How Predictive Analytics Can Reduce Employee Attrition at IBM

Section B Team 9: Alper Sayiner, Annabelle Nguyen,
Vatsal Sanghavi, and Sophia Rowland

Business Understanding

For a professional services company such as IBM, human capital is the most valuable asset and the most expensive cost of operations. In order to maintain its competitive advantage, IBM needs to hire the right talents and retain them, especially in this era when the workforce is very mobile. Each time an employee leaves the company, IBM incurs a cost of unfilled time in the vacant position, as well as the cost of time spent recruiting, interviewing, and training their replacement. These costs combined can be up to 1.5 - 2 times annual salary per employee[1]. Having the ability to predict which employees are most likely to leave IBM in the near future (in the next 6 months, for example), General Managers and HR Managers can prioritize their time and resources to engage with the at-risk employees, that is, those who are most likely to leave, before these employees put in their 2-week notices. While trying to keep employees who already put in their 2-week notices is a reactive practice, knowing who would be likely to leave and taking actions is a preventive practice. Therefore, we propose a predictive analytics model to predict employee attrition.

After predicting the probability of each employee leaving the firm, we will rank employees in descending order by their probability of leaving IBM and provide the list to senior leaders who are responsible for the retention effort. Because employees' data are confidential, we will request that the executives who have access to the list of at-risk employees keep this information confidential and that they not share the result of our analysis to employees. We will also provide a model that IBM can use to predict future employee attrition.

According to our research, current best practices in predicting employee attrition are machine learning methods such as Logistic Regression, Decision Tree, Random Forests, and Survival Analysis. To solve this business problem, companies also utilize various analytics services including IBM Watson, IBM Kenexa, Talent Analytics, and Visier, which use these methods.

[1] https://www.linkedin.com/pulse/20130816200159-131079-employee-retention-now-a-big-issue-why-the-tide-has-turned/

Data Understanding

Kaggle, a competition website for data miners, provided the data that our team used to analyze[2]. The data contains 1470 observations and 35 columns. The variables include demographic information about the employee, information about their position, and their current satisfaction with several aspects of that position (Exhibit 1). The dependent variable we looked at is 'Attrition', which is whether or not the employee will leave the company. To predict if an employee will leave IBM, we cleaned the data and built a prediction model from the 35 variables. Our data set consists of two types: factors and numbers (Exhibit 1). Factors are variables that have different levels or categories and, for our analysis, each one of these levels will be treated as a dummy variable. Numbers are continuous variables and will be treated as such.

We believe that this data set works best at predicting attrition within the next year and should be implemented on employees that have spent at least a year with IBM in order to yield the most accurate probabilities of leaving. There are several reasons for these suggestions. First, major occupational changes, such as promotions and pay increase, occur on a yearly basis. Second, there are several predictors in this data set, such as job satisfaction, that can only be accurately recorded after spending a year with the company.

There can be potential bias in the data set because certain variables such as job satisfaction, environment satisfaction, relationship satisfaction and work-life balance are self-reported scores. Whether an employee works over-time or not is also subjective because a "yes" for overtime may not necessarily mean that the employee works overtime every single week; it is likely an average over the past one year. Additionally, it is important to note that this is a simulated data set and hence will not reflect the true employee attrition trends at IBM.

[2] Data set available at: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Data Preparation

In preparing our data for analysis, we cleaned our data set so that it only contained the variables that are important to our analysis. First we dropped variables that didn't contain any useful information, which were EmployeeCount, EmployeeNumber, Over18, and StandardHours (Exhibit 2). The remaining variables are in Exhibit 3. After cleaning the data, we did some initial exploratory analysis and data visualization. After ensuring that the distribution of data for each predictor didn't include any outliers, we visually explored relationships among predictors and attrition. From this analysis, we found that the proportion of people that left the company wasn't different across gender (Figure 1), but was different across job satisfaction levels (Figure 2), stock options offered (Figure 3), work-life balance levels (Figure 4), and the number of companies worked at prior to IBM (Figure 5). The proportion of people that left the company was larger for those with the lowest job satisfaction, for those not offered any stock options, for those with bad work-life balance, and for those who had worked at more companies.

