Project Report - Group Project 2

1. Project Goal

The goal of this project is to build a decision support system for a company to determine if an employee is likely to leave the company or not, which will be referred to as employee attrition throughout the report. Employee attrition can be voluntary, such as retirement, or involuntary, such as termination. This decision support system will review a set of variables related to each employee to determine the probability that they will leave the company.

Historic measures of employee attrition are typically defined as employee separation. The terms can be used interchangeably. The Bureau of Labor Statistics keeps a running tally of attrition across various industries. As shown in Figure 1, the annual percentage of employees leaving companies (across all industries) has steadily increased from 40.30% to 44.30% between the years 2014 and 2018.

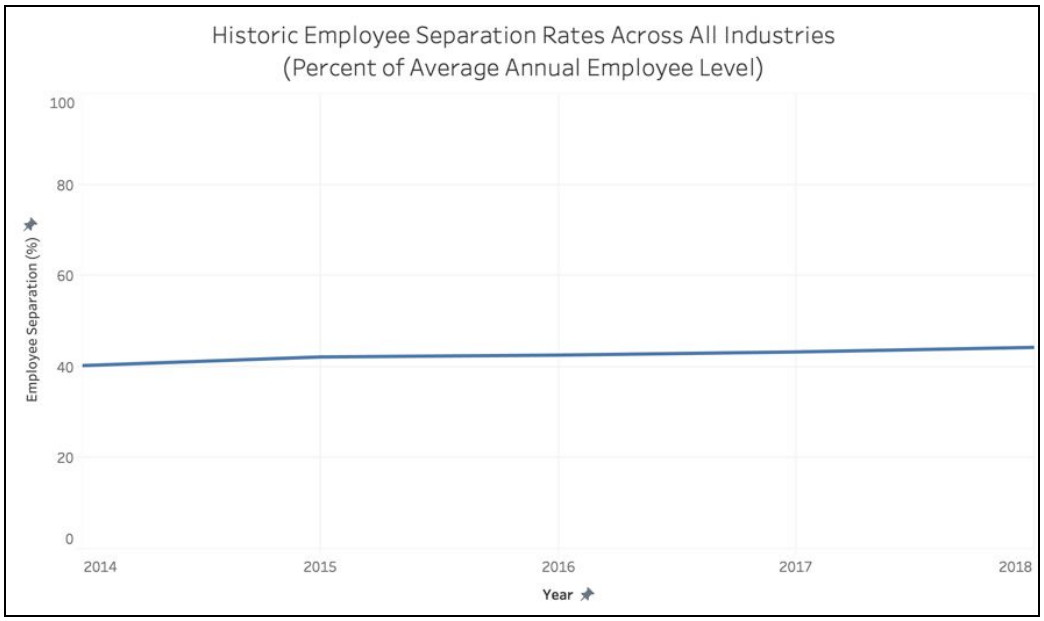


Figure 1: Historic historic levels of employee separation across all industries [1]

The data from the Bureau of Labor Statistics can also be broken down further to show which industries have experienced the highest and lowest levels of employee attrition. Figure 2 breaks down the employee separation rates for 2018. From this data, the team could see that the top three industries with high employee attrition were entertainment, leisure, and accommodation. On the side side of the spectrum, the three industries with the lowest levels of employee attrition were government, state/local education, and federal. These insights will be taken into account during our analysis and recommendations.

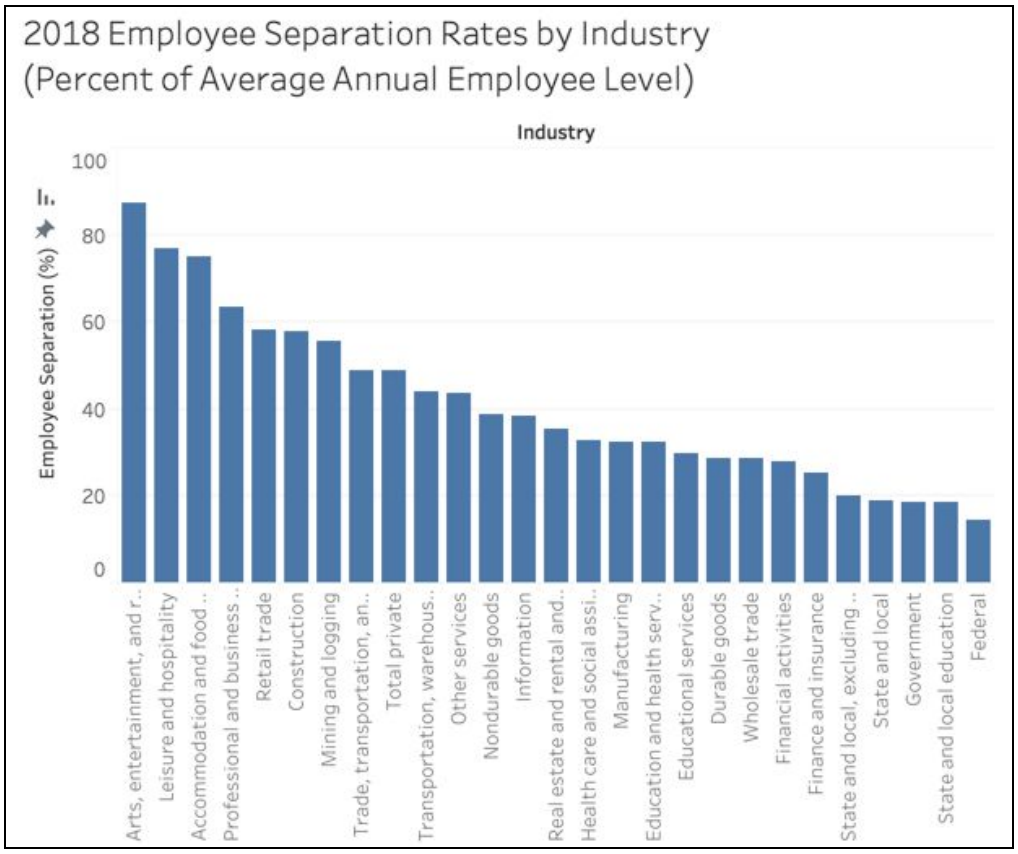


Figure 2: 2018 levels of employee separation broken down by industries [1]

In the project training data set given to us, we have information for one thousand employees, which includes their attrition status. We need to build a model to predict the probability of attrition for the other 470 employees. In this project, the team needs to understand

the attributes/variables given and build a classification model. Afterwards, the models can be tested and can provide estimate performance for that model. The subsequent sections will give a more detailed overview of the data cleaning, preparation, and modeling process.

