Employee Retention

ISM6136.020S22 Data Mining

Varsha Issrani

Helen Memoli

Ali Sadeghinia

Grayson Guo

Background of Problem (Helen Memoli)

Employee retention has been a concern as organizations have been competing for quality talent in a tight economy in recent times. According to Gallup, voluntary turnover costs the US business sector almost $1 trillion dollars annually. On top of that astounding figure, about fifty-two percent of those leaving feel their manager or organization could have done something to prevent their leave and it is not necessarily motivated by salary demands (McFeely and Wigert). Employee satisfaction is not only a key indicator of how long someone will stay at the company but also how productive and innovative their outputs will be. This has a direct impact on the bottom line of an organization. Regardless, employee turnover is inevitable. Sometimes though issues discussed in exit interviews can be amplified and purely anecdotal leading to some reactionary changes as opposed to verifiable data used to bridge gaps in an employee's satisfaction in their current roles.

There is a myriad of research illustrating when an employee has a deeper connection with the organization's mission and purpose there is a connection and/or hunger to want to succeed in the current environment (Carucci). It costs a company about 1.5 times a person's salary to replace a position. Coupled with the data showing employees and workforce units that are engaged with the company produce 17% higher productivity, 10% increase in customer satisfaction, and 17% increase in productivity; it is impossible for employers to not see the benefit in uplifting their employee's morale and connection to the company.

Motivation for Solving the Problem (Helen Memoli)

The calculation for turnover is pretty straightforward and is typically done monthly by the HR department. Turnover equals Total Number of Separations (voluntary and involuntary) divided by Average Number of Employees multiplied by 100. This simple formula is a great metric for determining an individual company's turnover rate and then comparing it to standards across industry lines (Holliday). As mentioned before, turnover is expected for various reasons. However, what would be the most beneficial would be a way to predict if an employee is at risk for leaving the company. If an intervention could be initiated; whether it be an increase in salary, decrease in volume of work, change in the department; the company would benefit greatly. The company would see a decrease in the amount of turnover. This would in turn lead to a decrease in the number of resources (either monetary or sweat equity) needed to replace vacant positions. In the age of technology, companies are monitoring their employee's emails and social media interactions to figure out if an employee is likely to leave. But what if there was an easier way and organizations can accurately predict if an employee is likely to leave.

Data mining strategies could also be used to determine if there is a trend in regard to certain company leaders and their management styles. For example, if people on Team A are reporting a markedly higher satisfaction and performance when compared to the rest of the company, the question could be asked what is the team's leadership doing to produce excellence. Conversely the same can be said for lower-performing teams. Companies that listen to their employees and enact the appropriate changes will see an increase in their bottom lines (Jaramillo). This data could also be used to see if there are teams that can benefit from new blood. If a team is too complacent with its structure, this can lead to a lack of innovation.

Our ideal data mining model should be able to answer the following questions:

1. Can it accurately predict if a particular employee will leave or not?
2. What are some of the top factors that influence an employee's decision to leave or stay?

3. Is there any relationship between 2 factors or more and their decision to stay?

Solutions to these questions will allow business leaders to make business decisions to help increase employee retention and decrease turnover costs.

Solution Methodology and Evaluation Metrics (Ali Sadeghinia)

Solution Methodology

Decision Trees and Neural Network models can help HR to accurately predict whether their employee is staying or leaving the company. It also shows the various thresholds at which the employees make these decisions. Under these algorithms, the dataset is going to be split into two: Training (80%) and Testing (20%).

Linear Regression modeling helps with the understanding of which independent variables that are numeric (0.48, 0.15, 1.0,…) have the most impact on whether an employee leaves or stays. To understand this better, a linear regression model between our only two numeric variables while setting the left variable as our dependent variable would help us understand which of the two (Satisfaction Level, Last Evaluation) have more impact on our dependent variable. K-means Clustering is also another great method to gain insights into employee behavior. Knowing the importance of our variables from the previous analyses, we will compare the two variables, 'Left' and 'Satisfaction Level', that are of most importance, to the rest of the variables. By doing so, we can understand the connections between these variables and help HR in many different ways.

Additionally, the outcome from the data of the different methods we use could also help HR understand the reasons why employees leave. This will result in more efficient communication and compensation; the company would be able to find a happy medium with their employees and finances. In other words, the company can understand who they would want to keep according to what they provide and compensate them enough that the company is not providing them with more than what is needed.

Evaluation Metrics

- **Two-Class Boosted Decision Tree and Two-Class Neural Networks:**

The confusion matrix and the ROC curve are two of the most important evaluation metrics. Precision, Recall, and Accuracy are all important indicators of how efficient our model is, but precision would be the most important of all metrics that would need to be maximized. This is due to the fact that the precision metric has a reverse relationship with the company's employee churn rate. The goal with these two models and the metrics from each is to increase the precision of our models while maintaining a desired accuracy for the company.

- **Linear Regression:**

The evaluation metrics that are of importance are to indicate which of our numeric variables is more important and has more impact on the employee's churn rate. Therefore, the variable with a higher Beta Coefficient in the formula will be the chosen variable for K-means Clustering.

- **K-Means Clustering:**

Lastly, we will use this method for visualization and gain a better understanding of the relationship between variables. The K-Means clustering model is typically hard to evaluate however, we can see if the model is effective if we can derive insights from the clusters that the model generated.Business leaders can get a good visualization and gain insights about their employees.

