Employee Retention Case Study

Jonathon Scroggins

Bellevue University

DSC 550

June 3, 2021

Continuity is an important aspect of any business. The ability to retain employees helps with the productivity of a business. Whenever an employee leaves a company, time and other resources must be used to make and post the job, go through the hiring process, and then train the new employees. That does not even include the concept of the learning curve for the new employee over the time that they would still be considered a new employee. This is only a small part of the negative results of not being able to retain employees in a company. Just a quick search in Google on “Employee Retention” will produce dozens of articles on the positive and negative aspects of employee retention. Therefore, my question that I explored for this project is, is there a way to predict if an employee might leave a company and therefore put into place a way to retain that employee or at least predict they were going to leave and put into place the process to fill that spot before it is too late. To explore the question of employee retention I used a dataset titled HR Employee Analytics found here https://www.kaggle.com/kmldas/hr-employee-data-descriptive-analytics. This dataset has the following variables:

1) Emp\_Id

2) satisfaction\_level - percent as a decimal

3) last\_evaluation - Time from last evaluation in years

4) number\_project - Number of projects employee is working on

5) average\_montly\_hours – Average hours worked last 3 months

6) time\_spend\_company - Commute time for employee

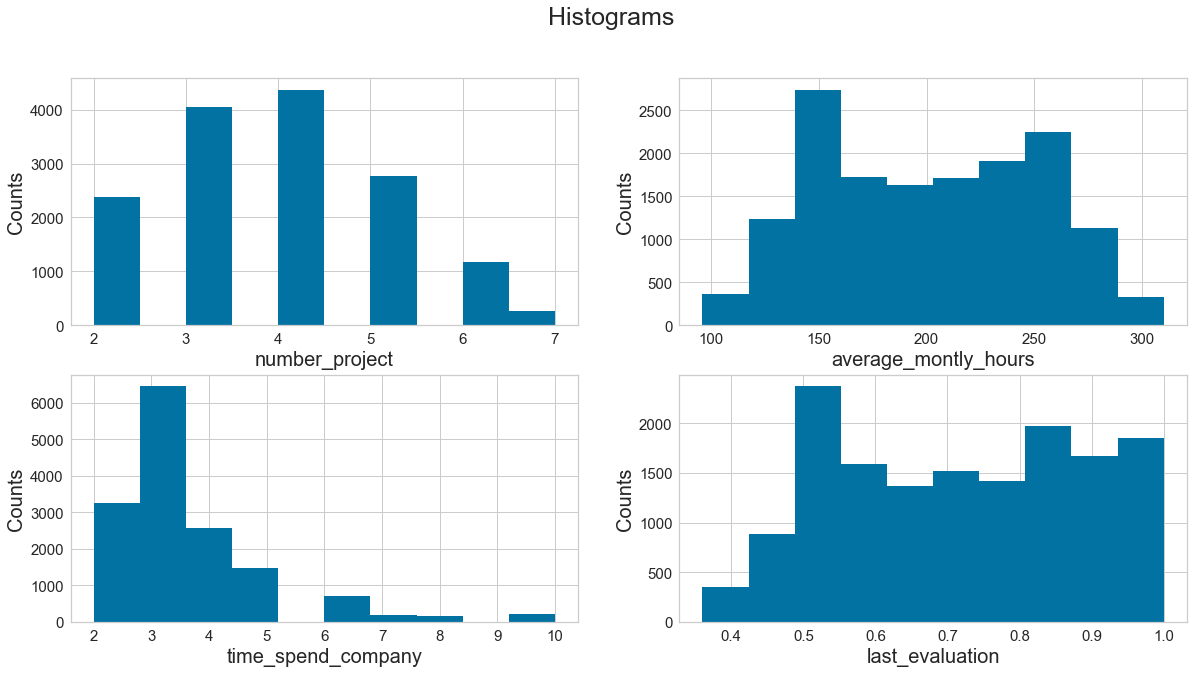
7) Work\_accident -

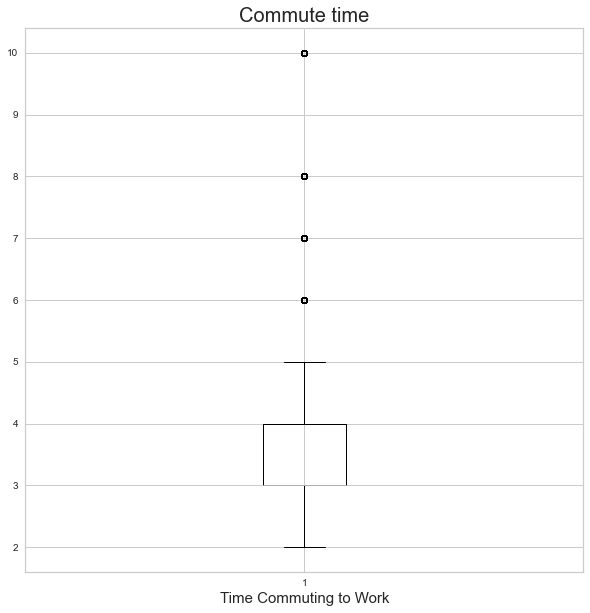
8) left - Whether or not the employee has left the company

9) promotion\_last\_5years -

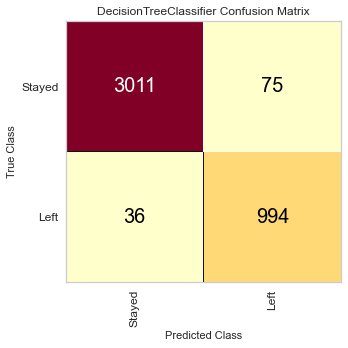
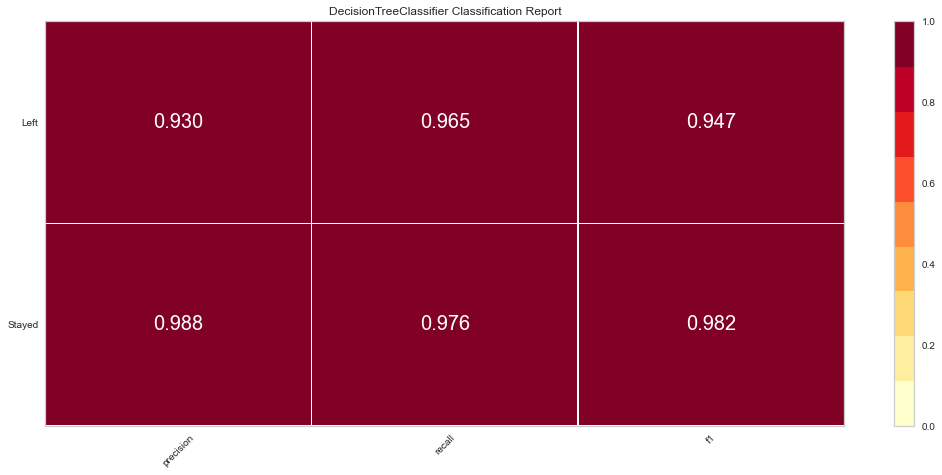
10) Department

11) salary

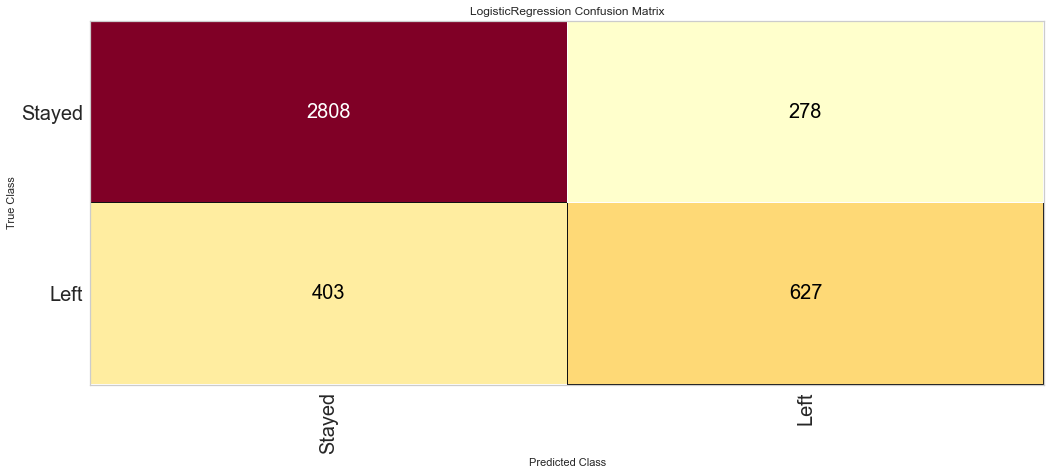
To start of the analysis, a graphical analysis of the data will explore the state the data set is in. This will see if there are any outliers that may be affecting the data and if there are any variables/features that may not be necessary for the analysis. The first set of graphs to look at are the histograms of some of the variables to see the spread of the data. 

