DATA SCIENCE FOR HUMAN RESOURCES

Predicting Employee Attrition

Image: peaksupport.io

INTRODUCTION

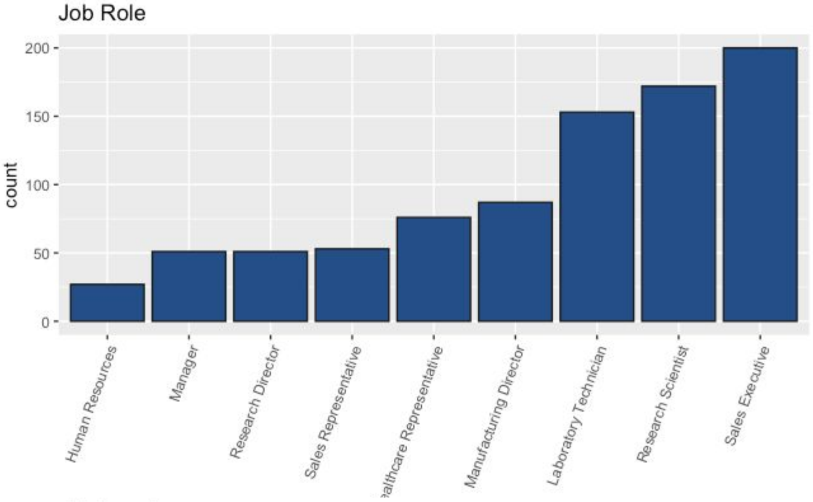
Provided with a dataset containing information about 807 employees:

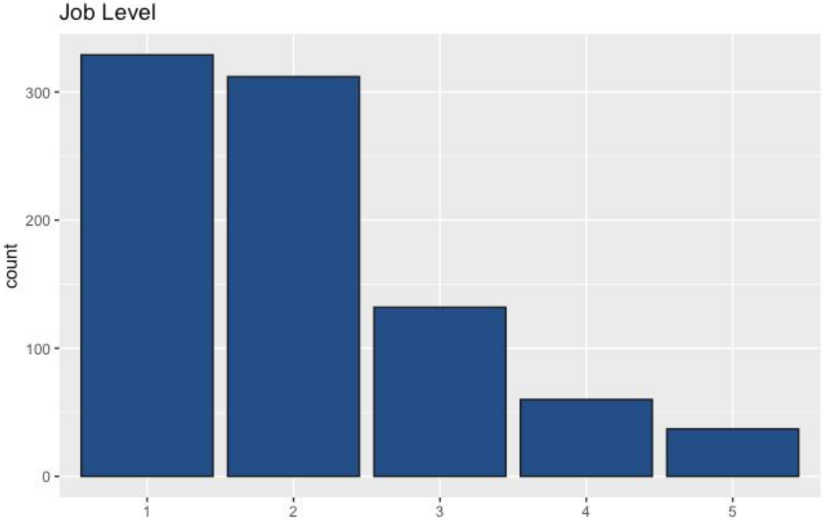
- 35 descriptive features (variables)
- 1 feature indicating attrition (yes/no)
- 9 job roles
- 5 job levels

Performed exploratory data analysis (EDA) to understand features.

Performed “feature selection” - Identified features containing the most relevant “information” related to attrition.

Fit multiple models and compared performance metrics.