Description of Dataset (Varsha Issrani)

Our team researched for a dataset that can satisfy the needs of what we wanted to achieve. We want to gain insights into the factors that cause employee turnover and help

companies find ways to decrease the costs of employee turnover. We looked at many online data repositories and found a dataset on Kaggle that we can use in our data mining model. It is an open data set that contains data for about 15,000 employees. The data is what typically the Human Resources department would collect.

Variables:

- satisfaction_level: Employee Satisfaction Level (%)
- last_evaluation: The last evaluation they received (%)
- number_of_projects: The number of projects they are working on
- average_monthly_hours: Average Amount of hours they work in a month
- years_at_company: Amount of years spent at the company
- work_accident: Whether they have had a work accident at the job
- left: Whether the employee left the company; 1: left, 0: not left
- Promotion_last_5years: Whether they have had a promotion in the last 5 years
- department: Name of the department they work in
- Salary: Whether the range of their salary is low, medium or high

These variables might not apply across all companies since every organization's structure varies from one on other. However, looking at these variables, a business can use variables similarly that has been highlighted above.

Data Source: https://www.kaggle.com/saurabh0/human-resources-employee-attrition

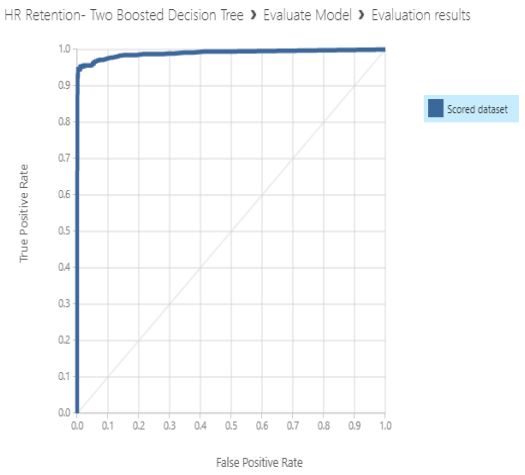
Comparison of Algorithm Models & Summary of Results

Varsha Issrani

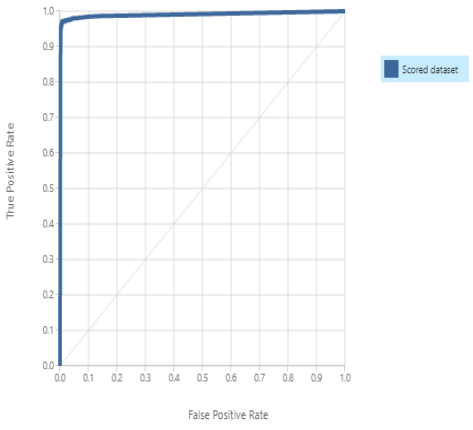
(Two-Class Boosted Decision Tree)

Experiment	Attributes	Accuracy	Precision	Recall	AUC
1	Random Seed: 40943519 Number of Trees: 20 Column Selected: Left	0.983	0.984	0.945	0.990
2	Random Seed: 40943519 Number of Trees: 50 Column Selected: Left	0.987	0.975	0.970	0.993

Experiment 1



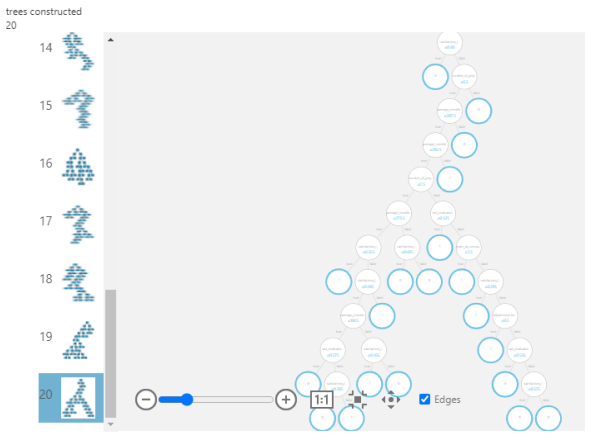
Experiment 2



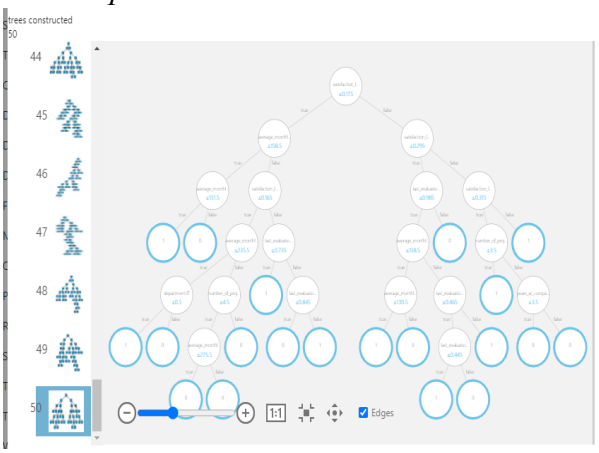
As shown above, both experiments show that the model in play is overall a good model to work with. Both experiments show a large AUC, therefore, the model is effective in distinguishing between negatives and positives. Experiment 2 which has more amount of trees is more close to 1 therefore a more number of trees produced increases the viability of the model. The recall for experiment 2 is higher than 1, but that caused the precision to decrease slightly which is always the case. Both experiments' accuracy is great. In my opinion, if the model has an accuracy of 90% or higher, then the model is effective. It must be noted that the user of this must define the threshold depending on their business objectives. The decision tree clearly illustrated the top reasons for employees to leave

