HR Analytics:
Can We Predict Attrition?

Predictive Modeling

August 16, 2019

Chris Henson | LaShay Fontenot
Rawini Dias | Chris Fitzgerald | Yikang Wang

Agenda

- Data Summary
- Overview
- Exploratory Data Analysis
- Methods
- Conclusion

Data Summary

The HR Analytics dataset consists of employee data for a company of ~4,400 people including results of an employee satisfaction survey and performance ratings

EMPLOYEE DATA

- Employee ID
- Weekly hours (calculated)
- Attrition
- Distance From Home
- In and Out time
- etc.

EMPLOYEE SURVEY DATA

- Employee ID
- Environment Satisfaction
- Job Satisfaction
- Work-life Balance

MANAGER SURVEY DATA

- Employee ID
- Job Involvement
- Performance Rating

4410 OBSERVATIONS OF 28 VARIABLES

Overview

Objectives

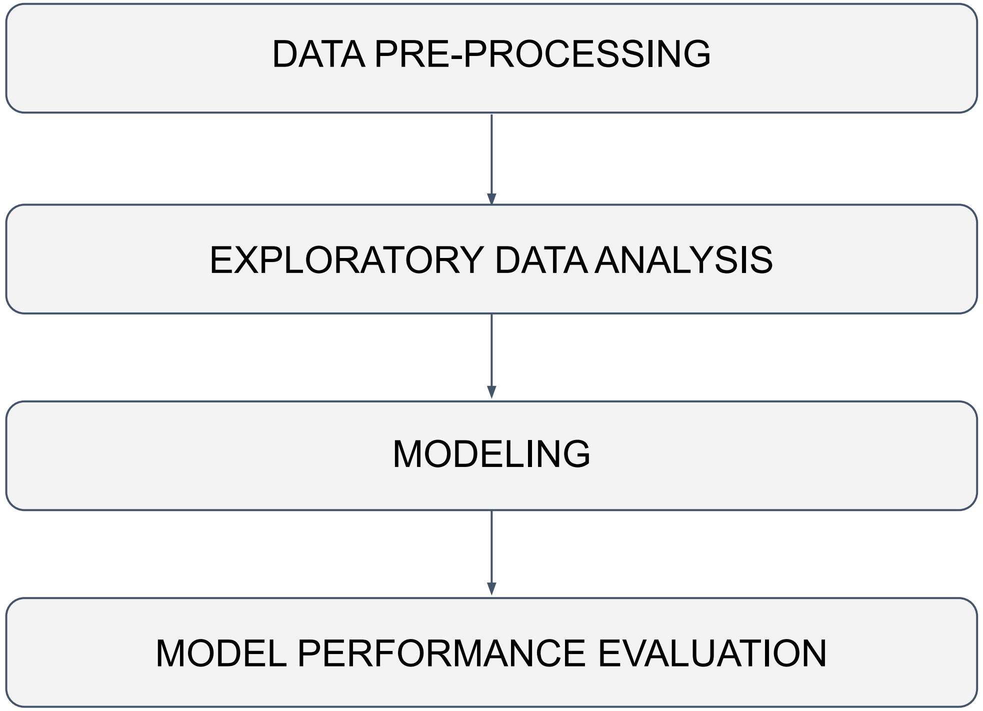
Understand

Determine significant predictors of attrition at the company

