JULY 2018

TEXAS
The University of Texas at Austin

MSBA - STATISTICS GROUP 2

IBM HR – EMPLOYEE ATTRITION

Vincent Kuo
Clay Mason
Stella Sun
Kimmy Zhuo
Candice Zuo

Contents

- Project Overview
- Data Dictionary
- Exploratory Analysis
- Models and Findings
- Summary

IBM HR Data Project Overview

- **Objective:** Create a model that accurately predicts which employees will leave the company
- **Source**: IBM Data Science Team [Kaggle.com - HR Analytics]

Data	Count	Note
Observations	1,470	
# of Variables	35	
# of Variables Removed	4	1. Employee Count, 2. Employee Number 3. Standard Hours, 4. Over 18
# of Variables with Null Values	0	
# of Variables with Numeric Values	23	8 are ratings on a scale (e.g. 1=low, 4= high)
# character columns converted to factors	8	1. Attrition 2. Business Travel 3. Department 4. Education Field 5. Gender 6. Job Role 7. Marital Status 8. Overtime Status

Data Dictionary

Variable	Type	Note
Attrition	String	Yes or No
Business Travel	String	Non_Travel Travel_Rarely Travel_Frequently
Department	String	HR R&D Sales
Education Field	String	Human Resources Life Sciences Marketing Medical Other Technical Degree
Gender	String	Male Female

Variable	Type	Note
Job Role	String	Healthcare Rep. Human Resources Laboratory Tech. Manager Manufacturing Dir Research Dir Research Scientist Sales Executive Sales Rep
Marital Status	String	Single Married Divorced
Overtime Status	String	Yes or No
Education	Integer - Scale	1 (Below College) 2 (College) 3 (Bachelors) 4 (Masters) 5 (Doctorate)