Job Role


Job Level

ATTRITION

No	Yes
730	140

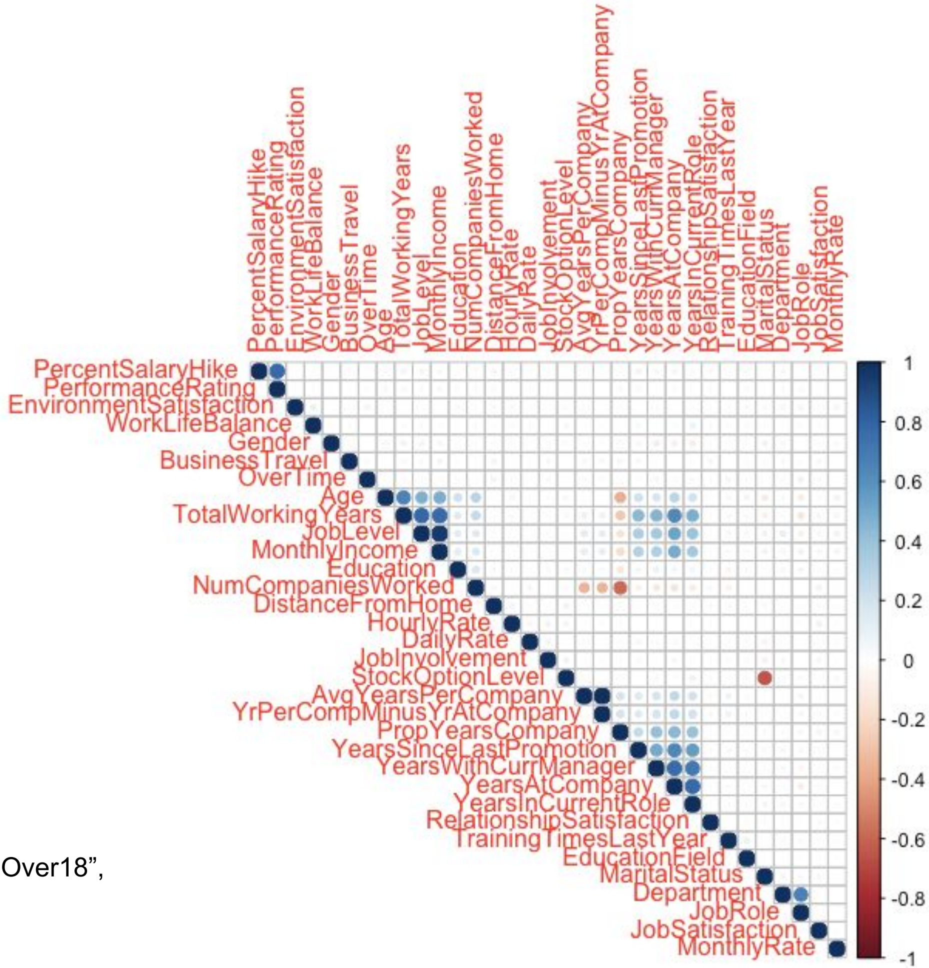
16.1% of employees in dataset quit their job.

We could achieve a classification accuracy of 83.9% just by classifying everyone as “No”

This indicates that accuracy might not be a terribly useful metric...