Predict

Build a model capable of predicting employee attrition for the company

Approach



DATA PRE-PROCESSING

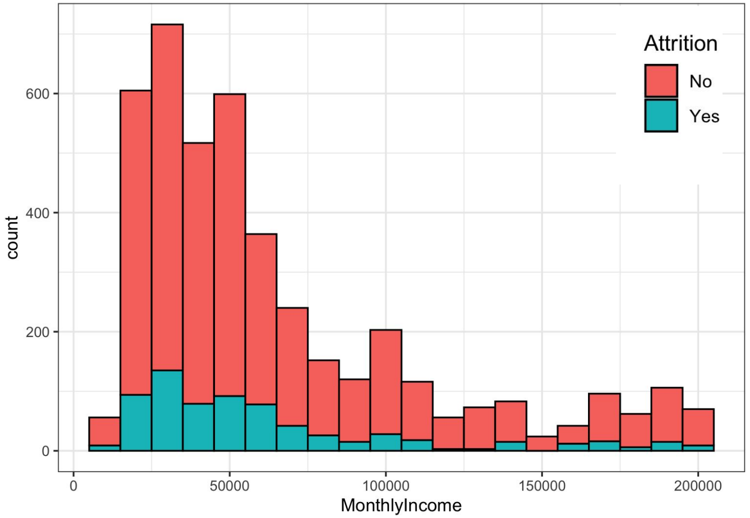
EXPLORATORY DATA ANALYSIS

MODELING

MODEL PERFORMANCE EVALUATION

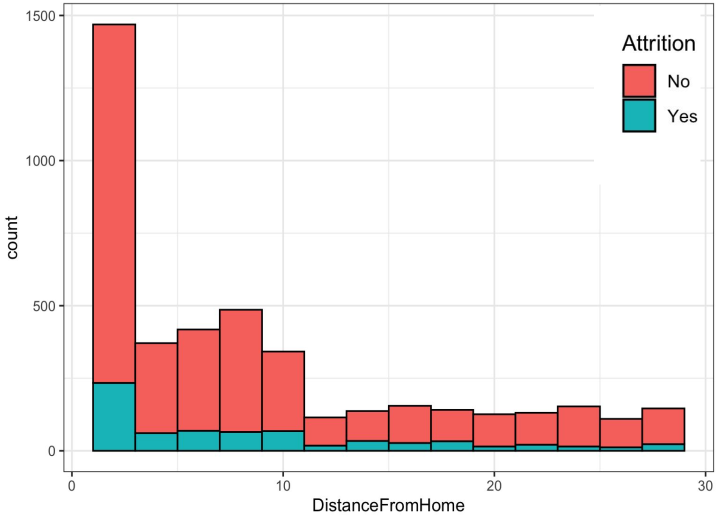
Exploratory Data Analysis

Distribution of Income & Attrition



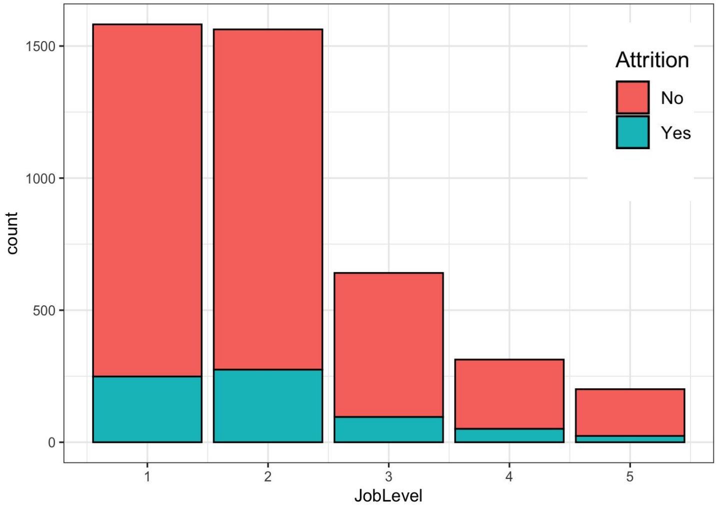
Of those employees that **left**, a majority had **lower** monthly incomes - because most employees have lower incomes.

Distribution of Distance & Attrition



Surprisingly, attrition appears **more** common among those that live **closer** to the job location - because most employees live close.

Distribution of Job Level & Attrition

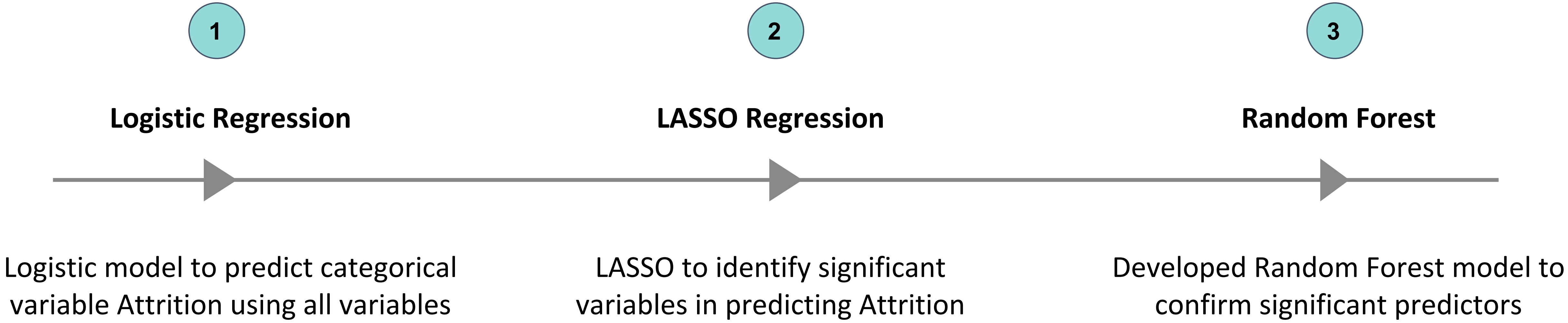


Attrition appears **least** common among those in the **highest** job levels - because there are very few employees at high levels.