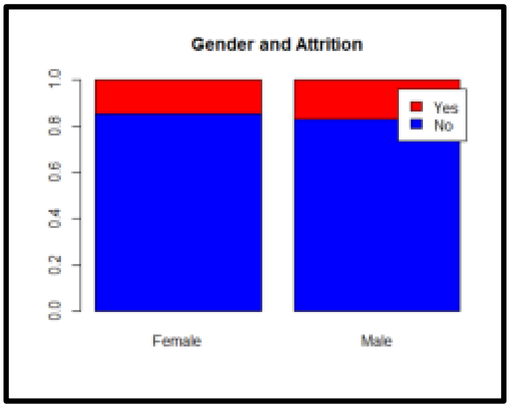


Figure 1. Gender and Attrition

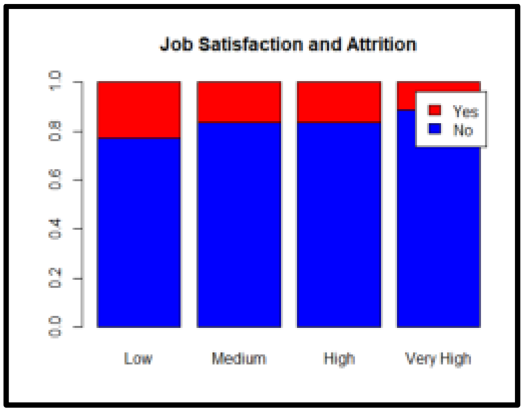


Figure 2. Job Satisfaction and Attrition

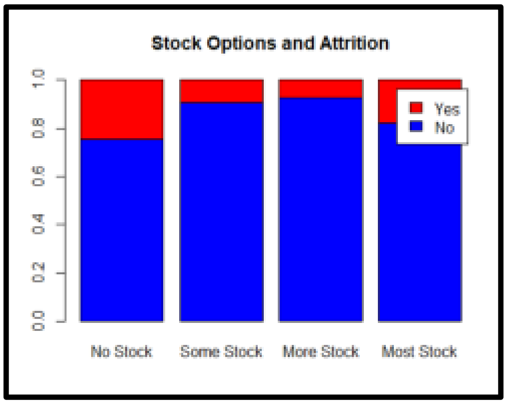


Figure 3. Stock Options and Attrition

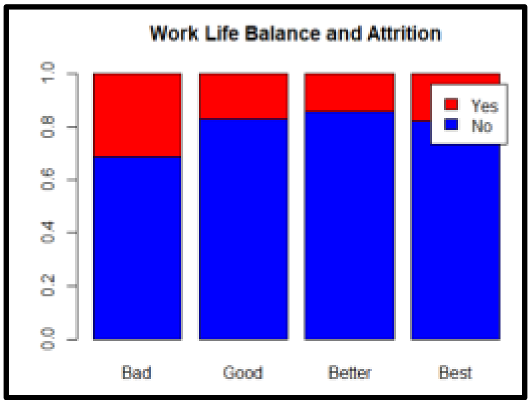


Figure 4. Work Life Balance and Attrition

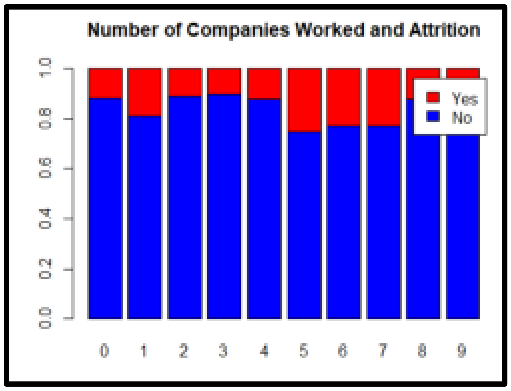


Figure 5. Number of Companies Worked and Attrition

Data Modeling and Evaluation

Because the business goal is to predict a certain outcome, we employed supervised learning algorithms. Since attrition is a binary outcome variable, we decided to try four classification models: logistic regression, logistic regression with interaction variables, decision tree, and decision tree with interaction variables. To determine which variables to include in our models, we began by using Lasso on all four models. For the logistic regression and decision tree models without interaction variables, Lasso suggested 16 variables. For the logistic regression and decision tree models with interaction variables, Lasso suggested 30 variables (Exhibit 4). Next, we examined the correlation matrix which visually shows the strength of the relationships between the continuous variables (Exhibit 5). Then, we cross checked Lasso's results by performing t-tests and chi-squared tests between these variables and attrition to identify significant relationships at the 95% confidence level. The t-test was used for numeric variables and the chi-squared test for categorical variables. After examining the correlation plot, t-tests, and chi-squared tests, we chose the common variables from those tests, and ended up with 14 of Lasso's initial set of variables for the models without interaction variables, and 28 of Lasso's initial set of variables for the models with interaction variables (Exhibit 4).

After determining which variables to include in each model, we created two data frames. Data Frame 1 included the 14 variables for the models without interaction variables. Data Frame 2 included the 28 variables for the models with interaction variables (Exhibit 4). Next, using Out of Sample R^2 and

ROC Curves, we compared the logistic regression model without interactions, the logistic model with interactions, the tree model without interactions, the tree model with interactions, and a null model using the chosen variables in the data set. The ROC Curves (Figure 6) shows that the tree model without interactions performs very poorly in comparison to all other models. Measures of model accuracy in prediction while creating the ROC Curve (Figure 7) show that tree models were not as accurate as the logistic models. Finally, we performed the 10-folds cross validation and created the box plot of each model's performance for Out of Sample R^2 (OOS R^2; Figure 8). The logistic models had the highest OOS R^2, but the logistic model with interactions had a much larger spread of R^2 value. Through these various tests of accuracy, we decided to move forward with our logistic model without interactions due to its consistent high performance across the different performance metrics.

