STA545: Statistical Data Mining: Final Project Presentation

Analysis of IBM's HR Dataset

Arun Krishnamurthy   Ashwin Vijayakumar

Department of Computer Science and Engineering
State University of New York, Buffalo

December 1, 2017

Motivation

The idea behind this study is to look at the different factors that lead to attrition and to come up with a model that, with some degree of accuracy, will be able to predict if a given employee is going to quit.

Overview of the Dataset Used

This is a dataset created by IBM's Watson Analytics group. It has 35 different attributes on 1470 employees.

Age	Attrition	BusinessTravel
Department	DistanceFromHome	Education
EmployeeCount	EmployeeNumber	EnvironmentSatisfaction
HourlyRate	JobInvolvement	JobLevel
JobSatisfaction	MaritalStatus	MonthlyIncome
NumCompaniesWorked	Over18	OverTime
PerformanceRating	RelationshipSatisfaction	StandardHours
TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance
YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
DailyRate	EducationField	Gender
JobRole	MonthlyRate	PercentSalaryHike
StockOptionLevel	YearsAtCompany	

Table: Attribute Names

Attrition Rate

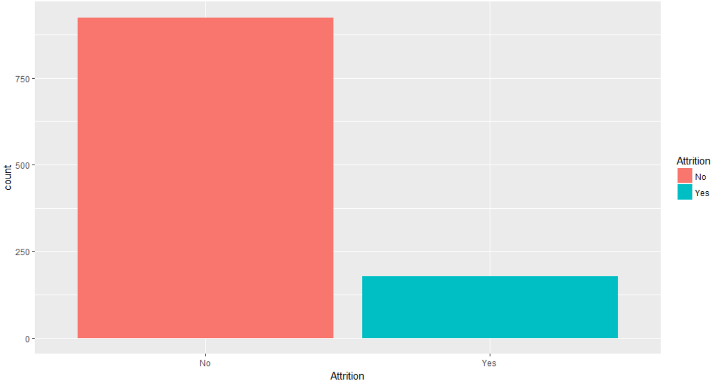


Figure: Rate of attrition

Note

About 16.13% of the employees leave the company.

Influence on Attrition

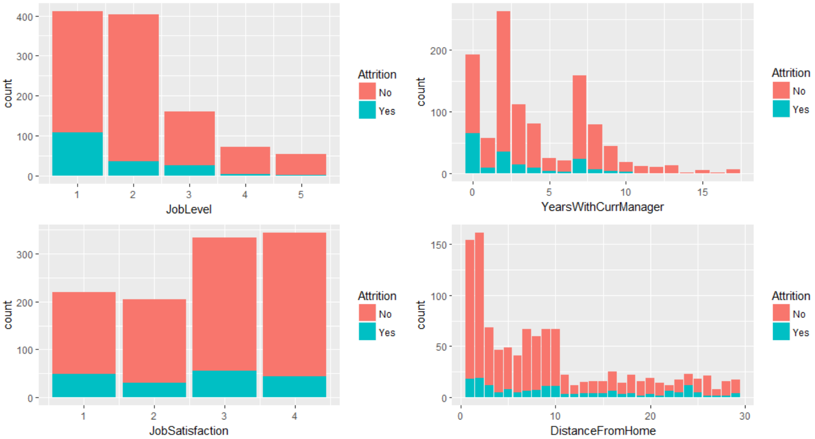


Figure: How different variables influence attrition

Data Cleaning

The variables that were removed after initial exploration of data

1. EmployeeCount
2. EmployeeNumber
3. StandardHours
4. Over18

Feature Engineering

Based on domain knowledge and intuition, we are adding two new features that may help improve the prediction.

1. TenurePerJob:
   (TotalWorkingYears / NumCompaniesWorked)
2. YearsWithoutChange:
   (YearsInCurrentRole - YearsSinceLastPromotion)

Methods Used

The different models that are used to compare accuracy

1. Logistic Regression
2. k-Nearest Neighbour
3. Recursive Partitioning
4. Random Forests
5. Support Vector Machine
6. Extreme Gradient Boosting

Recursive Partitioning

1. Recursive partitioning creates a decision tree that strives to correctly classify members of the population by splitting it into sub-populations based on several dichotomous independent variables.
2. The process is termed recursive because each sub-population may in turn be split an indefinite number of times until the splitting process terminates after a particular stopping criterion is reached.

Recursive Partitioning

1. Advantages
    1. Generates clinically more intuitive models that do not require the user to perform calculations.
    2. Allows varying prioritizing of misclassifications in order to create a decision rule that has more sensitivity or specificity
2. Disadvantages
    1. Does not work well for continuous variables
    2. May overfit data

Random Forests

1. Random Forests improve variance by reducing correlation between trees, this is accomplished by random selection of feature-subset for split at each node
2. Forests give results competitive with boosting and adaptive bagging, yet do not progressively change the training set
3. parameters = number of trees, number of bagged variables

Random Forests

1. Advantages
    1. Combines weak learners to give a strong one
    2. Doesn't generally overfit because of the law of large numbers
2. Disadvantages
    1. Low interpretability

XGBoost

XGBoost is short for “Extreme Gradient Boosting”, where the term “Gradient Boosting” is proposed in the paper Greedy Function Approximation. It’s a scalable system for learning tree ensembles.

$$
\begin{aligned}
\text{obj}^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant
\end{aligned}
$$