Modeling Approach

Multiple predictive methods were used to reach our objectives

Objectives	Determine significant predictors and build a strong predictive model for Attrition



1

Logistic Regression

Logistic model to predict categorical variable Attrition using all variables

2

LASSO Regression

LASSO to identify significant variables in predicting Attrition

3

Random Forest

Developed Random Forest model to confirm significant predictors

Logistic Regression

Variable of interest, attrition, is categorical → Classification problem

Method

Model:
$\Pr(Y = \text{Yes} \mid X = (x_1, x_2, \ldots, x_p))$
where:
Y = Attrition
X = Characteristics specific to each employee considered, such as age, monthly income, job level, etc.

Classification Rule:
Guess Yes if: $\Pr(Y = \text{Yes} \mid X) > 0.5$

Results

Misclassification Error = 0.15
Accuracy = 0.85

	Predicted: NO	Predicted: YES
Actual: NO	1043	38
Actual: YES	150	59

61% of the people who were predicted to leave the company actually left.

87% of the people predicted to stay actually stayed.

LASSO Method

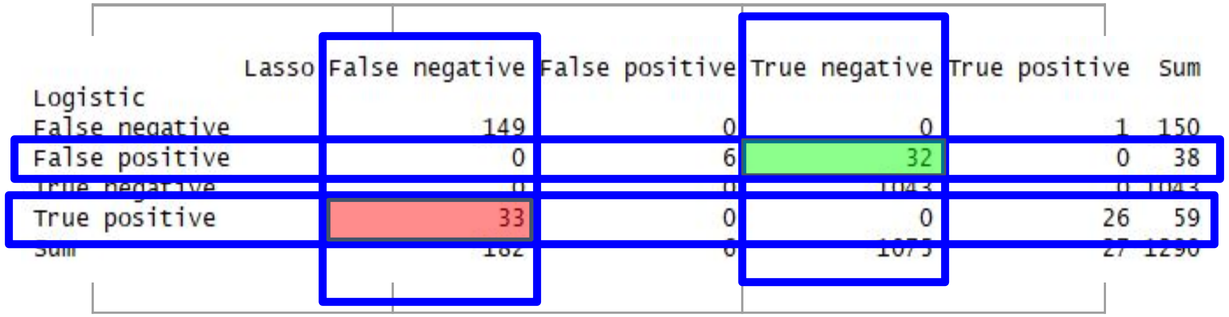
Use the LASSO method to determine which regression coefficients are most significant

Method

- Minimize RSS with penalty to coefficients under the L1 norm
- Cross validate across a range of lambda
- Consider the interpretability of which coefficients are selected and compare accuracy

Results

No difference in misclassification error



Lasso Logistic	False negative	False positive	True negative	True positive	Sum
False negative	149	0	0	1	150
False positive	0	6	32	0	38
True negative	0	0	1043	0	1043
True positive	33	0	0	26	59
Sum	182	6	1075	27	1290

