Predicting if an Employee will Leave a Company with the HR Dataset

Morgan Hardin, mhardin5@bellarmine.edu

ABSTRACT
The dataset chosen for the analysis is an HR dataset and it consists of data the shows if an employee left a company along with 9 other variables that could have led to their departure. The dataset also contains information on the employee's satisfaction, last evaluation, number of projects worked on, average monthly hours, time spent at the company, work accidents, promotions in the last five years, department, and salary. The dataset was chosen because it contained a mix of numerical and categorical variables with no missing values. The goal is to see what variables have a positive correlation with an employee leaving and training a machine learning logistic regression model to predict if they will leave a company based on the 9 variables. With this model, a company can predict if an employee will leave based on these variables before the employee actually leaves.

I. INTRODUCTION

The HR dataset was 1 file consisting of 10 columns and 14,999 rows. There were no null entries in the dataset and there were 8 numerical columns and 2 categorical columns. The focus was on if an employee left a company or not and what variables correlate to leaving. This dataset was chosen because it had a lot of entries which helped give a better overview on how all of these variables impact an employee leaving the company. It is very interesting and different than the other datasets since each column was important in determining what causes a person to leave. The dataset focuses on 9 different variables that all impact if an employee left. This dataset shows just how much goes into determining an important decision such as quitting one's job. The HR dataset was found on Kaggle and the link can be found here: https://www.kaggle.com/datasets/jixiangruyi/predict-employee-left. The target is to determine if an employee will leave the company based on the 9 numerical and categorical variables.

II. BACKGROUND

The HR dataset helps companies see their turnover rate and what factors lead to an employee leaving their company. A company's turnover rate is the percentage of employees that leave in a specific time period. Every company has a turnover rate and employees want to be at a company where they can continue to grow and develop their career. The HR department keeps track of this information and analyzes the data to determine what plays into an employee leaving to try and prevent them from leaving before it is too late. This dataset was collected to analyze what variables play into an employee leaving to help a company understand why they left. This is an important dataset because it helps both the company and the employees. It helps the company understand what areas they need to improve to help an employee stay while helping the employee understand more about the company and where they stand. By predicting if an employee will leave, it can help a company lower their turnover rate.

III. EXPLORATORY ANALYSIS

The HR dataset consists of 1 file with 10 columns and 14,999 rows. Within the dataset, there are no null entries in any of the rows or columns. There are 8 numerical columns and 2 categorical columns. For all of the columns, 0% of the data is missing. The columns 'satisfaction_level' and 'last_evaluation' are numerical, float64 data types. They are discrete data types since there is a specified range from 0 to 1 as a decimal percentage. They are also ratio data since there is an absolute zero with a natural order. The columns 'number_project', 'average_monthly_hours', and 'time_spend_company' are numerical, int64 data types. They are continuous data types since there is a continuous range from 0 on. They are also ratio data since there is an absolute zero with a natural order. The columns 'Work_accident', 'left', and 'promotion_last_5years' are numerical, int64 data types. They are discrete data types since there is a specified range from 0 to 1. They are also ratio data since there is an absolute zero with a natural order. The column 'Department' is a categorical, object data type. It is a nominal data types since there is no order. The column 'salary is a categorical, object data type. It is an ordinal data types since there is order.

Table 1: Data Types

Variable Name	*Data Type*	*Missing Values*
satisfaction_level	Float64 (Numerical)	0%
last_evaluation	Float64 (Numerical)	0%
number_project	Int64 (Numerical)	0%
average_monthly_hours	Int64 (Numerical)	0%
time_spend_company	Int64 (Numerical)	0%
work_accident	Int64 (Numerical)	0%
left	Int64 (Numerical)	0%
promotion_last_5years	In64 (Numerical)	0%
department	Object (Object)	0%
salary	Object (Object)	0%

Table 2: Summary Statistics for HR Dataset – Numerical Columns

Variable Name	*Count*	*Mean*	*STD*	*Min*	*25%*	*50%*	*75%*	*Max*
satisfaction_level	14999	0.612834	0.248631	0.09	0.44	0.64	0.82	1
last_evaluation	14999	0.716102	0.171169	0.36	0.56	0.72	0.87	1
number_project	14999	3.803054	1.232592	2	3	4	5	7
average_monthly_hours	14999	201.050337	49.943099	96	156	200	245	310
time_spend_company	14999	3.498233	1.460136	2	3	3	4	10
work_accident	14999	0.144610	0.351719	0	0	0	0	1
left	14999	0.238083	0.425924	0	0	0	0	1
promotion_last_5years	14999	0.021268	0.144281	0	0	0	0	1