2. Overview of Data, Including Data Exploration and Analysis

As stated in the previous section, the training data set given to the team had a total of one thousand employee entries. Each entry had thirty-six total variables documented. The variables were both numeric and factors. The target variable for the project was determined to be "Attrition", which would be a factor with two levels of "yes" and "no", which would be predicted by the classification model. The remaining thirty-five variables were found to be nine factors and twenty-five numeric variables. Of the twenty-five numeric variables, an additional eight variables were converted to factors prior to building a model.

These eight numeric variables were converted to factors because the model would otherwise treat changes in these variables as identical incremental steps. However, it is not appropriate to treat certain variables in that manner. For example, the variable "Education" was entered with values from "1" to "5", which correspond to "Below College" to "Doctor". It is not appropriate to treat each step up in the level of education as an identical increase. This is also applicable for the variables "environment satisfaction", "JobInvolvement", "job satisfaction", "performance rating", "JobInvolvement", "WorkLifeBalance", and "StockOptionLevel".

Once the types of data were converted, the next steps were to review the cleanliness of the data. This involves checking for missing values and the normality/skewness of the data. The team started with reviewing missing values and found that the data set given to us had no

missing values. Therefore, no steps were needed to determine if the values were missing at random, nor was it required to impute any missing values.

The first variable the team reviewed was the target variable, “Attrition”, to further understand the current state of the company in regards to industry averages. The team found that the attrition rate for the company was 16.7%, as shown in Figure 3. This is on the low end of the spectrum from the previous historical data review. However, the team will need to acquire more information from the company to determine the time-frame of the data collection for the company.

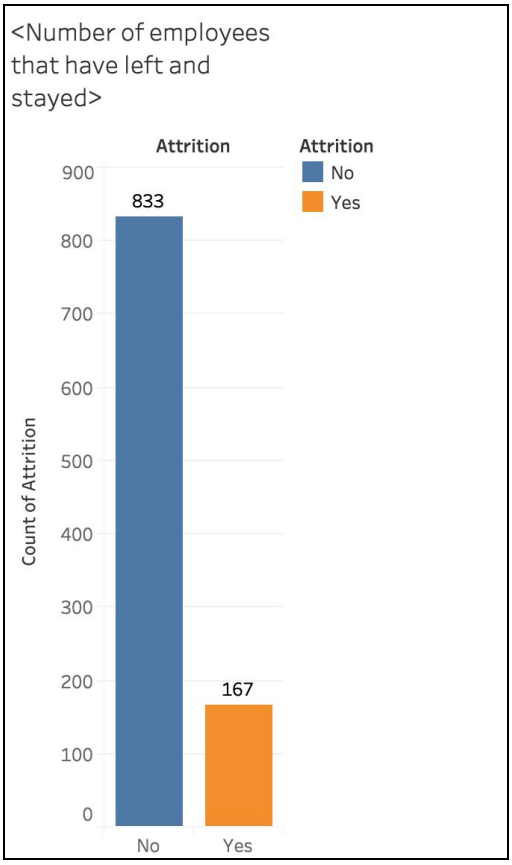


Figure 3: Number of employees that have left and stayed at Company XYZ

The next step in the data exploration was to review the normality and skewness of the data to see if any transformation were required. To accomplish this, the team decided to plot out all of the variables via histograms, since the number of variables was manageable to plot out individually. While looking for normality/skewness issues, the team was also looking to see if any patterns or trends could be identified between employees that have left the company and those that have not prior to running the models. The following figures and charts below will depict some of the exploration and discoveries the team found.

The first few variables the team reviewed were “Gender”, “Age”, and “Years at Company”. This is shown in Figures 4 through Figure 7. The team could not identify any apparent trends in the initial pass through on these variables. The distribution appears to be very similar between employees that left the company and those that stayed with the company. There was a slightly higher number of employees that left the company in the first couple of years; however, this would be further reviewed with the models were created.

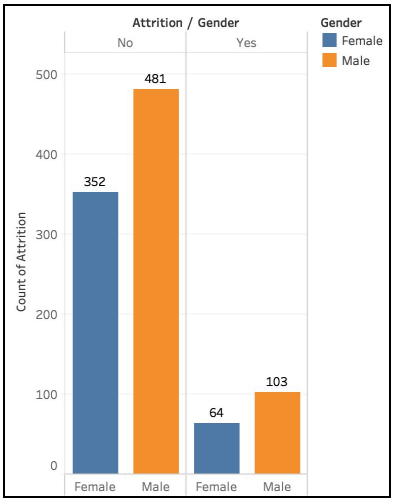


Figure 4: We could see that in the company XYZ, there are more male employees churn to other company than the female. The male employees churn rate is 60% higher than the female churn rate.