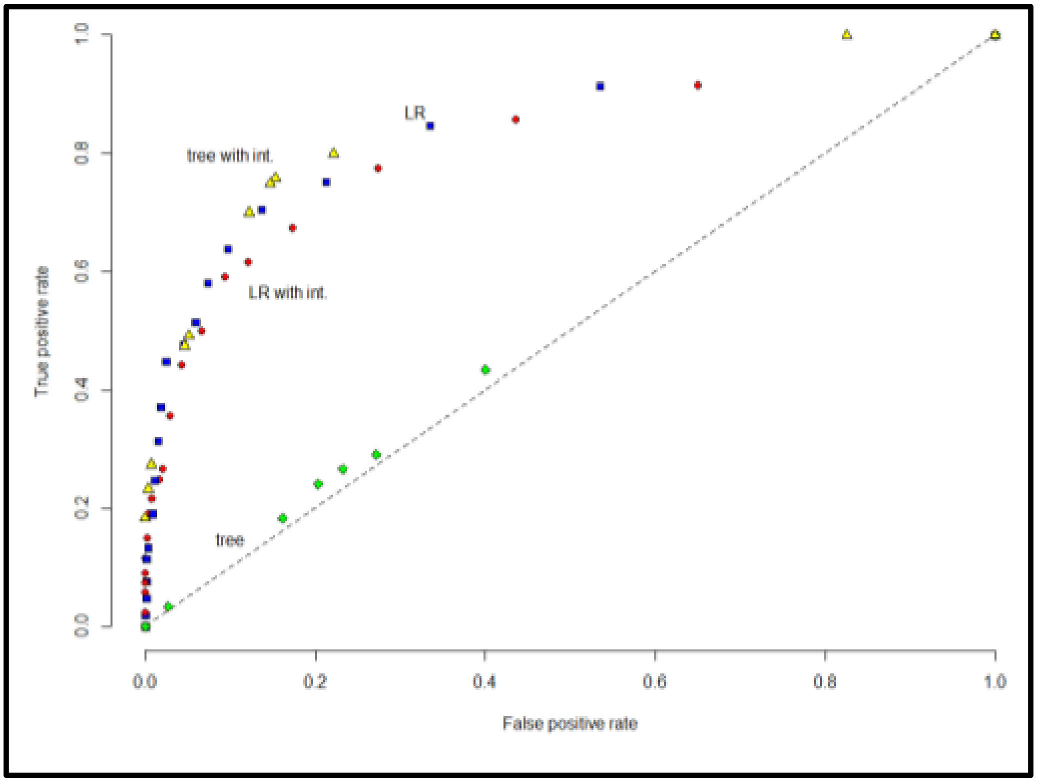


Figure 6. ROC Curves

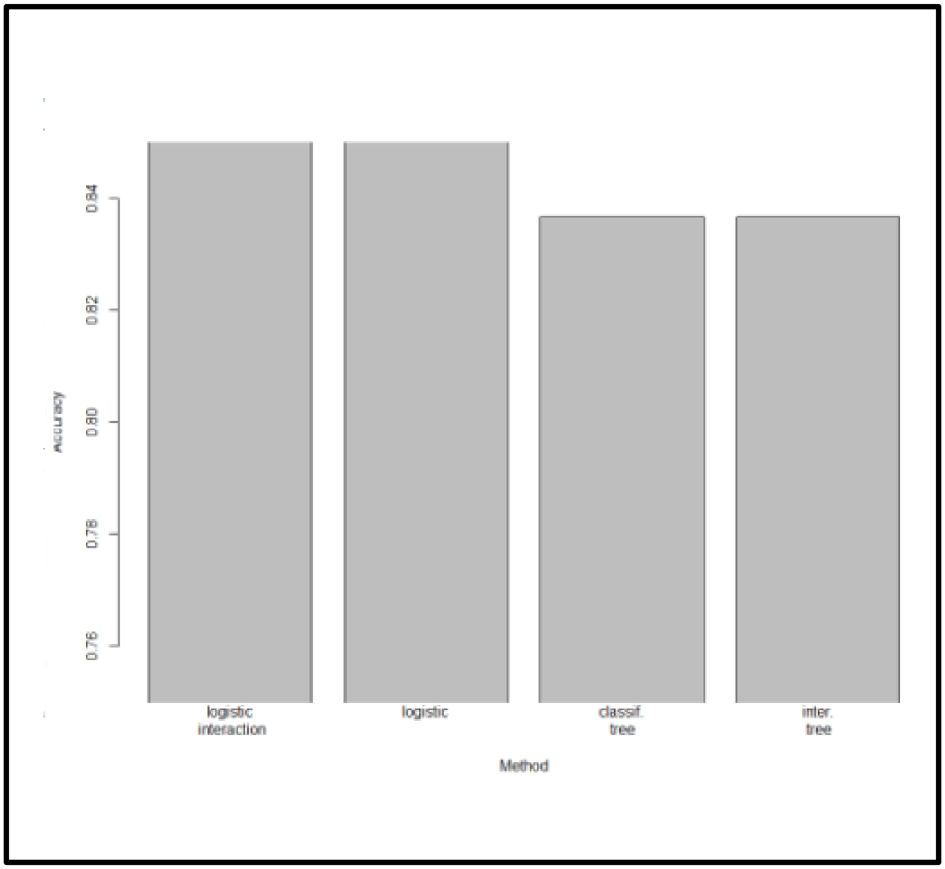


Figure 7. ROC Accuracy Bar Chart

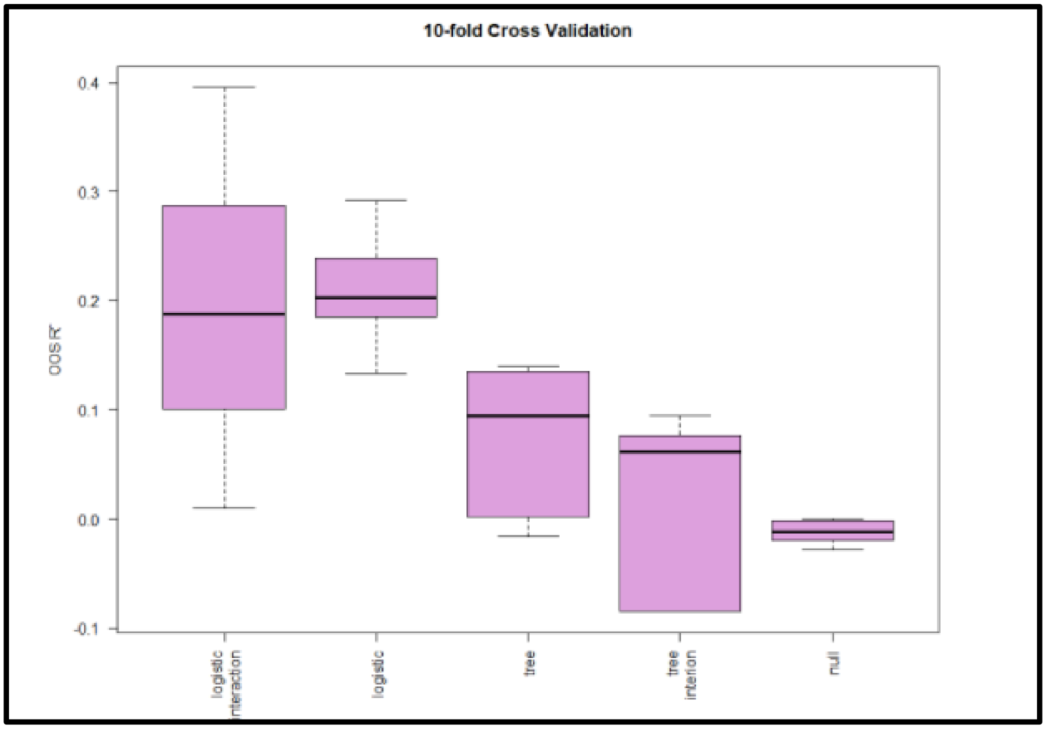


Figure 8. 10-Folds Cross Validation

After selecting the logistic regression model without interactions, we ran the model on Data Frame 1 to see whether all of the coefficient estimates are significant or not. Eliminating insignificant coefficients at the 95% significance level, we narrowed down to 10 variables out of 14 that yielded significant coefficient estimates and re-ran the logistic regression on Data Frame 1 using these 10 variables (Exhibit 6).

Not relying too heavily on automated tools, we evaluated from a business perspective how logical it was to include the chosen variables in the regression. We expect that number of companies worked is an indicator of how prone an employee is to change jobs. Being single allows more freedom and flexibility to switch jobs/careers or relocate and explore new opportunities. A lot of business travel and working overtime reduce work-life balance, leading to attrition. A job role as a laboratory technician might involve repetitive work, and monotony likely leads to employee attrition. Older employees may find it more difficult to both relocate and learn new skills at a new job, causing them to stay longer at their current jobs. Employees in the Research and Development department differ from other departments in terms of having longer project durations, leading to more commitment. Those given stock option will feel more attached to the company and more dedicated to their work. Literature on human capital management has shown that satisfaction and engagement are one of the main reasons employees stay in their jobs. Lastly, mid-level employees might be less likely to leave the company since they can soon be promoted to a senior position, having a higher opportunity cost than the entry-level employees if they left.

It is interesting to note that salary is not part of our final regression model because once controlled for other factors like job satisfaction and job level, salary is not a contributing member to affect attrition. Similarly, years in current role, years with current manager, years since last promotion, environment satisfaction and relationship satisfaction are not predictors of attrition because job satisfaction, job level and department would control for them in the regression and predict attrition probabilities. With work-life balance being a self-reported score, other measures like business travel frequency and working overtime become better indicators of work-life balance and hence serve as predictors of attrition. Therefore, we believe these variables make business sense to be included in the model.

Our initial data visualizations and subsequent hypothesis is confirmed by this regression summary (Exhibit 6). Number of companies worked, marital status (single), high business travel, working overtime and job role (laboratory technician) yield positive attrition coefficients. On the other hand, age, working in research and development, having stock option, a high job satisfaction and a mid-level position have negative coefficients. This shows that our modeling and evaluation is in line with our business understanding of the problem.