Results:

Experiment 1



Experiment 2



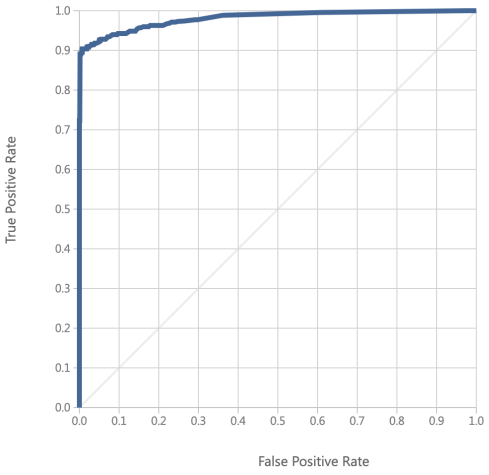
Experiment 2 which produced 50 trees provides a better illustration to clearly predict what are the circumstances for employees to leave or stay. The results of both experiments show that satisfaction level is the main deciding factor for an employee's decision to leave or stay. In experiment 1, if the satisfaction level was less than 11%, they will leave. If not, then the number of projects comes into play. If the satisfaction level was more than 11% but the number of projects exceeded 6, then the employee will also leave and so on. If an organization were to apply this model in their human resources function, they can accurately predict at what levels employees are likely to leave and make effective business decisions surrounding that. Employee satisfaction may vary from one person to another. What one employee values may not be the case for another employee. The experiment looked at all variables in comparison to experiment 1. Therefore, the greater number of trees would make the model more effective.

Helen Memoli (Two-Class Boosted Decision Tree)

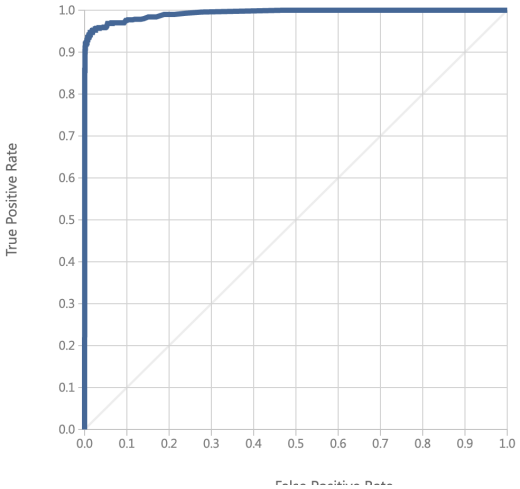
Experiment	Attributes	Accuracy	Precision	Recall	AUC
1	Column Select: leave Max. # of Leaves: 5 Min. # of Sample/Leaf Node: 10 Learning Rate: .2 No. of Trees: 10	.966	.950	.902	.969
2	Column Select: leave Max. No. of Leaves: 25 Min. No. of Sample/Leaf Node: 150	0.979	0.986	0.921	0.993

	Learning Rate: .2 No. of Trees: 50				

Experiment 1

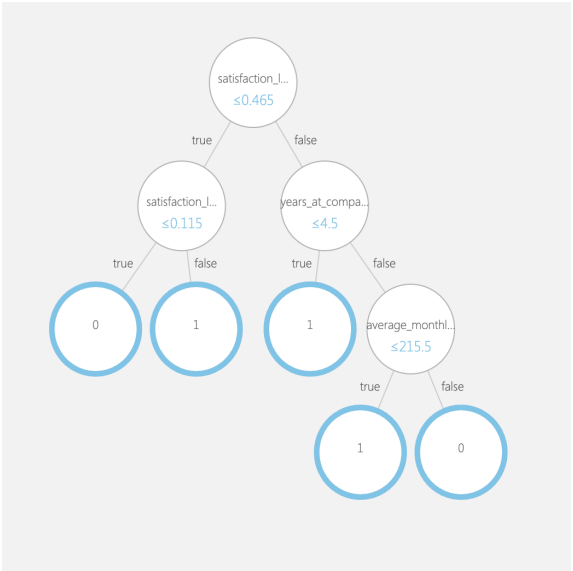


Experiment 2

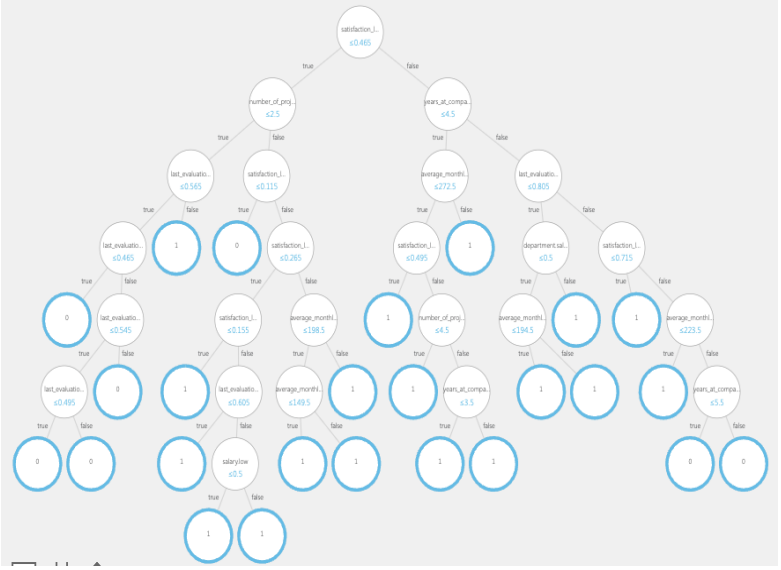


Results

Experiment 1



Experiment 2



While conducting the experiments above, there are a few factors to take into consideration. First, the accuracy of the two is in similar ranges. However, by reducing the number of leaves per node, the number of samples, and the number of trees produced; there is a decrease in the precision, accuracy, recall, and area under the curve. Based on those results in the evaluation phase, experiment 2 would be the better Two-Class Decision Tree Model to