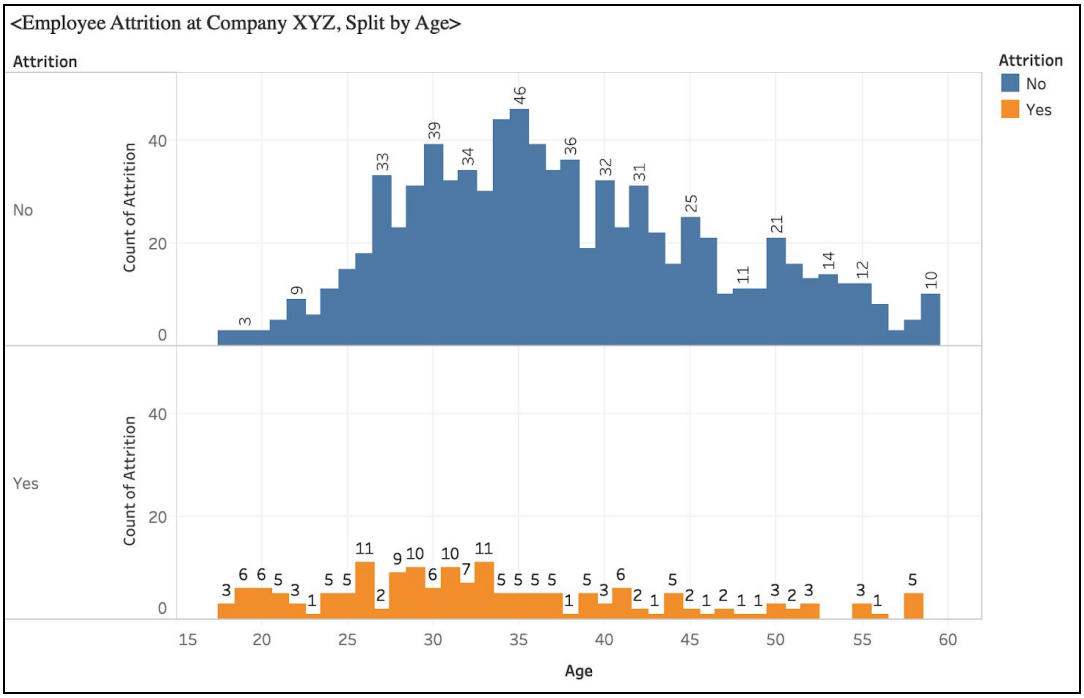


Figure 5: Shows that there are more employees aged from 26 to 33 churn to other

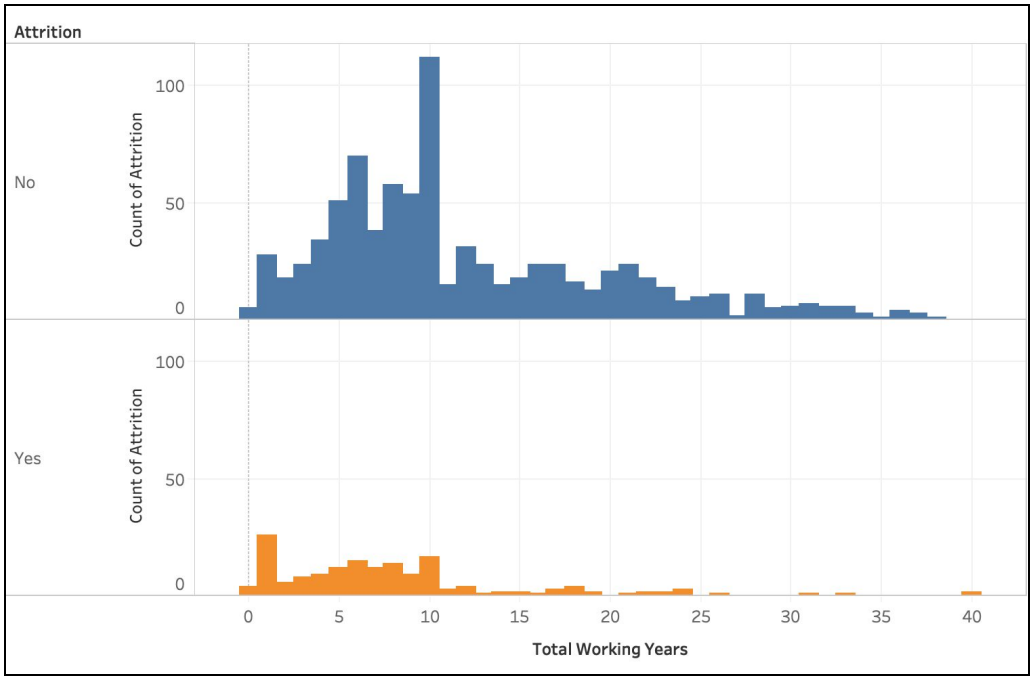


Figure 6: Shows that employees have more likely to leave the company in their first year working there and after their 10 years working in the company, they are less likely to leave the company.