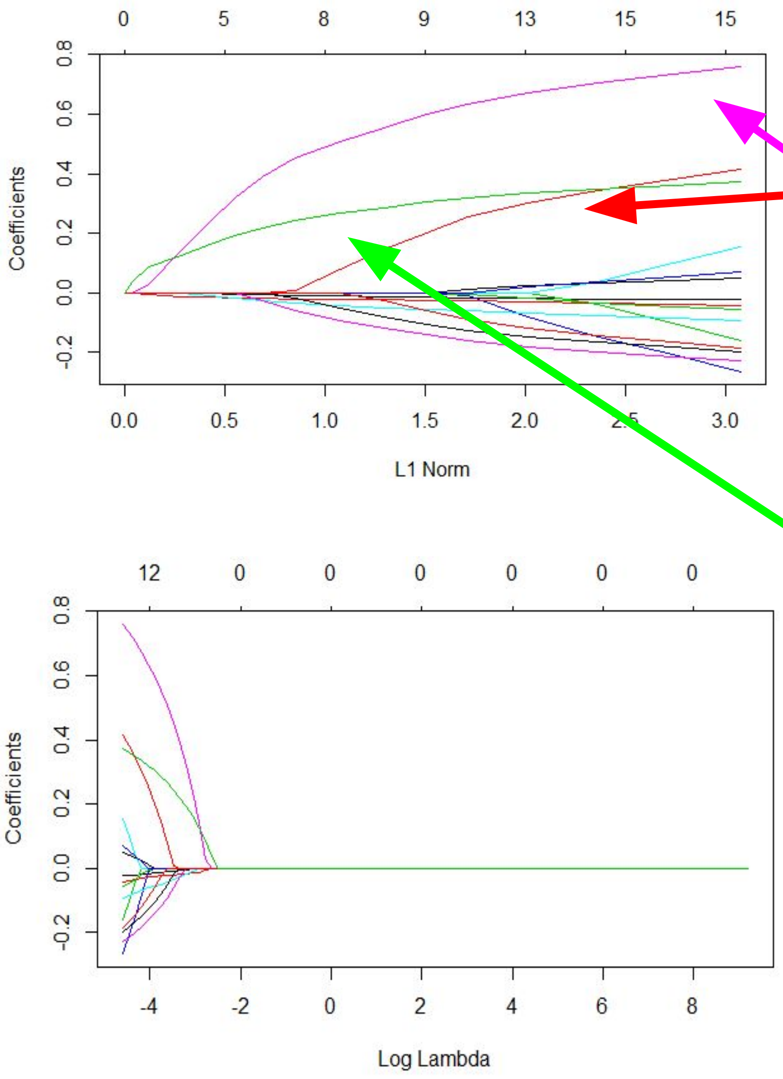
82% (up 21%) of the people who were predicted to leave the company actually left.

Bottom Line? For a small tradeoff in predicting who stays, we made a large gain in predicting who leaves

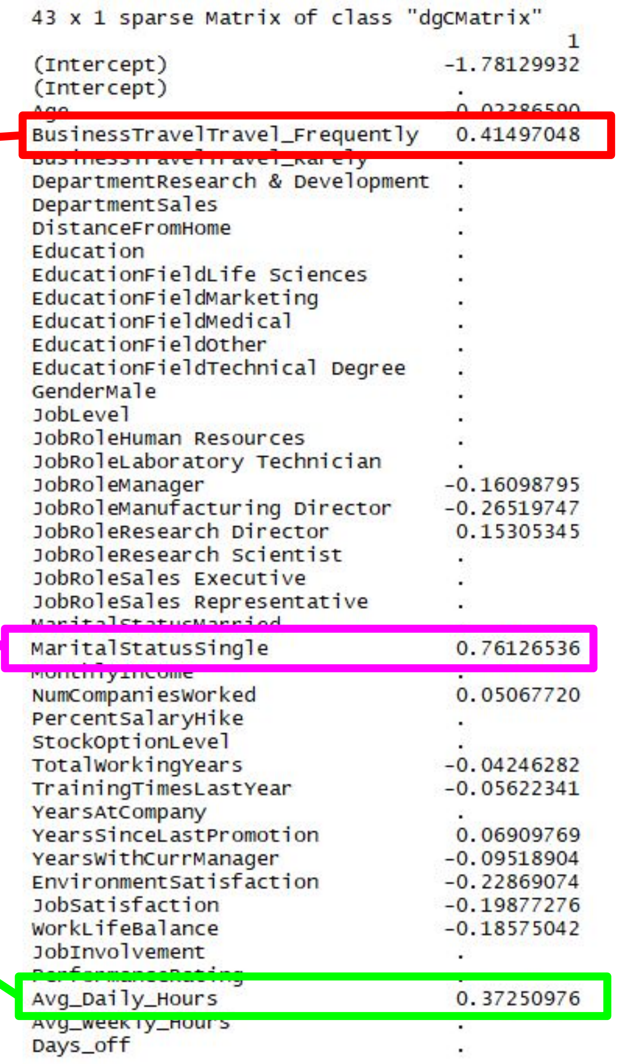
86% (down 1%) of the people who were predicted to stay with the company didn’t leave

LASSO Method (cont.)

