Predicting Salary

Using Random Forest Regression

Sam Hopkins

Introduction

- This model uses regression to predict salaries
- Key questions:
  - How do gender, education level, time in the workforce, and age affect salary?
  - Which groups tend to make more?
  - To what extent can these things be used to predict salary? Is any one predictor more significant than the others?
- This project is important because it will inform us how various factors may aid or inhibit one's ability to earn a certain salary, revealing pay gaps between different groups.

State of the Art

- Many attempts have already been made to predict salaries. R^2 values above 0.90 are typical for well-thought-out models.
- https://www.kaggle.com/code/yosefibrahim/salary-prediction
  - This notebook uses linear regression, decision trees, and random forests to predict salary.
- https://www.kaggle.com/code/aqua55s/salary-prediction
  - This notebook uses linear regression to predict salary, introducing ridge regression at the end.

Materials and Methods

Dataset and Pre-Processing

https://www.kaggle.com/datasets/mohithsairamreddy/salary-data/code

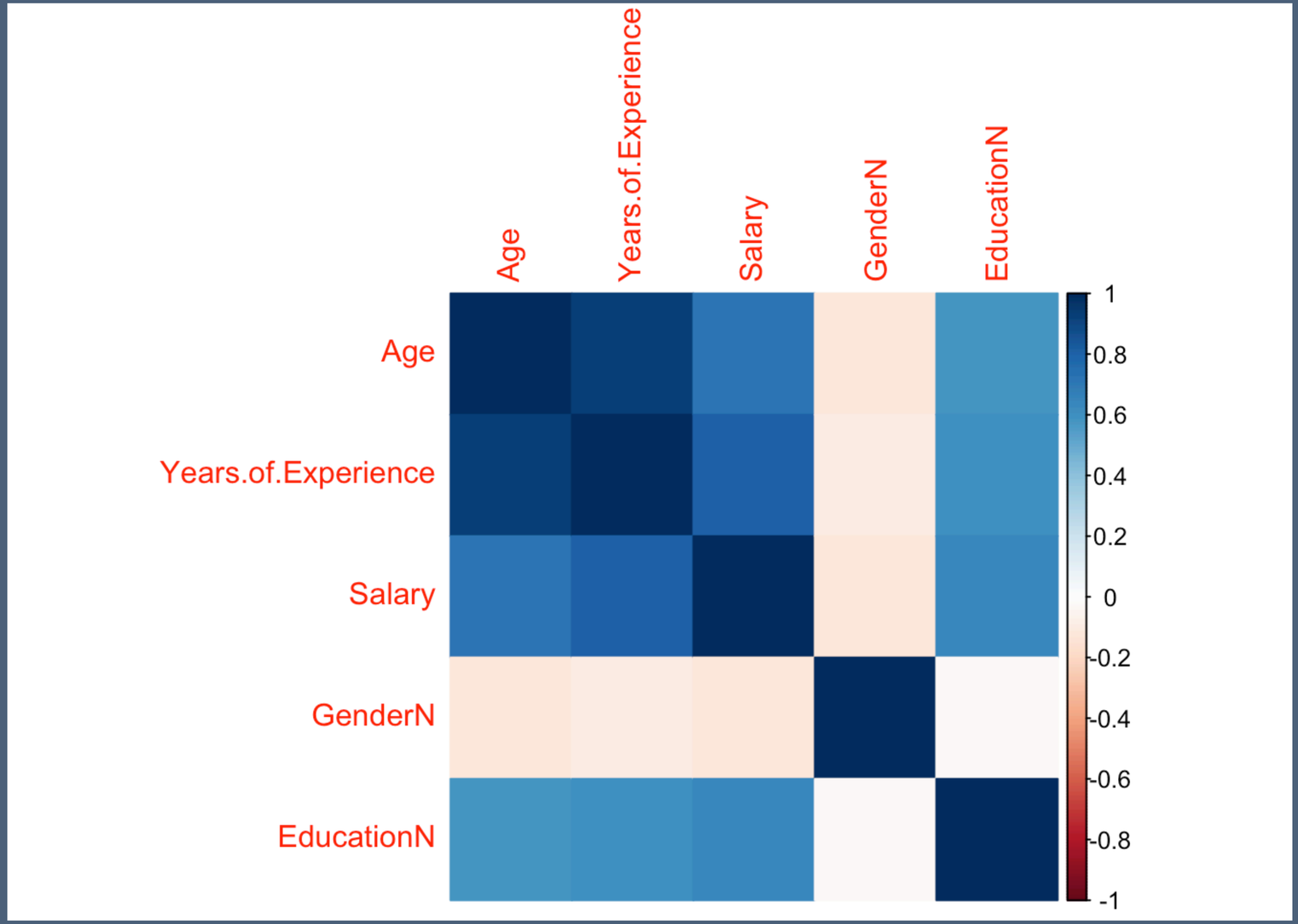
- Compiled from publicly available sources: surveys, job posting sites.
- Contains salary, education level, age, number of years of experience, gender, and job title of 6704 individuals. Last updated 1 year ago.
- Removed 10 NA values using na.omit().
- Eight values for education level: phD, PhD, Bachelor's, Bachelor's Degree, Master's, Master's Degree, High school, and blank/none. Consolidate to 0-4 with like types combined.
- Assign 0 or 1 to gender. Convert new columns GenderN and EducationN to factors using as.factor().