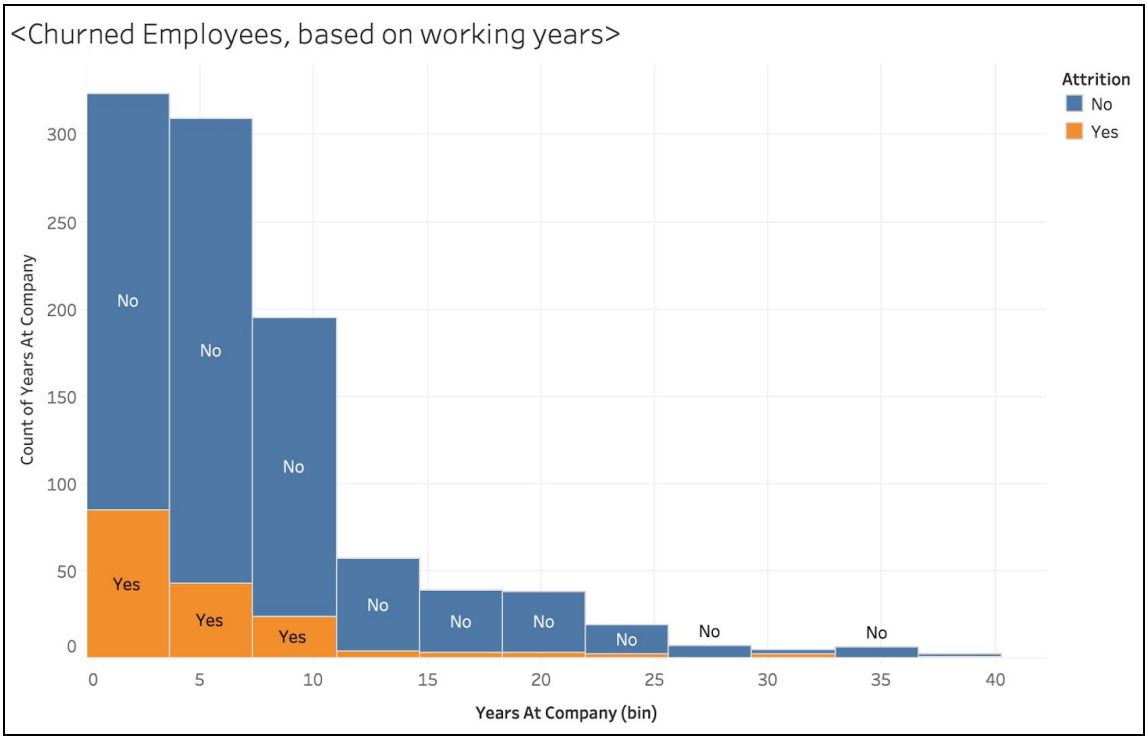


Figure 7: Additional view of the number of employees that have left the company, based on working years

The next set of variables that the team reviewed were "Department", "Education", "Marital Status", and "Stock Option". These values and comparisons are shown in Figure 8 to Figure 12. We can see from Figure 8 that the majority of the company's churned employees have come from the research and development department. However, we do not have the information to state if this is due to it being the largest department or some other reason. Figure 9 tells a similar story in regards to the job position. Laboratory technicians appear to be the position with the highest turnover rate, followed by a sales executive and research scientist. The distribution of male and females are approximately equal. This is further reinforced by Figure 10, which shows the education background of the highest churned employees are found in the life sciences and medical fields.

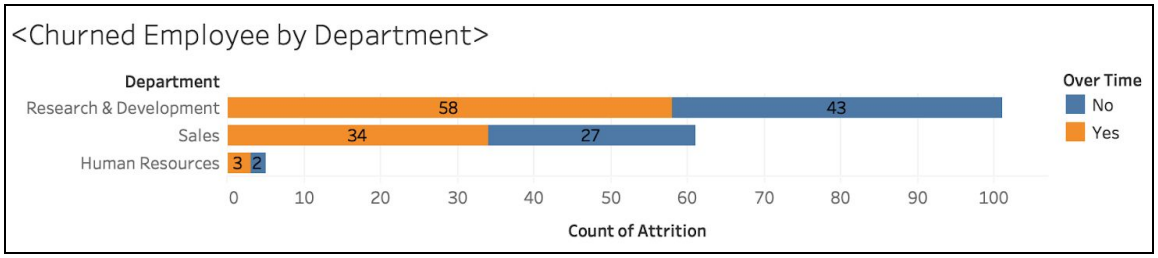


Figure 8: Shows that the research and development department has more employees leave the company.

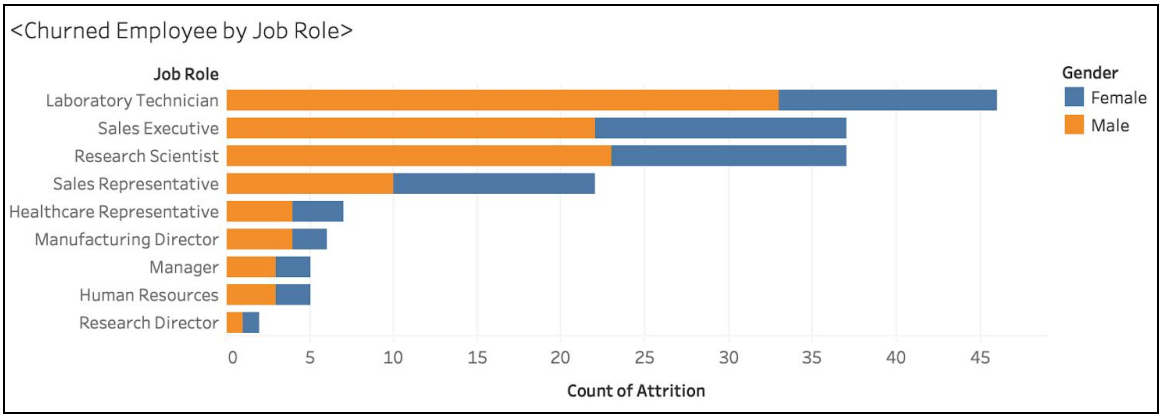


Figure 9: Shows only the felt employee in the training dataset, we could see that there is a higher number of employees who work as a laboratory technician, sales executive, and research scientist left the company. Meanwhile, comparing with the female employee, male employees have a higher number to leave the company.

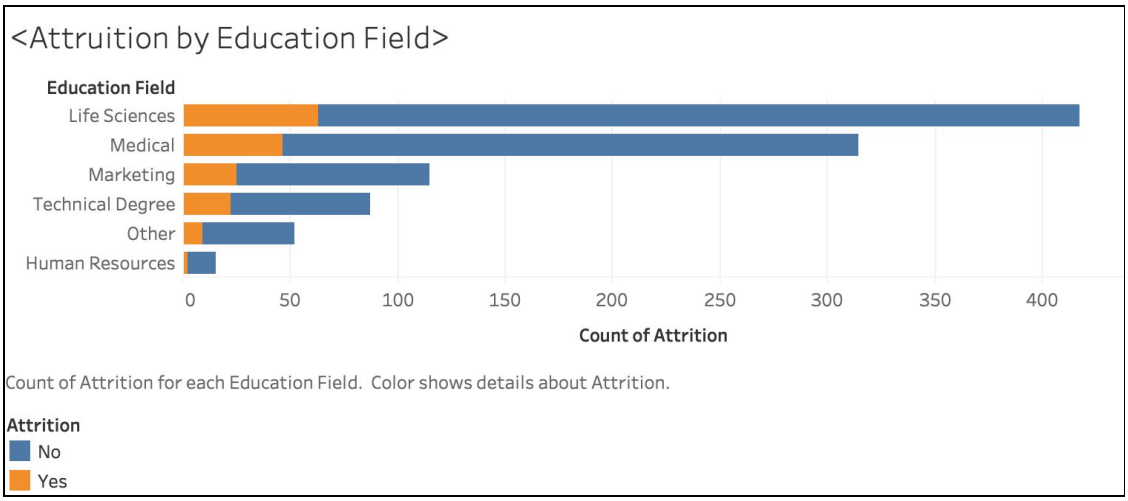


Figure 10: Shows that employee who works in the life sciences has the most number of employees left the company.