EDA

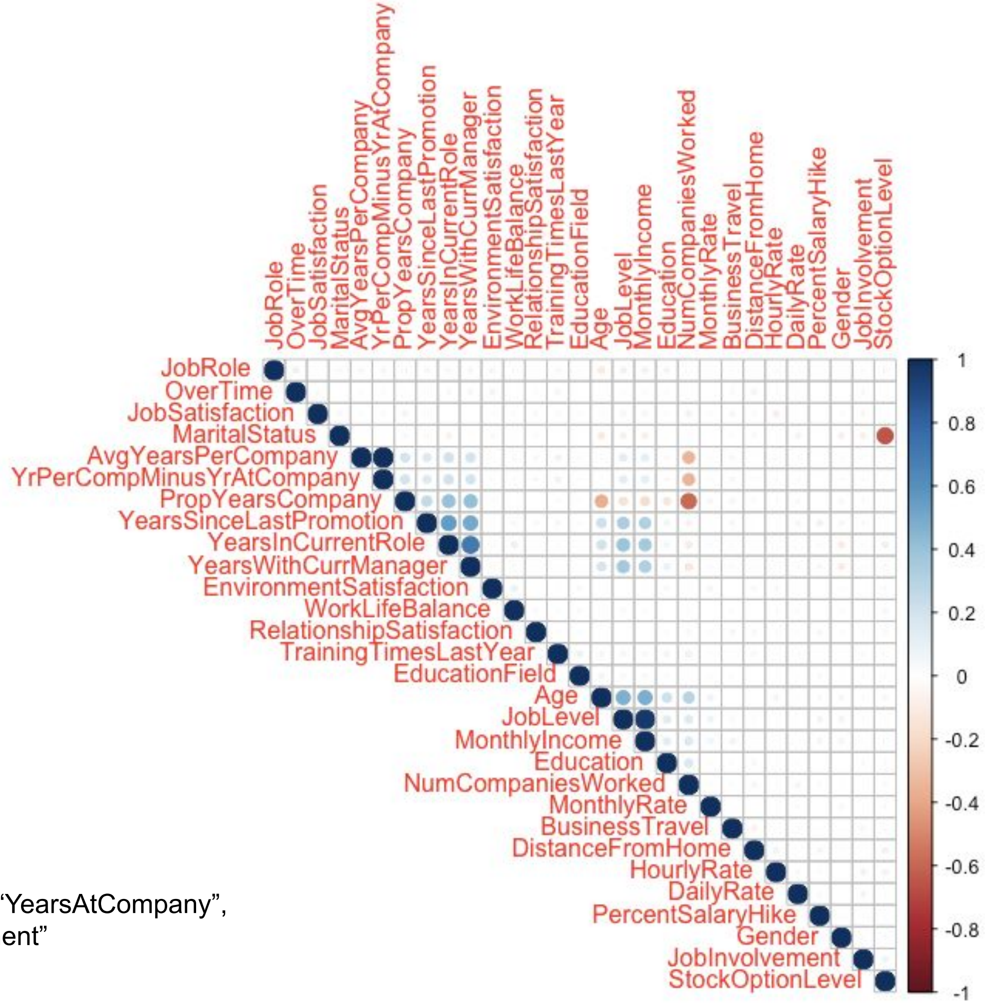
CORRELATION MATRIX



Removed “ID”, “EmployeeNumber”, “Over18”, “StandardHours”, “EmployeeCount

EDA

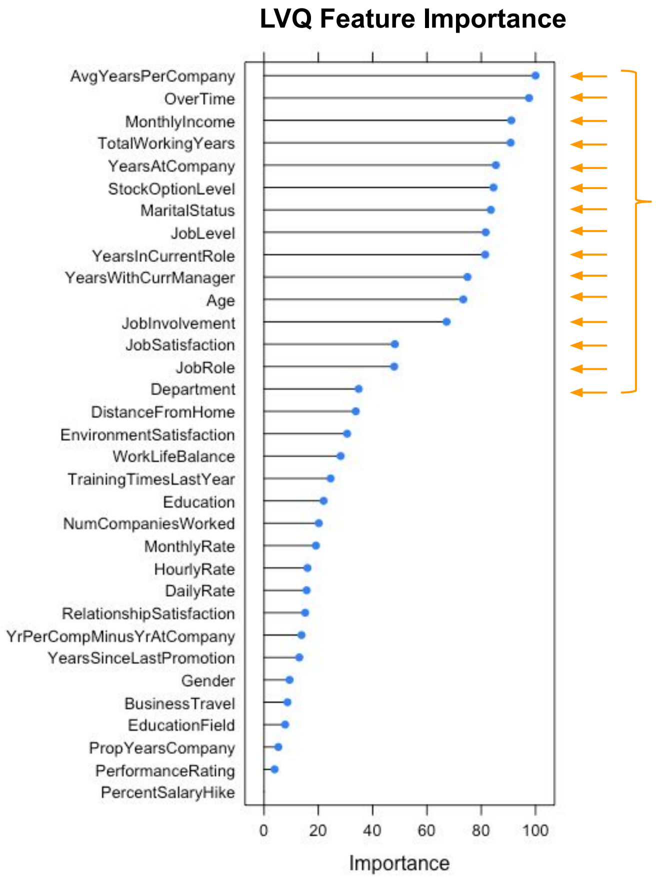