Further Pre-Processing — Interaction Terms

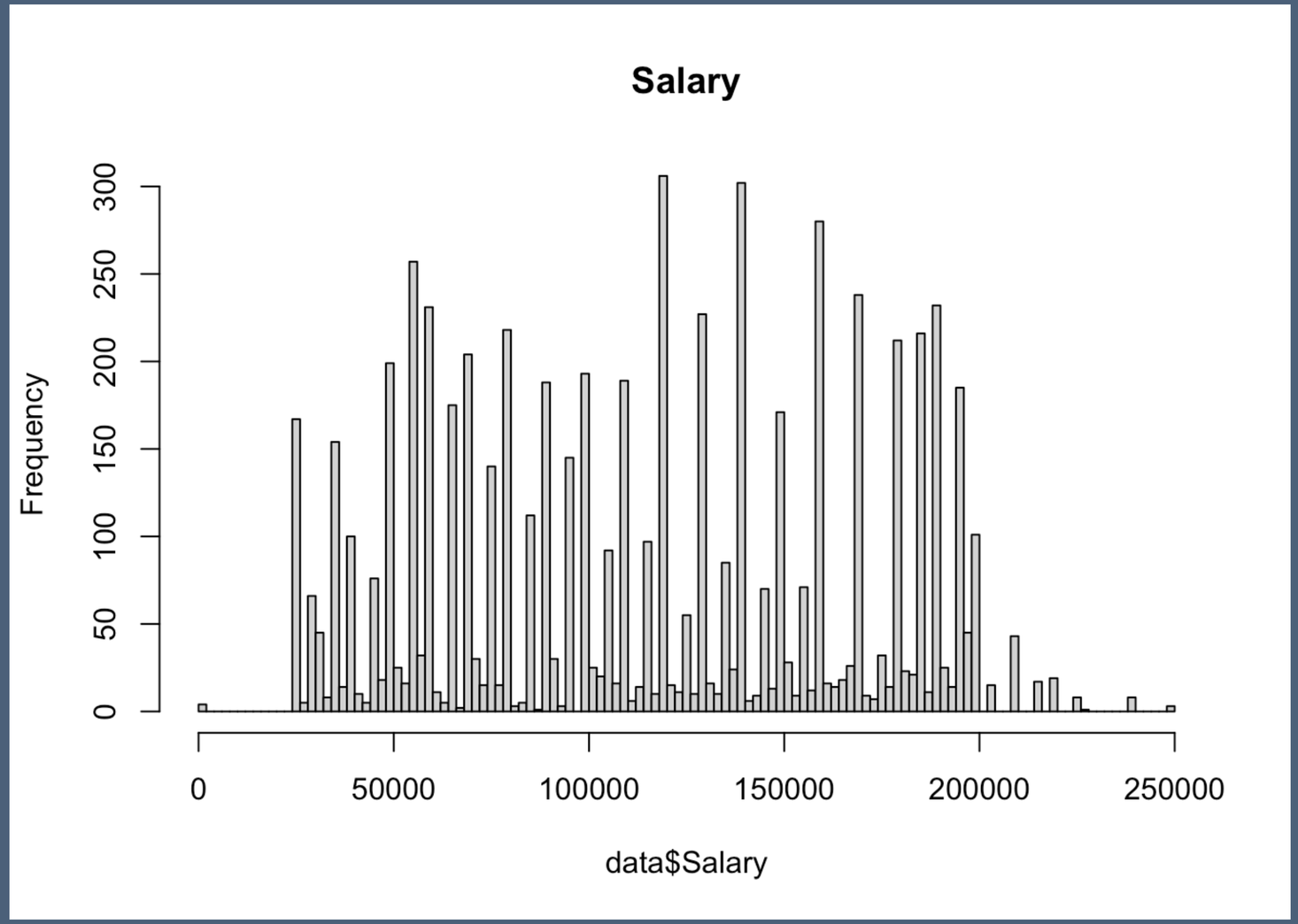
- Tried a few interaction terms — none had a significant impact on results
  - data$rtAge <- log(data$Age)
  - data$rtYears <- (data$Years.of.Experience)^0.25
- Attempts to address non-linear relationships (age and salary, years exp. and salary)
  - data$eduYears <- log(as.numeric(data$EducationN) + mean(data$Years.of.Experience)/2) * data$Years.of.Experience
- Attempts to address negative effect on years of experience of higher education level (more education = less time in the workforce)