Table 3: Correlation Table for HR Dataset – Numerical Columns

Variable Name	satisfaction_level	last_evaluation	number_project	average_monthly_hours
satisfaction_level	1.000000	0.105021	-0.142970	-0.020048
last_evaluation	0.105021	1.000000	0.349333	0.339742
number_project	-0.142970	0.349333	1.000000	0.417211
average_monthly_hours	-0.020048	0.339742	0.417211	1.000000
time_spend_company	-0.100866	0.131591	0.196786	0.127755
work_accident	0.058697	-0.007104	-0.004741	-0.010143
left	-0.388375	0.006567	0.023787	0.071287
promotion_last_5years	0.025605	-0.008684	-0.006064	-0.003544

Table 3 Continued: Correlation Table for HR Dataset – Numerical Columns

Variable Name	time_spend_company	work_accident	left	promotion_last_5years
satisfaction_level	-0.100866	0.058697	-0.388375	0.025605
last_evaluation	0.131591	-0.007104	0.006567	-0.008684
number_project	0.196786	-0.004741	0.023787	-0.006064
average_monthly_hours	0.127755	-0.010143	0.071287	-0.003544
time_spend_company	1.000000	0.002120	0.144822	0.067433
work_accident	0.002120	1.000000	-0.154622	0.039245
left	0.144822	-0.154622	1.000000	-0.061788
promotion_last_5years	0.067433	0.039245	-0.061788	1.000000

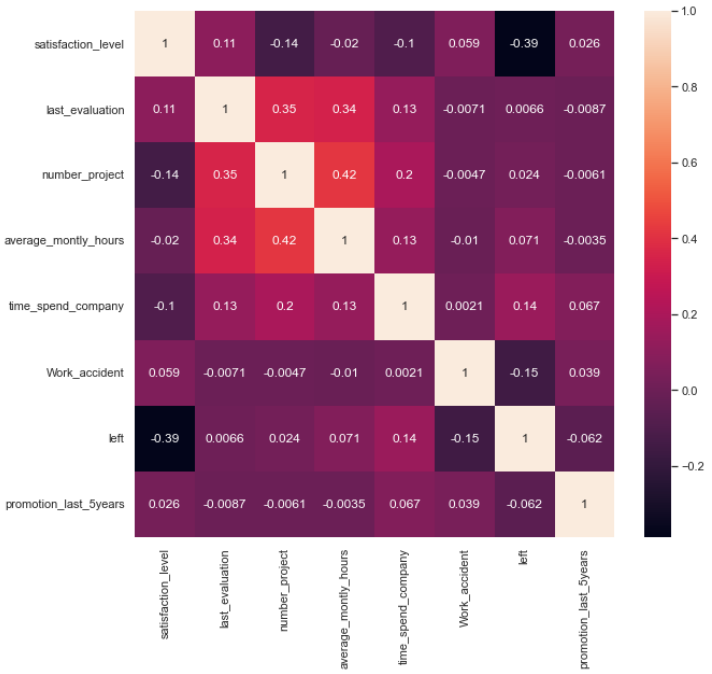

Figure 1: Correlation Matrix

Figure 1: This correlation matrix shows that the 'satisfaction_level', 'Work_accident', and 'promotion_last_5years' all have a negative correlation on if an employee left or not. This is an interesting find since in the original hypothesis, it was believed that salary would have a negative impact on an employee leaving. It makes since that a promotion might cause an employee to leave because they may feel stuck at their job and want to keep moving up to get a higher salary, but they cannot do that if they are not promoted which could lead them to leave their job. It was also found that the more projects an employee has, the more average monthly hours they have. These have a positive correlation between one another which could indicate that they need to put more hours in to finish their projects. It is also interesting to see that the number of projects has a somewhat weak, but positive correlation on 'last_evaluation'. This makes sense because if an employee performed well on their projects, then they are more likely to get a better evaluation score. All of the other categories did not have much of a significant impact on one other, so the most important were discussed.