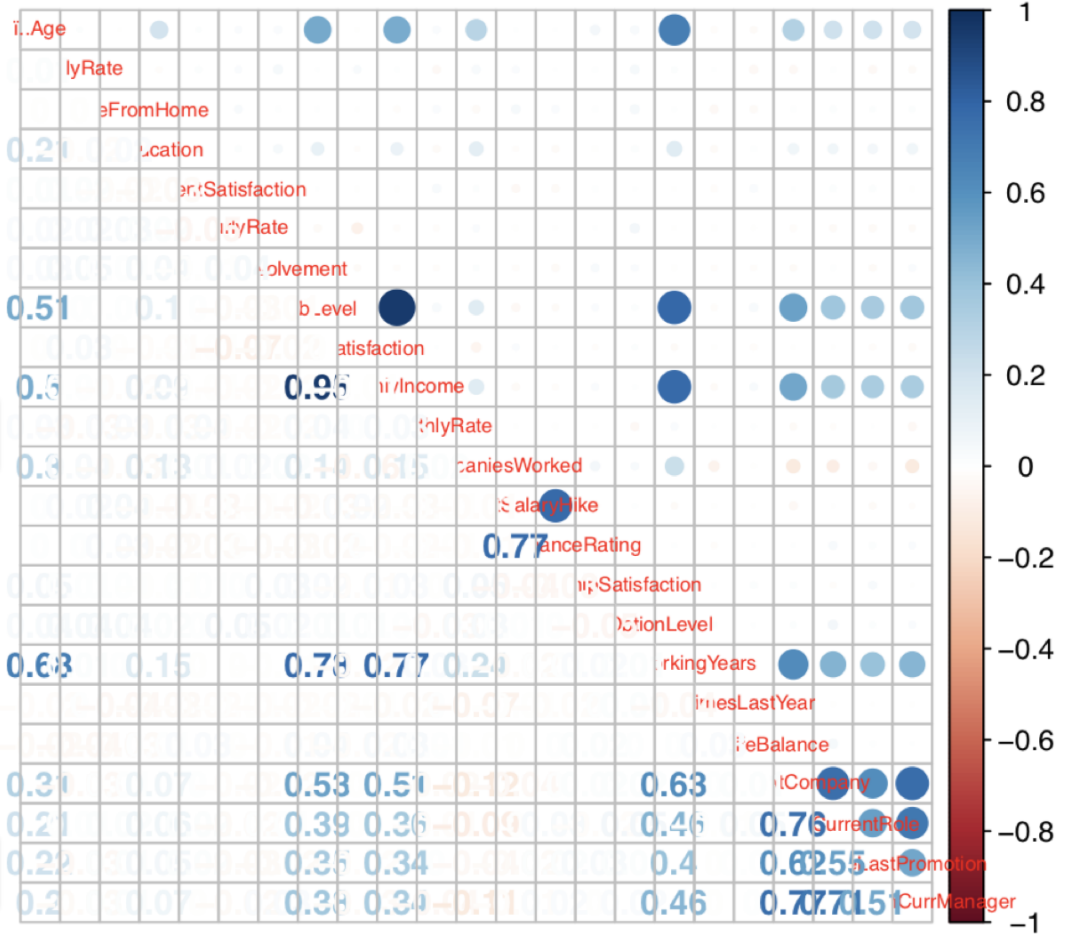
Data Dictionary

Variable	Type	Note
Environment Satisfaction	Integer – Scale	1(low) – 4 (high)
Job Satisfaction	Integer – Scale	1(low) – 4 (high)
Job Involvement	Integer – Scale	1(low) – 4 (high)
Job Level	Integer – Scale	1(low) – 5 (high)
Performance Rating	Integer – Scale	3(low) – 4 (high)
Relationship Satisfaction	Integer – Scale	1(low) – 4 (high)
Stock Option Level	Integer – Scale	0(low) – 3 (high)
Work Life Balance	Integer – Scale	1(low) – 4 (high)
Age	Integer	
Daily Rate	Integer	
Distance From Home	Integer	
Hourly Rate	Integer	
Monthly Income	Integer	
Monthly Rate	Integer	
Num Companies Worked	Integer	
Percent Salary Hike	Integer	
Total Working Years	Integer	
Training Times Last Year	Integer	
Years At Company	Integer	
Years In Current Role	Integer	
Years Since Last Promotion	Integer	
Years With Curr Manager	Integer	

Variable	Type	Note
Employee Count	Integer	REMOVED
Employee Number	Integer	REMOVED
Over 18	String	REMOVED
Standard Hours	Integer	REMOVED

Exploratory Analysis Correlation Matrix

Variable Relationship	Correlation
Monthly Income & Job Level	95.0%
Total Working Years & Job Level	78.2%
Total Working Years& Monthly Income	77.3%
Performance Rating & Percent Salary Hike	77.4%
Education & Total Working Years	14.8%
Education & Num Companies Worked	12.6%
Education & Percent Salary Hike	9.5%
Work Life Balance & Age	-2.1%
Num Companies Worked & Job Satisfaction	-5.6%
Training Times Last Year & Over Time	-7.9%
Num Companies Worked & Years With Curr Manager	-11.0%
Num Companies Worked & Years At Company	-11.8%