Correlations, Distributions and Relationships

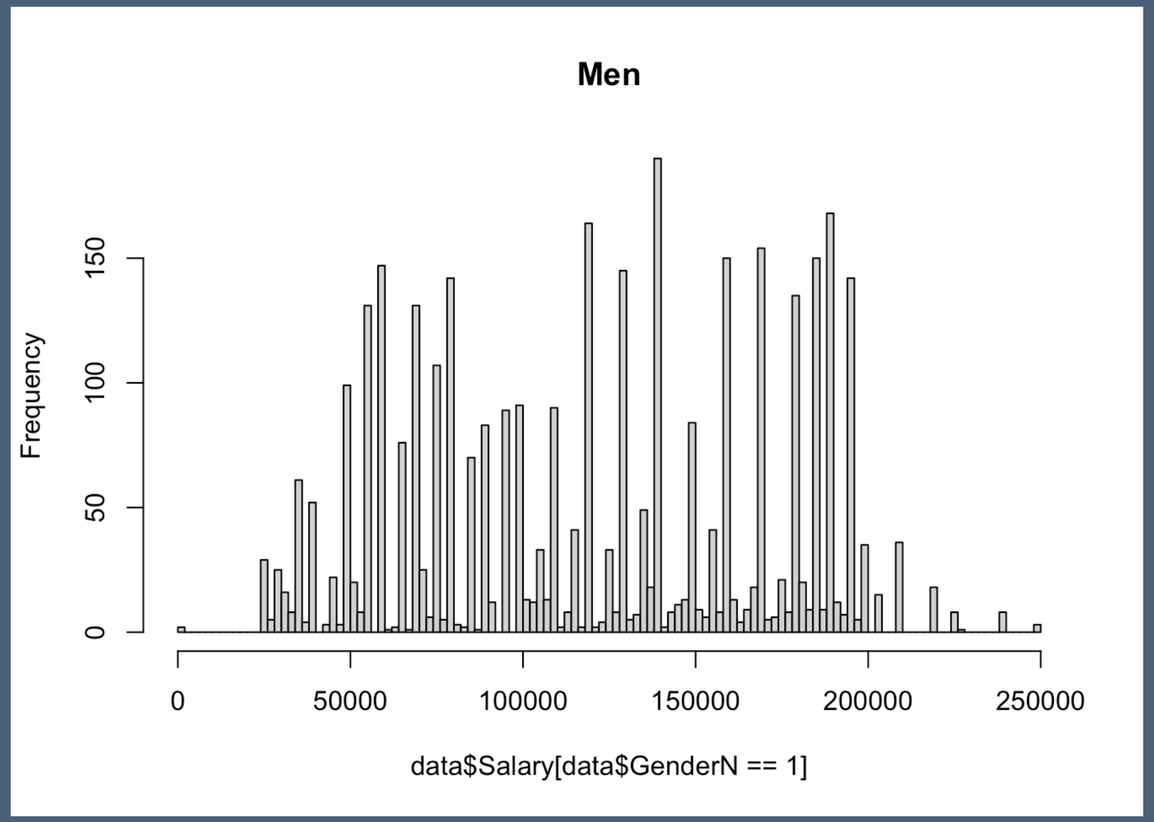
Correlation Matrix

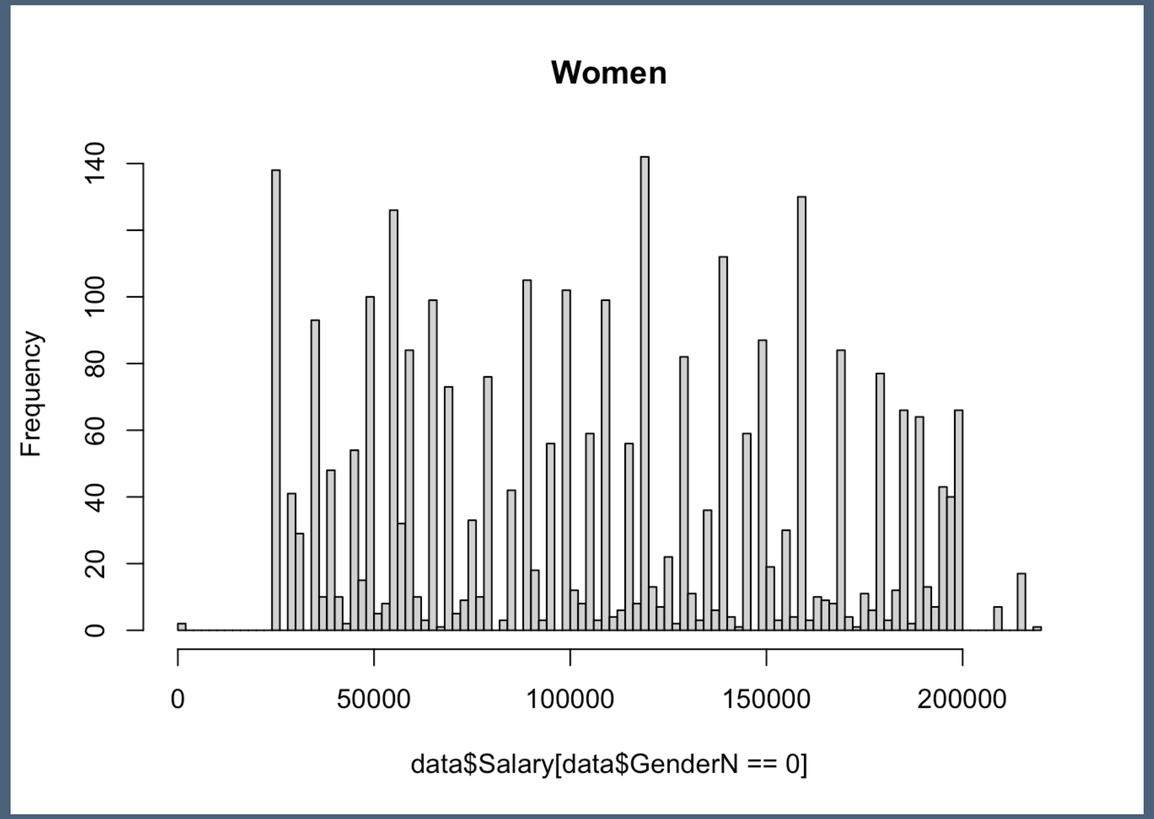


Full Distribution

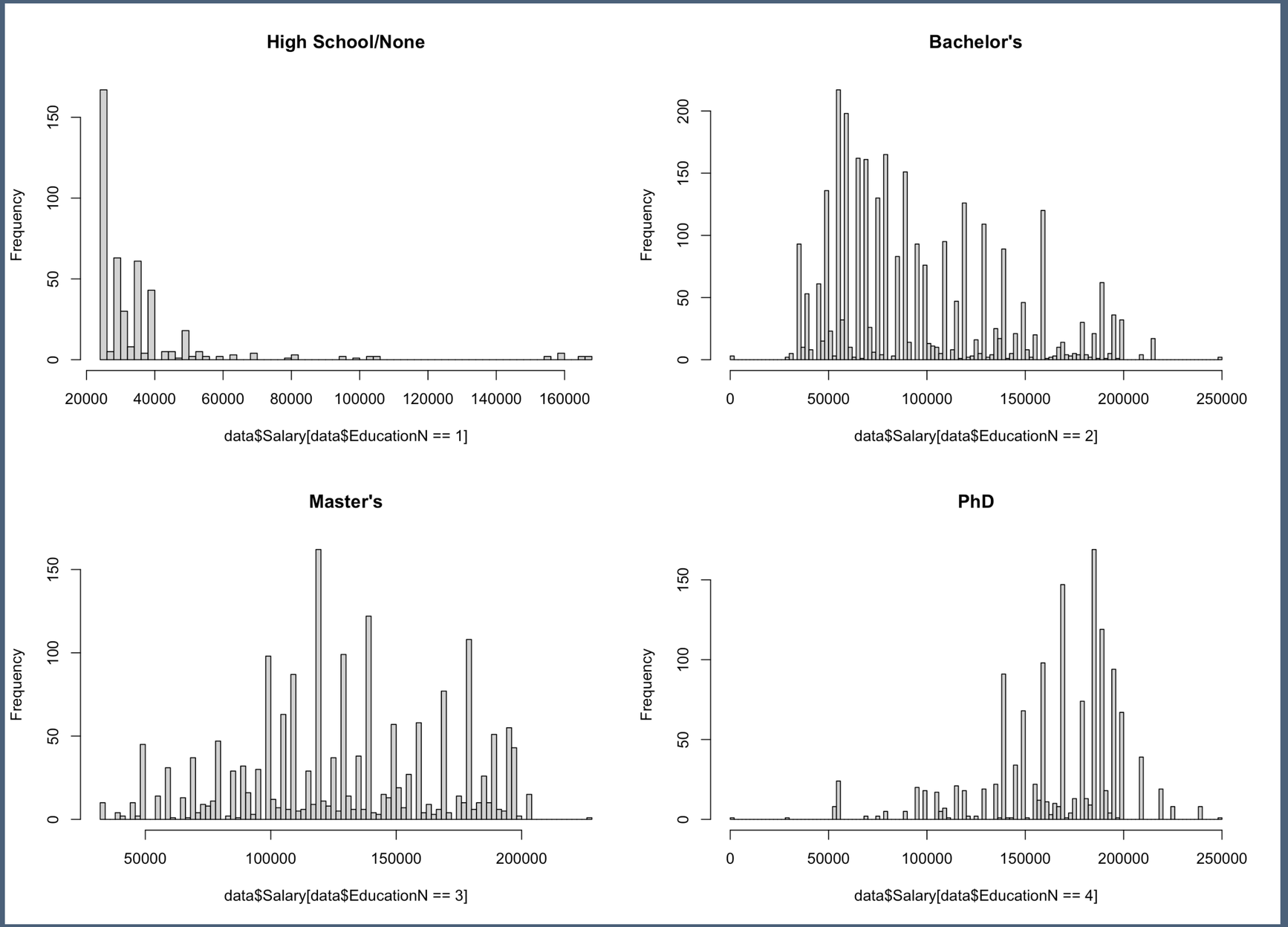


Gender


Men
Frequency
0
50
100
150
0
50000
100000
150000
200000
250000
data$Salary[data$GenderN == 1]


Women
Frequency
0
20
40
60
80
100
140
0
50000
100000
150000
200000
data$Salary[data$GenderN == 0]

Education Level



Age and Years of Experience

Methods and Evaluation

- Random forest is clearly best model. Also tried:
  - General linear model
  - Ridge Regression
  - Decision Tree
- Evaluating using R^2 and RSE
- Dataset is split 80/20 at random indices each time program is run
  - 5-fold cross validation is used for each of the models, using random indices to select each fold