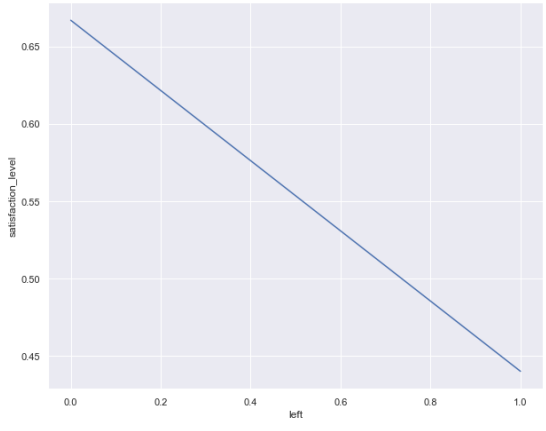

Figure 2: Line Plot of Left vs Satisfaction

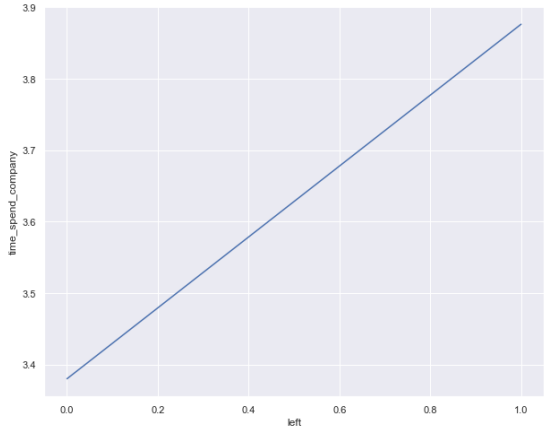

Figure 3: Line Plot of Left vs Time Spent at Company

Figures 2 and 3: These line plots show that there is a linear correlation between if an employee left and their satisfaction level, and if an employee left and if the time they spent at the company. These graphs show that the lower the satisfaction level, the more likely a person will leave their job. The second graph shows that the less time an employee spends at a company, the more likely they will leave the company.

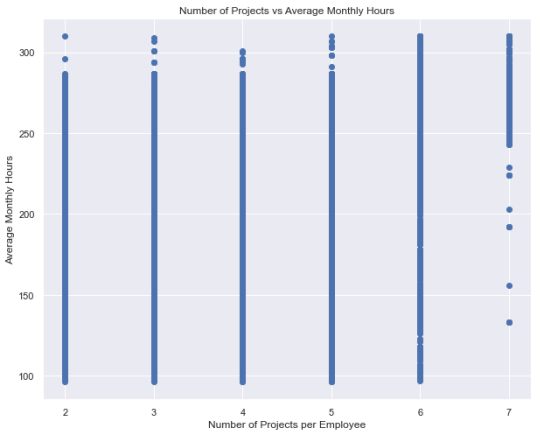

Figure 4: Scatter Plot of Project Number vs Monthly Hours

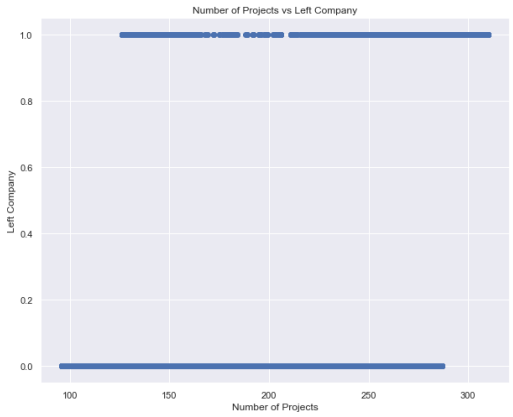

Figure 5: Scatter Plot of Project Number vs Left

Figures 4 and 5: These scatter plots show that the more projects an employee has, the more hours they will put in monthly. This can be seen from the majority of the plots being toward the bottom for the number of projects from 2 to 5. At 6, the majority begins to be found at the top of the hours range. At 7, there are a few outliers, but the majority put in between 240 and 320 hours a month. From the bottom scatter plot of Number of Projects vs Left Company, it is clear that the more projects an employee has, the less more likely they will leave the company. This can be seen from the majority of the projects between 75 and 275 all leaving the company. If a person left the company, none had fewer than 125 projects and there was a little bit of a gap in the middle where some employees stayed, even if they had an average number of projects. From this, it is clear that an employee will likely leave if they have more projects to work on and more hours to put in monthly.