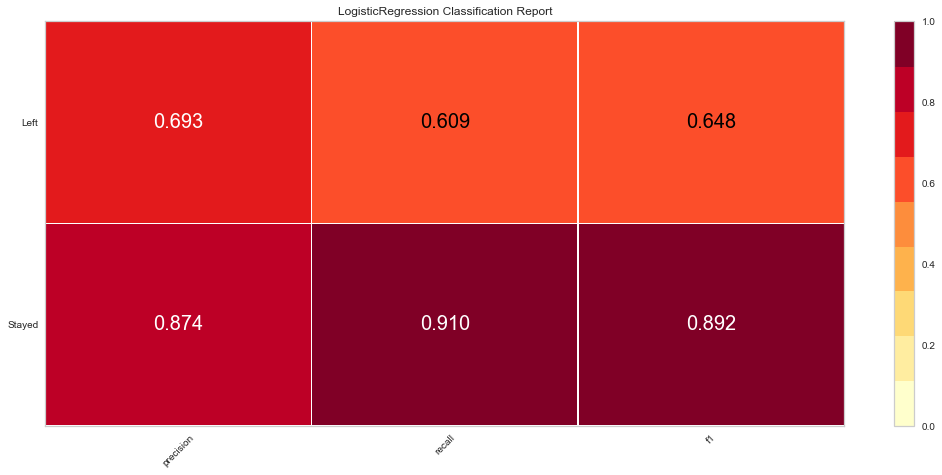
Looking at the histograms for these variables the one that stands out the most is the distribution of the time\_spend\_company variable. This variable is the time spent commuting to the company and it is very skewed to the left so that variable will need to be explored further to make sure there are not any outliers in the data. To do that a box plot works well. 

Looking at the boxplot, it does look like there are outliers from 6-10 so those data points will be dropped. After doing more graphical analysis with the data, the rest of the variables looked good so now we can look and see if any of the variables/features are unnecessary for the analysis and drop those as well. To build a good model, the only variable that is unnecessary in this analysis is the Employee ID variable, so that will be dropped. We also check the data set for any null values and there are none in this set so we can move on to choosing and building a model.

For this analysis, we will run a series of models and see how well they can predict the employee retention. We will be using the left variable as the target variable for the model since we are trying to predict whether an employee left or stayed with the company. We will also need to convert the variables department and salary from categorical variables so we will set up dummy variables for each department and salary level. After we have all of our variables ready run the model, we can first split the model into a training set and a testing set. For this analysis, I chose to split the data with 70% (9601 data points) of the data in the training set and 30% (4116 data points) of the data in the testing set. For this analysis I also chose to run three models, a decision tree classifier, a logistic regression model and a random forest classifier. For the decision tree classifier model we got good scores all around. Here is a look at the confusion matrix and classification report for this model.  

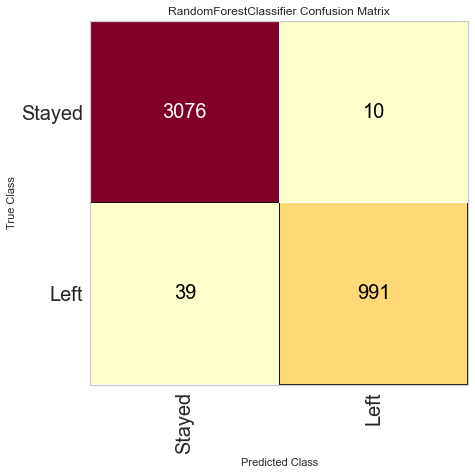
The accuracy of this model was 97% and the confusion matrix score was 97% and all the other scores as you can see were high as well.