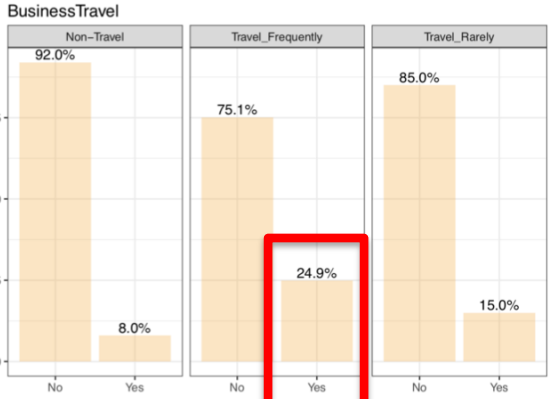


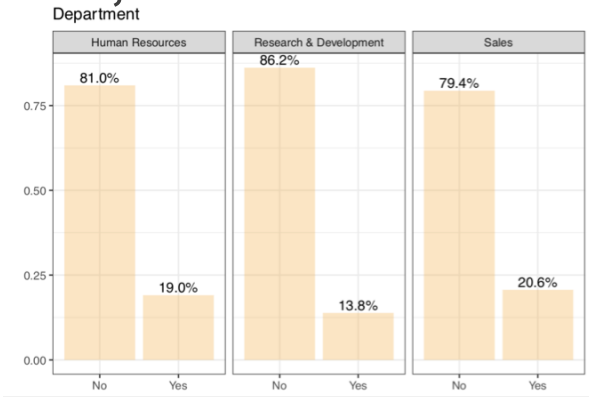
Exploratory Analysis – Population

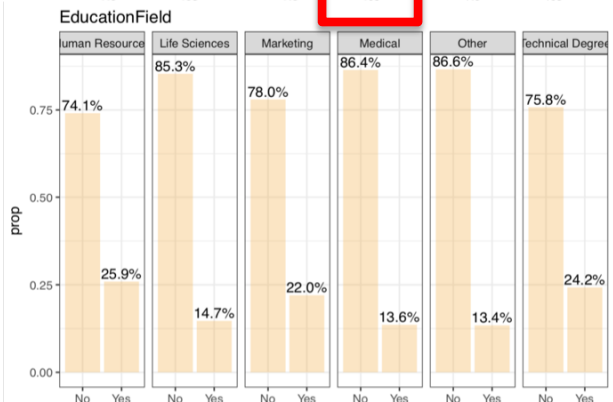
Variable		% of Sample	No	Yes
	Overall Sample	**100%**	**84%**	**16%**
Business Travel	**No Travel**	**10%**	92%	8%
	Rarely	**71%**	75%	25%
	Travel Frequently	**19%**	85%	15%
Department	**Human Resources**	**4%**	81%	19%
	Research & Development	**65%**	86%	14%
	Sales	**30%**	79%	21%
Education Field	**Human Resources**	**2%**	74%	26%
	Life Sciences	**41%**	85%	15%
	Marketing	**11%**	78%	22%
	Medical	**32%**	86%	14%
	Other	**6%**	87%	13%
	Technical Degree	**9%**	76%	2%
Gender	**Female**	**40%**	85%	15%
	Male	**60%**	83%	17%
Job Role	**Healthcare Representative**	**9%**	93%	7%
	Human Resources	**4%**	77%	23%
	Laboratory Technician	**18%**	76%	24%
	Manager	**7%**	95%	5%
	Manufacturing Director	**10%**	93%	7%
	Research Director	**5%**	98%	3%
	Research Scientist	**20%**	84%	16%
	Sales Executive	**22%**	83%	17%
	Sales Representative	**6%**	60%	40%