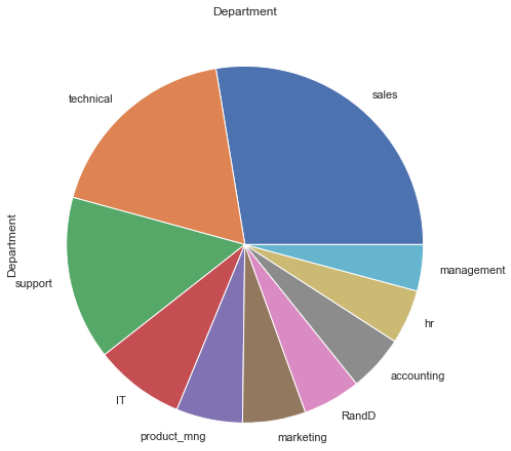

Figure 6: Pie Plot of Department

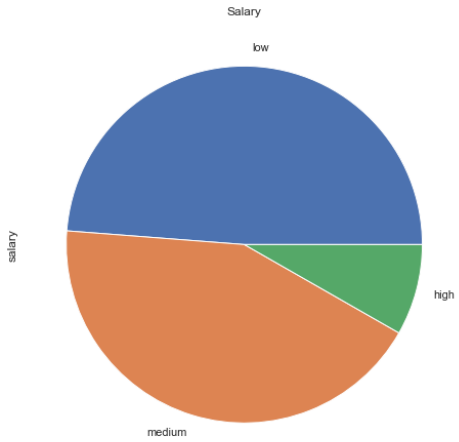

Figure 7: Pie Plot of Salary

Figure 6 and 7: These pie plots show that there is mostly an even distribution between the smaller departments, but sales, technical, and support are the biggest departments. The pie plot for salary shows that there is about an even distribution between low and medium salaries, but there are a lot fewer high salaries. This all makes sense because the bigger departments will likely have a lower pay because the departments are bigger and don't necessarily require

as much education and experience. For example, the sales department needs more people and does not require as much experience, so it does not pay as much. The accounting department requires education and skill and is a smaller department, so it is likely to pay more. So, the bigger the department, the more likely the pay will be low or medium. The smaller the department, the more likely the pay will be higher. With this being said, this is from these pie plots and depends on lots of other variables that are not listed or in this dataset.