The next several figures dive deeper into the factors that may account for attrition, "Marital Status" and "Stock Option Level". From Figure 11, the team found that employees tending to stay with the company have a higher proportion of being married. This logically makes sense, because an employee will be less likely to leave a company if they have other family members and dependents to support. Figure 12 takes a look at "Stock Option Level", and this shows that employees leaving the company typically have a higher probability of being a "0" or "1" stock option level. This also makes sense, if they are shorter term employees, not at a level to earn stock options, or not satisfied with their stock options, so end up leaving the company.

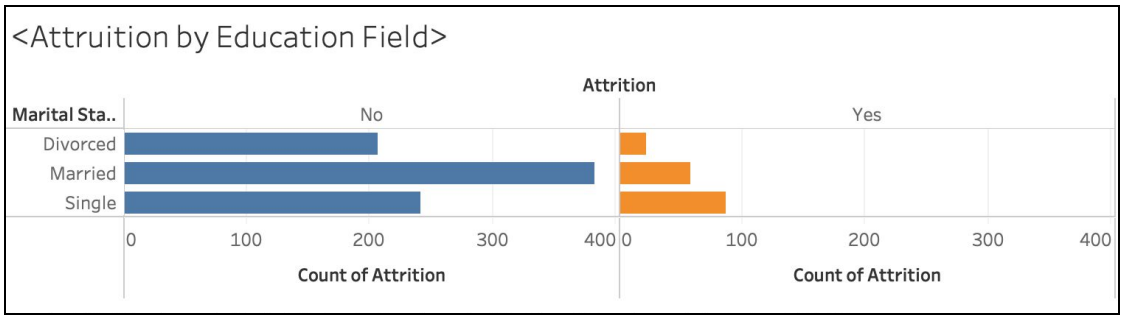


Figure 11: shows that the single employees left the company more than divorced and married employees.

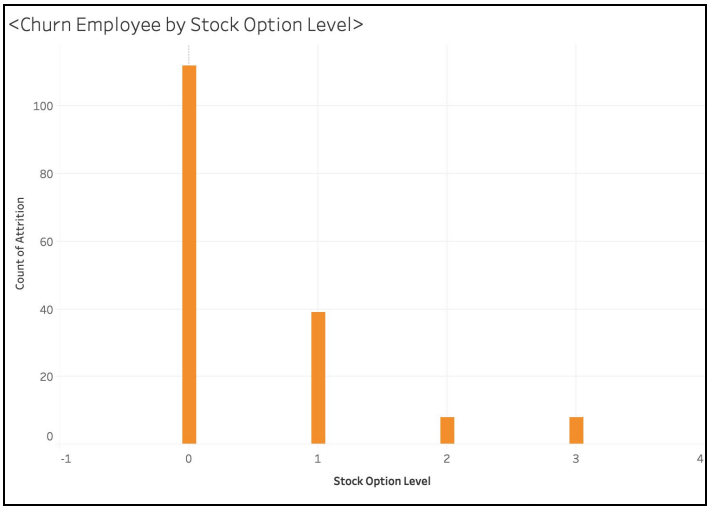


Figure 12: Only shows the churned employees from the training dataset

This section on data exploration and analysis allowed the team to get a better handle on the data set as a whole prior to moving forward into the modeling stage. We made sure to clean up any variables that are needed, as well as make some hypotheses on what may or may not be causing employees to leave the company. This will further be proven or disproven as the data is entered into various classification models to determine the one with the best performance for predicting employee attrition rates. Further charts and analyses can be found in the team's R Markdown file for "Data Exploration".

3. Details of Modeling Strategy and Estimation of Model Performance

The target variable in this model is a binary variable, therefore the model must be a classification model. Selecting the correct classification model can be very challenging since there are multiple options to choose from. We selected four models (Random Forest, GLM, LDA, and SVM) to attempt and decided to compare their accuracy and then go with the one with the highest accuracy. A summary of the results is shown below in Table 1. A description of each model will be described in the subsequent paragraphs.

Model <fctr>	on_Training_set <dbl>	on_Testing_set <dbl>
Random Forest	0.8441310	0.8447489
GLM	0.8373880	0.8264840
LDA	0.8911652	0.8584475
SVM	0.8655570	0.8401826

Table 1: Summary of the model performances attempted on the attrition training data

The first model attempted was Random Forest, which is a form of decision tree bagging. It will take a random subset of variables at each node of each tree to make a decision about the split point. The two hyperparameters that can be tuned in this model are the number of trees and the number of variables to attempt at each node within the tree.