Figure: Objective function

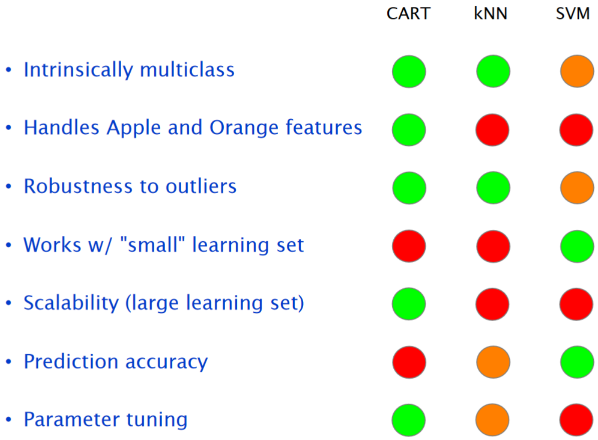
$$
\begin{aligned}
\text{obj}^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + constant
\end{aligned}
$$

Figure: Objective function with loss = MSE

XGBoost

1. Advantages
    1. Quick computation . Atleast 10 times as fast as existing gradient based boosting algorithms
2. Disadvantages
    1. Low interpretability

Model Comparison



	CART	kNN	SVM
• Intrinsically multiclass	🟢	🟢	🟠
• Handles Apple and Orange features	🟢	🔴	🔴
• Robustness to outliers	🟢	🟢	🟠
• Works w/ "small" learning set	🔴	🔴	🟢
• Scalability (large learning set)	🟢	🔴	🔴
• Prediction accuracy	🔴	🟠	🟢
• Parameter tuning	🟢	🟠	🔴

Accuracies

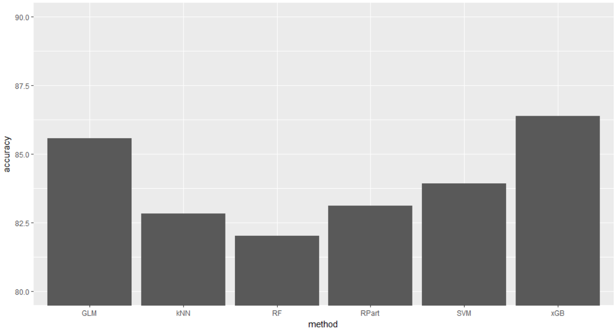


Figure: Accuracy level for different methods

Note

Highest is xGB with 86.38% accuracy.

Receiver Operating Characteristic (ROC) curves

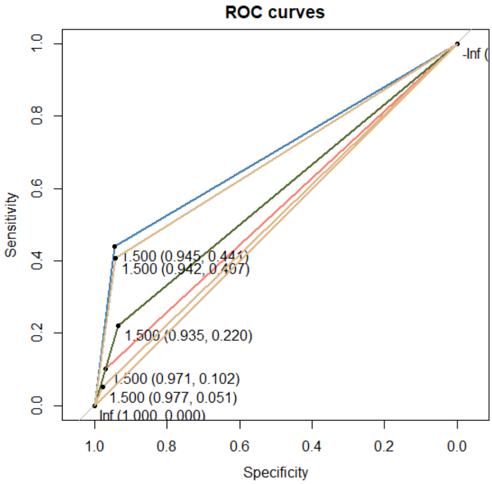


Figure: ROC curves

LR	0.6742
kNN	0.5141
RPart	0.5362
RF	0.5777
SVM	0.5
xGB	0.6927

Table: AUC

Calculation of variable importance for XGB

The five most important variables for the XGB fit were:

1. OverTime
2. Salary related variables (MonthlyIncome, DailyRate, MonthlyRate)
3. TotalWorkingYears
4. YearsWithCurrManager
5. StockOptionLevel

Attrition Probability by Job Role

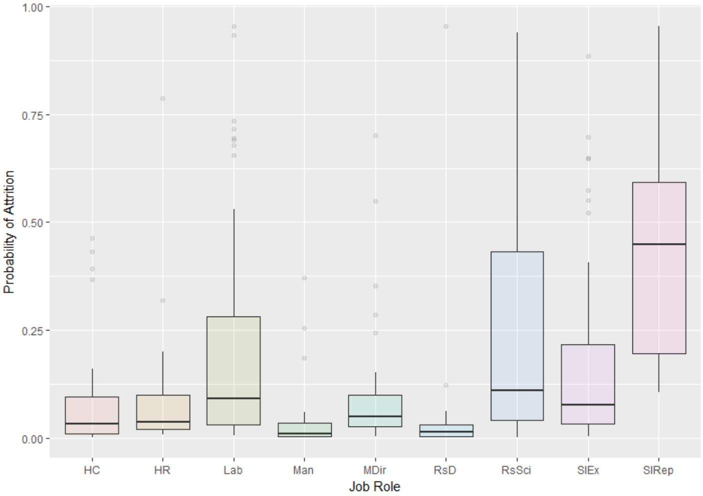
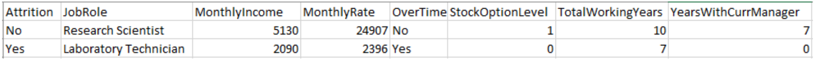


Figure: Probability of attrition based on job role

Predicting on New Data

When new data is presented to the model, the behavior is as expected.



Attrition	JobRole	MonthlyIncome	MonthlyRate	OverTime	StockOptionLevel	TotalWorkingYears	YearsWithCurrManager
No	Research Scientist	5130	24907	No	1	10	7
Yes	Laboratory Technician	2090	2396	Yes	0	7	0

Figure: New Data

Recommendations

1. Thorough audit of the sales department to ascertain why about 50% of Sales Representatives are at risk of quitting.
2. Explore why Research Scientists are quitting. Increase recruitment in this area so as to not fall behind competition in R&D
3. Reconsider the overtime policies of the company
4. Data shows that salary is a highly important variable to attrition. The recommendation is to match salaries for important employees when there are competing offers.
5. Reduce shuffling of managers between teams as a change of manager causes attrition.