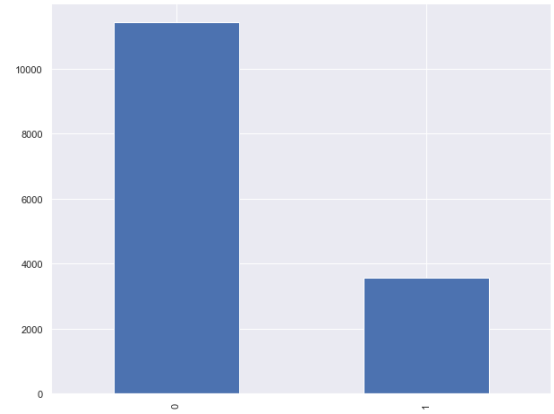

Figure 8: Bar Plot of Left

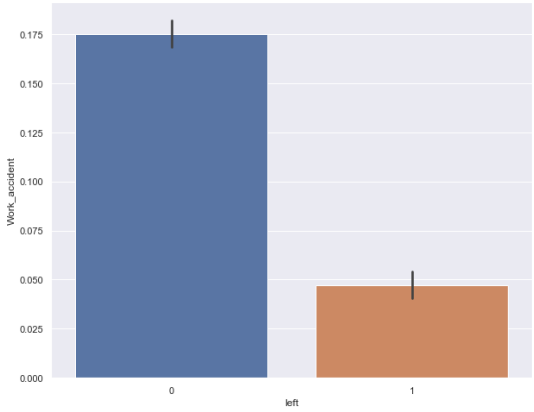

Figure 9: Bar Plot of Left vs Work Accident

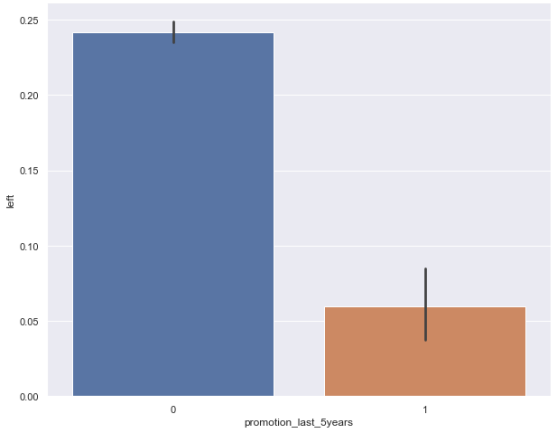

Figure 10: Bar Plot of Promotion vs Left

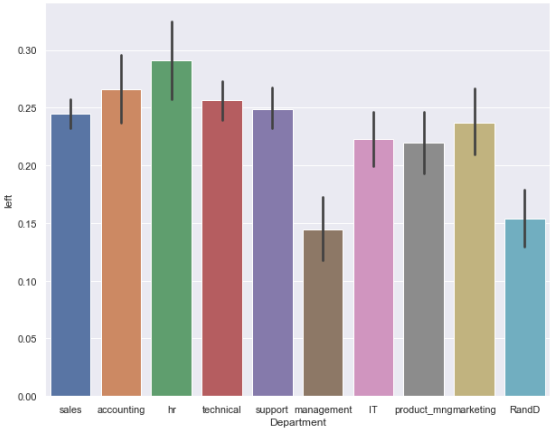

Figure 11: Bar Plot of Department vs Left

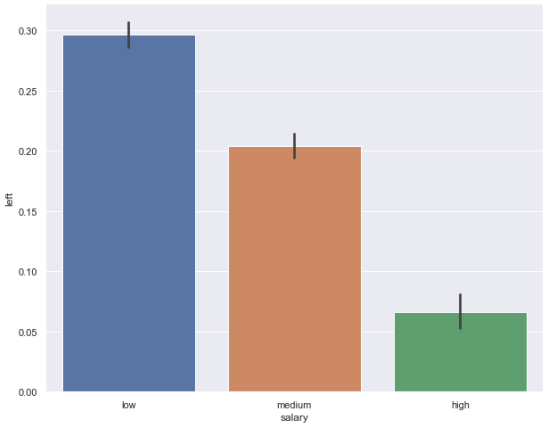

Figure 12: Bar Plot of Salary vs Left

Figure 8, 9, 10, 11, and 12: These bar graphs show that the values are skewed and there are more 0's than 1's in the dataset. This means that more people stayed at the company in this dataset than left. It also shows that the majority of people who had work related accidents and had not been promoted in the last 5 years left the company. The employees in the HR, accounting, and technical departments were the ones who left their job the most. It also states that low salaries were the ones that left their jobs the most. This could mean that employees in HR, accounting, and technical left their jobs due to their salaries or not being promoted.

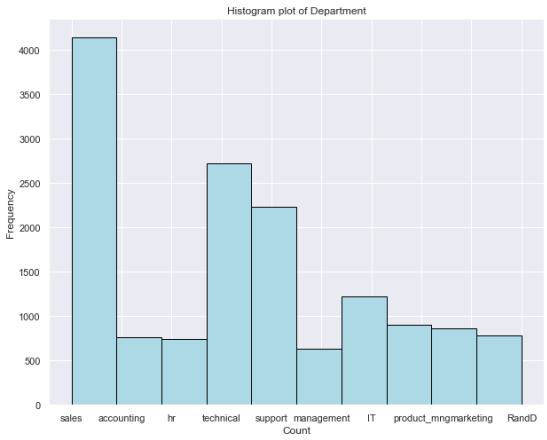

Figure 13: Histogram Plot of Department

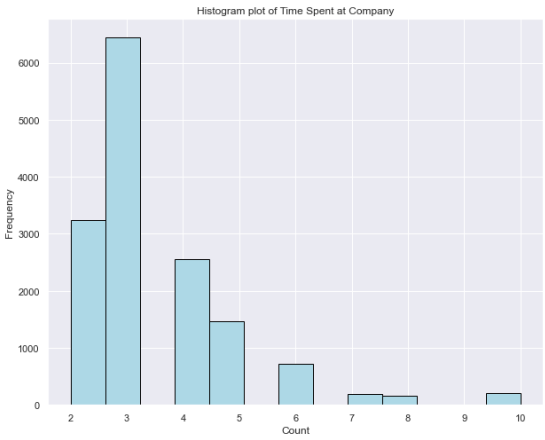

Figure 14: Histogram Plot of Time at Company

Figures 13 and 14: These histograms show that sales, technical, and support are the biggest departments. This matches the pie plot with these departments being the biggest. The sales department has about 4250 employees, the technical department has around 2600 employees, and the support department has about 2250 employees. The other histogram also shows that most employees have been with the company around 3 years. The graph is right skewed, so most of the employees fall between the 2 and 5 year range. This could mean that most people leave the company after about 5 years of being employed, so it is likely that they are coming from either the sales, technical, or support departments since they have the most employees.