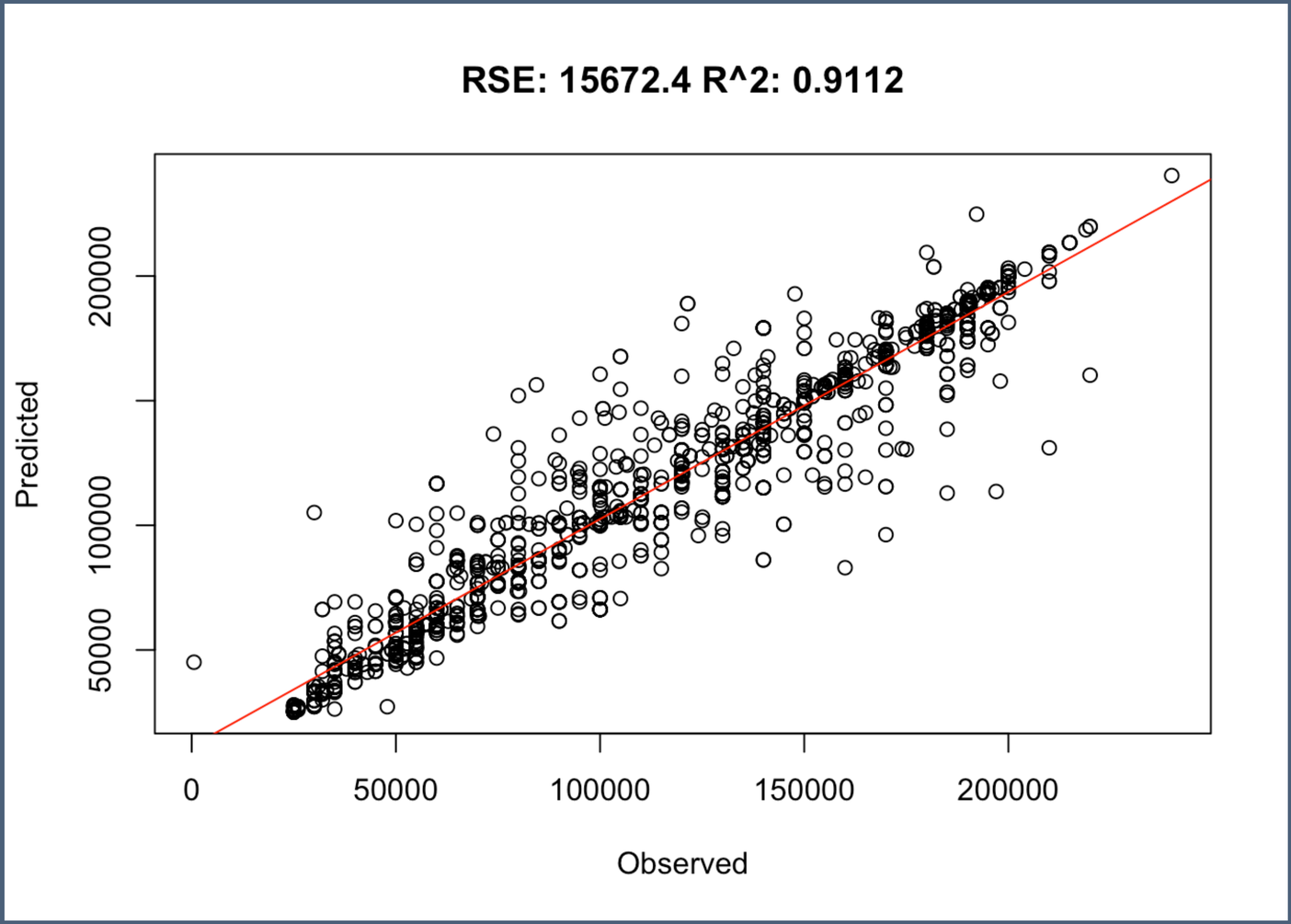
Results

Comparison Table for Models Tested

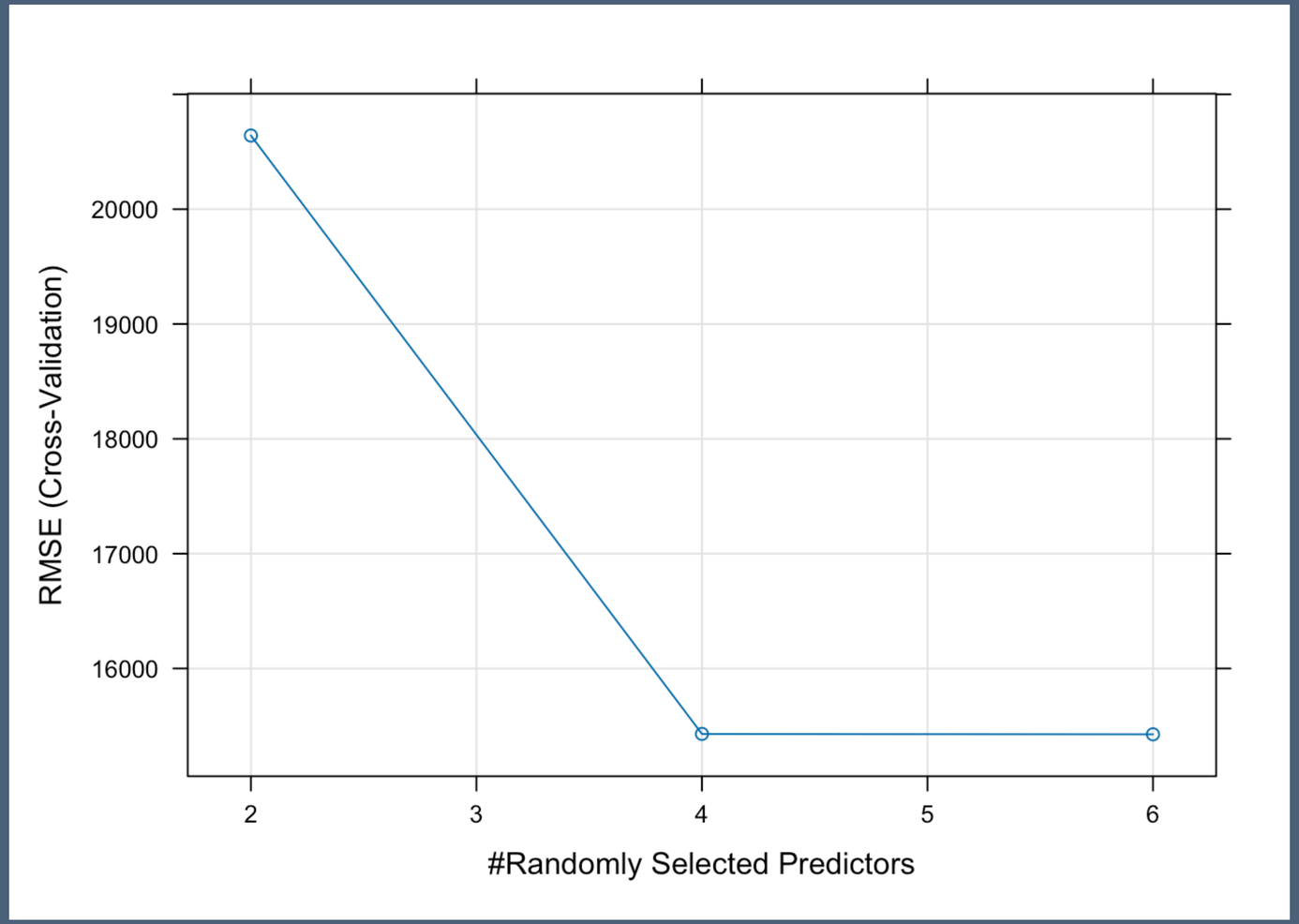
		Model Performance	
methods	forms	R2s	RSEs
General Linear Model	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7241793	27698.92
Ridge Regression	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7131128	28175.24
Decision Tree	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7530028	26143.16
Random Forest	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7949237	23821.54
Random Forest	Salary~rtAge+Years.of.Experience+EducationN+GenderN	0.7964567	23732.34
Random Forest	Salary~Age+rtExp+EducationN+GenderN	0.7986722	23602.82
Random Forest	Salary~Age+Years.of.Experience+eduExp+GenderN	0.8429496	20846.43
CV Random Forest	Salary~Age+Years.of.Experience+eduExp+GenderN	0.9115498	15644.49
CV Random Forest	Salary~Age+Years.of.Experience+GenderN+EducationN	0.9120391	15601.16
CV GLM	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7241793	27626.48
CV Decision Tree	Salary~Age+Years.of.Experience+GenderN+EducationN	0.6832623	29604.79
CV Ridge Regression	Salary~Age+Years.of.Experience+GenderN+EducationN	0.7243554	27617.66

