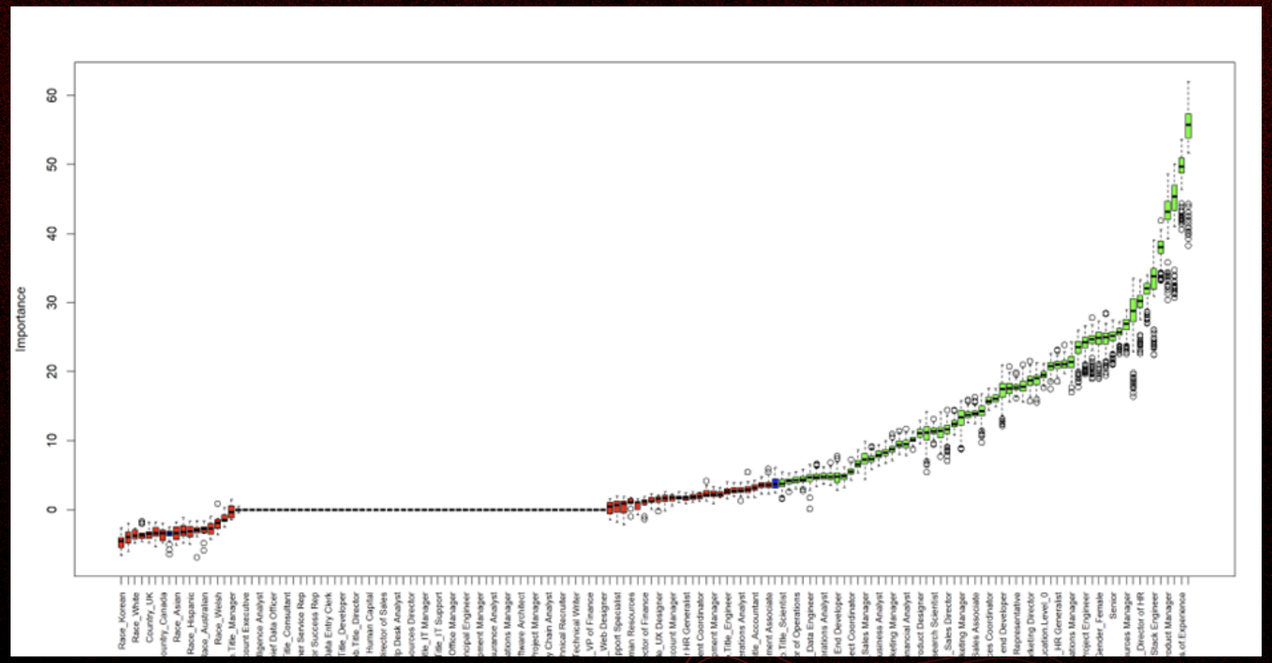
Scaling

	Age	Years.of.Experience
1	32	5.0
2	28	3.0
3	45	15.0
4	36	7.0
5	52	20.0

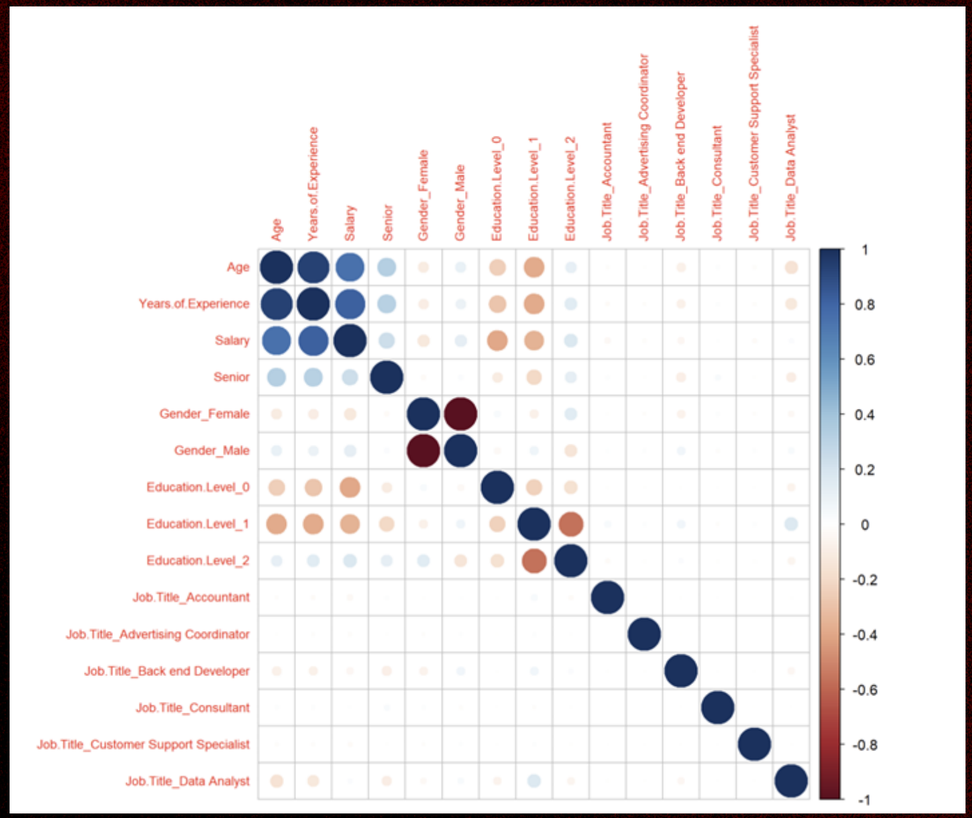
	Age	Years.of.Experience
1	-0.2577865185	-0.53527068
2	-0.7730592801	-0.85611524
3	1.4168499566	1.06895210
4	0.2574862430	-0.21442613
5	2.3185772893	1.87106350