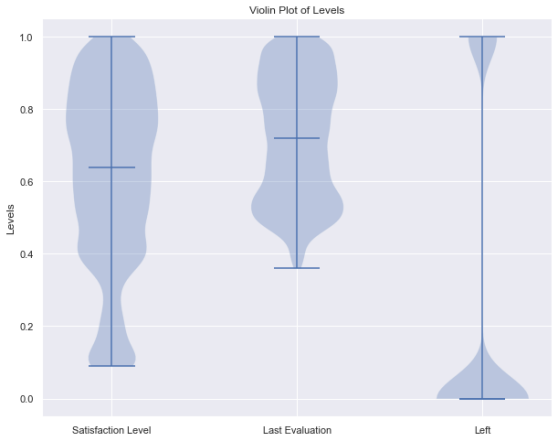

Figure 15: Violin Plot of Satisfaction Levels, Last Evaluation, and Left

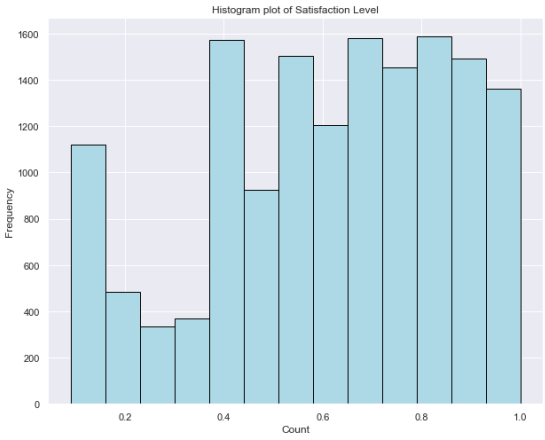

Figure 16: Histogram Plot of Satisfaction Levels

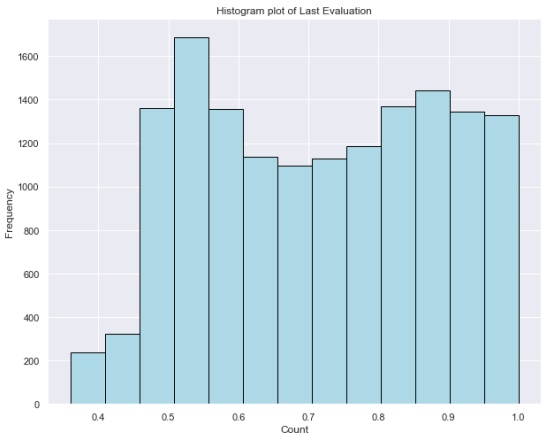

Figure 17: Histogram Plot of Last Evaluation

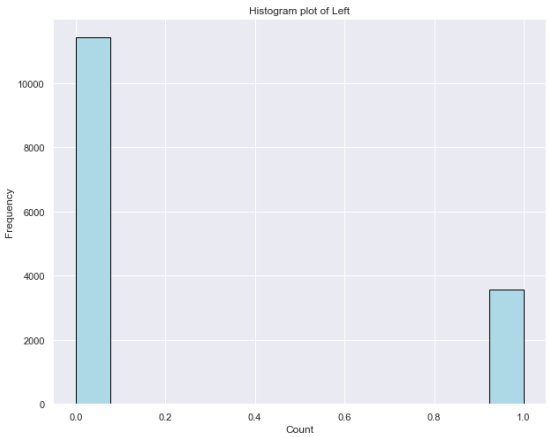

Figure 18: Histogram Plot of Left

Figures 15, 16, 17, and 18: The violin and histogram plots show that an employee's satisfaction level matches roughly to what their last evaluation level was. The higher their satisfaction level is, the higher their last evaluation level was. This means that they are less likely to leave the company. If the satisfaction level was low, it might mean that their last evaluation level was low and that they are more likely to leave the company. This can be seen with both the violin plots and the left skewed histograms for both satisfaction level and last evaluation compared to leaving the company. This means that the higher the level, the less likely they will leave the company.

IV. METHODS

The HR dataset did not require much preparation for the model to be trained. The data needed to be split into independent and dependent variables in order to be trained and tested. The 'satisfaction_level', 'last_evaluation', 'number_project', 'average_monthly_hours', 'time_spend_company', 'Work_accident', 'promotion_last_5years', 'Department', and 'salary' columns are all independent variables that were put into the x variable. The 'left' column is the dependent variable that was put into the y variable. From here, the categorical data was encoded using the pandas get dummies. After this, the model is set to be trained using logistic regression. This trains the data to predict if an employee will leave a company or not based on the 9 independent variables. The model was trained using various sizes for the training and testing sets in order to determine which split gave the highest accuracy score for the model.

A. Data Preparation

The HR dataset was prepared to be trained with logistic regression by splitting the dataset into independent and dependent variables. The 'satisfaction_level', 'last_evaluation', 'number_project', 'average_monthly_hours', 'time_spend_company', 'Work_accident', 'promotion_last_5years', 'Department', and 'salary' columns are all independent variables that were put into the x variable. The 'left' column is the dependent variable that was put into the y variable. Pandas get dummies was used to handle the categorical variables 'Department', 'salary', and 'left' and the first column was dropped to preserve space. The 'salary_low' is labeled as 100, 'salary_medium' as 010, and 'salary_high' as 000. The 'Department_RandD' is labeled as 0100000000, 'Department_accounting' as 0010000000, 'Department_hr' as 0001000000, 'Department_management' as 0000100000, 'Department_marketing' as 0000010000, 'Department_product_mng' as 0000001000, 'Department_sales' as 0000000100, 'Department_support' as 0000000010, 'Department_technical' as 0000000001, and 'Department_it' as 0000000000. The 'left' column is labeled 0 for the employee staying at the company and 1 for the employee leaving the company. Every column will be used for the training model so none of them will be dropped. The data is now set to be trained using logistic regression.