UPDATED CORRELATION MATRIX

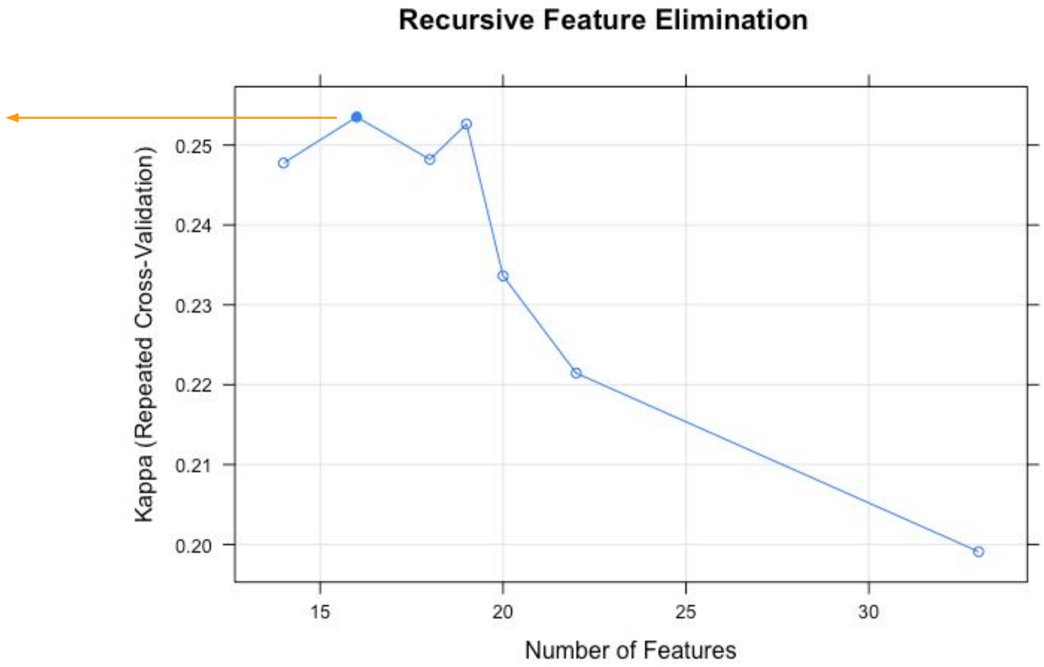


Removed “TotalWorkingYear”, “YearsAtCompany”, “PerformanceRating”, “Department”