use. As seen in the decision trees produced, experiment 2 also provides more variables into why an employee leaves their position. In experiment 1, 4 metrics are used to determine the likelihood of someone leaving; conversely, there are 24 possible parameters used in experiment 2. If the HR team was to employ the decision tree model, they should select enough leaves on the nodes that would provide a holistic view of the reasoning for someone to leave. In experiment 2, it can be verified by the decision trees that satisfaction in the role contributes to employee retention.

Ali Sadeghinia (Linear Regression and K-Means Cluster)

Linear Regression (R-Studio)

Firstly, it is best if linear regression is used to find and choose a numeric variable that allows for a better understanding of the data and resolution for the problem. Using linear regression allows for finding beta coefficients (the slope in the formula), which then helps with the understanding of which variables are causing the most changes in the outcome (Churn rate of employees).

Using R-Studio, we can simply execute this task with a few lines of code. After importing the data and trying different variations and different variables in the equation, the best fit that helps with making a decision on which variable is going to be used in the K-means Clustering method, the following lines of codes was the final and determinant of that:

```
left.out.satlast = lm(mydata$left ~ mydata$satisfaction_level +
mydata$last_evaluation)
summary(left.out.satlast)
```

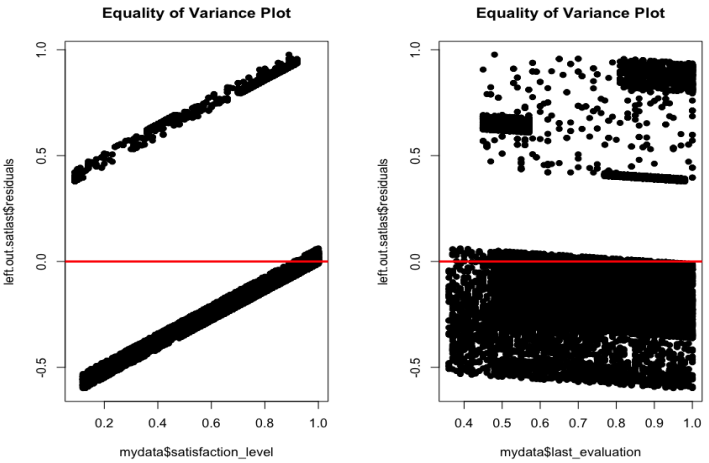
Console (Output):

```
Call:
lm(formula = mydata$left ~ mydata$satisfaction_level +
mydata$last_evaluation)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59809 -0.25673 -0.12630  0.02361  0.97684

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.56577    0.01523  37.140  < 2e-16 ***
mydata$satisfaction_level -0.67393    0.01295 -52.060  < 2e-16 ***
mydata$last_evaluation     0.11915    0.01880   6.336 2.42e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.392 on 14996 degrees of freedom
Multiple R-squared:  0.1531,   Adjusted R-squared:  0.153
F-statistic:  1355 on 2 and 14996 DF,  p-value: < 2.2e-16
```

Although the R-squared and Residual standard error do not have a very promising value when it comes to how good of an analysis regression is for this set of data, the highlighted values are more of what our goal of understanding the data is for.

