HR Dataset

By: Morgan Hardin

Background

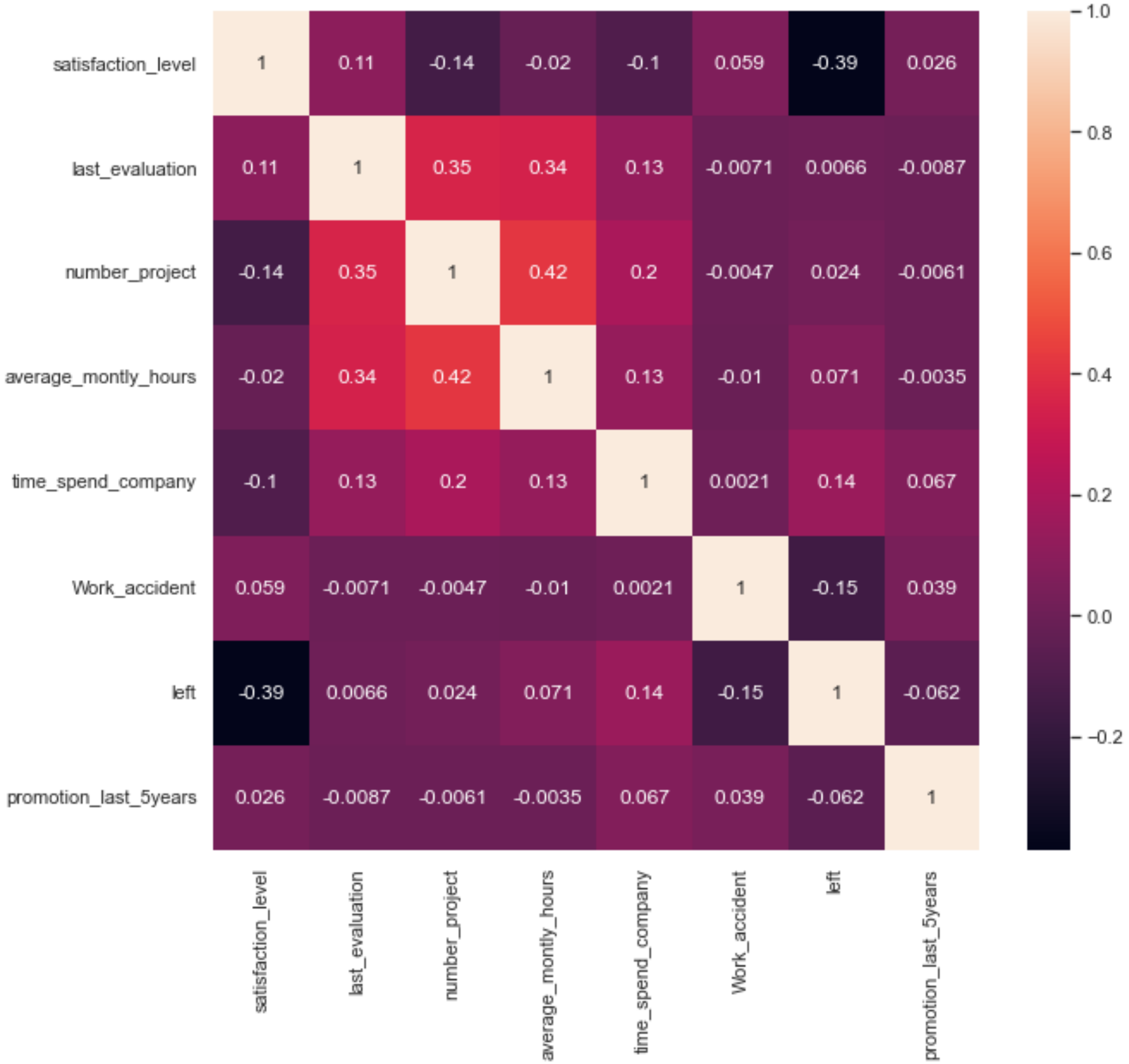
- **Goal: Predict if an employee will leave a company or not**
- **Hypothesis: Satisfaction, Hours Worked, and Salary have the most impact on an employee leaving**
- HR dataset helps companies see their turnover rate and what leads to an employee leaving their company
- Turnover Rate: percentage of employees that leave in a specific time period
- 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
- Dataset was collected to analyze what variables play into an employee leaving to help a company understand why they left
- Helps the company understand what areas they need to improve to help an employee stay
- Helps 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.

Introduction

- 1 file with 10 columns and 14,999 rows
- Found on Kaggle: https://www.kaggle.com/datasets/jixiangruyi/predict-employee-left
- No null entries
- 8 numerical columns
- 2 categorical columns
- 2 Float data types
  - Decimal between 0 and 1 to represent percentage
- 6 Integer data types
- 2 Object data types
- Independent variables were:
  - ‘satisfaction_level’, ‘last_evaluation’, ‘number_project’, ‘average_monthly_hours’, ‘time_spend_company’, ‘work_accident’, ‘promotion_last_5years’, ‘Department’, and ‘salary’
- Dependent variable:
  - ‘left’ --> 0 means Stayed and 1 means Left

Correlation Matrix

- ‘satisfaction_level’, ‘Work_accident’, and ‘promotion_last_5years’ have negative correlation with an employee leaving
- Different than original hypothesis that salary would negatively impact if they left
- Number of projects and hours worked also have positive correlation
- All other categories were too close to 0 to show an impact