EDA

AUTOMATED FEATURE SELECTION





*The 16 features identified here were used to train the following models

PREDICTING ATTRITION

KNN MODEL

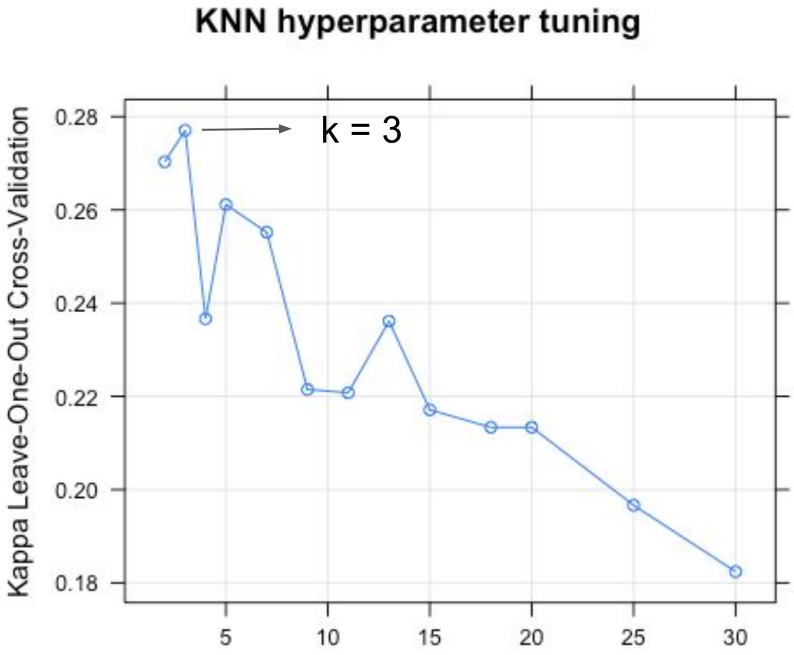
CONFUSION MATRIX

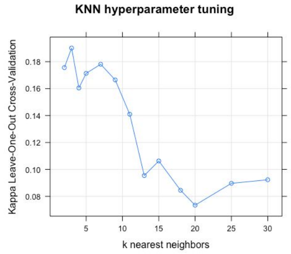
predicted

	No	Yes
No	711	70
Yes	19	70

PERFORMANCE METRICS

METRIC	VALUE	
Overall Accuracy	0.8977	
Sensitivity	0.9740	✓
Specificity	0.5000	✗





PREDICTING ATTRITION

LOGISTIC REGRESSION

CONFUSION MATRIX

		No	Yes
predicted	No	713	73
	Yes	17	67

PERFORMANCE METRICS

METRIC	VALUE	
Overall Accuracy	0.8966	
Sensitivity	0.9767	✓
Specificity	0.4786	✗

PREDICTING ATTRITION

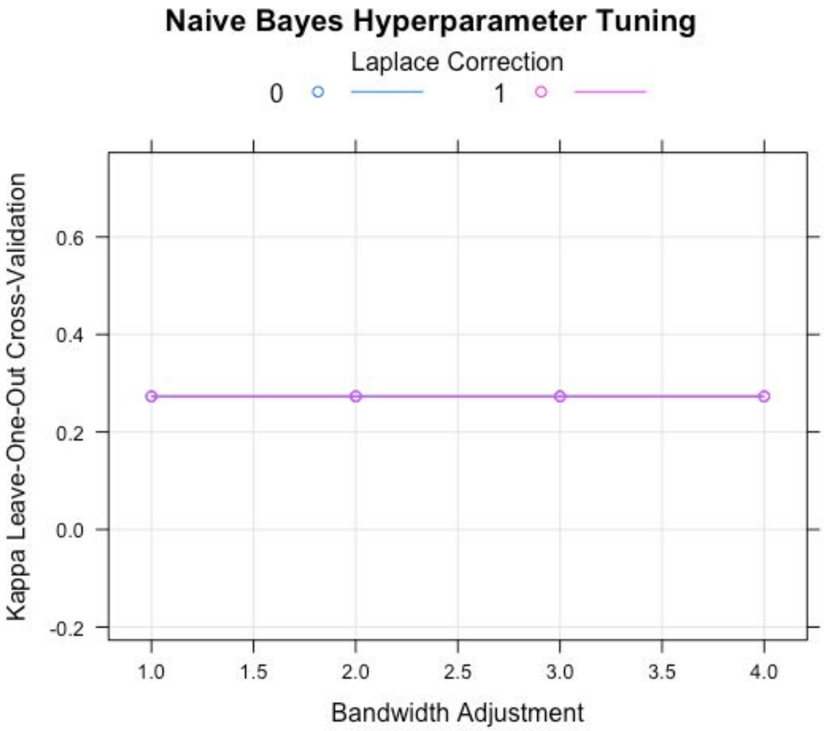
NAIVE BAYES MODEL

CONFUSION MATRIX

		No	Yes
predicted	No	520	38
	Yes	210	102

PERFORMANCE METRICS

METRIC	VALUE	
Overall Accuracy	0.7149	
Sensitivity	0.7123	✓
Specificity	0.7286	✓



PREDICTING ATTRITION

BOOSTED CLASSIFICATION TREE

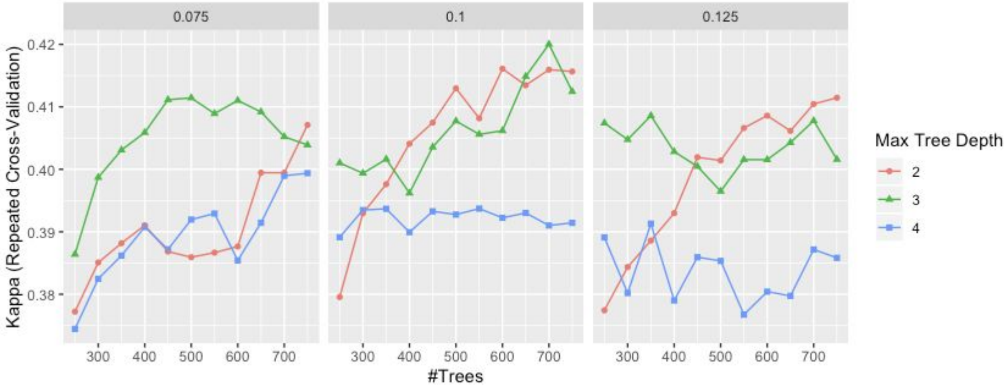
CONFUSION MATRIX

		No	Yes
predicted	No	725	46
	Yes	5	94

PERFORMANCE METRICS

METRIC	VALUE	
Overall Accuracy	0.9414	