Preprocessing and Cleaning

Feature Selection

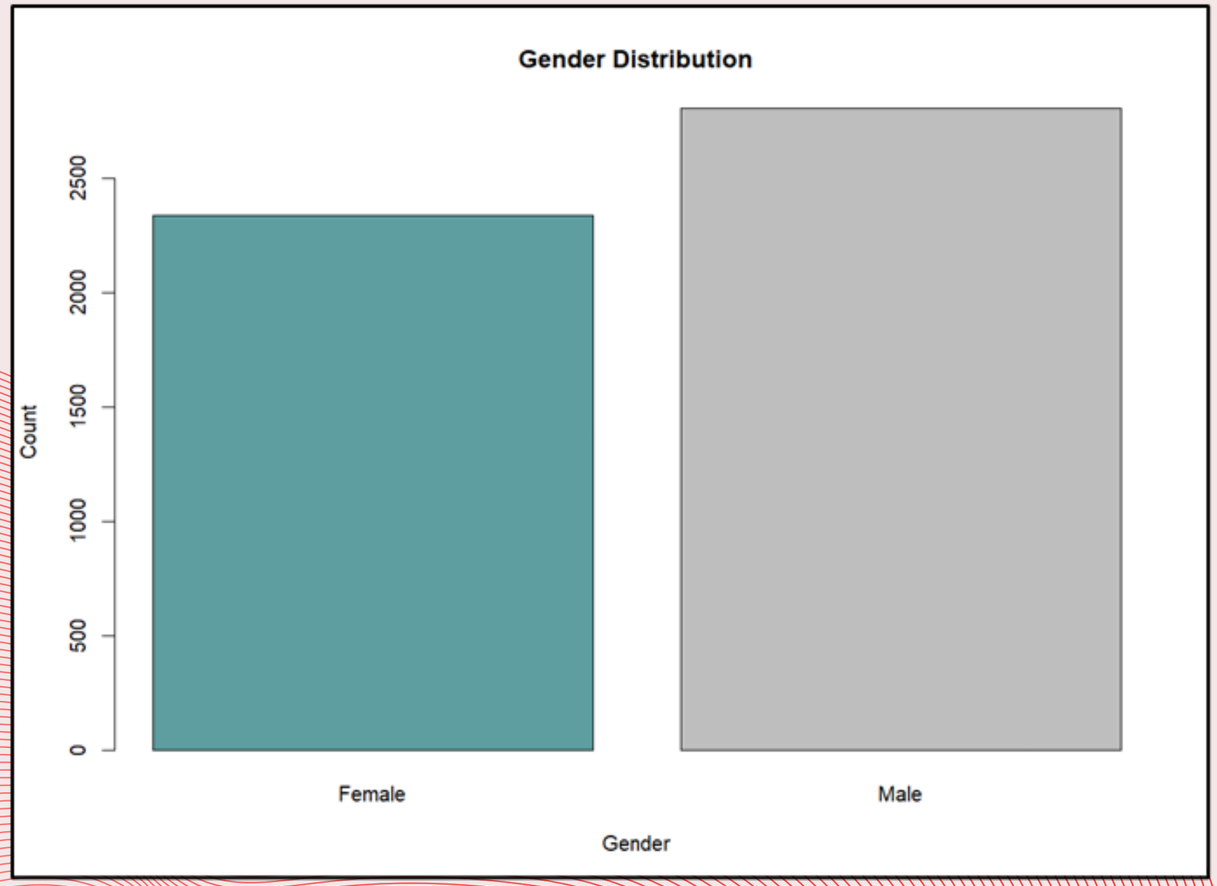


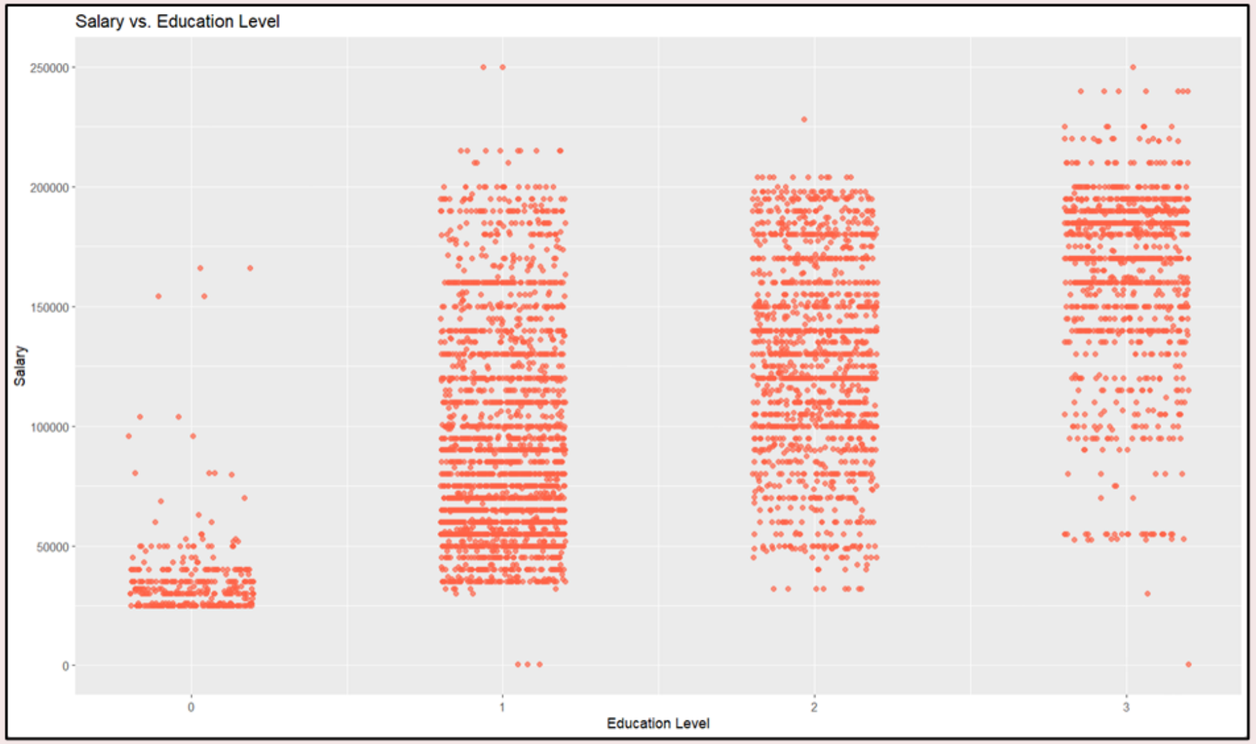
Boruta



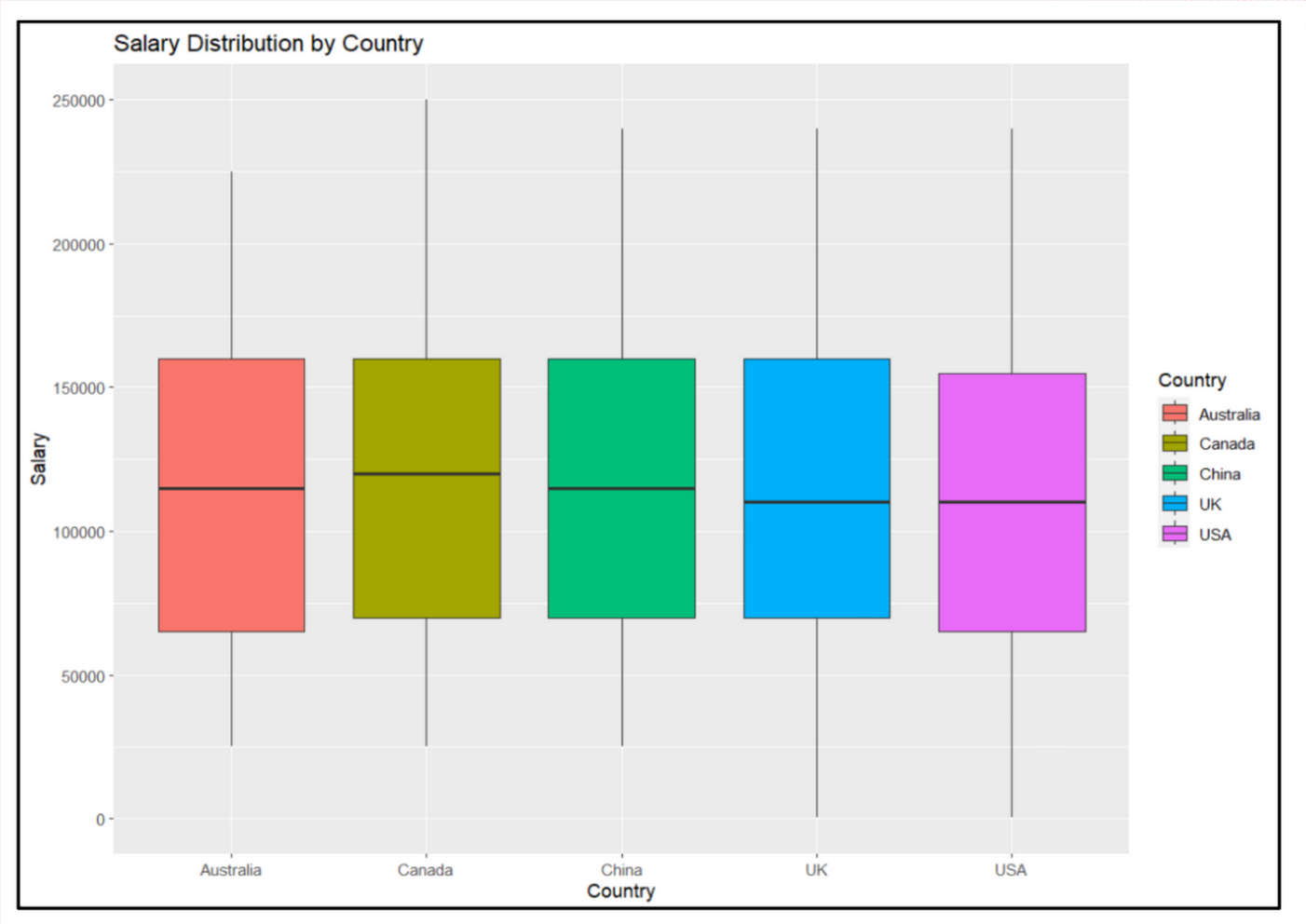
Correlation Matrix

VISUALIZATION