B. Experimental Design

The HR dataset logistic regression model was trained four different times in order to determine the highest accuracy score for predicting if an employee will leave a company. It was trained with an 85/15 split, an 80/20 split, and a 70/30 split. Each of these was also trained with a different random state of either 42 or 20. The independent and dependent variables were already prepared into x and y, so these were split into x_train, x_test, y_train, and y_test. The training variables consisted of either 85, 80, or 70 percent of the data while the testing variables consisted of either 15, 20, or 30 percent of the data. The model was trained and the highest accuracy score was chosen.

Table 4: Experiment Parameters

Experiment Number	Parameters	Accuracy Score
1	85/15 Split for Train and Test with a Random State of 42	78%
2	85/15 Split for Train and Test with a Random State of 20	80%
3	80/20 Split for Train and Test with a Random State of 42	78%
4	80/20 Split for Train and Test with a Random State of 20	79%
5	70/30 Split for Train and Test with a Random State of 42	78%
6	70/30 Split for Train and Test with a Random State of 20	79%

Table 4 shows the 6 experiments that were tested to determine the highest accuracy score for the logistic regression model. It shows an 85/15 split, 80/20 split, and a 70/30 split, each with a random state of either 42 or 20. The accuracy scores are all very close together and pretty high for the model, but the highest accuracy score was 80% with an 85/15 split and a random state of 20. With this being said, the other models were still very accurate being either 1 or 2 percent away from the 80%.

C. Tools Used

The tools used were Anaconda v23.7.4, Anaconda Navigator v2.5.0, Jupyter Notebook v6.4.8, Python v3.11.5, Matplotlib v3.7.2, Numpy v1.24.3, Pandas v2.0.3, Seaborn v0.12.2, and Scikit-Learn v1.3.0. Anaconda, Anaconda Navigator, Jupyter Notebook, and Python were chosen because of the environment and necessary tools to write and execute code to process and analyze data. The specific tools like Matplotlib, Numpy, Pandas, Seaborn, and Scikit-Learn were used to import the dataset, explore and analyze the data, create graphs to visualize the data, handle categorical variables, and train a logistic regression model. All of these tools are essential in predicting if an employee left a company and analyzing the data and results.

V. RESULTS

A. Classification Measures/ Accuracy measure

For the logistic regression model on the HR dataset, the R2 Score, Mean Squared Error, and RMSE was calculated to show how well the model was trained. A confusion matrix and classification report was also calculated to find the accuracy of the model.

Table 5: R2 Score, MSE, and RMSE

R2 Score	-0.1279503105590063
Mean Squared Error	0.20177777777777778
RMSE	0.4491968140779471

This table shows that the R2 Score is not good since it is negative and not between 0.5 and 1. This means that the model does not fit the data very well. With this being said, the Mean Squared Error value is pretty low meaning that the accuracy will be high for the dataset. The RMSE is also extremely low meaning that the model makes accurate predictions and the data is fit well. The R2 Score does not match the findings of the MSE and RMSE values, but this could be because the model fits the 0's better than the 1's so the model struggles with the 1's causing the line of best fit is skewed to try to fit these 1 values.

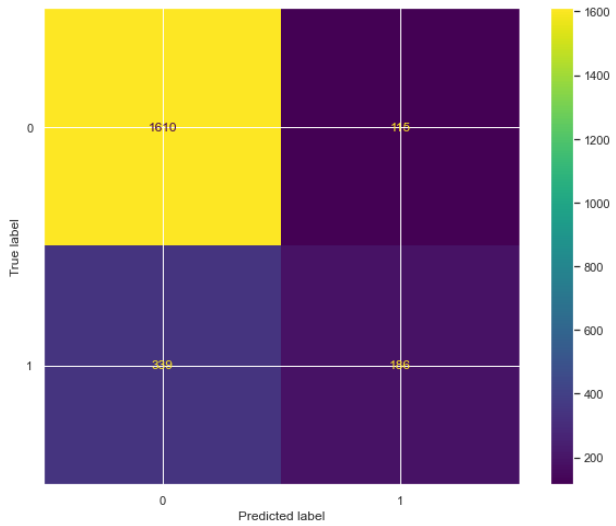

Figure 19: Confusion Matrix Heat Map

Table 6: Classification Report

	Precision	**Recall**	**Support**	**Support**
0	0.83	0.93	0.88	1725
1	0.62	0.35	0.45	525
Accuracy			0.80	2250
Macro Average	0.72	0.64	0.66	2250
Weighted Average	0.78	0.80	0.78	2250

This shows that the model trained pretty well with 80% being the highest test case with a test size of 0.15 and a random state of 20. The model could have trained better, but the distribution of 0's and 1's was skewed and there were more than 3 times as many 0's as there were 1's. There were also 1610 true negatives, 339 false negatives, 115 false positives, and 186 true positives. Therefore, it did a better job of predicting if an employee would stay than it did for an employee leaving.