Sensitivity	0.9932	✓
Specificity	0.6714	✓

HYPERPARAMETER TUNING



PREDICTING ATTRITION

COMPARING FINAL MODELS

NAIVE BAYES

CONFUSION MATRIX

		No	Yes
predicted	No	520	38
	Yes	210	102

PERFORMANCE METRICS

METRIC	VALUE
Overall Accuracy	0.7149
Sensitivity	0.7123
Specificity	0.7286

BOOSTED CLASSIFICATION TREE

CONFUSION MATRIX

		No	Yes
predicted	No	725	46
	Yes	5	94

PERFORMANCE METRICS

METRIC	VALUE
Overall Accuracy	0.9414
Sensitivity	0.9932
Specificity	0.6714

PROFILE OF A QUITTER

LIKELY ATTRIBUTES

A male who:

- Is a sales representative
- Works overtime
- Is single
- Spends an average of 2 years at each company he has worked
- Has low work-life balance
- Has low job satisfaction
- Is early in his career

DATA SCIENCE FOR HUMAN RESOURCES

Predicting Salary

Image: peaksupport.io

MODEL SELECTION

FULL MODEL VS. BACKWARD/FORWARD/STEPWISE SELECTION

STEP 1:

SELECT INITIAL MODEL (10-FOLD CV)

MODEL	MEAN RMSE
Full	$1081.40
Backward	$1055.92
Forward	$1055.92
Stepwise**	$1055.20

STEP 2:

SELECT FINAL MODEL (10-FOLD CV)