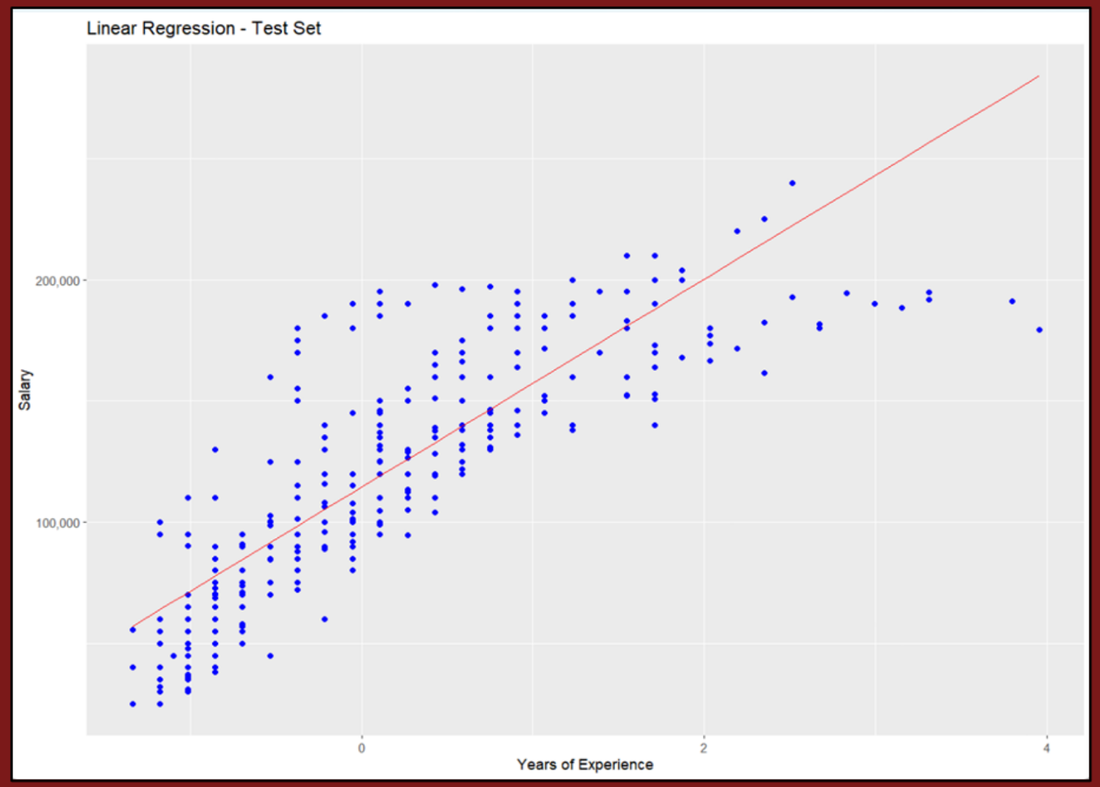




VISUALIZATION

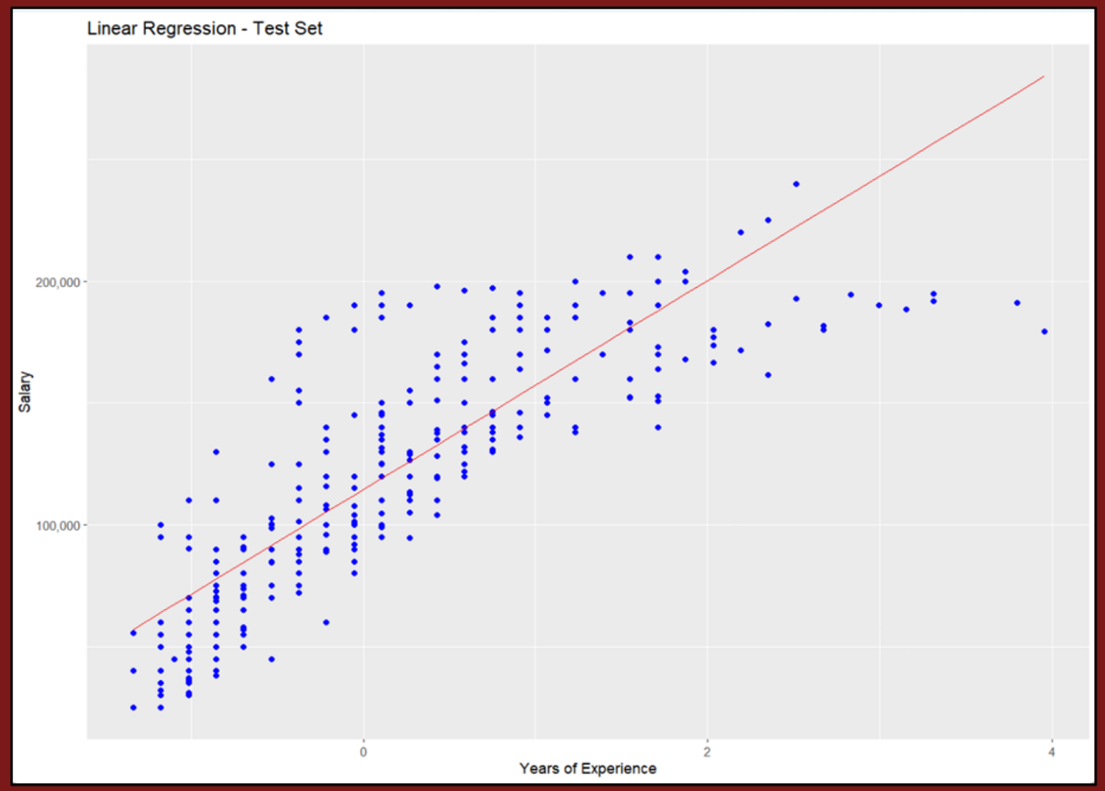


Single Linear Regression



Before Cross-Validation

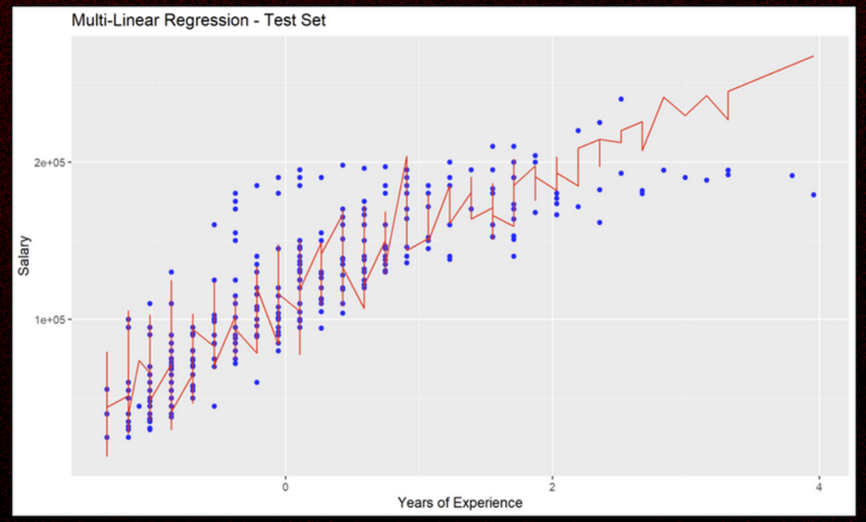
0.6936