Final Model

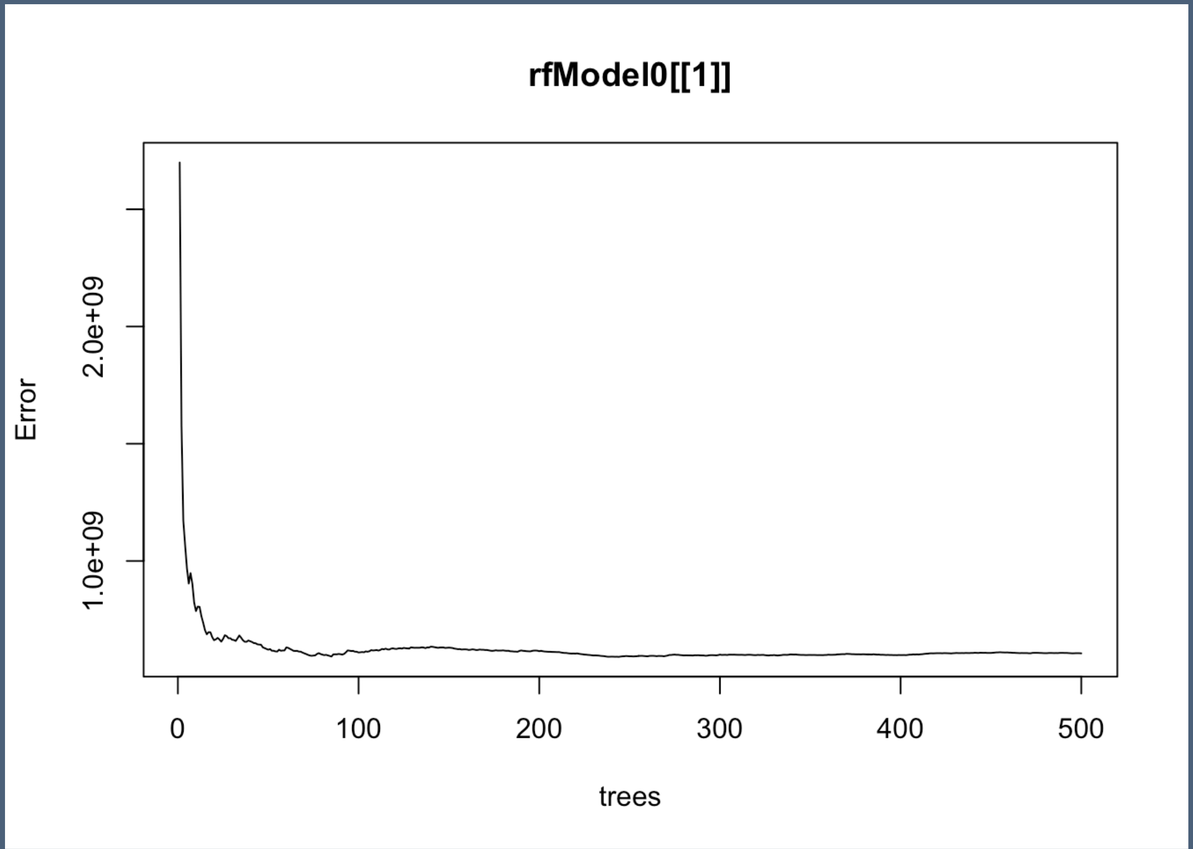
Best model: cross-validated random forest using Age, Years.Of.Experience, EducationN, and GenderN