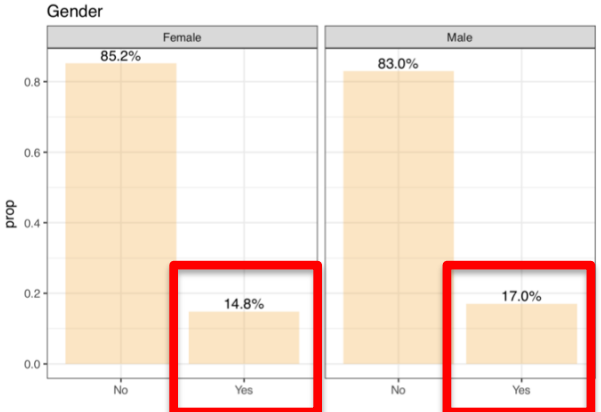
Exploratory Analysis –

Business Travel, Department, Education Field, Gender



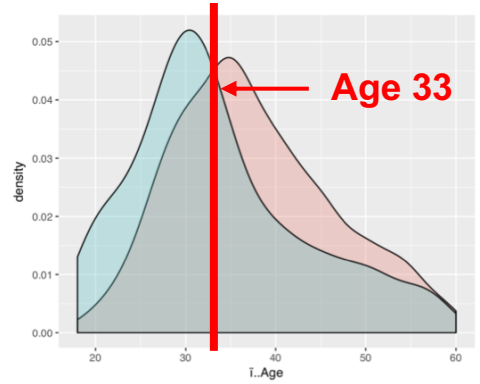


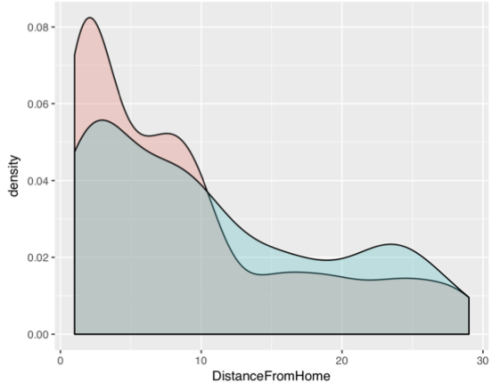


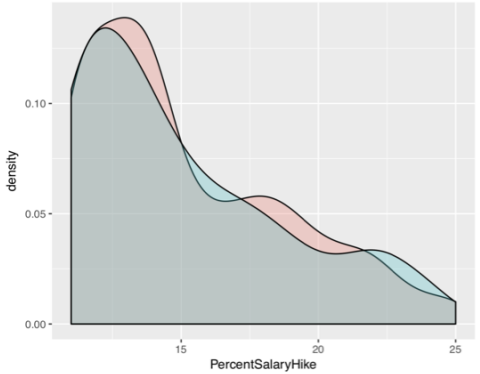


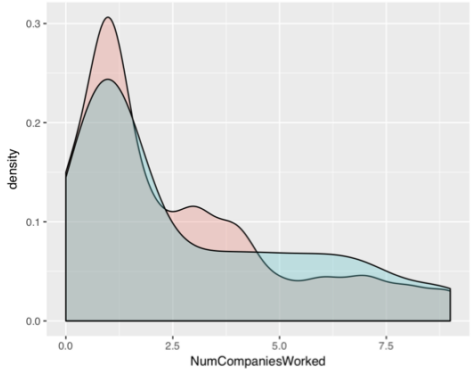
Exploratory Analysis –

Age, Distance From Home, % Salary Hike, # of companies worked



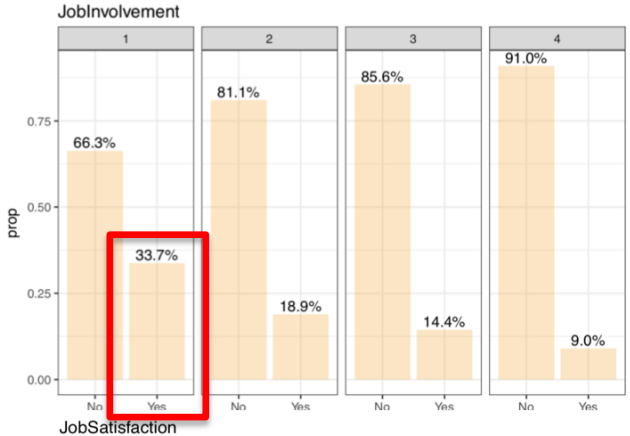