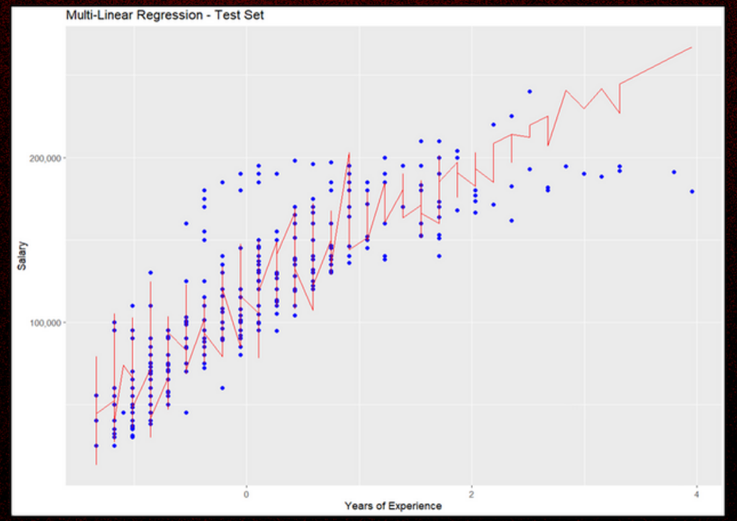
After Cross-Validation

0.6937

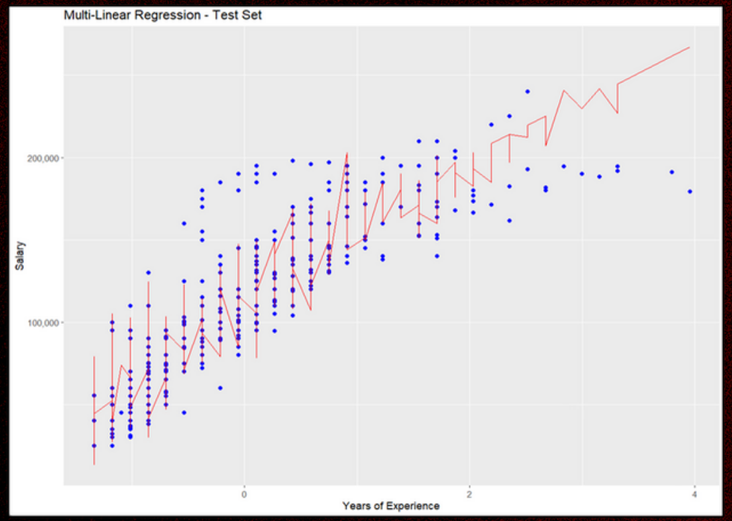
Multi-Linear Regression



Before Cross-Val and Tuning
0.8040



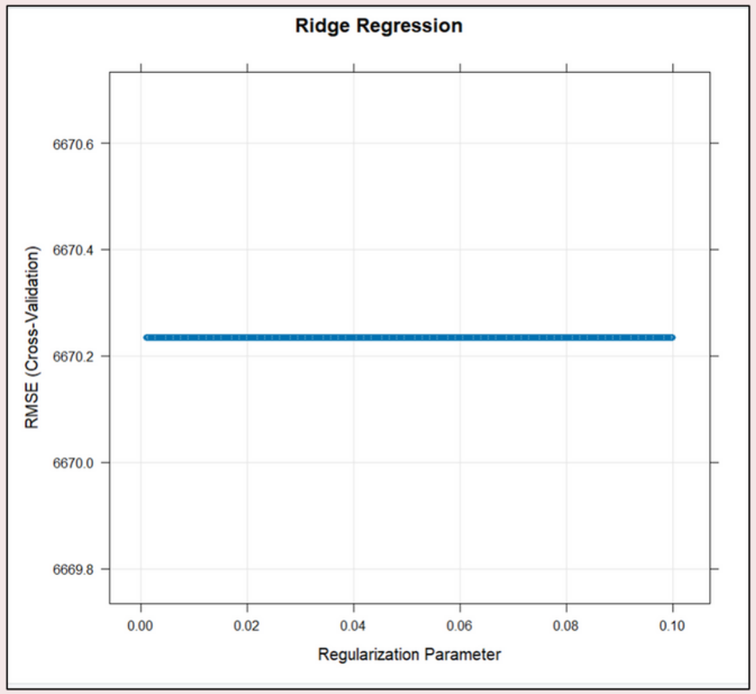
After Cross-Validation