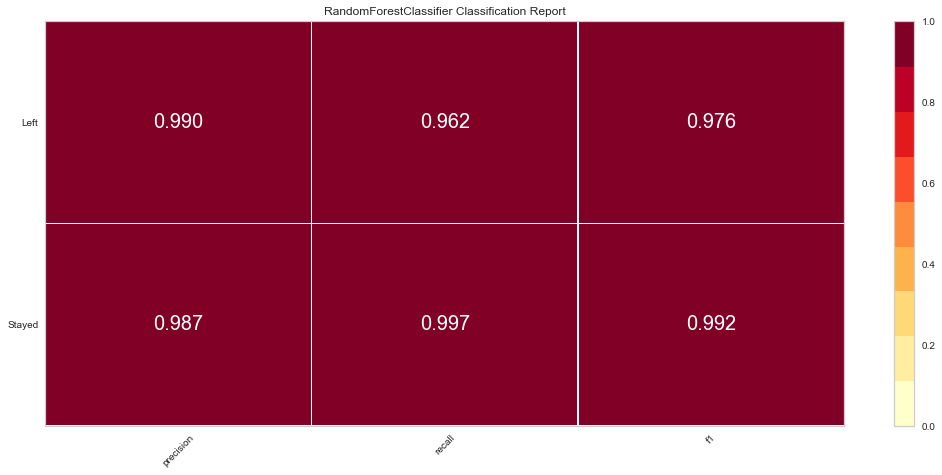
When we look at the scores for the logistic regression, we see that the confusion matrix score is 83%, which is still good but not as good as the previous model. 

The classification report metrics are also not as good for the logistic regression model. 

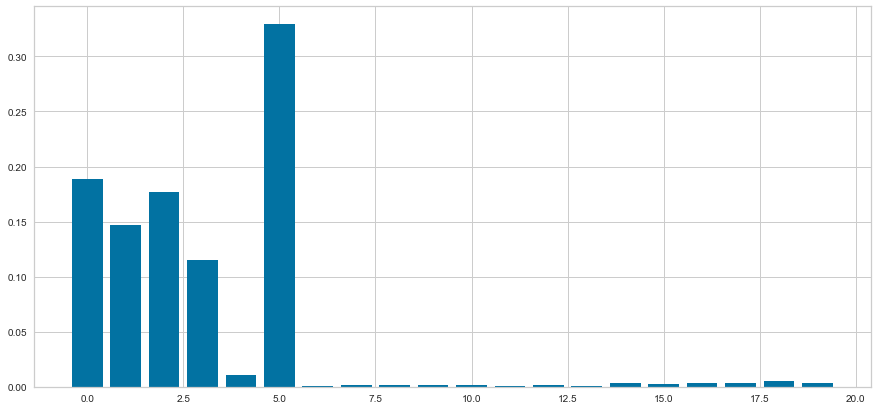
This would tell me that though the logistic regression model is still a good model, it is not the model I would use for this data set at this time.

The last model to look at is the random forest classifier. This model had the highest confusion matrix score of 99%.