Cross Validation



Coefficient Selection



```
43 x 1 sparse Matrix of class "dgCMatrix"
                                          1
(Intercept)                     -1.78129932
(Intercept)                      .
Age                              0.02386590
BusinessTravelTravel_Frequently  0.41497048
BusinessTravelTravel_Rarely      .
DepartmentResearch & Development .
DepartmentSales                  .
DistanceFromHome                 .
Education                        .
EducationFieldLife Sciences      .
EducationFieldMarketing          .
EducationFieldMedical            .
EducationFieldOther              .
EducationFieldTechnical Degree   .
GenderMale                       .
JobLevel                         .
JobRoleHuman Resources           .
JobRoleLaboratory Technician     .
JobRoleManager                  -0.16098795
JobRoleManufacturing Director   -0.26519747
JobRoleResearch Director         0.15305345
JobRoleResearch Scientist        .
JobRoleSales Executive           .
JobRoleSales Representative      .
MaritalStatusMarried             .
MaritalStatusSingle              0.76126536
MonthlyIncome                    .
NumCompaniesWorked               0.05067720
PercentSalaryHike                .
StockOptionLevel                 .
TotalWorkingYears               -0.04246282
TrainingTimesLastYear           -0.05622341
YearsAtCompany                   .
YearsSinceLastPromotion          0.06909769
YearsWithCurrManager            -0.09518904
EnvironmentSatisfaction         -0.22869074
JobSatisfaction                 -0.19877276
WorkLifeBalance                 -0.18575042
JobInvolvement                   .
PerformanceRating                .
Avg_Daily_Hours                  0.37250976
Avg_Weekly_Hours                 .
Days_off                         .
```

Random Forest p=28

Method

Number of Trees: 500
Accuracy: 0.976

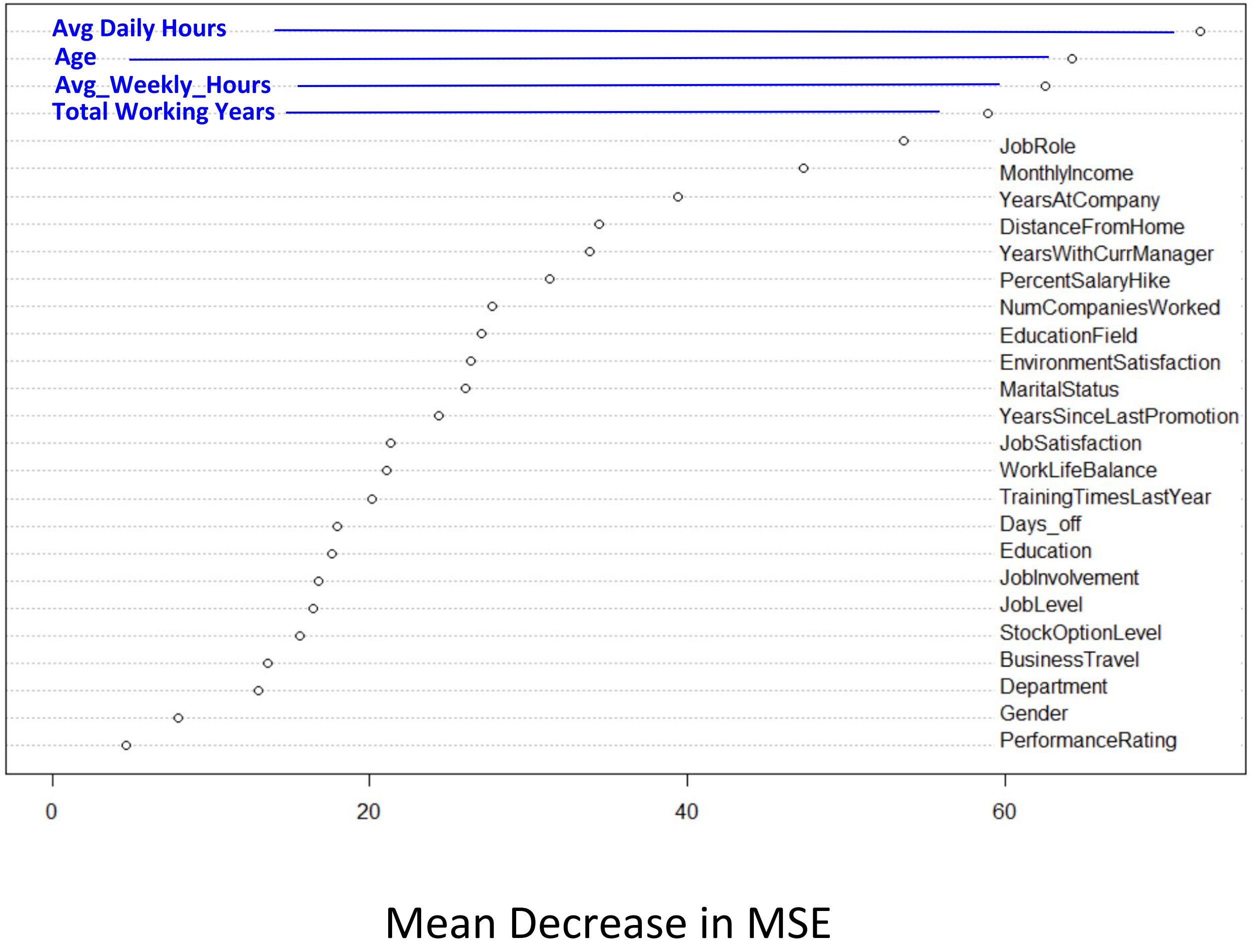
	Predicted: NO	Predicted: YES
Actual: NO	1075	6
Actual: YES	25	184