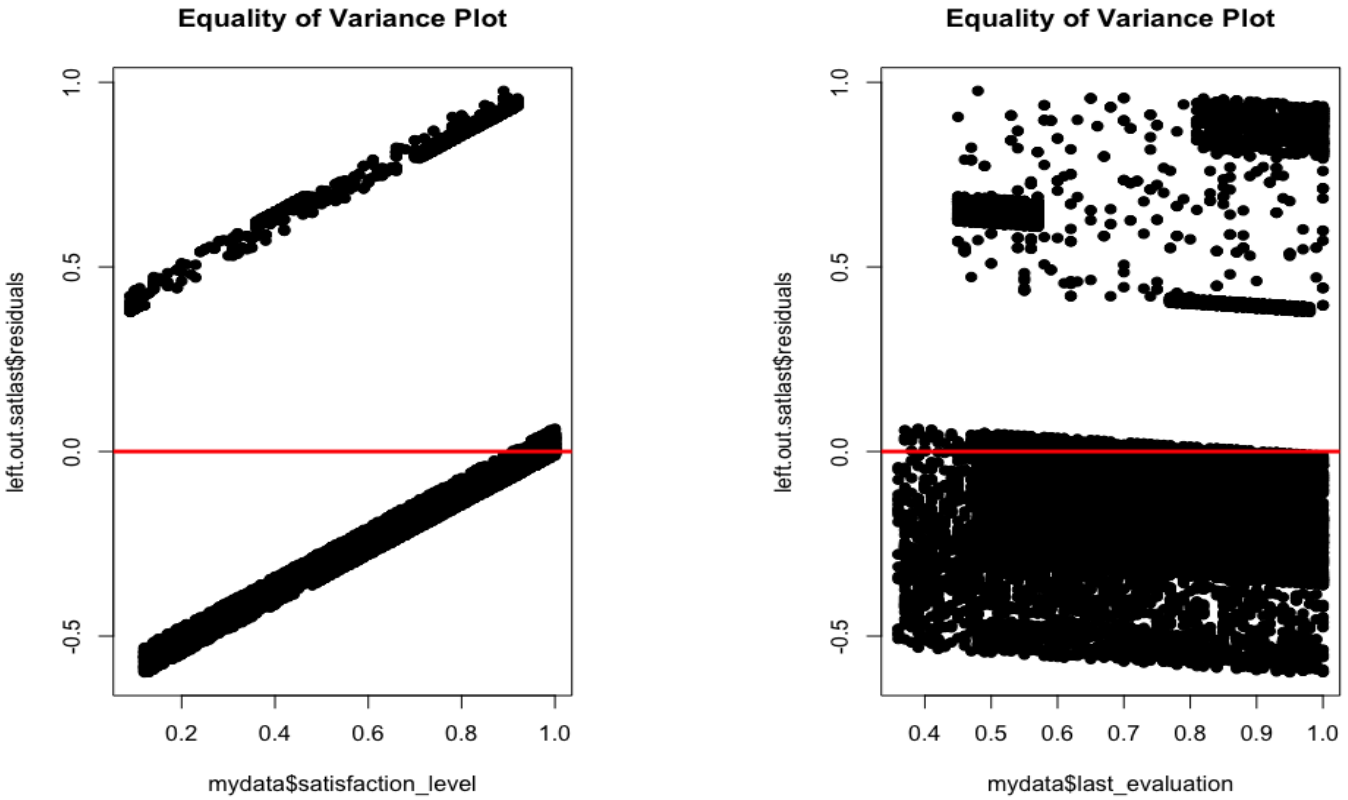
The Beta Coefficient of satisfaction level is by far higher than the last evaluation numerically. This means that satisfaction levels are affecting employees' decision on leaving or staying at their job/company much more than the last evaluation.

The Equality of Variance plot for satisfaction level also shows us the relationship between employees churn rate much more linear than the plot for the last evaluation. The linearity of the data points is also an indication of why using satisfaction level gives a better understanding of the underlying reasoning.



The data point at the top of the plots are the employees who have left their job/company, and reversely the data points at the bottom are employees who have stayed.

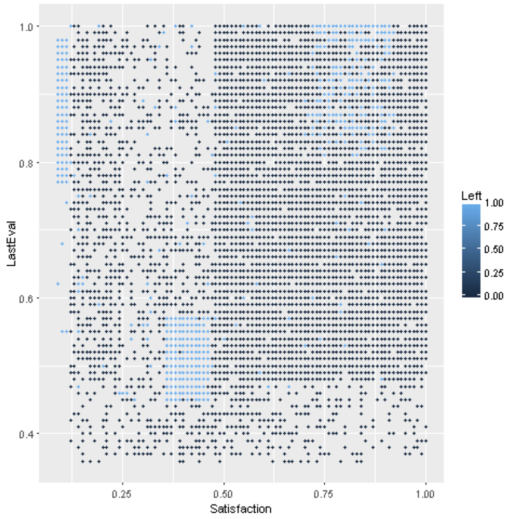
As it can be seen, the employees who have stayed are more likely to have a shorter distance from our residual line. But the employees who have stayed, with a long distance from the residual line at 0 on the vertical axis, need more observation and research to understand the reasoning better.



(data points also show why our R-squared and residual error values are not satisfactory)

Thus, we will use the satisfaction level and the left variable as our set variables for K-means clustering and do a deeper analysis of the effects of each of the remaining variables in our dataset. The plots are magnified for better visualization and comparison next to each other.

K-Means Cluster



Satisfaction Level, Last Evaluation, and Left

First visual graph of the data with 3 specific variables from the data: Satisfaction Level, Last Evaluation, and Left (Dark Blue/Stay = 0, Sky Blue/Leave = 1). Although the data was not of the sort to find centers for the centroids of the data points, it can still be useful solely because of the visualization aspect of the graph.

As it can be seen the dark blue dots are more so crowded on the right side of the chart if it was vertically split in half. While there is also a small crowd of sky blue dots on that side as well, it can still be explained as to why people left their job with high satisfaction levels.

