Attrition for IBM Employees

Team C41: Keena Desai, Peter Law, Yash Kanoongo, Crystal Li, Sai Kiran Reddy

Business Understanding

Attrition is defined as when an employee resigns or retires from a company. Training a new employee is a costly and long process, it is in a company's best interest to decrease employee attrition. Through exploring the IBM dataset, we would like to analyze factors that would be highly suggestive/predictive of an employee leaving the company. The analysis can give the HR department and management some insights into some of the reasons that employees decided to leave the organization. They can then implement any changes that might incentivize employees to continue to work with the company. Moreover, this model can also help change hiring practices, by anticipating the qualities in individuals that would make them more likely to stay with the company for a longer period of time and hiring accordingly. We will be making use of unsupervised models such as PCA and K-means to gain an understanding of the variables in our dataset, and supervised models such as regression modelling, random forest, lasso and post lasso and neural network to establish which variables best explain attrition within IBM.

Data Understanding

The dataset was created by IBM employees and was downloaded from Kaggle. The dataset is fictional and that data does not actually represent any actual IBM employees. The dataset consists of 1470 observations and 35 variables. This is supervised learning as we have identified the target outcome as attrition. We anticipate the variables that relate to compensation (ex: percent salary hike, monthly income), overtime work, age, and tenure with the organization will be the most useful. One thing we noticed was that majority of the Attrition values were “No” in the dataset, however, we were not as concerned about it as the difference between the “No” and “Yes” values were not extreme.

Data Preparation

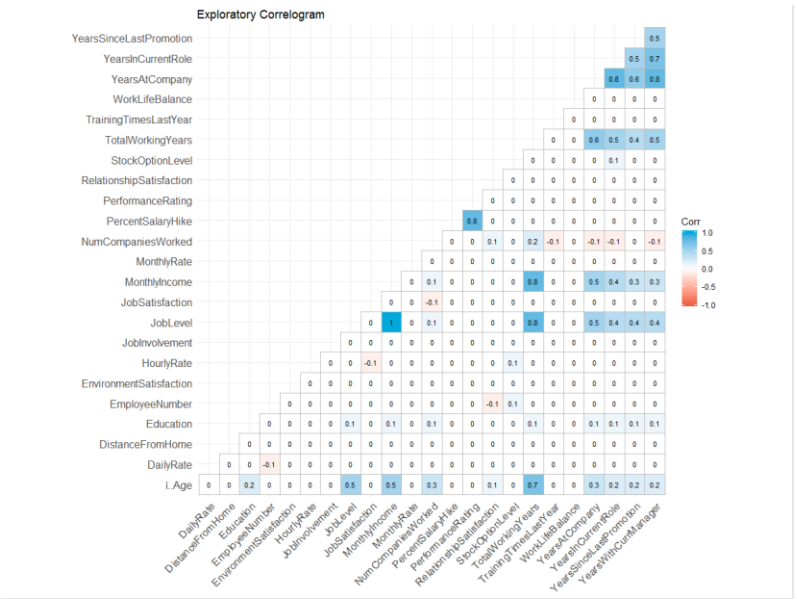
In terms of data preparation, there were no missing values in the dataset. There were, however, many categorical variables that were converted into dummy variables. Furthermore, there were three variables "Attrition", "OverTime" and "Gender" that were coded as "yes" or "no" and had to be converted to a 0 and 1. There were also three variables "Over18", "Standard Hours" and "Employee Count" that all had the same value for each observation, so we decided to remove these variables. Furthermore, we also had to rename certain variables due to issues with spacing and hyphens. In order to perform cross validation and get out of sample results we split up the data into a train and test set. We divided this data based on 80% sample data and 20% holdout data. In order to check for balance, we took the means of cases for where attrition was 1 in both the train and test set to see if they were close together. As shown below the means were relatively similar, so we can assume balance.

mean(IBMdata2$Attrition==1	mean(IBMdata2.holdout$Attrition==1)
0.1632653	0.1530612