96.8% of the people who were predicted to leave the company actually left

97.7% of the people who were predicted to stay with the company didn't leave

Results

Variable Importance Plot



Random Forest p=14

Method

Number of Trees: 500
Accuracy: 0.978

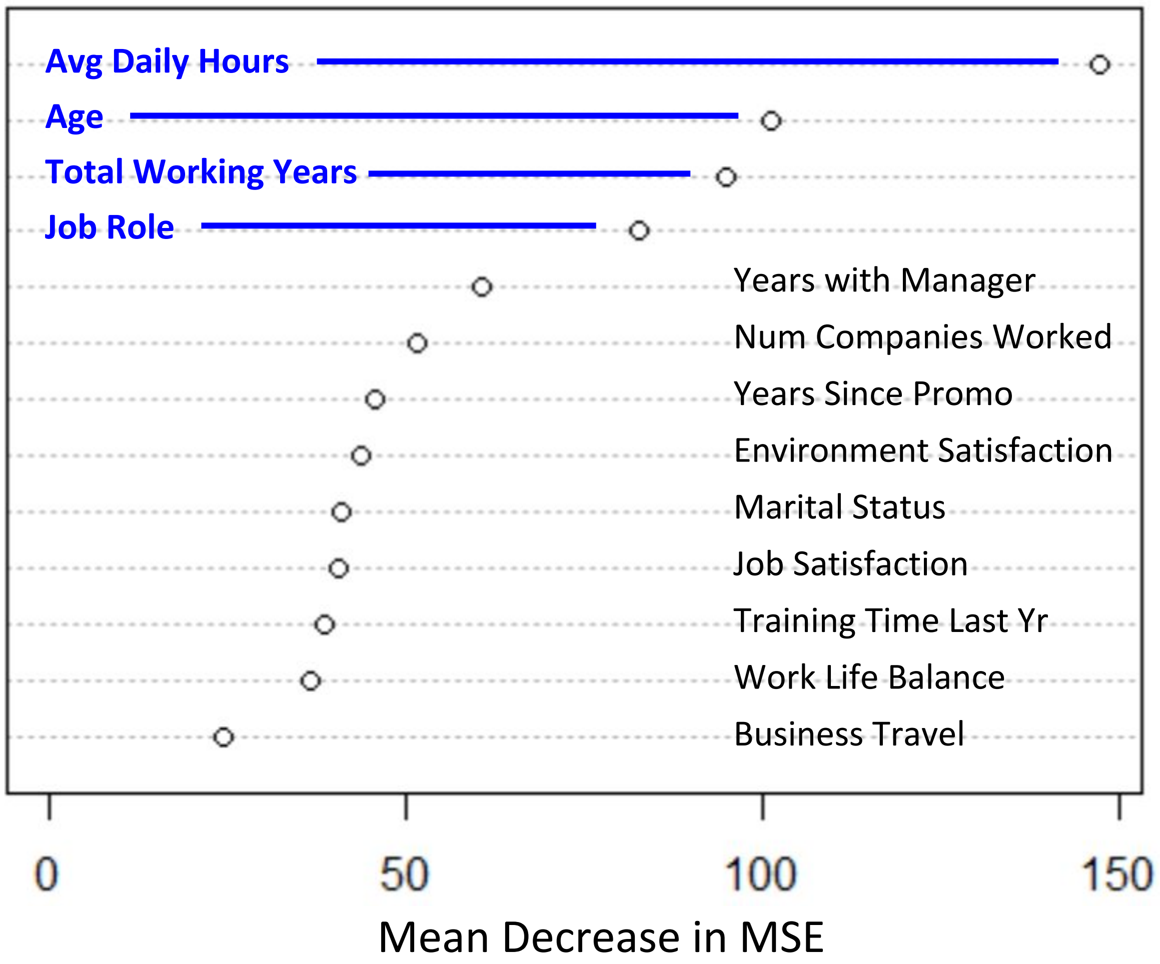
	Predicted: NO	Predicted: YES
Actual: NO	1075	6
Actual: YES	22	187