Deployment

After deciding on an accurate model and picking which variables to include, we will provide IBM with two deliverables: a logistic regression model to predict attrition probability for future employees, and the probability of attrition for current IBM employees.

First, we created a model using Data Frame 1 to predict attrition probability for future employees. Based on the signs and magnitudes of the regression coefficient estimates (Exhibit 6), we determined that there are several ways for IBM to reduce attrition in their high-risk employees. First, the company could reduce the amount of overtime hours put in by employees. Second, IBM could promote high-performing entry-level employees faster as well as revisit their promotion criteria. Third, IBM could reduce employee travel by teleconferencing. Fourth, IBM could strive to improve overall job satisfaction by seeking the causes of employee dissatisfaction and brainstorming possible solutions. Finally, offering all employees in the company stock options would increase their investment into IBM's success and thus increase their incentives to stay at the company.

Next, we predicted which of IBM's current employees were likely to leave within a year. We divided the set of employees into current and former. We then created multiple and training and predicting datasets. Each predicting subset consisted of a set of unique current employees. The corresponding training subset consisted of the remaining current employees and all former employees. We ensured that for every subset, the training:predicting ratio was 9:1. This allowed us to divide the data systematically, where we made predictions for each current employee without using a model that they influenced, thereby avoiding overfitting.

After predicting attrition on all current employees, we narrowed down on 32 employees who had a predicted attrition probability greater than 0.5. Predicted attrition rates of all current employees are available as an additional file (output of the R code), titled “Deliverable.csv,” which will be presented to IBM General Managers and HR Managers. Of these 32 employees, 2 are in Human Resources, 17 are in Research and Development, and 13 are in Sales (Exhibit 7). In addition, of the 32 at-risk employees, 25 work overtime (Exhibit 8), 24 were offered no stock in IBM (Exhibit 9), and 20 were entry level (Exhibit 10).

These results confirm our initial hypotheses, corroborate our results from our initial data visualizations, and replicate our results from the logistic regression model. Overall, since the costs of replacing an employee can be up to 1.5 - 2 times annual salary, as long as IBM explores solutions with a price tag below this threshold, they will be saving money that could be used to improve overall infrastructure and employee satisfaction.

Although the predictive analytics model provides valuable insights and helps IBM’s General Managers and HR Managers prioritize their retention efforts, there are a few drawbacks that they should be aware of. First, they should avoid relying too heavily on this regression with the risk of ignoring the welfare of employees at a low risk of leaving. Second, if the predicted attrition probability for each employee is not kept confidential, it will be a self-fulfilling prophecy because employees who know they are predicted to leave will actually decide to leave IBM based on this information.

Conclusion

Employee attrition can cost companies up to 1.5 - 2 times annual salary per employee, so IBM has plenty of incentives to predict at-risk employees. From a data set of employee attrition, demographics, and job characteristics, we predicted which of IBM's current employees are at risk of leaving and created a model to predict attrition for future employees. From the results of these methods, we recommend that IBM reduce employee overtime, recognize good work through promotions (especially at the entry level), reduce employee travel, work to eliminate overall job dissatisfaction and offer stock option to all employees. This investment in improving retention not only reduces cost, but also demonstrates IBM's willingness to invest in their employees.

Appendix

Exhibit 1. Data Set Information

Variable	Definition	Type	Values
Age	Employee age	Numeric	18 - 60
Attrition	Will the employee leave the company; Target Variable	Factor	0: ‘No’ 1: ‘Yes’
BusinessTravel	How often the employee must travel for role	Factor	‘Non_Travel’ ‘Travel_Rarely’ ‘Travel_Frequently’
DailyRate	How much the employee makes per day	Numeric	102 - 1499
Department	The department the employee works in	Factor	‘Human Resources’ ‘Research and Development’ ‘Sales’
DistanceFromHome	How far the employee lives from their workplace, assumed to be in miles	Numeric	1 - 29
Education	Employee’s highest level of education	Factor	1: ‘Below College’ 2: ‘College’ 3: ‘Bachelor’ 4: ‘Masters’ 5: ‘Doctor’
EducationField	What field the employee’s education was in	Factor	‘Human Resources’ ‘Life Sciences’ ‘Marketing’ ‘Medical’ ‘Other’ ‘Technical Degree’
EmployeeCount	Number of employees	Numeric	1
EmployeeNumber	An unique id number for each employee	Numeric	1 - 2068
EnvironmentalSatisfaction	How satisfied the employee is in their current environment	Factor	1: ‘Low’ 2: ‘Medium’ 3: ‘High’