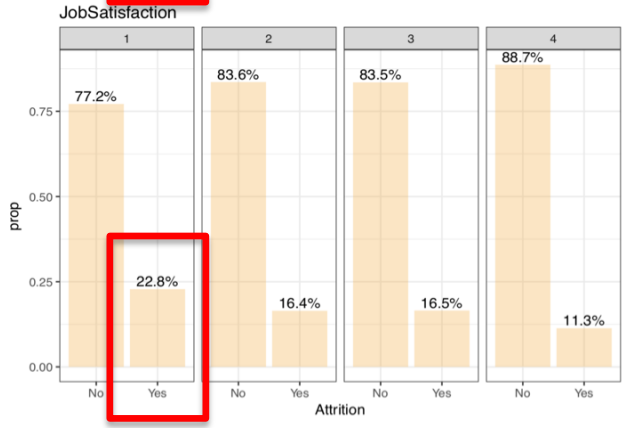


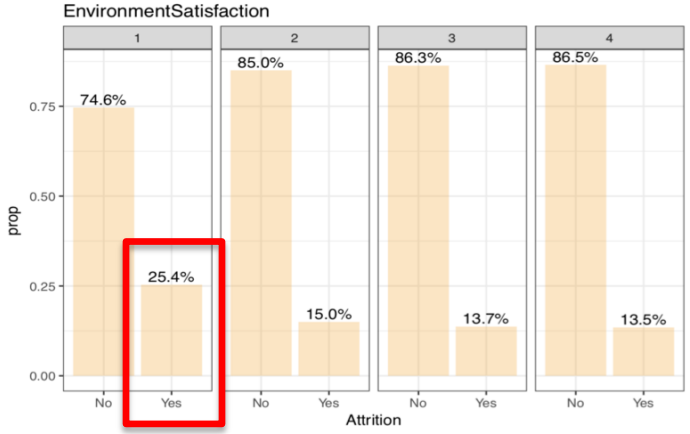


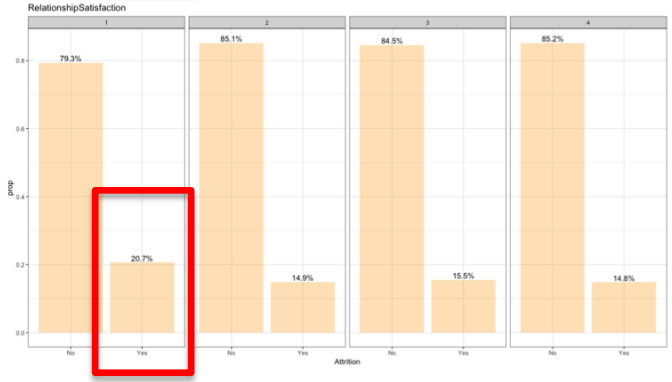
Exploratory Analysis –

Job Involvement. Environment Sat., Job Sat., Relationship Sat









Analysis and Findings – Model Overview

- Logistic Regression
- Boosting
- Random Forest
- Tree
- K Nearest Neighbors
- Naïve Bayes
- Linear Discriminate Analysis