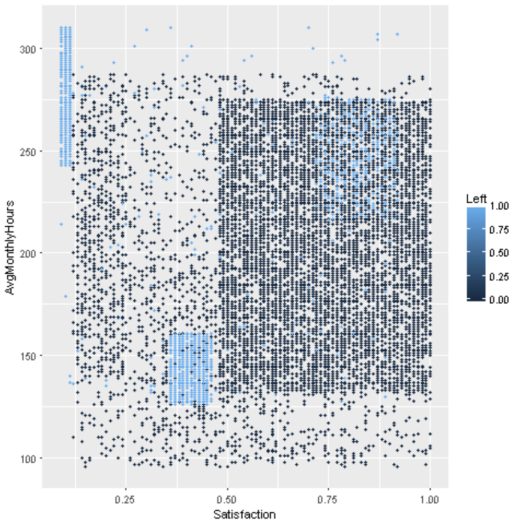
Satisfaction Level, Average Monthly Hours, and Left

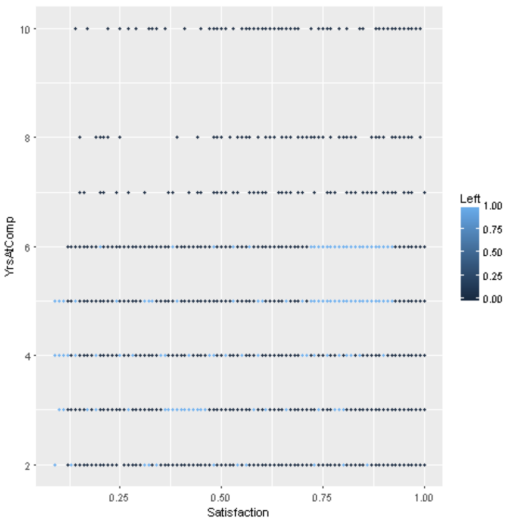
This graph also looks similar to the previous graph with similar clusters of the data points. It could help explain why there are 3 clusters of sky blue data points: 1. Very low satisfaction, high avg. hours; 2. Low satisfaction, low avg. hours; 3. High satisfaction, high avg. hours.

Category 1. Burned out employees, employees who need more time outside of work and their satisfaction at work is directly affected by their work hours.

Category 2. Employees that require more hours to earn more money, or need more hours to become more satisfied.

Category 3. Employees who found a better job opportunity, a promotion in their rank in another company, employees who lied to keep a good connection with their job.





Satisfaction Level, Years at Company, and Left

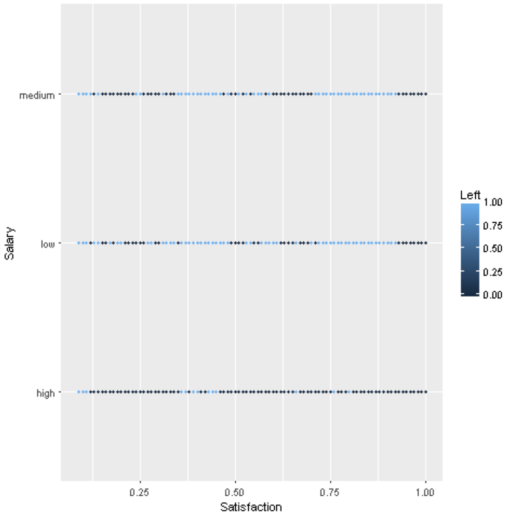
Some small clusters can be seen in this graph as well. However, they can be divided into 3 clusters again: Category 1. Low Satisfaction, Fewer Years; 2. High Satisfaction, Few Years; 3. Any Satisfaction, Many Years.

Category 1. Employees that think they are not valued/trusted enough because of their younger reputation.

Category 2. Employees who had a great medium length experience with the company and were able to improve in that time and find a better position in another company.

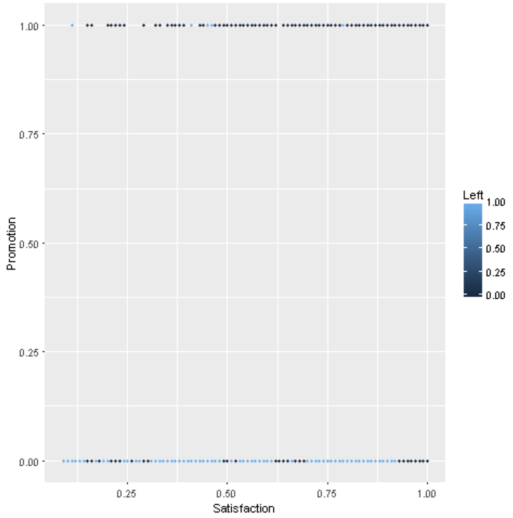
Category 3. Employees that are either happy and comfortable staying in the company or the ones who are only comfortable staying at the company satisfaction aside. This shows that there is a much higher chance for employees to stay no matter their satisfaction level as long as they stay with the company for longer years.

Satisfaction Level, Salary, and Left



We will only analyze the three different salary categories: 1. Low Salary; 2. Medium Salary; 3. High Salary (each split into 2 subcategories of Low satisfaction and High Satisfaction)

1.1. Employees who are unsatisfied with their job/company and their low salary have a great impact on them. 1.2. Employees who are satisfied with their job/company but require more salary. 2.1. Employees who need more of both, 2.2. Employees who might need better pay. 3.1. Employees who need more satisfaction from their job even though their salary is high. 3.2. Employees who might have retired or found another opportunity where they are more satisfied.



Satisfaction Level, Promotion, and Left

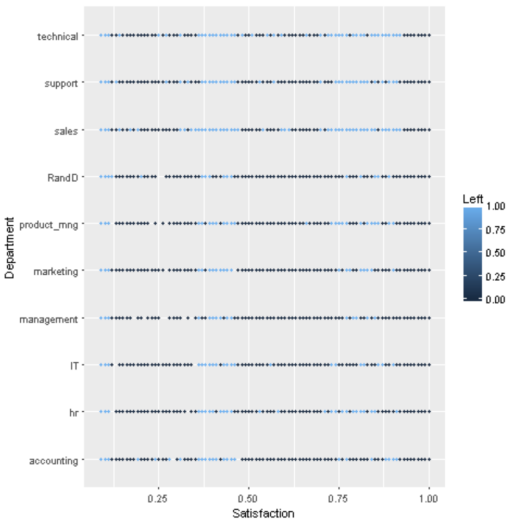
Promotion in this dataset were one of the more clear indicators of both the churn rate and satisfaction of the employees: 1. No promotion; 2. One Promotion.