			4: ‘Very High’
Gender	Employee Gender	Factor	‘Female’ ‘Male’
HourlyRate	How much the employee makes per hour	Numeric	30 -100
JobInvolvement	How involved the employee is in their current role	Factor	1: ‘Low’ 2: ‘Medium’ 3: ‘High’ 4: ‘Very High’
JobLevel	Employee’s management level	Factor	1: ‘Entry Level’ 2: ‘Mid Level’ 3: ‘Senior Level’ 4: ‘Executive Level’ 5: ‘Senior Executive Level’
JobRole	Employee’s current job role	Factor	‘Healthcare Representative’ ‘Human Resources’ ‘Laboratory Technician’ ‘Manager’ ‘Manufacturing Director’ ‘Research Director’ ‘Research Scientist’ ‘Sales Director’ ‘Sales Representative’
JobSatisfaction	Employee’s satisfaction with current job	Factor	1: ‘Low’ 2: ‘Medium’ 3: ‘High’ 4: ‘Very High’
MaritalStatus	Employee’s marital status	Factor	‘Divorced’ ‘Married’ ‘Single’
MonthlyIncome	Employee's monthly income	Numeric	1009 - 19999
MonthlyRate	Employee’s month rate	Numeric	2094 - 26999
NumCompaniesWorked	The number of companies the employee worked at prior to IBM	Numeric	0 - 9

Over18	Is the employee over 18	Factor	‘Y’
OverTime	Does the employee recieve overtime	Factor	‘No’ ‘Yes’
PercentSalaryHike	How much has their salary increased since they started at IBM, expressed in percentage	Numeric	11 - 25
PerformanceRating	Employee’s latest performance rating	Factor	1: ‘Low’ (unused in data) 2: ‘Good’ (unused in data) 3: ‘Excellent’ 4: ‘Outstanding’
RelationshipSatisfaction	Employee’s satisfaction with relationships at work	Factor	1: ‘Low’ 2: ‘Medium’ 3: ‘High’ 4: ‘Very High’
StandardHours	Employee’s standard hours for a 2-week period	Numeric	80
StockOptionLevel	Set of stock options offered to employee	Factor	0: ‘None’ 1: ‘Some’ 2: ‘More’ 3: ‘Most’
TotalWorkingYears	How many years has the employee been working in general	Numeric	0 - 40
TrainingTimeLastYear	How many times did the employee attend a training event late year	Numeric	0 - 6
WorkLifeBalance	Employee’s perception of their work-life balance	Factor	1: ‘Bad’ 2: ‘Good’ 3: ‘Better’ 4: ‘Best’
YearsAtCompany	How many years has the employee been with IBM	Numeric	0 - 40
YearsInCurrentRole	How many years has the employee been in	Numeric	0 - 18

	their current role		
YearsSinceLastPromotion	How many years since the employee’s last promotion	Numeric	0 - 15
YearsWithCurrManager	How many years has the employee worked for their current manager	Numeric	0 -17

Exhibit 2. Variables Dropped During Data Preparation Step

Variable	**Reason for Dropping**
EmployeeCount	Column contains 1 for every employee
EmployeeNumber	Unique ID number for each employee
Over18	Column contains ‘Y’ for every employee
StandardHours	Column contains 80 for every employee

Exhibit 3. Variables Included in Analysis

Variables	**Types**	**Levels**
Age	Numeric	NA
Attrition	Factor	2
BusinessTravel	Factor	3
DailyRate	Numeric	NA
Department	Factor	3
DistanceFromHome	Numeric	NA
Education	Factor	5
EducationField	Factor	6
EnvironmentalSatisfaction	Factor	4
Gender	Factor	2
HourlyRate	Numeric	NA

JobInvolvement	Factor	4
JobLevel	Factor	5
JobRole	Factor	9
JobSatisfaction	Factor	4
MaritalStatus	Factor	3
MonthlyIncome	Numeric	NA
MonthlyRate	Numeric	NA
NumCompaniesWorked	Numeric	NA
OverTime	Factor	2
PercentSalaryHike	Numeric	NA
PerformanceRating	Factor	2
RelationshipSatisfaction	Factor	4
StockOptionLevel	Factor	4
TotalWorkingYears	Numeric	NA
TrainingTimeLastYear	Numeric	NA
WorkLifeBalance	Numeric	4
YearsAtCompany	Numeric	NA
YearsInCurrentRole	Numeric	NA

Exhibit 4. Suggested Variables

Method	**Suggested Variables**
Lasso without interaction variables	Age, BusinessTravel_Travel Frequently, Department_Research & Development, *DistanceFromHome(excluded),* JobLevel_2, JobRole_Laboratory Technician, JobRole_Sales Representative, JobSatisfaction_4, MaritalStatus_Single, *MonthlyIncome* (excluded), NumCompaniesWorked, OverTime_Yes, StockOptionLevel_1, TotalWorkingYears, YearsInCurrentRole, YearsWithCurrManager
Lasso with interaction variables	Age, TotalWorkingYears, Age:StockOptionLevel_1, BusinessTravel_Travel Frequently:DistanceFromHome, BusinessTravel_Travel Frequently:JobRole_Sales Representative, BusinessTravel_Travel Frequently:MaritalStatus_Single, *DailyRate:StockOptionLevel_1(excluded)*, Department_Research & Development:JobLevel_2, Department_Sales:MaritalStatus_Single, *Department_Research & Development:MonthlyIncome(excluded)*, DistanceFromHome:MaritalStatus_Single, DistanceFromHome:OverTime_Yes, Education_3:JobRole_Sales Representative, EducationField_Technical Degree:MaritalStatus_Single, Gender_Male:OverTime_Yes, JobLevel_2:YearsAtCompany, JobLevel_2:YearsInCurrentRole, JobRole_Laboratory Technician:MaritalStatus_Single, JobRole_Sales Representative:MaritalStatus_Single, JobRole_Laboratory Technician:OverTime_Yes, JobRole_Sales Representative:OverTime_Yes, JobRole_Sales Representative:WorkLifeBalance_4, JobSatisfaction_4:WorkLifeBalance_3, MaritalStatus_Single:OverTime_Yes, MaritalStatus_Single:PercentSalaryHike, NumCompaniesWorked:OverTime_Yes, OverTime_Yes:PercentSalaryHike, PercentSalaryHike:TotalWorkingYears, TrainingTimesLastYear:YearsInCurrentRole,and WorkLifeBalance_3:YearsWithCurrManager.