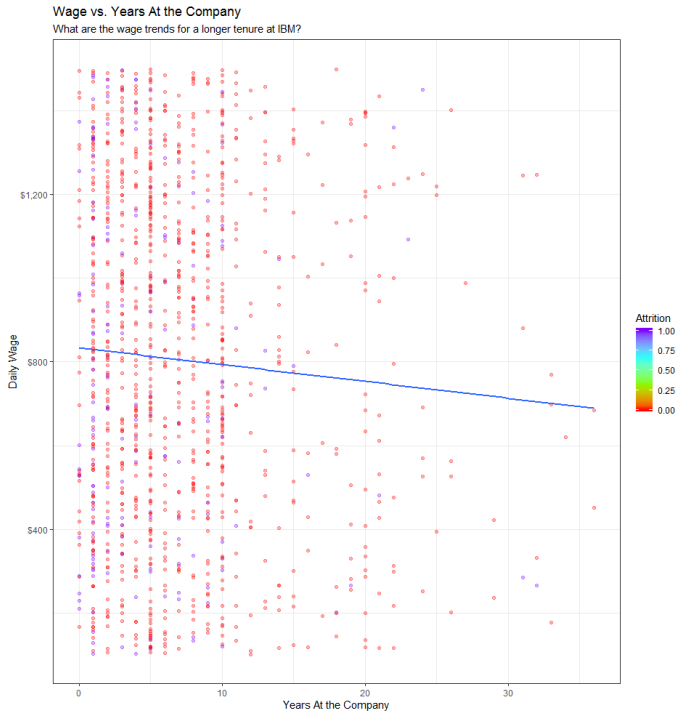
Data Exploration

A. Correlation Plot

We created a correlation matrix to preliminarily understand the correlations between the different variables in our dataset. From the matrix, we observe that total working years is positively correlated with the years with current manager and years since last promotion. We also observe some correlation between monthly income and years in current role and years at the company. The correlation plot can be observed below



B. Data Visualization



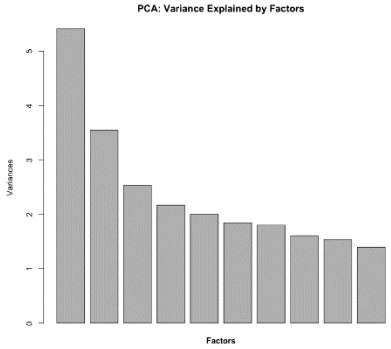
In order to further explore the relationships between the various variables in our dataset, we identified certain variables that might intuitively explain some of the Attrition within IBM. The boxplot is broken into two parts in order to develop some intuitive reasoning behind the reasons and timing of Attrition in IBM. From the first boxplot, we can see that the median avg daily salary of an employee, peaks at year 7 when working under the same manager and is observed to be highly variable thereafter. From the second boxplot in the figure, we expectedly notice higher attrition within the first 5 years of working under the same manager, possibly in anticipation of declining or highly variable salaries in the future. Alternatively, if an employee is under the same manager for 7+ years, it could also possibly mean that he/she is not getting promoted, explaining

the decline in median salaries. Therefore, in order to further confirm our intuition, we expanded the scope to all employees, not just the ones managed by the same manager over several years. Observing the trend of the scatter plot, we can observe that Daily Wage of employees, on average, is expected to drop, the longer they stay at IBM. This explains the high Attrition Rate in the first 5 – 10 years (as shown by the purple color points in the figure). It appears that the employees, after spending some years within the organization, tend to realize that their salaries will decline and their value to the organization will decrease, which results in them leaving to pursue other opportunities.

C. Unsupervised Modelling:

Principal Component Analysis

We looked at PCA to find patterns and extract the latent features from our dataset.



Principal Component (PC)	Percentage of Variation Explained	Cumulative Variation
PC1	10.61%	10.61%
PC2	6.95%	17.55%
PC3	4.98%	22.53%
PC4	4.25%	26.78%

The histogram and the table above indicate that the drop off in variation between principal components is not very drastic, which hints at the interpretation that PCA might not be the most accurate unsupervised model. As seen above, our first four PC's explain only 26.78% variation in our model.

<u>First Principal Component</u>

Original Features	Loading
JobLevel	-0.3788066
MonthlyIncome	-0.3730580
TotalWorkingYears	-0.3707950
YearsAtCompany	-0.3397165

The first principal component explains the relationship between Salary and Work Experience. There is a high negative value for Monthly Income, Total Working Years, and Years at Company.

This is likely because of the observed negative relation between wages and years of work experience. We did not consider the other Principal Components as they do not explain significant variation in our data.