0.8063



After Parameter Tuning
0.8042

Ridge Regression

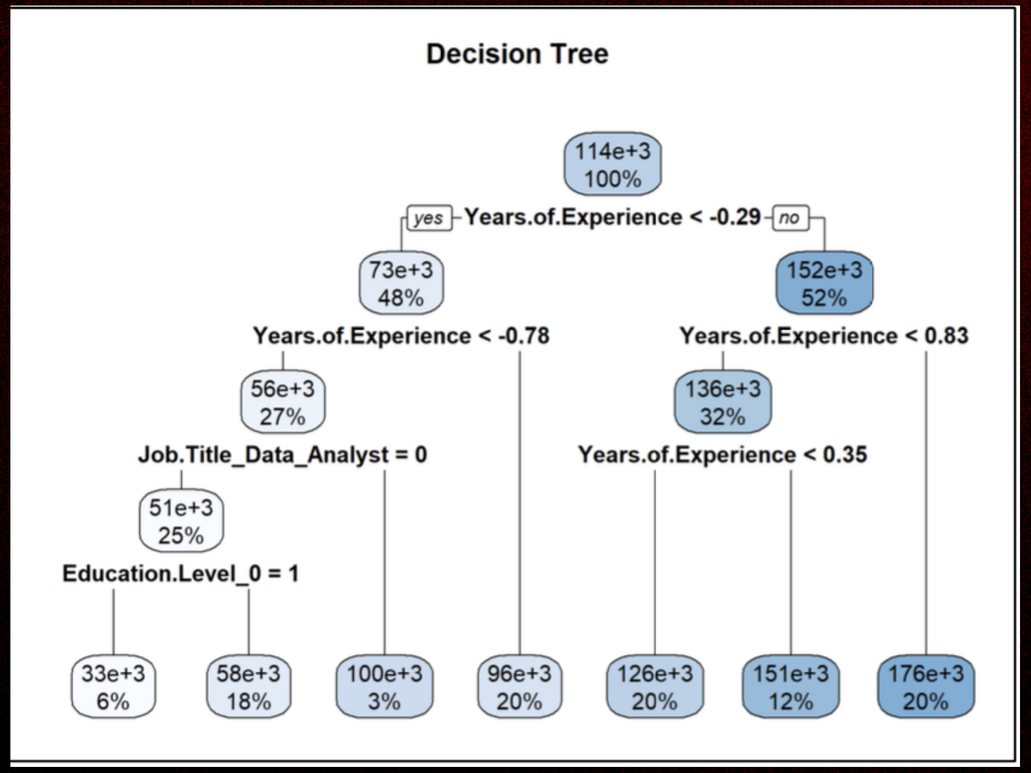
- Not completely different model but optimized version of Linear Regression
- Helps to avoid overfitting