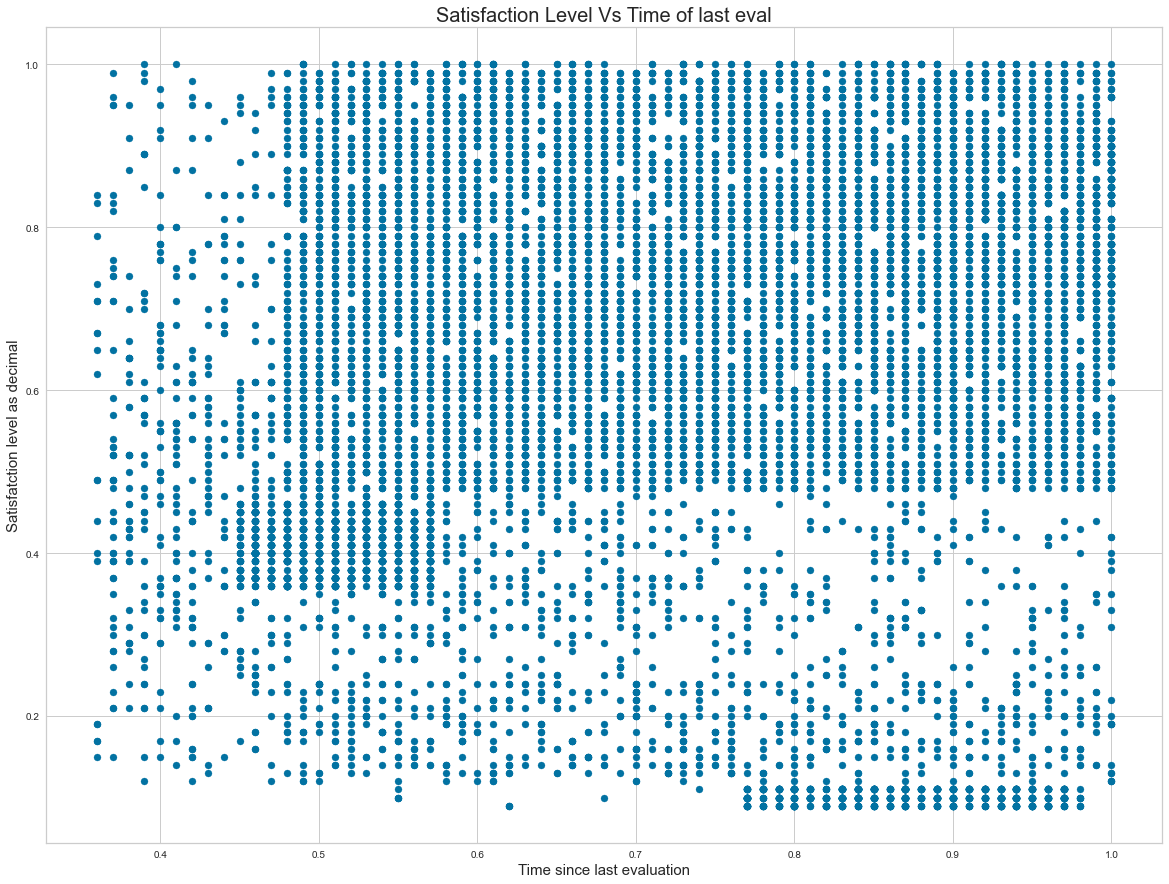




We can see here that the other scores for the random classifier are strong as well.

Now that we have run the models and see that whether an employee stays or leaves the company can be predicted at a significant percentage, we can look at the models and see which features played the most significant roles in the model. For this step we will look at the model that performed the best and see if there is a feature or features that stand out as the most significant to predict the target variable. When looking at the model’s feature importance we can see that the feature that was the most significant was the satisfaction level. In fact, in all the models, the satisfaction level of the employee was the most significant feature for the retention of an employee. Here is a look at the feature importance for the random forest classifier: 

Since that feature was the most significant in all the models, I also decided to do a linear regression model to determine what features were the most significant for predicting the satisfaction level of the employee. When running this model, we see that the most significant features that influence the satisfaction level of the employee are the time spent commuting to work and the time between evaluations. Neither of those features had a large influence, however, they were the most influential feature for the satisfaction level of the employee which from earlier was the most significant feature to determine if an employee will be retained.

Based on the graphical analysis and the models, what conclusions can be made and what steps can be taken to help make sure an employee wants to stay with the company? I think the most significant take away is that a company should put forth an effort to keep the satisfaction level of the employee up. The higher the satisfaction level, the more likely the employee will be retained. Also, based on the results of this project, to keep the satisfaction level of the employee up, the company should investigate the time between evaluations and be mindful of the commute of the employees. As far as the time between evaluations is concerned, my recommendation would not necessarily be to change the amount of time between the evaluations, but to understand what the time between evaluations does to the satisfaction level. Glancing at the graph for the satisfaction compared to the time between evaluations one can see there is a positive relationship between the two variables.

The longer between evaluations, the higher the satisfaction level of the employee and the more likely they will stay. What that might mean is that, if the company does evaluations at the same time for all employees, that might cause a higher chance for employee turnover, so a company would want to be more prepared for that turnover to ease the transition. The longer between the evaluation the more likely the employee will stay. It may also suggest making there be more time between evaluations, but not so much time as to sacrifice efficiency. I am not suggesting doing away with evaluations, but an adjustment to the time might be something to explore. The other feature that was a more significant influence on the satisfaction level of the employee was the commute time for the employee. To investigate this feature is a little more difficult for a company. Without providing housing there is not much a company can do to limit the commute of an employee. However, a company might want to explore the concept of working for home if it is something feasible for the company. That would decrease the commute time for most employees. An analysis of what a work from home program can do for the company is a next step that I would recommend based on the analysis of this data. There is a chance that would continue to keep the satisfaction of employees up and keep the retention percentage up as well. What that would do for the productivity and bottom line for the company is something that would need to be explored as well before offering it companywide.

Employee retention is an aspect of a company that can help a company look attractive to potential employees and potential investors. If the employees are happy and have bought into the company it will show in their willingness to stay with the company. That can translate to the consumer as a business that will take care of its people and therefore take care of its consumers as well. Based on this analysis keeping the satisfaction level of an employee up will have the most significant influence on whether an employee will stay or leave. Surprisingly, the factors that influenced the employee’s satisfaction level was not the number of hours worked or the number of projects they were working on. The most significant influencers were the amount of time they spent commuting to work and the time between evaluations. As I mentioned earlier, these would be the two places I would continue investigating to make sure to keep the employee retention level up and possibly increase the percent of employees that stay with the company for a long time.