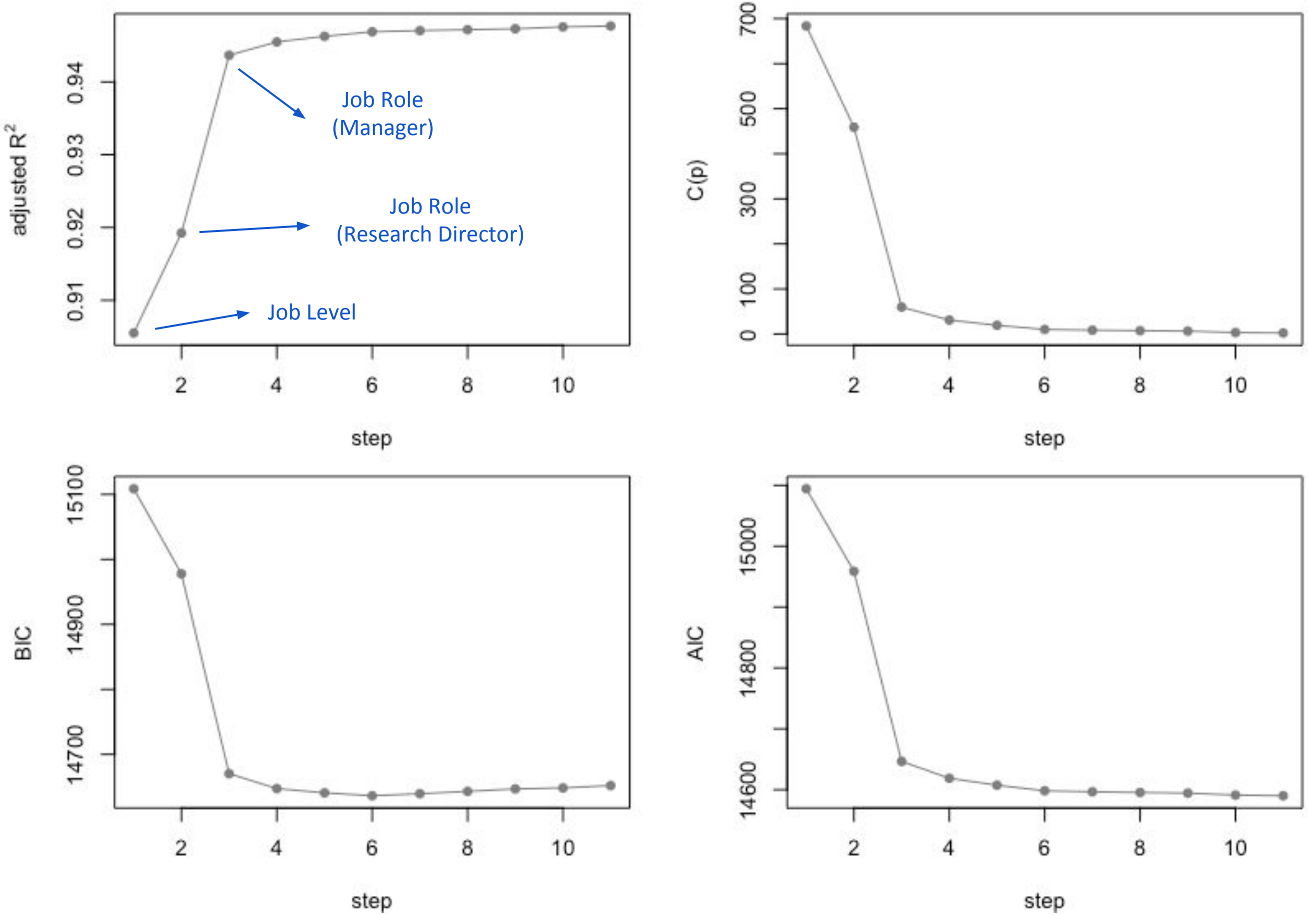
MODEL	MEAN RMSE
Stepwise**	$1055.20
w/o Gender	$1055.54
w/o Distance From Home	$1055.44
w/o Either	$1055.83

SELECTED FEATURES

VARIABLE	ESTIMATE	P-VALUE	VIF
INTERCEPT	6390.26	< 0.0001	NA
Job Level	3196.96	< 0.0001	3.91
Res. Director	939.2	< 0.0001	1.54
Manager	924.76	< 0.0001	1.76
Working Years	255.61	< 0.001	3.45
Lab Tech	-127.58	< 0.001	1.16
Travel Rarely	117.27	< 0.01	1.01
Man. Director	80.11	< 0.05	1.13
Monthly Rate	-63.88	0.075	1.01
Proportion Yrs Company	-103.65	< 0.05	1.37
Yrs Since Promotion	105.11	< 0.05	1.60
Daily Rate	63.48	0.077	1.01
⚠ Gender - F	-56.68	0.11	1.02
Distance From Home	-54.96	0.13	1.01