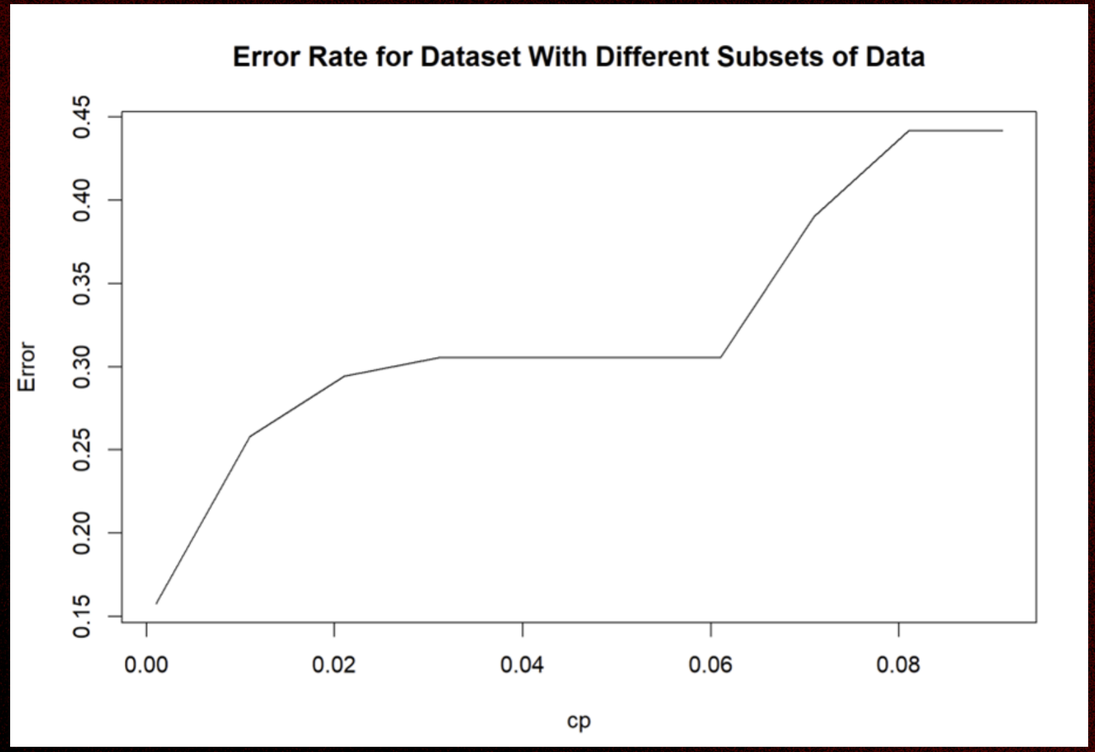
lambda	RMSE	Rsquared	MAE
0.0010	6670.234	0.9876128	5116.734
0.0012	6670.234	0.9876128	5116.734
0.0014	6670.234	0.9876128	5116.734
0.0016	6670.234	0.9876128	5116.734
0.0018	6670.234	0.9876128	5116.734
0.0020	6670.234	0.9876128	5116.734
0.0022	6670.234	0.9876128	5116.734
0.0024	6670.234	0.9876128	5116.734
0.0026	6670.234	0.9876128	5116.734
0.0028	6670.234	0.9876128	5116.734
0.0030	6670.234	0.9876128	5116.734
0.0032	6670.234	0.9876128	5116.734
.		.	
.		.	
.		.	
0.0506	6670.234	0.9876128	5116.734
0.0508	6670.234	0.9876128	5116.734

Decision-Tree

It is used for both classification and regression tasks.