K-Means

We also looked at k-means and found only 17% variation explained with four clusters. We then looked at how the clusters related to attrition and found the following:

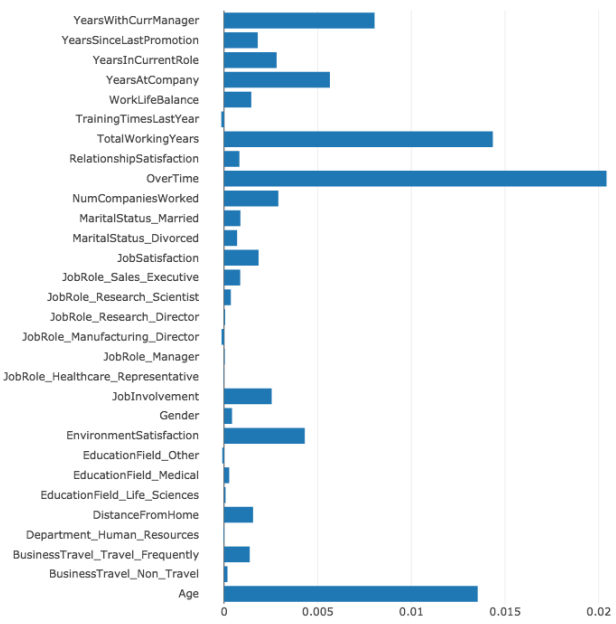
Cluster	IBMdata2$Attrition == 1
1	0.15555556
2	0.23417722
3	0.16612378
4	0.04477612

Both our unsupervised models point to the fact that no cluster of variables explain a large amount of variation in our dataset. Since the variation is evenly distributed across variables in our dataset, clustering methods are not very suitable for our dataset.

Modeling

As this is a classification problem, we chose to build several supervised models and test the out of sample R2 for these models.

A. Random Forest



The Random Forest that we ran used 500 trees with 10 variables split at each iteration. The random forest performs better than the null model but is not as good as the logistic regression. However, the random forest helps explain the significance of the variables by using the percentage increase in mean squared errors. The comparison of all the variables is shown on the left with their

Mean Square Error Percentage. We can see that older aged employees, employees working over-time and long serving employees are the ones likely to quit.

B. Logistic Regression

We chose to use logistic regression since it can provide us with the probability of attrition, and this is a classification problem. We used both forwards and backwards stepwise function in order to get a logistic regression model. The step model selected 30 variables and received an AIC of 737.35. Some notable variables that we saw that had a large coefficient and statistically significant p value include: OverTime, Job Role Healthcare Representative and Job Role Research Scientist.

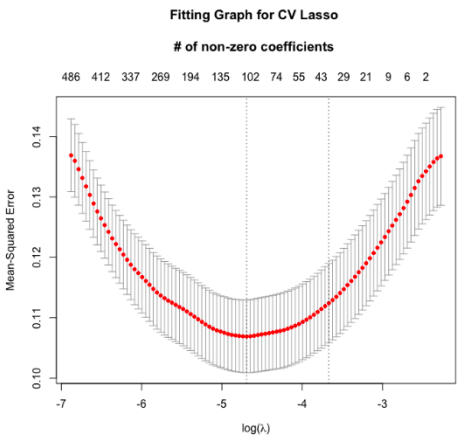
C. Decision Tree

We built a decision tree using the variables from the step wise logistic regression. However, with all the statistically significant variables of the logistic regression, the performance of the decision tree is worse than the null model. The ROC curve for decision tree can be misleading and the residual sum of squares and the R-squared values show us that the performance of the decision tree is sub optimal.

D. Logistic Regression with interactions

We used second degree interactions with all the possible pairs and performed a logistic regression to predict the attrition rate. The model performs worse than the null model. The number of variables used is very high and hence, the model suffers from too many correlations without causations. The bad performance combined with the complication of too many variables to evaluate the model does not make this model a good one.