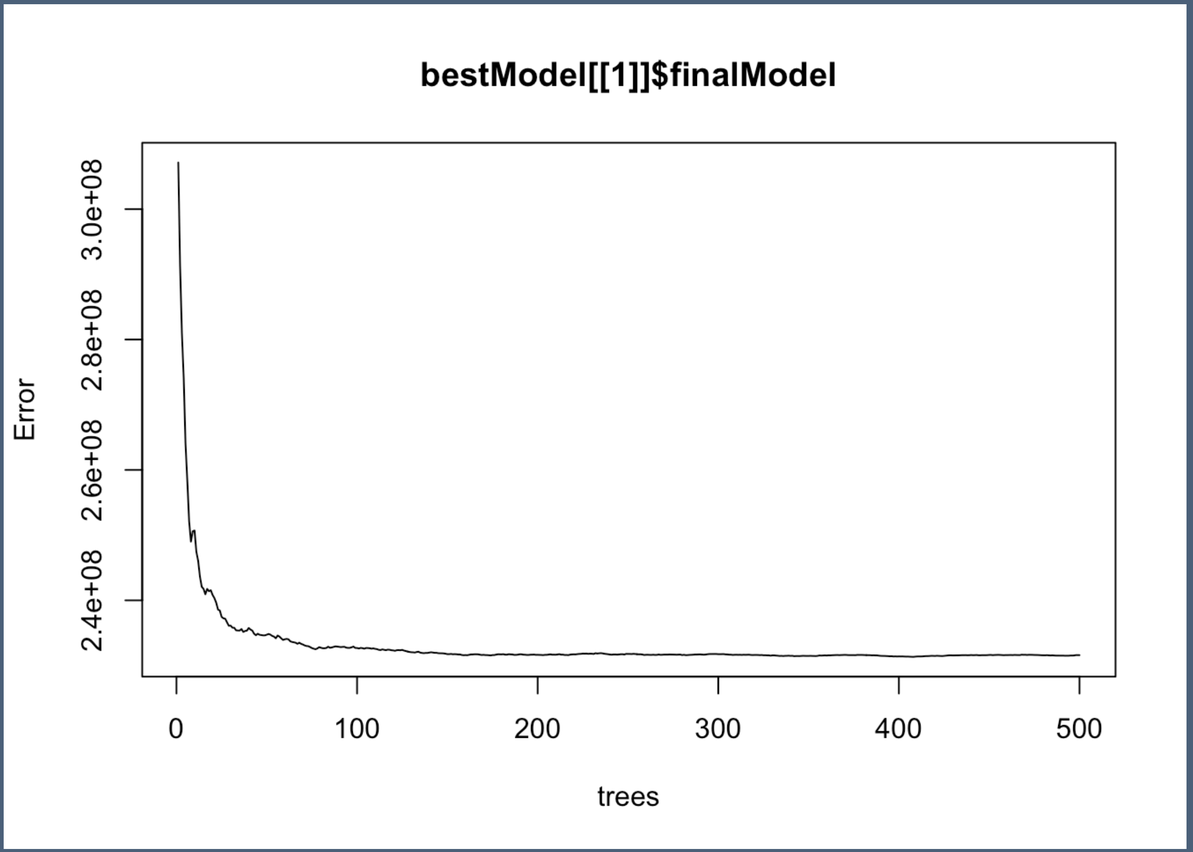


Random Forest CV — #Predictors vs RMSE



#Trees vs Error: Random Forests





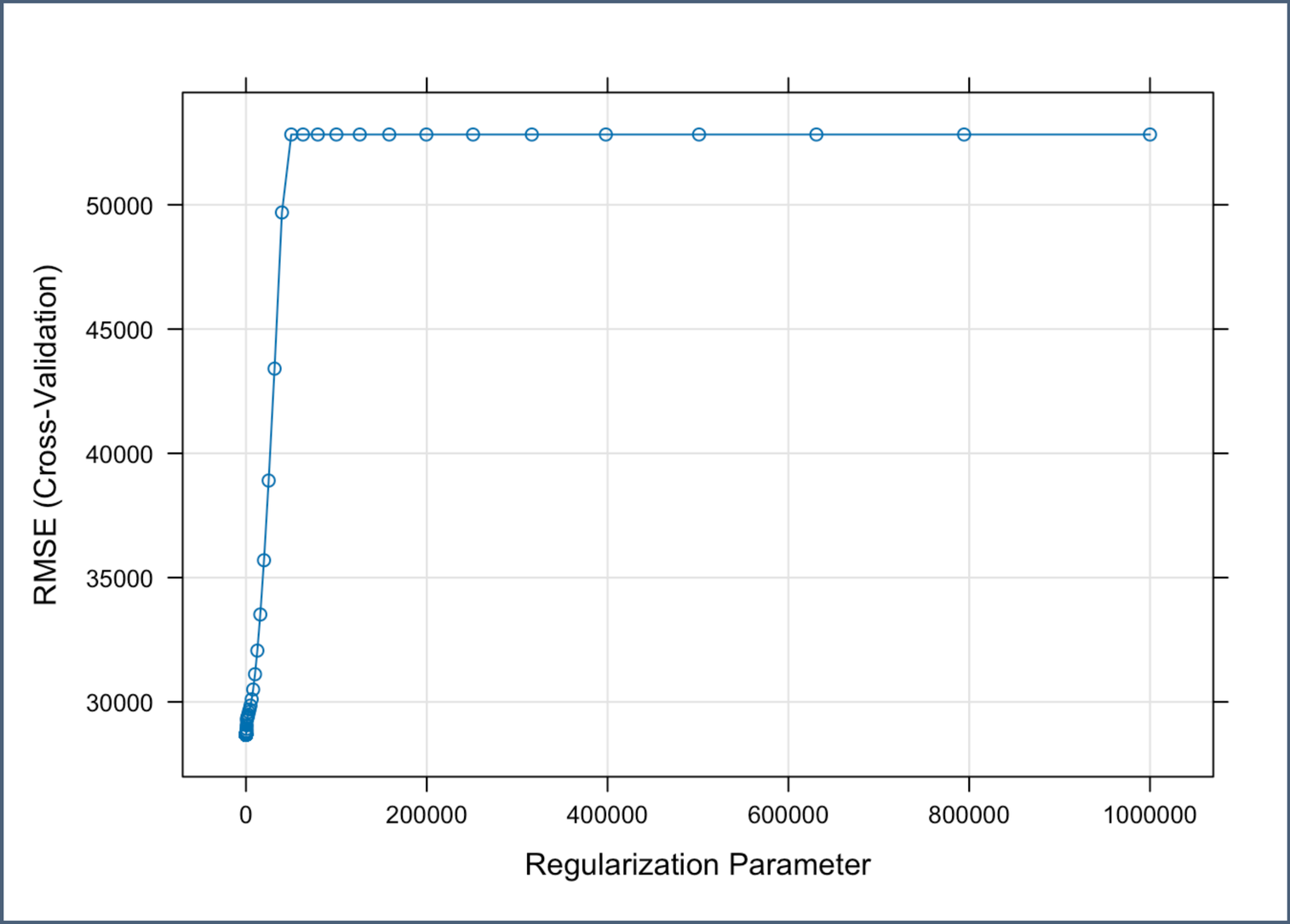
Coefficients: Linear Models

- Cross-validation does not affect the general linear model

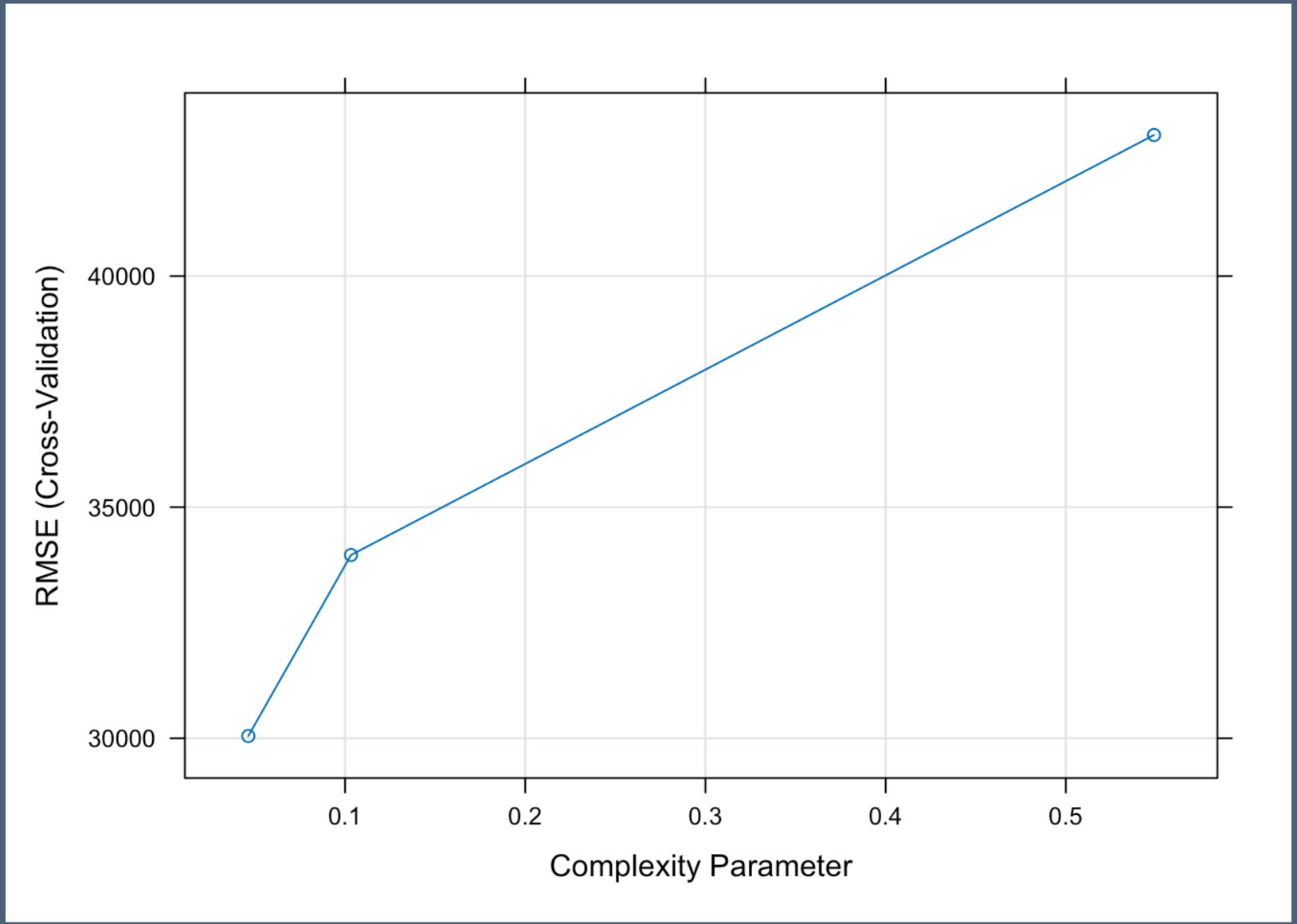
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         74500.6     3920.2  19.004  < 2e-16 ***
Age                 -2080.5      147.7 -14.082  < 2e-16 ***
Years.of.Experience  8137.7      187.1  43.483  < 2e-16 ***
GenderN1             5875.9      794.9   7.392 1.67e-13 ***
EducationN2         36080.7     1635.9  22.055  < 2e-16 ***
EducationN3         47861.0     1784.0  26.828  < 2e-16 ***
EducationN4         60075.8     1972.9  30.451  < 2e-16 ***
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         74500.6     3920.2  19.004  < 2e-16 ***
Age                 -2080.5      147.7 -14.082  < 2e-16 ***
Years.of.Experience  8137.7      187.1  43.483  < 2e-16 ***
GenderN1             5875.9      794.9   7.392 1.67e-13 ***
EducationN2         36080.7     1635.9  22.055  < 2e-16 ***
EducationN3         47861.0     1784.0  26.828  < 2e-16 ***
EducationN4         60075.8     1972.9  30.451  < 2e-16 ***
```

Ridge Regression CV: Lambda vs RMSE



Decision Tree CV: Complexity vs RMSE



Observations

- Error decreases with more trees, with more predictors in RF (to a point).
- Error increases with complexity the decision tree and penalty in ridge regression. This dataset and/or problem seem prone to overfitting; the random forest model is working well because it reduces overfitting by generalizing.
- Linear model shows all predictors are equally significant
- Linear model does not improve with cross validation; is already optimal with one fit, likely due to the size of the dataset or stability of the model.

Conclusions and Future Work

- Model is very good at predicting someone's salary based on their age, years of experience, education level, and gender, with an R^2 of 0.91 and MSE of 15672.4.
- These things, when taken together, explain 91% of the variance in someone’s salary, with an average prediction error of $15,672.40.
- This dataset only contains salaries up to $250K. While that is high, the size of the dataset is 6704, and there should be a handful people with higher salaries than that in a randomly selected sample of American citizens. It would be interesting to try to create a similar kind of model using a dataset containing a wider range of salaries.
- This dataset includes a job title column, but it is too varied and too vague in its specifications to fit it into this problem. Looking at a dataset including the sector in which someone works and the level of their position in a more interpretable way could significantly improve the performance, since jobs in certain sectors, and higher ranking jobs, tend to pay more.
- The results could also be different with a larger dataset containing more variables.