B. Discussion of Results

Overall, the model trained pretty well with a high accuracy score with an 85/15 split with a random state of 20. The other accuracy scores were very close, but this pairing gave the highest accuracy score of 80%. Although the MSE and RMSE was low, the R2 score was negative, meaning that the model fit somewhat well, but there is room for improvement. The model could be trained better with more 1's to even out the skew of their being more 0's than 1's. This could have led to the model training at only 80% and the R2 Score being negative while the MSE and RMSE was low. For this dataset, the "worst" model had an accuracy of 78% with multiple of the experiments leading to that score. This model was pretty even across the board for the accuracy scores, so it could be improved with the addition of more employees leaving a company to make the data less skewed.

C. Problems Encountered

The biggest problem faced was finding a good dataset that had enough columns and numerical and categorical data. This dataset seemed fitting for logistic regression since it was predicting if an employee would leave a company, so the results being either 0 or 1. There were also issues with finding the best accuracy score. No matter what split and random state was changed, the model continued to fit between 78 and 80 percent. This issues led to the belief that the model accuracy would not worsen or improve due to the data. Because of the data being skewed with more 0's than 1's, the model could not improve above 80%. These issues were easily fixed but took a while to find the right dataset and analyze the results in a way that made sense.

D. Limitations of Implementation

The main limitation of this model was the skewed number of 0's and 1's. According to the classification report, the model consisted of 1,725 0's and only 525 1's. This shows that there were about 3 times as many 0's as there were

1's. This definitely prevented the accuracy score from improving because the F1-Score was 88% for 0's, but only 45% for 1's. If the 1's had a higher score, then the average of the accuracy score would improve significantly. Other models may work better that deal with skewed data. Possibly fixing the skew would also help the accuracy score improve.

E. Improvements/Future Work

To improve the model, getting more data where employees leave a company would help raise the accuracy score. With this being said, it also makes sense that the number of employees who leave a company would be less than the number of employees who stay at a company. It would also be worth considering more variables that could be taken into account while predicting if an employee will leave. This could possibly be work ethic, number of hours worked each day, or number of sick days taken. This dataset might also do better with a different model that deals with skewed data. Finding a different dataset that is more evenly distributed would also help the accuracy to improve.

VI. CONCLUSION

Overall, an employee's satisfaction level and if they received a promotion in the last 5 years played the most important role in determining if an employee would leave a company or not. All of the independent variables were extremely helpful in this dataset and showed a correlation between leaving or not. The satisfaction level showed that the higher an employee's satisfaction, then the higher their last evaluation was and the less likely they will leave the company. It was also found that the more projects an employee has, the more hours they will be putting in monthly and the more likely they are to leave the company. This all makes sense because they are putting lots of hours in and very overwhelmed so their satisfaction level drops and can cause their evaluation to fall, meaning they are more likely to leave the company. It was also discovered that most employees who leave have not been at the company long and have been there between 2 and 5 years. This makes sense that the range is up to 5 years because if an employee had not been promoted in 5 years, they will likely leave the company. Also, accidents while at work did play a small roll, but not a huge one. It indicated that work accidents have a negative correlation with leaving so if a person had an accident, they will likely leave. The departments who had the most employees leave was hr, accounting, and technical. It is likely that these departments do not get paid as much or their satisfaction levels were low. Finally, the salary played a pretty big role in indicating if a person would leave. It was found that the lower the salary, the more likely an employee will leave. The model also trained pretty well at 80% and did a decent job of predicting if an employee would leave or not. The model did better at predicting 0's, if an employee stayed, then 1's, if they left. This might be because there are more 0's than 1's in the dataset, leaving it skewed. This model could be more accurate if there was a more even distribution of 0's and 1's so it can train more accurately to predict if an employee would leave the company. In conclusion, the hypothesis was correct and the model trained pretty well at 80% accuracy. This was a decent model that was not bad but could still be improved. The accuracy score being only 80% shows that it could be improved but is still a good model. Although it needs more work, the model does a pretty good job of predicting if an employee will stay or leave a company.

REFERENCES

WeiZhang. "Predict_employee_left." Kaggle, 18 Oct. 2023,https:// www.kaggle.com/datasets/jixiangruyi/predict-employee-left.