Exhibit 5. Correlation Matrix

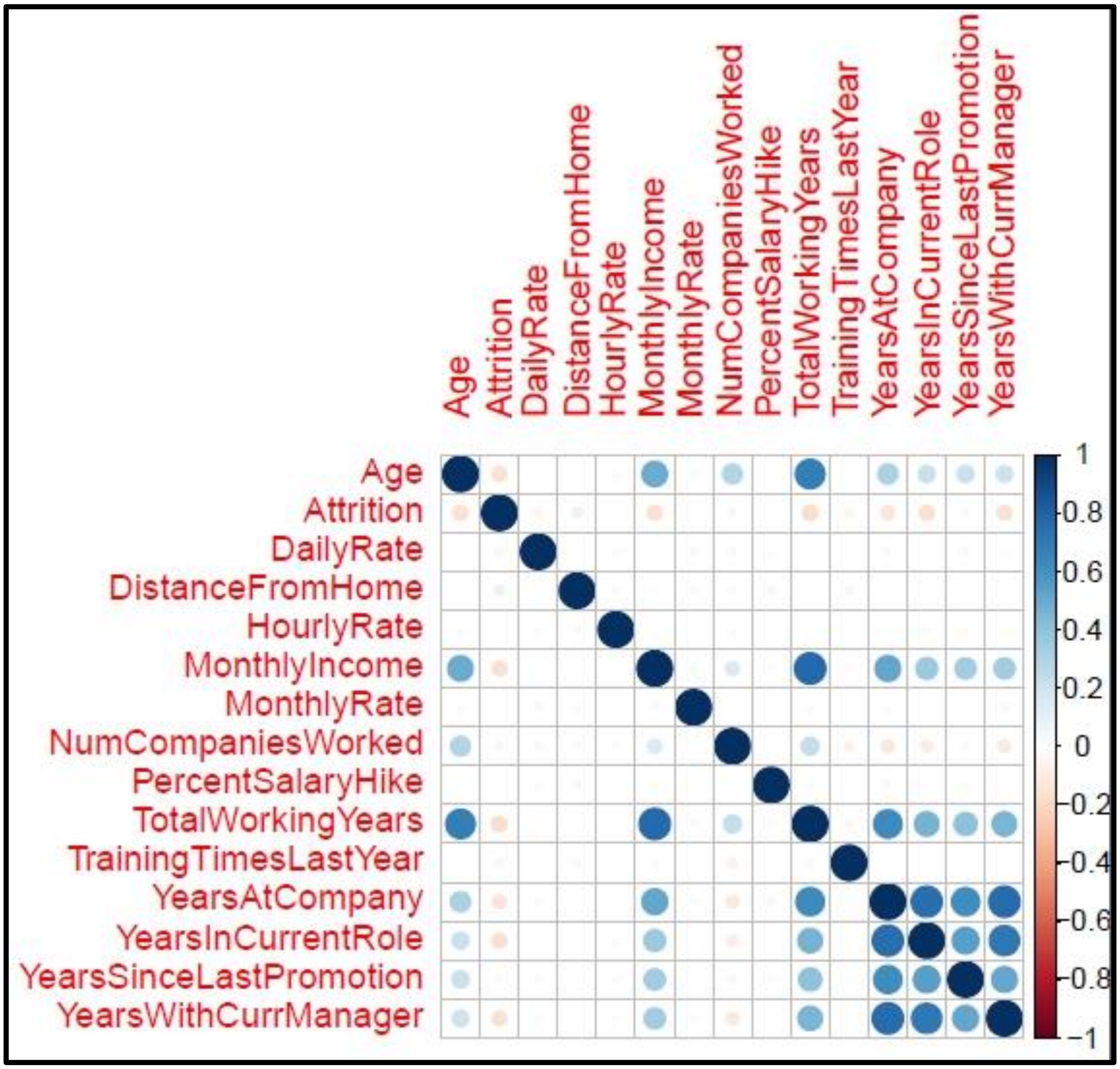
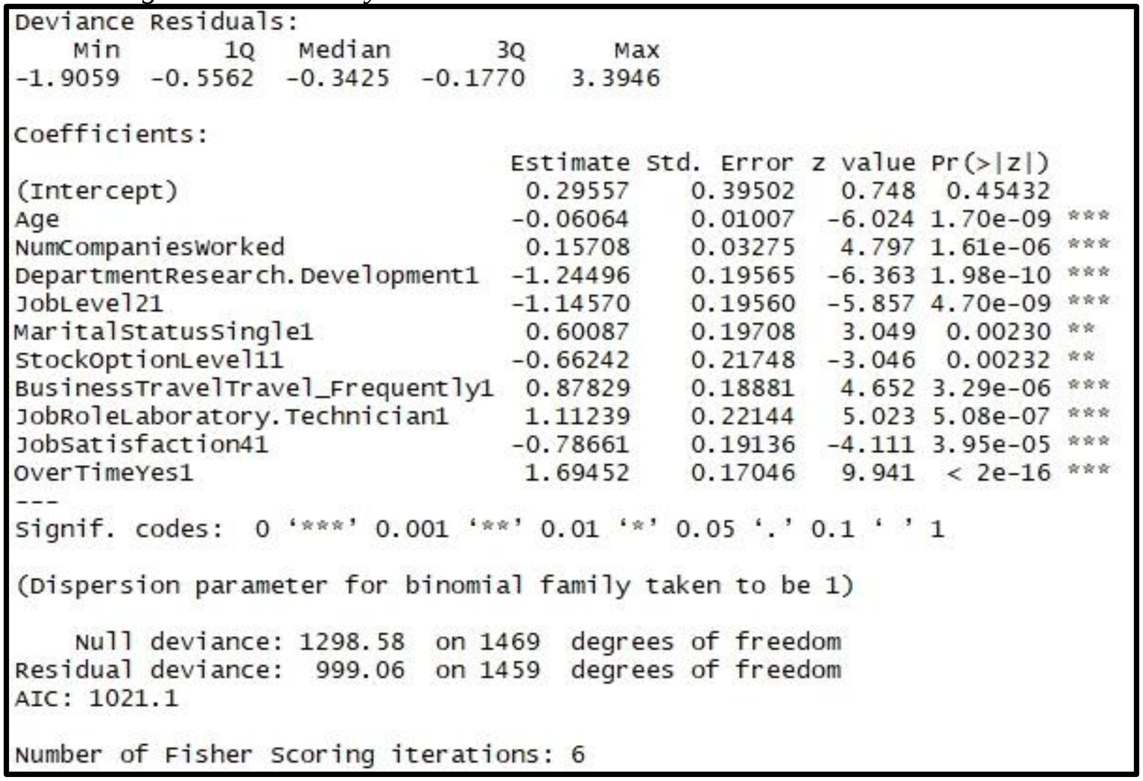