96.9% of the people who were predicted to leave the company actually left

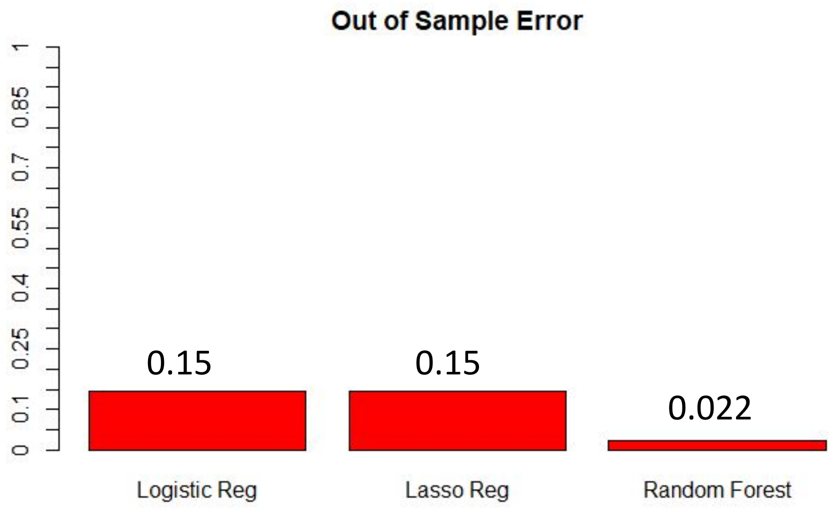
98% of the people who were predicted to stay with the company didn't leave

Results

Variable Importance Plot



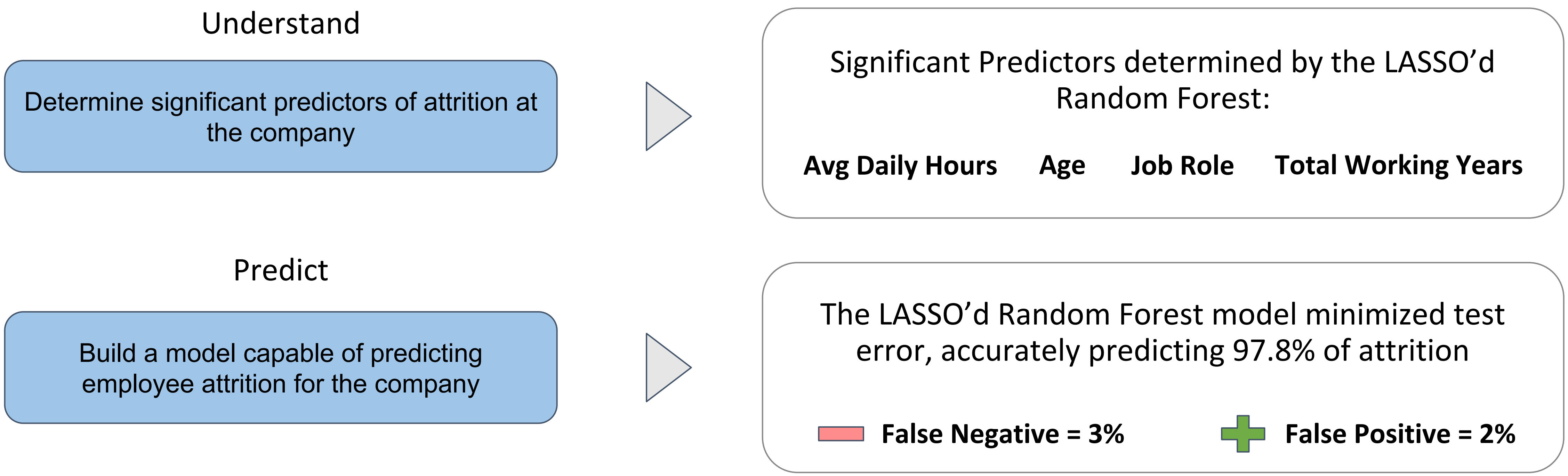
Model Performance Evaluation



	Logistic Regression	**Lasso Regression**	**Random Forest**
People who were predicted to **leave** the company that actually **left**	61%	82%	96.9%
People who were predicted to **stay** that actually **stayed**	87%	86%	98%

Conclusion

After running various models and cross validating results, Random Forest produced the most desirable results