MODEL SELECTION

STEPWISE MODEL FIT STATISTICS



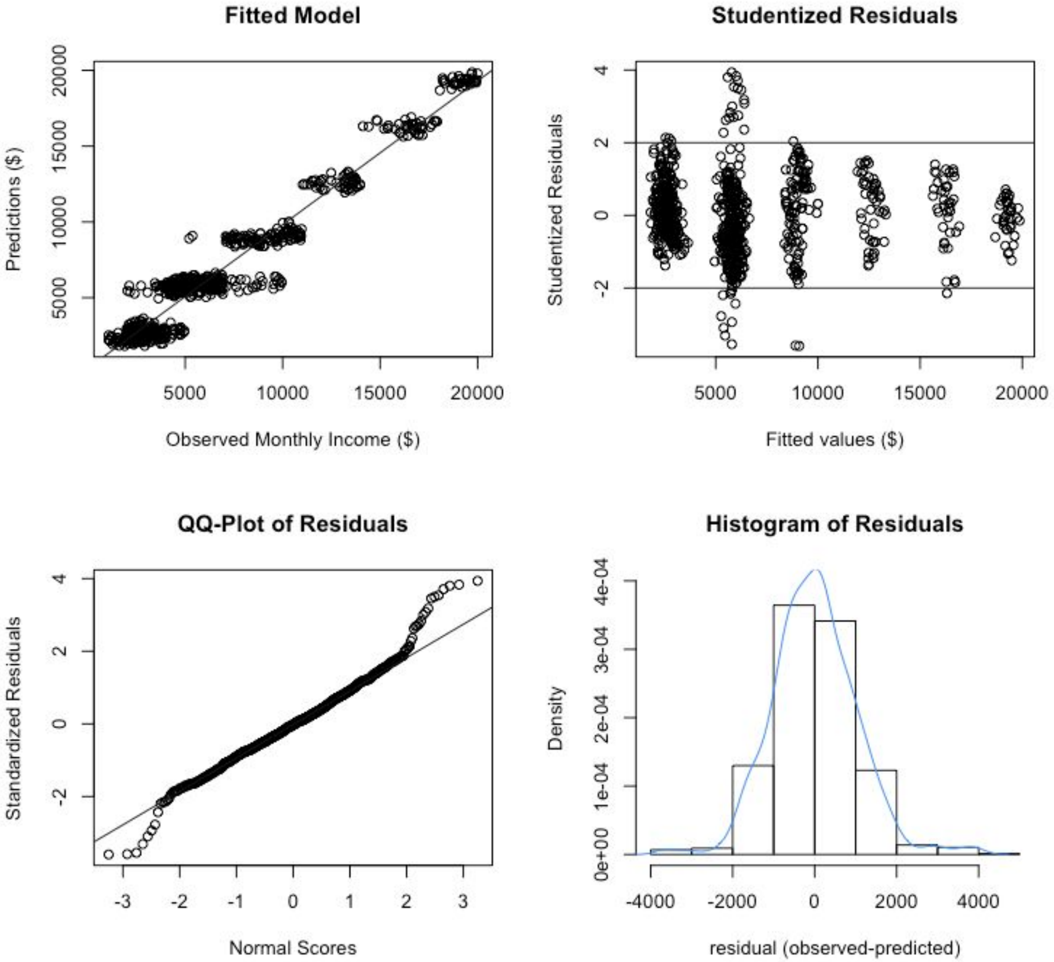
Job Role (Manager)

Job Role (Research Director)

Job Level

FINAL MODEL

DIAGNOSTICS



Fitted Model

Studentized Residuals

QQ-Plot of Residuals

Histogram of Residuals