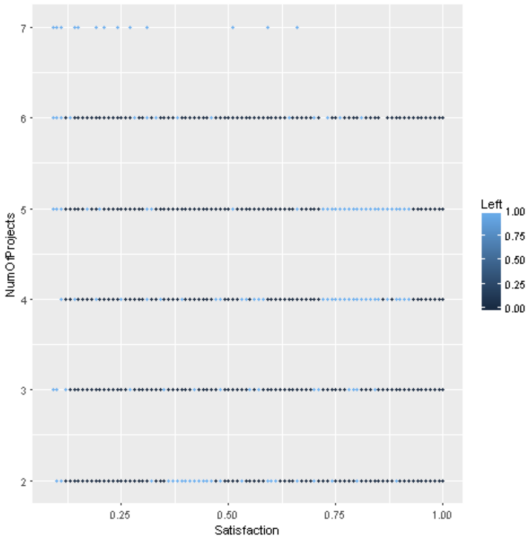
Category 1. More likely for employees to leave if they receive no promotion, or receive a promotion in their position and job from another company.

Category 2. Very unlikely for the employees who received one promotion from their company to leave. However, those who left might have found another job where they are more satisfied all around.

Satisfaction Level, Department, and Left

This is mainly the comparison of each department/field in the industry and their churn rate considering their satisfaction level. Some departments can see that the churn rate goes up as the satisfaction levels go up, while others' churn rate decreases as their satisfaction levels go up. An increase in both satisfaction and churn rate could be an indication that the department does not compensate the employees well, OR the industry is a stepping stone in most employees' careers, OR maybe an indication of a possible saturation in the department and the industry.





Satisfaction Level, Number of Projects, and Left

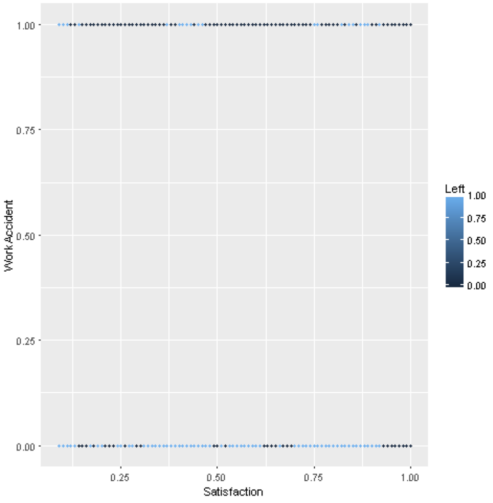
Mostly random structure for the data points except for one section where employees are given 4-5 projects and are highly satisfied. It is likely that this could be an indication of the employees finding a better opportunity after receiving more experience in their corresponding field and industry.

Although this cluster of sky blue data points is not as significant as the previously mentioned one, it is worth mentioning that low satisfaction with fewer number of projects high churn rate could be an indication of employees who do not feel trusted and valuable to be given responsibility, OR the company is not capable of assigning responsibilities, thus the employees are unhappy.

Satisfaction Level, Work Accident, and Left

Although this plot seems like a really good indicator of employee churn rate, it is not much of an indicator/important factor for an employee to stay or leave. In some cases, however, it might.

However, based on the trend from this graph, it can be hypothesized that accidents at work could increase one's commitment to their job and company. OR maybe employees who are more committed to their job and company are more sensitive when it comes to recognizing what is considered an accident and what is not.



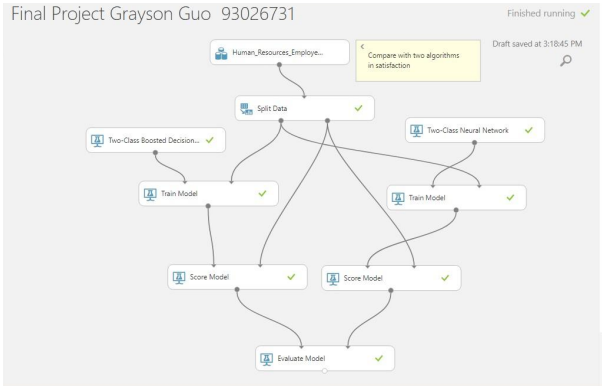
*The R script for the k-means cluster models and plots can be found at the end of the paper below conclusion.