Objectives

Understand

Determine significant predictors of attrition at the company

Predict

Build a model capable of predicting employee attrition for the company

Conclusions

Significant Predictors determined by the LASSO’d Random Forest:

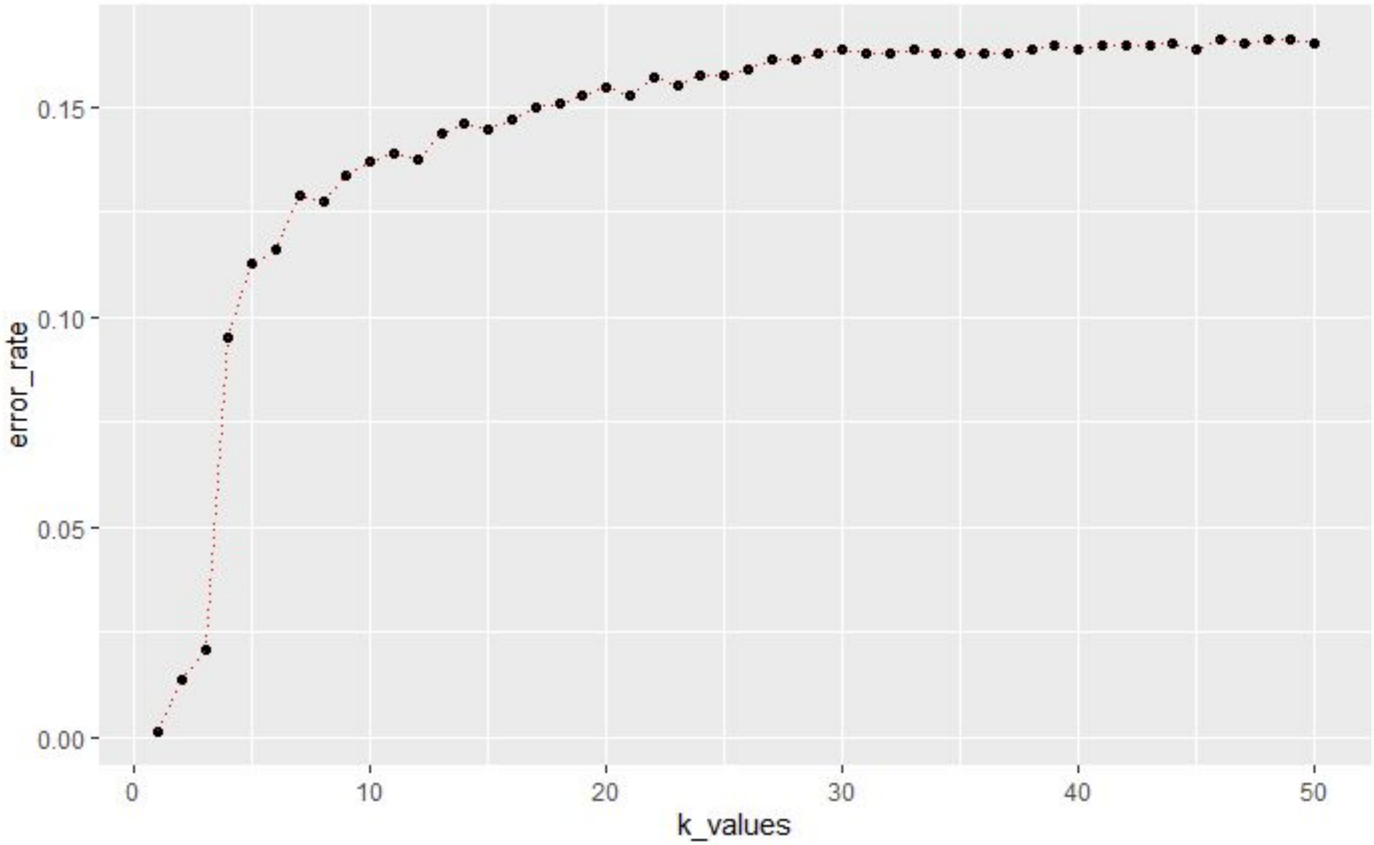
Avg Daily Hours **Age** **Job Role** **Total Working Years**

The LASSO’d Random Forest model minimized test error, accurately predicting 97.8% of attrition

🟥 **False Negative = 3%** ✚ **False Positive = 2%**

Questions?

K-nearest neighbors algorithm



- K value ↑
  Error rate ↑

  Accuracy: 0.9984615

  Misclassification error: 0.001538462

- Mix of categorical and continuous variables: cannot be scaled appropriately to use KNN.