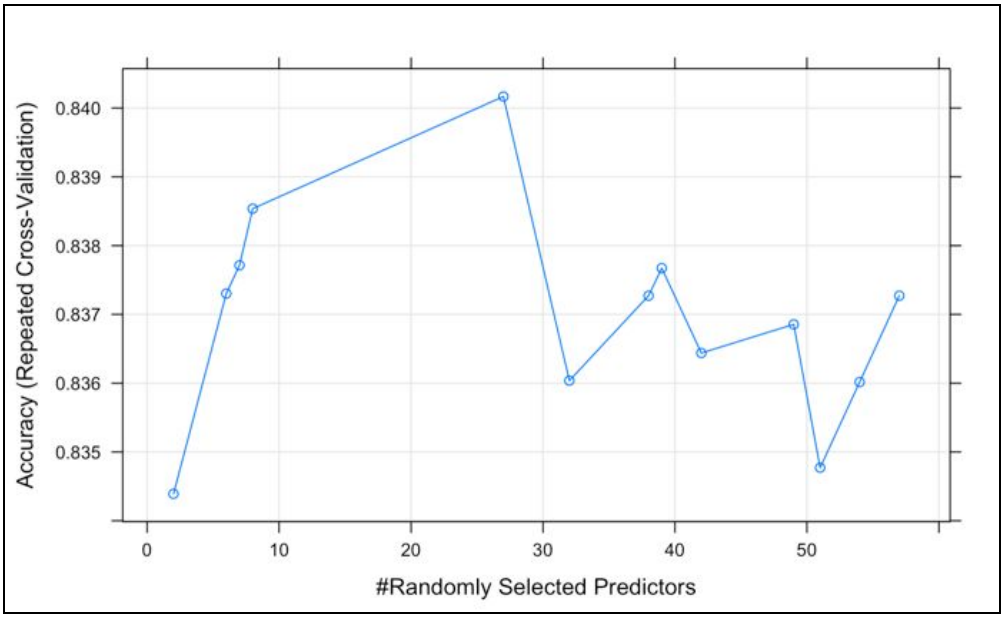


Figure 13: Accuracy of Random Forest model with differing "mtry" values

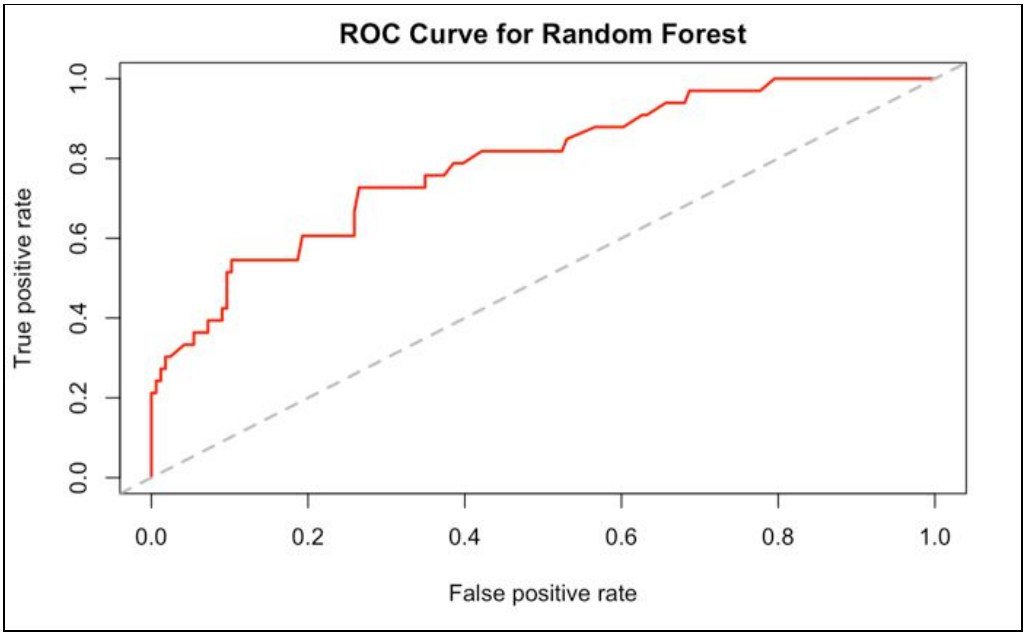


Figure 14: AUC curve for the previously described random forest model. AUC value of 0.778.

From Figure 13 and Figure 14 (above), it is shown that the random forest model had the highest accuracy rating when using twenty-eight randomly selected variables per node. This resulted in an AUC value of 0.844 for the specific tree shown. Additionally, in Figure 15 it shows the top ten variables determined to be the most important for creating the model. From this chart, it is clear that "MonthlyIncome", "OverTimeYes", and "Age" were the top three variables of importance. This intuitively makes sense when considered in real-world applications. Employees want a higher income, which comes from overtime. Additionally, younger and older employees are more likely to leave the company.

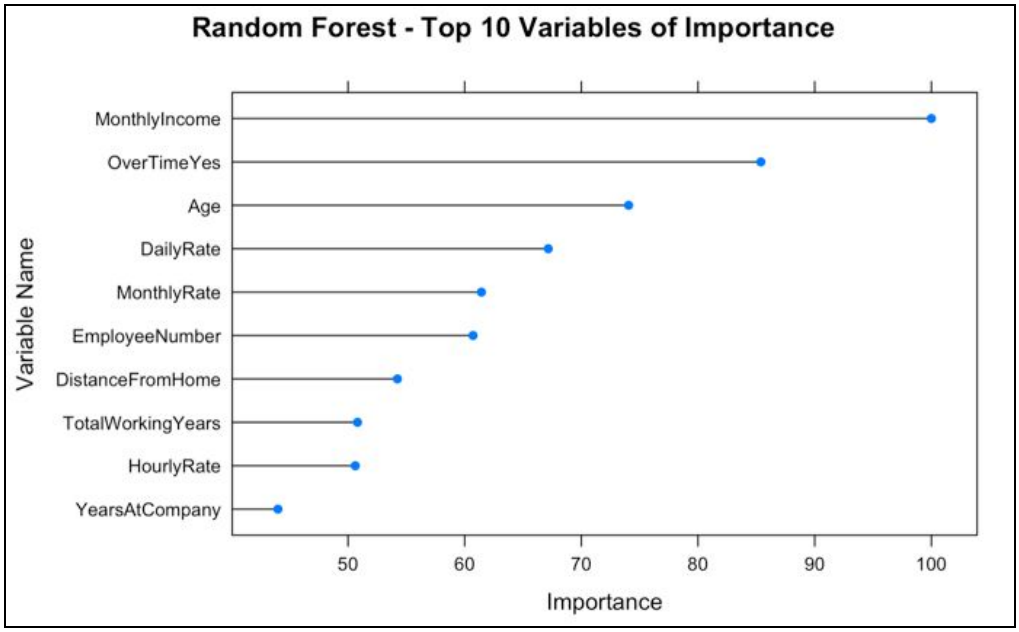


Figure 15: Top ten variables of importance from the random forest model.