E. Lasso

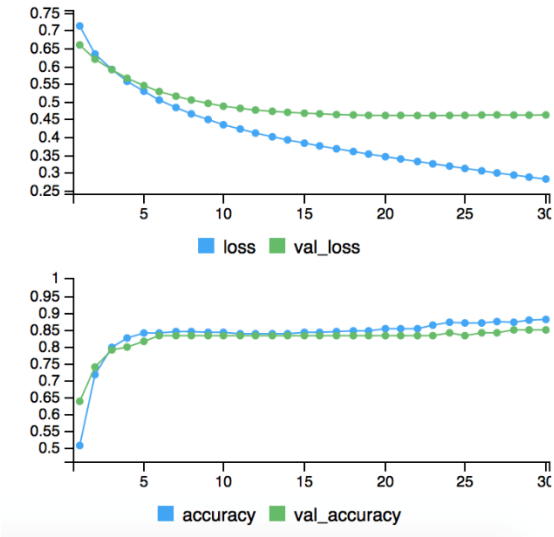


Lasso is a good way to help us select variables. We chose to run a lasso on an interaction model with minimum MSE optimization selection of 99 variables. The out of sample R^2 performance was 0.35 and was achieved using aggressive variable selection(feature.min). The downside of this model is that it selected too many variables, and apart from 'JobInvolvement' all the other 98 variables are interaction terms, which means the influence of the variables are confounded.

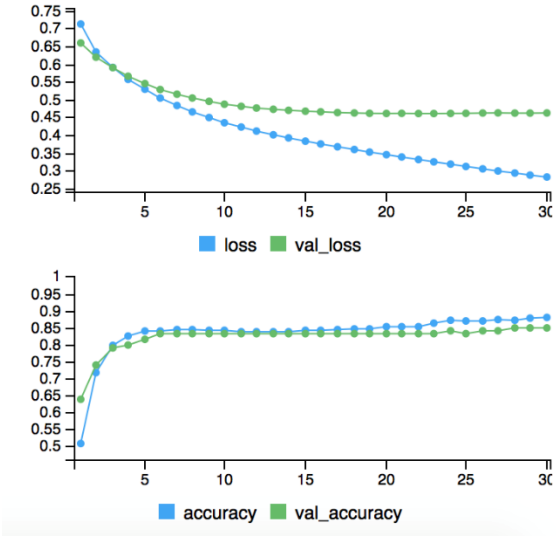
F. Post Lasso

The Post Lasso performance is better with variable selection than Lasso by theory, conservative and aggressive approaches. The aggressive approach under Post Lasso performs the best and has the highest R2. We calculated the lambda min and lambda 1se and created features for both the cases. Using these lambda thresholds, we could analyze the performance metric of interest which is the R2 along with all the other models. Post Lasso with aggressive variable selection has the best overall performance with an out of sample measure of 0.401.

G. Neural Network



Apart from the basic data mining models above, we also chose to apply advanced Neural Network model. The biggest advantage of using this model is that it generates a high prediction accuracy of the employee attrition level. However, it is hard to know how the model weighs each feature, which is the essence of our business problem. In our case, the prediction accuracy is 85.7%, which is much higher than other basic models. After adding regularizer or the dropout, the accuracy doesn't chance much. The graph above shows the training error after iteration and the bottom is showing the prediction accuracy for the holdout sample.



Evaluation

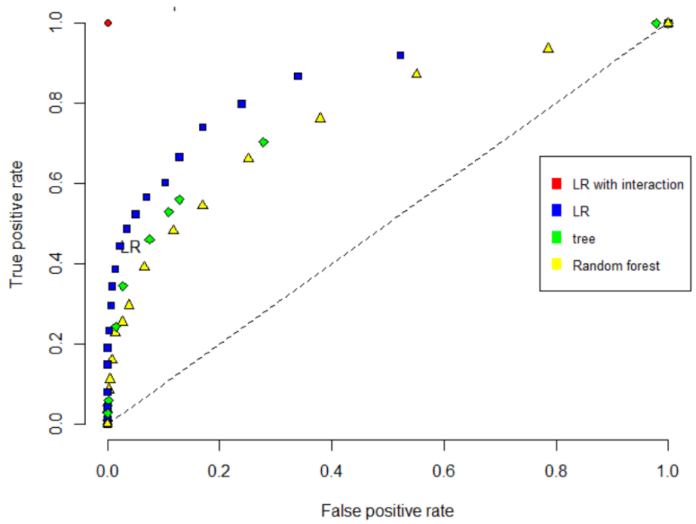
We are using R-Squared as a performance metric to compare different models and methods from variable selection among the models. The null model shows that the R-squared value is approximately equal to