R-Square: 0.7586



After Cross-Validation and Parameter Tuning

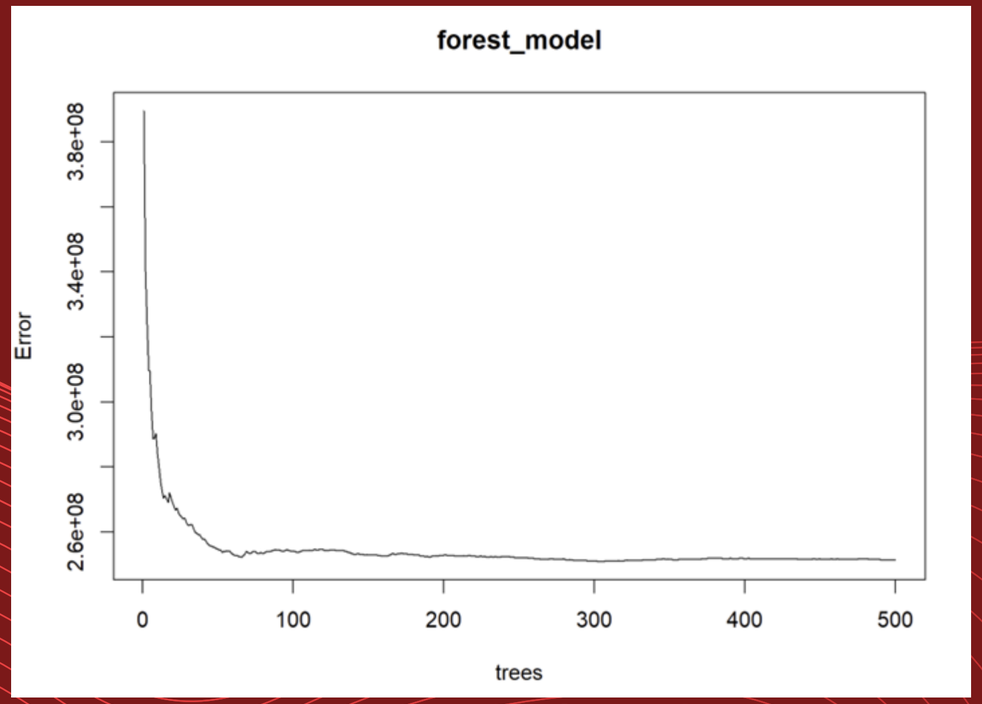
0.8423

Random Forest

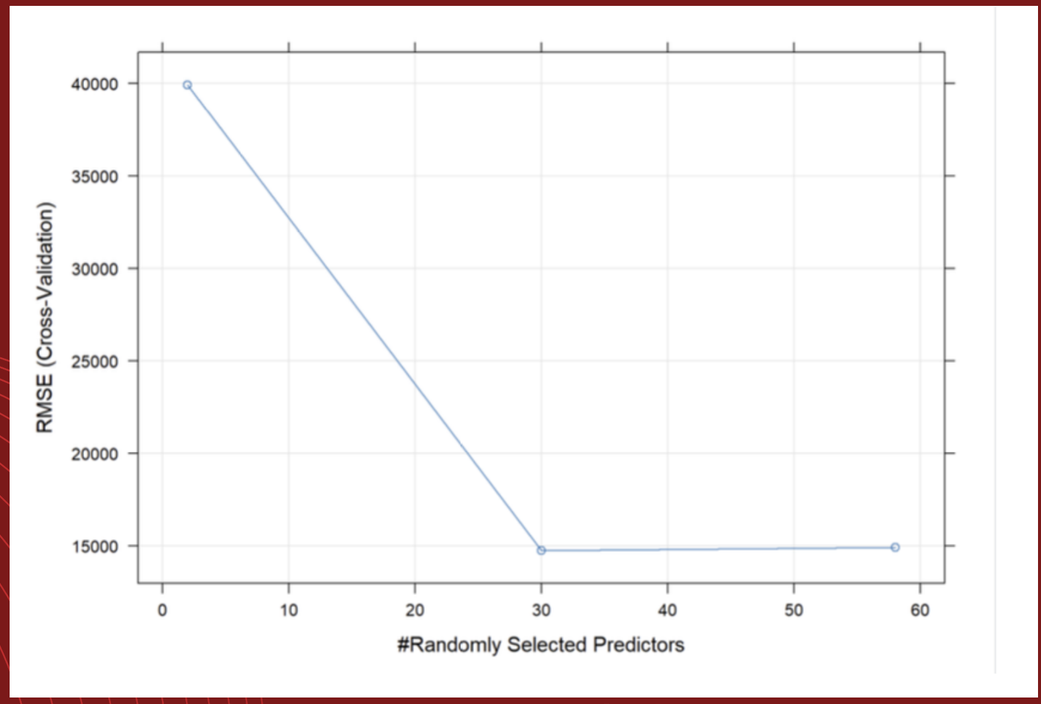
```
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 4120, 4118, 4118, 4118, 4118
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    39926.70  0.6878333  34428.373
  30    14760.77  0.9209926   9040.680
  58    14936.03  0.9191628   8469.611

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 30.
```

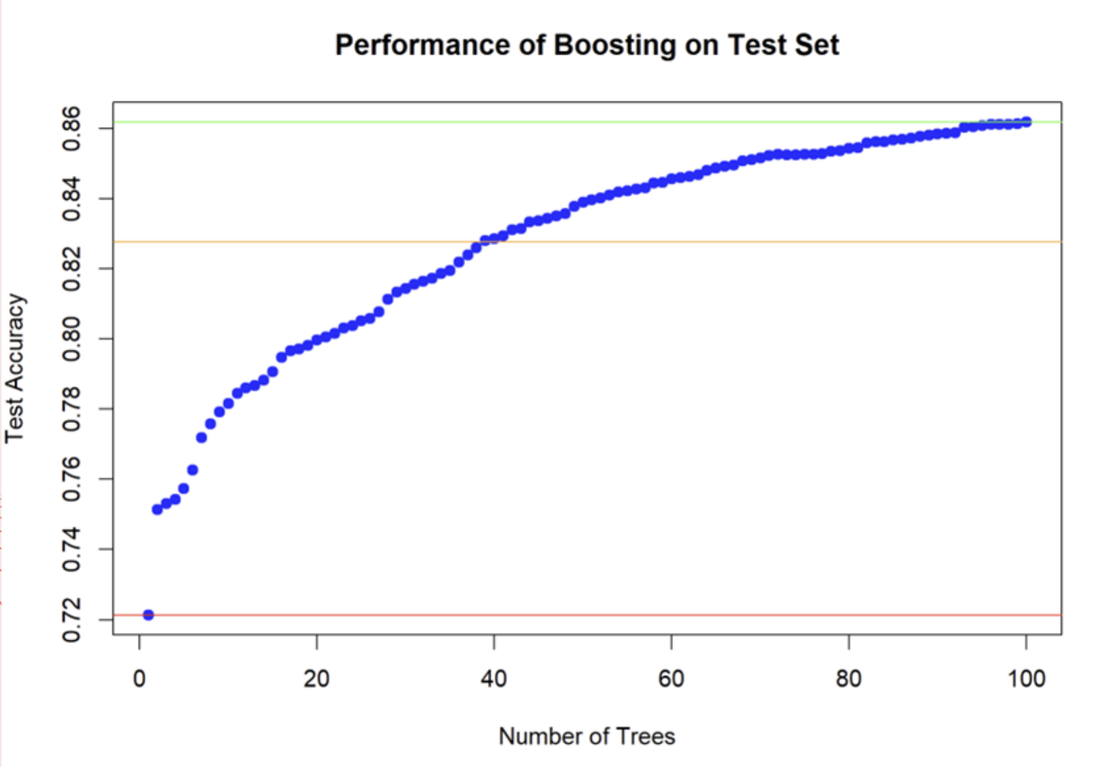


Before Cross-Val and Tuning
0.9292



After Cross-Validation
0.9209

Gradient Boosting




```
Resampling results across tuning parameters:

  interaction.depth  n.trees  RMSE      Rsquared   MAE
  1                   50      25618.31  0.7699290  20413.90
  1                  100      23855.97  0.7963693  18593.72
  1                  150      22979.59  0.8099564  17686.30
  1                  200      22447.51  0.8180469  17112.95
  2                   50      23585.08  0.8018308  18380.80
  2                  100      21736.38  0.8297480  16507.04
  2                  150      20816.13  0.8433217  15520.99
  2                  200      20220.27  0.8519939  14915.87
  3                   50      22357.54  0.8208725  17134.93
  3                  100      20617.75  0.8462423  15311.73
  3                  150      19730.58  0.8590463  14444.50
  3                  200      19233.72  0.8659492  13932.60
  4                   50      21575.72  0.8326400  16325.62
  4                  100      19854.80  0.8573546  14563.12
  4                  150      19046.79  0.8686129  13772.78
  4                  200      18551.30  0.8752625  13291.94

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning parameter 'n.minobsinnode' was held constant at a value of 10
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were n.trees = 200, interaction.depth = 4, shrinkage = 0.1 and n.minobsinnode = 10.
```