The next model we attempted to run on the training data was generalized linear regression model. Figure 16 and Figure 17 show the results of this model below. It was found to have an AUC value of 0.862, and the confusion matrix threshold was set at 0.60 for the results shown in the figure below. This is a method for controlling the amount of Type 1 and Type 2 errors. Depending on the application, this type of error can be extremely expensive if a company is going to invest a lot of money in employees that may end up leaving anyway. "OvertimeYes", "EnvironmentSatisfaction", and "JobSatisfaction" were found to be the top three most important variables according to the model.

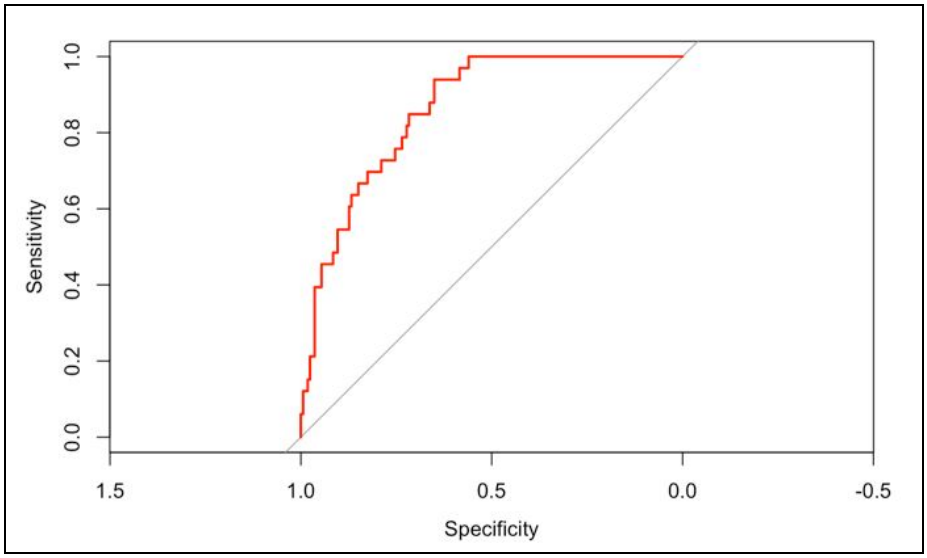
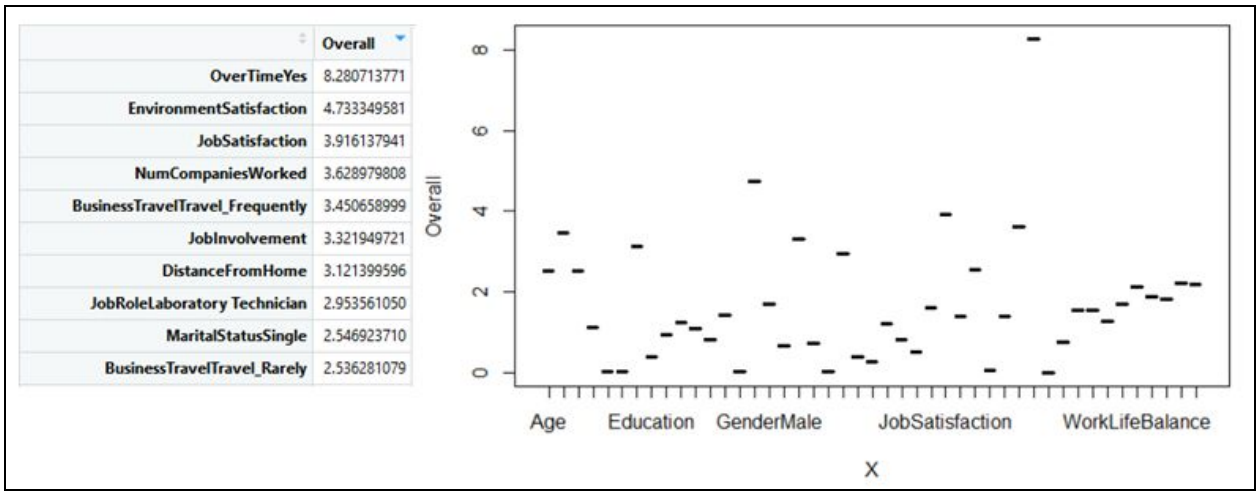


Figure 16: AUC curve for GLM. Resulting AUC value was found to be 0.83.

	True	
Predicted	No	Yes
0	153	18
1	13	15

Figure 17: Confusion matrix with a threshold value of 0.60 to control Type 1 and Type 2 errors.



	Overall
OverTimeYes	8.280713771
EnvironmentSatisfaction	4.733349581
JobSatisfaction	3.916137941
NumCompaniesWorked	3.628979808
BusinessTravelTravel_Frequently	3.450658999
JobInvolvement	3.321949721
DistanceFromHome	3.121399596
JobRoleLaboratory Technician	2.953561050
MaritalStatusSingle	2.546923710
BusinessTravelTravel_Rarely	2.536281079

Figure 18: Variables of importance from the GLM Model

The third model that the team attempted was linear discriminant analytics (LDA). LDA is similar to principal component analysis (PCA) in the method it is using to classify the target