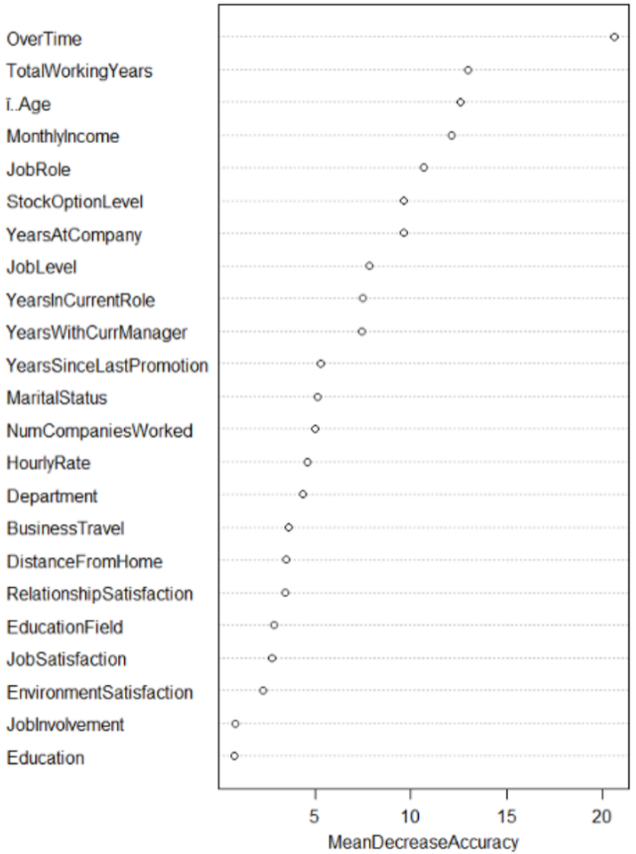
Analysis and Findings – Model Selection

- Unbalanced Attrition Sample
  - 84% No, 16 % Yes
- Error Measurement
  - % of People who leave IF we predict they will *leave*
  - % of People who leave IF we predict they will *stay*

Analysis and Findings –

Models: Tree, Boosting, and Random Forest

- Tree
  - Size = 3, NO: p>0.65
  - criterion 1 = .41, criterion 2 = .11
- Boosting
  - CV on lambda, d = 1 or 2
  - criterion 1 = .76, criterion 2 = .14
- Random Forest
  - m =sqrt(p), ntree = 500
  - criterion 1 = .78, criterion 2 = .14



Analysis and Findings –

Other Models Explored: knn, lda, nb

- K Nearest Neighbors
  - 24 predictors; k=5;No: p>0.6
  - criterion 1: 0.667; criterion 2: 0.141
- Linear Discriminant Analysis
  - 25 predictors; No: p>0.35
  - criterion 1: 0.821; criterion 2: 0.126
- Naïve Bayes
  - 30 predictors; No: p>0.1
  - criterion 1: 0.655; criterion 2: 0.136

Analysis and Findings –

Best Model: Logistic Regression

- stepwise selection: 18 variables left

		Actual	
		NO	YES
Predicted	NO	364	55
	YES	2	20

```
log.fit = glm(Attrition ~ OverTime + JobRole + JobInvolvement +
                MaritalStatus + JobSatisfaction + EnvironmentSatisfaction +
                BusinessTravel + DistanceFromHome + YearsInCurrentRole +
                YearsSinceLastPromotion + TrainingTimesLastYear + ï..Age +
                NumCompaniesWorked + RelationshipSatisfaction + WorkLifeBalance +
                YearsWithCurrManager + YearsAtCompany + TotalWorkingYears, data = train_data, family="binomial")
```

Model Summary (1-3 of 7)

Model	Model Formula (Shows Excluded Variables)		Accuracy	% of People - LEAVE If Prediction - WILL LEAVE	% of People - LEAVE If Prediction - WILL STAY
Logistic Regression	'Attrition~. - Daily Rate - Department - Education - Education Field - Gender - Hourly Rate	- Job Level - Monthly Income - Monthly Rate - Percent Salary Hike - Performance Rating - Stock Option Level	**0.871**	**0.909**	**0.131**
Linear Discriminant Analysis	'Attrition~. - Monthly Income - Performance Rating	- Education - Department - Gender'	0.871	0.821	0.126
Boosting	'Attrition~. - Performance Rating	- Gender - Department'	0.859	0.760	0.135

Model Summary (4-7 of 7)

Model	Model Formula (Shows Excluded Variables)		Accuracy	% of People - LEAVE If Prediction - WILL LEAVE	% of People - LEAVE If Prediction - WILL STAY
Random Forest	'Attrition~. - Monthly Rate - Performance Rating - Training Times Last Year	- Gender - Work Life Balance - Percent Salary Hike - Daily Rate'	0.871	0.679	0.136
knn	'Attrition ~ . - Job Level - Gender - Daily Rate	- Department - Percent Salary Hike'	0.844	0.667	0.141
Naïve Bayes	'Attrition~.-		0.850	0.655	0.136
Tree	'Attrition~.-		0.841	0.414	0.110

Questions?