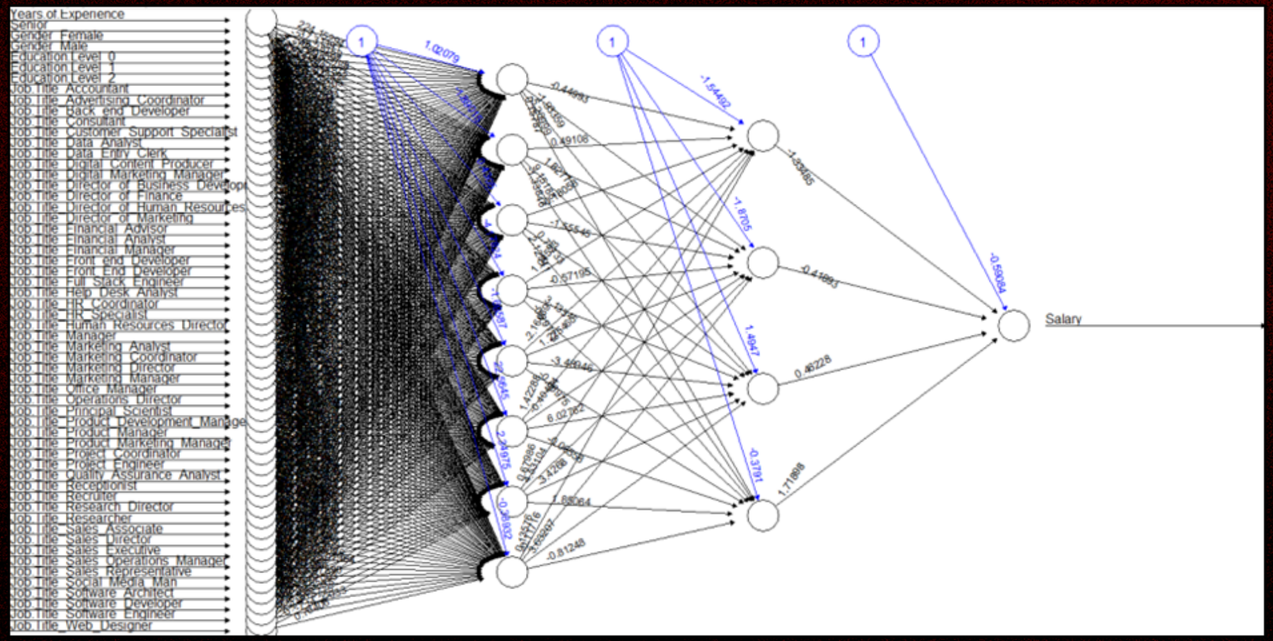
Before Cross-Val and Tuning

0.8661

After Cross-Val and Tuning

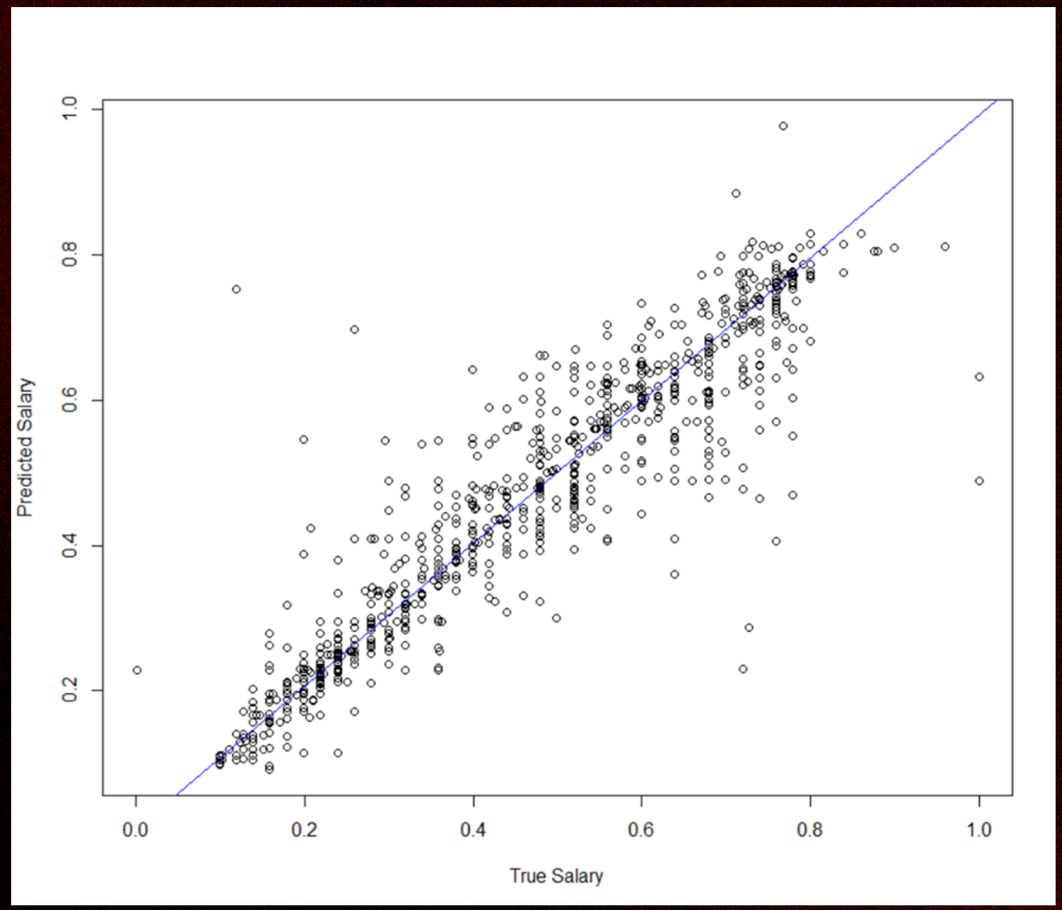
0.8752

Neural Network



SSE.train : 6.131883

SSE.test: 4.030059



Conclusion

The model that gave the most accurate results for our dataset was Ridge Regression

Model Name	SSE	MAE	R^2	RMSE	Runtime
Single Linear Regression	4.732	24314.04	0.693	30318.66	0.006
Multi Linear Regression	2.348	17806.05	0.804	23959.75	0.038
Ridge Regression	1.78811	5116.734	0.987	6670.234	2.133
Decision Tree	1.60836	14724.77	0.872	19827.94	0.186
Random Forest	5.515	9040.680	0.920	14760.77	13.19s
Gradient Boosting	1.322	13291.94	0.875	18551.30	0.5959s
Neural Network First NN	4.085	0.0495	0.881	0.0727	51.339s
Neural Network Second NN	4.030	0.0433	0.897	0.0675	8.452m

Thank you for listening