binary variable. LDA estimates the probability that a new set of inputs belongs to every class by utilizing the distribution of predictors alongside Bayes' theorem. The output class is the one that has the highest probability. This model was the one that was found to have the highest accuracy of all four models. The accuracy on training set was 0.89 and 0.85 on the testing subset. As a result, we decided to pursue LDA as the model of choice for our classification model.

```
Call:
lda(Attrition ~ ., data = training)

Prior probabilities of groups:
       No       Yes
0.8497791 0.1502209

Group means:
         Age BusinessTravelTravel_Frequently BusinessTravelTravel_Rarely DailyRate DepartmentResearch & Development DepartmentSales
No  37.88215                       0.1802426                  0.7279029  827.9133                       0.6880416       0.2790295
Yes 32.70588                       0.2745098                  0.6470588  786.5588                       0.6274510       0.3529412
    DistanceFromHome Education EducationFieldLife Sciences EducationFieldMarketing EducationFieldMedical EducationFieldOther
No          8.693241   2.89948                   0.4246101              0.1074523            0.3206239          0.05199307
Yes         9.558824   2.77451                   0.3235294              0.1176471            0.3039216          0.06862745
    EducationFieldTechnical Degree EmployeeNumber EnvironmentSatisfaction GenderMale HourlyRate JobInvolvement JobLevel
No                       0.08145581       667.4922               2.762565  0.5649913   65.81802       2.809359 2.142114
Yes                      0.16666667       719.9706               2.441176  0.5882353   62.90196       2.519608 1.578431
    JobRoleHuman Resources JobRoleLaboratory Technician JobRoleManager JobRoleManufacturing Director JobRoleResearch Director
No              0.02599653                    0.1594454    0.083188908                   0.12131716              0.058925477
Yes             0.01960784                    0.2745098    0.009803922                   0.04901961              0.009803922
    JobRoleResearch Scientist JobRoleSales Executive JobRoleSales Representative JobSatisfaction MaritalStatusMarried
No                  0.2045061             0.2131716                 0.03466205       2.807626            0.4610052
Yes                 0.2647059             0.2254902                 0.11764706       2.460784            0.3039216
    MaritalStatusSingle MonthlyIncome MonthlyRate NumCompaniesWorked OverTimeYes PercentSalaryHike PerformanceRating
No            0.2911612      6842.355    14029.59           2.766031   0.2547660          15.15078          3.159445
Yes           0.5588235      4484.441    13743.40           3.068627   0.5980392          15.37255          3.156863
    RelationshipSatisfaction StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
No                  2.771231        0.8041594         11.916811              2.755633        2.769497       7.275563
Yes                 2.500000        0.4215686          7.607843              2.617647        2.774510       4.774510
    YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
No            4.396880               2.112652            4.310225
Yes           3.078431               1.862745            3.019608
```

Figure 19: Output from the LDA model is shown above

4. Insights and Conclusions

Based on the classifier model we selected, the test data for 470 employee were ran through the model. The resulting predictions show 59 out of 470 employees likely to churn. This corresponds to an attrition rate of approximately 12.6%, which is relatively close to the historic attrition rate captured by the company's original training data set. Additionally, the top four most important variables were found to be "OvertimeYes", "EnvironmentSatisfaction", "JobSatisfaction", and "BusinessTravelFrequent".

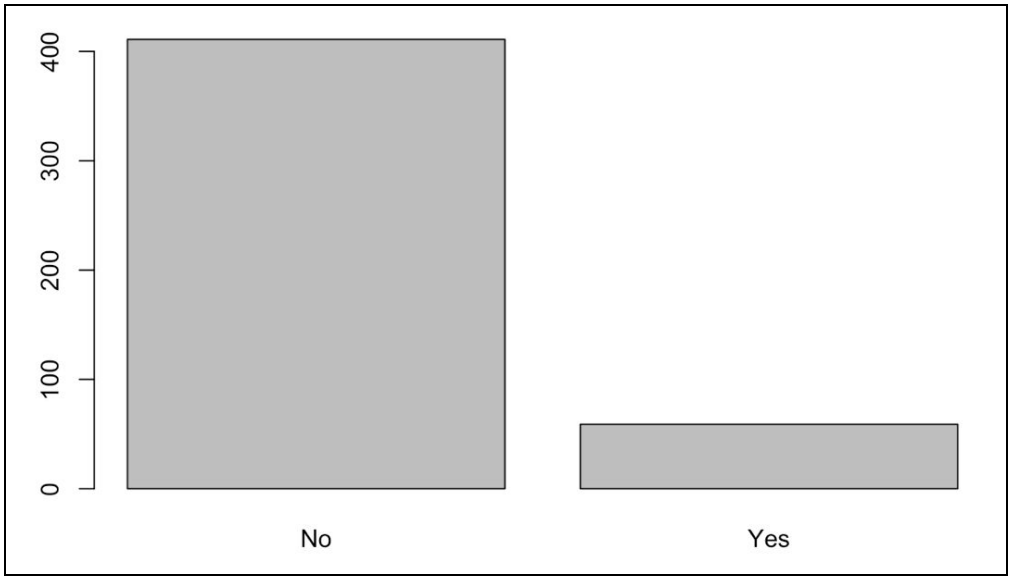


Figure 20: Output of classification of the 470 testing samples for employee entries. 59 employee predicted to churn.

Our recommendations would be to address these four variables of importance. The first variable of importance would be "OvertimeYes". To reduce the impact of this variable and reduce attrition the team recommends several items to address this. First, the company can create greensheets to see which departments are in significant need of additional employees and/or reallocate employees around the company. Second, the company can remove mandatory overtime or increase the pay for overtime. Lastly, the team recommends designating weekends/days that employees will know for sure they will not be required to work. This allows employees to plan ahead for activities with their family, which should also help in the next two variables of importance, job and environment satisfaction.

The second set of variables of importance would be "JobSatisfaction" and "EnvironmentSatisfaction". To reduce the impact of this variable and reduce attrition the team recommends making employees comfortable at the work environment as if they were home. Paying attention to what they need and provided it to them while working will not cost the organization that much and it will return great results for them. For example, providing a resting area so employees can use during break time. That area shouldn't be a normal rest are it should

contain almost everything the employee might need, they should conduct a survey and ask employees what they need. Additionally, the company could offer parties and/or free lunches for employees if certain metrics are met. This will help improve both environmental and job satisfaction.

The third variable of importance would be “BusinessTravelFrequent”. To reduce the impact of this variable and reduce attrition the team recommends lowering the requirements for allowing employees to travel. By allowing employees at all levels of the company to travel, it will allow the burden of travel to be diminished on the few current employees. Newer employees or trainees can be utilized for customer and/or supplier visits. Additionally, the company could take advantage of technological advancements to reduce the amount of travel required by employees. Skype, FaceTime, and numerous other services can be used to perform customer or supplier visits instead of on-site visits. An added benefit of this is that it should reduce the time and cost investment of these trips. What may have taken several days of travel can be completed in a few short hours via an online meeting.

References

[1] 2019. Bureau of Labor Statistics, Economic News Release, Retrieved from: <https://www.bls.gov/news.release/jolts.t16.htm>