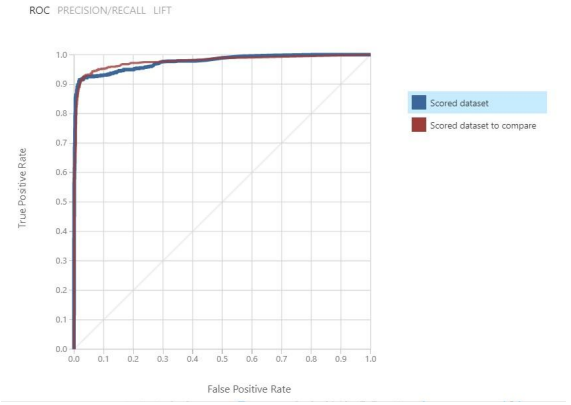
Grayson Guo

(Two-Class Boosted Decision Tree & Two-Class Neural Network)

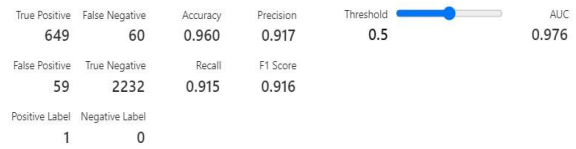
Experiment #	Attributes	Accuracy	Precision	Recall	AUC
DT	Random Seed: 93026731 Number of Trees: 20 Column Selected: Satisfaction	0.960	0.917	0.915	0.976
NN	Default Column Selected: Satisfaction	0.961	0.934	0.900	0.978

Analysis





This time the ROC curve looks normal, both of them are pretty good.They all have high Accuracy precision and Recall values, which means that both algorithms are good models. It is obvious that the red curve has a larger area from the look of the ROC, which means that the Two-Class Neural Network model is even better. The AUC of NN is 0.978, which is higher than the AUC of DT, 0.976. The Recall in NN is the only value that is lower than DT, in this case though, the company wishes to figure out the job satisfaction because NN has higher AUC and Accuracy, the Two-Class Neural Network model is preferred.



From the perspective of organizational behavior, employee job satisfaction in the workplace is the most important indicator because job satisfaction directly affects life satisfaction. Improving job satisfaction is the most effective way to prevent employees from flowing out and attract other employees to flow in further.

Conclusion and Recommendations (Team)

Based on the results of all of our experiments, the following insights for our initial questions:

1. There is a way to accurately predict at what threshold an employee will leave or stay in the company
2. The job satisfaction level is the most important factor that influences an employee's decision to leave or stay in this company.

3. A low job satisfaction level influences an employee's decision to leave this company while a high satisfaction level will most likely cause the employee to stay.
4. There is a group of employees that leave the company regardless of the factors we have outlined. Some of these factors may include a change of location, change of careers, or a better opportunity somewhere else.
5. All renditions of the algorithms have been proven to be effective models which show that the quality of data is good and can be integrated into any algorithm.
6. Salary was one of the main components of the data that determined whether an employee's satisfaction levels are high or low, which then eventually would lead them to leave their job
7. Additionally there were trends that could be observed with each of the other variables while having satisfaction level and left variables fixed for the k-means clustering models. Many of which could be integrated into each other and give us a better understanding of what other factors might be involved in an employee's decision to leave the company even with a higher recorded satisfaction level.

The various data mining models that our group has experimented with can be easily integrated into any type of business structure. It must be noted that the variables will change depending on how that company operates. For example, the number of projects might not be applicable to a particular company. Metrics such as the importance given to work/life balance or training metrics can also be inputted as variables into these models. It may also be beneficial to norm salary based on a position (eg, IS jobs may pay more than sales jobs that are commission-based, etc.) Regardless of what variables, the overall viability of these models to give organizations insights on their employees can be seen which enables business leaders to make effective business decisions surrounding their employees. Additionally there were trends that could be observed with each of the other variables while having satisfaction level and left variables fixed for the k-means clustering models. Many of which could be integrated into each other and give us a better understanding of what other factors might be involved in an employee's decision to leave the company even with a higher recorded satisfaction level.

R Script from Azure ML

```
library(ggplot2)

Scores <- maml.mapInputPort(1)

Left <- Scores$left;
Satisfaction <- Scores$satisfaction_level;
LastEval <- Scores$last_evaluation;
Promotion <- Scores$promotion_last_5years;
Department <- Scores$department;
Salary <- Scores$salary
NumOfProjects <- Scores$number_of_projects
AvgMonthlyHours <- Scores$average_monthly_hours
YrsAtComp <- Scores$years_at_company
WorkAccident <- Scores$work_accident


ggplot(Scores,aes(Satisfaction,LastEval,color=Left))+
```

```
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,AvgMonthlyHours,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,YrsAtComp,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,Salary,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,Promotion,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,Department,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,NumOfProjects,color=Left))+
geom_point(size=1);
ggplot(Scores,aes(Satisfaction,WorkAccident,color=Left))+
geom_point(size=1);

maml.mapOutputPort("Scores");
```

Works Cited

Carucci, Ron. “Balancing the Company’s Needs and Employee Satisfaction.” *Harvard Business Review*, no. November, 2019. *Harvard Business Review*, https://hbr.org/2019/11/balancing-the-companys-needs-and-employee-satisfaction.

Holliday, Marc. “What Is Employee Turnover & Why It Matters for Your Business.” *NetSuite*, 14 January 2021, https://www.netsuite.com/portal/resource/articles/human-resources/employee-turnover.shtml. Accessed 28 February 2022.

Jaramillo, Sanitago. “Four Lessons From Companies That Get Employee Engagement Right.” *Forbes*, 22 June 2018, https://www.forbes.com/sites/forbeshumanresourcescouncil/2018/06/22/four-lessons-from-companies-that-get-employee-engagement-right/?sh=573caa5221bd. Accessed 28 February 2022.

McFeely, Shane, and Ben Wigert. “This Fixable Problem Costs US Businesses $1 Trillion.” *Gallup*, 13 March 2019, https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx. Accessed 28 February 2022.