zero and we compare our models with the null model as the baseline. The ROC curve shows that Logistic Regression has the best True Positive and False Negative Rates. Furthermore, it also shows the best R-squared value before performing k-fold cross validation and Lasso. However, Lasso and Post Lasso help us with better variable selection and boosts the R-squared, especially aggressive variable selection in 10-fold cross validation. In order to understand statistical significance of variables on the attrition rate, the random forest performed with 500 tree iterations helps us evaluate the MSE percentages. However, the R-squared value is not as high as Logistic regression, so we can use the MSE percentages (%IncMSE) to evaluate statistical significances. The model suggests that Age, higher Number of working years, and OverTime are significant influencers of attrition rate.

R2 as a performance evaluation metric.

PostLasso Aggressive	PostLassoConservative	PostLassoTheory	LassoAggressive	LassoConservative	LassoTheory	logistic.interaction	logistic	tree	null	RandomForest
0.4012946	0.3576926	0.2849117	0.3598047	0.2375556	0.1784237	-1.714308	0.3679109	0.06874832	-5.397381e-05	0.2354653

ROC curve

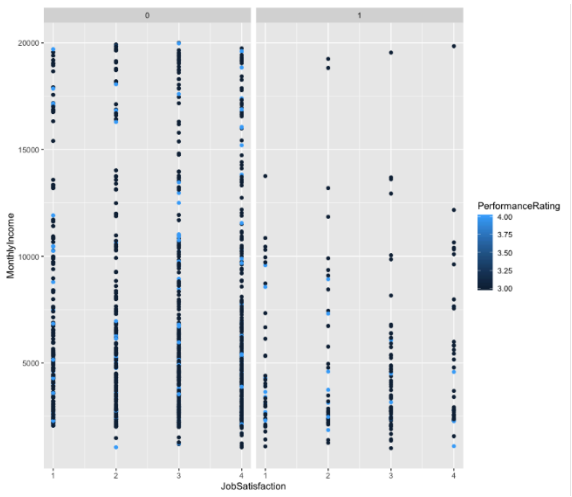


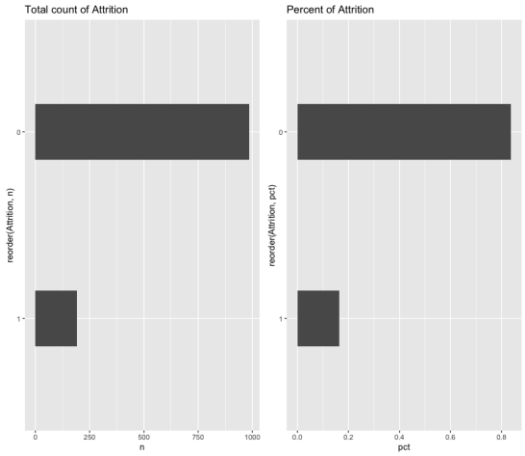
Deployment

The resulting model would be the most useful to IBM's human resource department. Given that the logistic regression model was the best predictor for attrition within IBM, we suggest IBM to pay close attention to two variables explaining a significant amount of attrition in the organization. The most important being Overtime. Through our model, we observed that employees that worked overtime have the highest likeliness to leave the organization. Alternatively, it was also observed that no business travel was associated with strong negative correlation with Attrition, explaining that the employees preferred not having to travel for work. Our models can thus give IBM some insight into the attributes of their employees and can assist them in building policies keeping these attributes in mind. Some concerns and ethical consideration need to be kept in mind when using our models as a basis for decision making in the organization. Our performance metric, R^2 is observed to be low across models. Thus, while our models should help IBM understand some of the reasonings behind why employees leave their organization, there is potential for confounding factors that may be playing a large but unexplained role leading employees to leave the company. Finally, when it comes to data associated to personal information, there are always ethical and data privacy concerns to keep in mind.

Appendix

Various other visualizations





Keena Desai - Business Understanding, Data Understanding, Data Preparation, Modeling

Peter Law- Deployment, Presentation

Yash Kanoongo- Data Exploration, Presentation, Deployment

Crystal Li- Modeling, Evaluation

Sai Kiran Reddy- Modeling, Evaluation