Exhibit 6. Final Regression Summary



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9059  -0.5562  -0.3425  -0.1770   3.3946

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       0.29557    0.39502   0.748  0.45432
Age                              -0.06064    0.01007  -6.024 1.70e-09 ***
NumCompaniesWorked                0.15708    0.03275   4.797 1.61e-06 ***
DepartmentResearch.Development1  -1.24496    0.19565  -6.363 1.98e-10 ***
JobLevel21                       -1.14570    0.19560  -5.857 4.70e-09 ***
MaritalStatusSingle1              0.60087    0.19708   3.049  0.00230 **
StockOptionLevel11               -0.66242    0.21748  -3.046  0.00232 **
BusinessTravelTravel_Frequently1  0.87829    0.18881   4.652 3.29e-06 ***
JobRoleLaboratory.Technician1     1.11239    0.22144   5.023 5.08e-07 ***
JobSatisfaction41                -0.78661    0.19136  -4.111 3.95e-05 ***
OverTimeYes1                      1.69452    0.17046   9.941  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.58  on 1469  degrees of freedom
Residual deviance:  999.06  on 1459  degrees of freedom
AIC: 1021.1

Number of Fisher Scoring iterations: 6
```

Exhibit 7. Predicted Employee Attrition by Department

Department	**Predicted Attrition**
Human Resources	2
Research & Development	17
Sales	13
Total	**32**

Exhibit 8. Predicted Employee Attrition for by Overtime

Overtime	**Predicted Attrition**
No	7
Yes	25
Total	**32**

Exhibit 9. Predicted Employee Attrition by Stock Option

Stock Option	**Predicted Attrition**
0: ‘None’	24
1: ‘Some’	4
2: ‘More’	4
Total	**32**

Exhibit 10. Predicted Employee Attrition by Job Level

Job Level	**Predicted Attrition**
Entry Level	20
Mid Level	2
Senior Level	8
Executive Level	1
Senior Executive Level	1
Total	**32**

Project Workload Allotment

	Alper Sayiner	Annabelle Nguyen	Sophia Rowland	Vatsal Sanghavi
Business Understanding		X		
Data Understanding			X	
Data Preparation	X	X	X	X
Modeling	X	X	X	X
Evaluation	X	X	X	X
Deployment	X	X	X	X
R Script	X	X	X	X