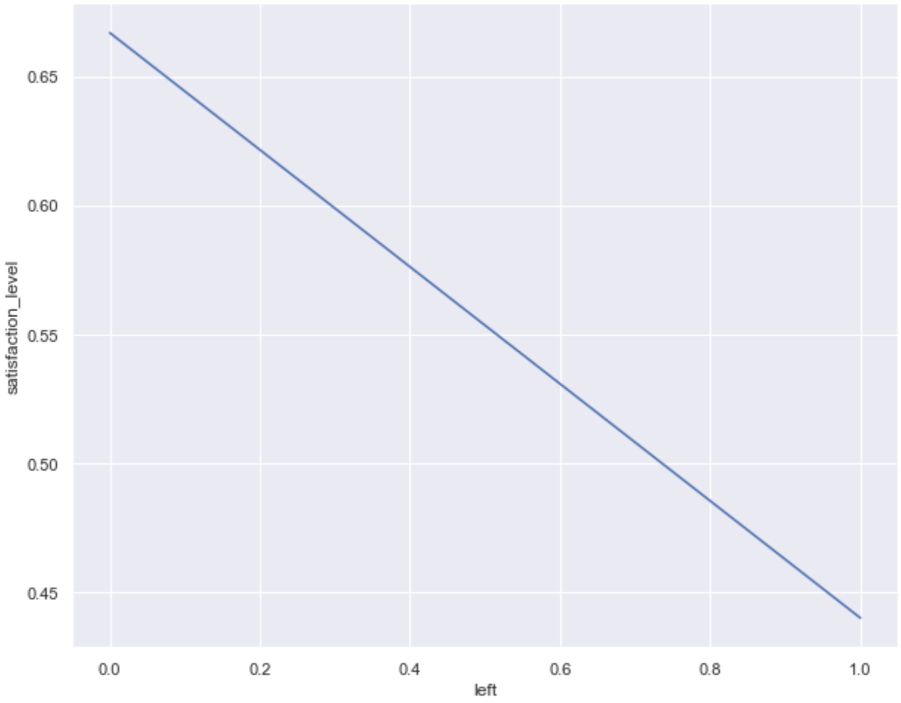


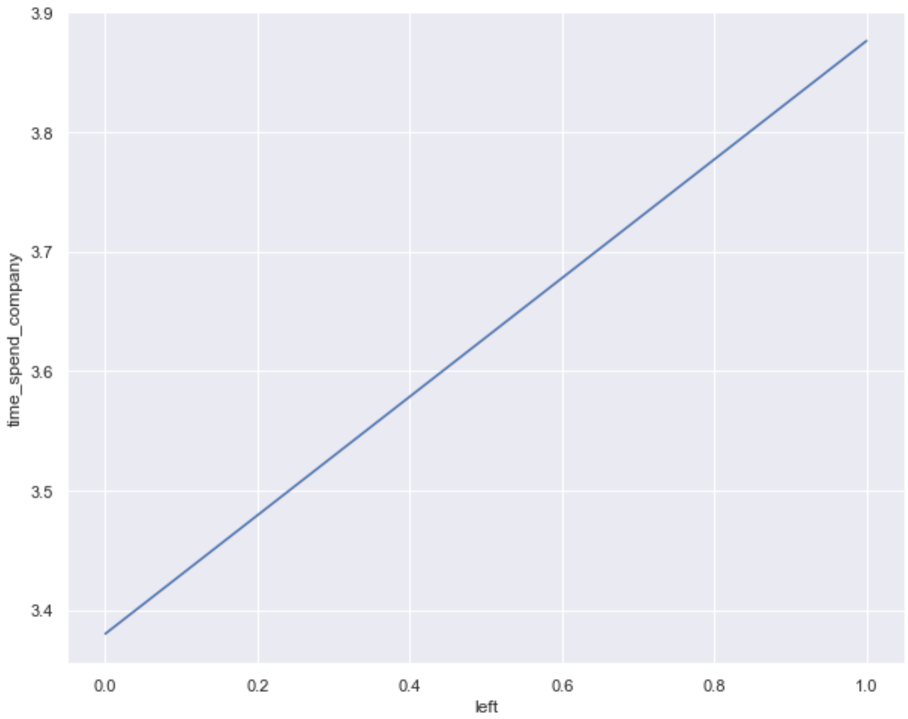
	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years
satisfaction_level	1	0.11	-0.14	-0.02	-0.1	0.059	-0.39	0.026
last_evaluation	0.11	1	0.35	0.34	0.13	-0.0071	0.0066	-0.0087
number_project	-0.14	0.35	1	0.42	0.2	-0.0047	0.024	-0.0061
average_montly_hours	-0.02	0.34	0.42	1	0.13	-0.01	0.071	-0.0035
time_spend_company	-0.1	0.13	0.2	0.13	1	0.0021	0.14	0.067
Work_accident	0.059	-0.0071	-0.0047	-0.01	0.0021	1	-0.15	0.039
left	-0.39	0.0066	0.024	0.071	0.14	-0.15	1	-0.062
promotion_last_5years	0.026	-0.0087	-0.0061	-0.0035	0.067	0.039	-0.062	1

Line Plot of Satisfaction and Time Spent vs Left

- Higher the satisfaction level, the less likely an employee will leave

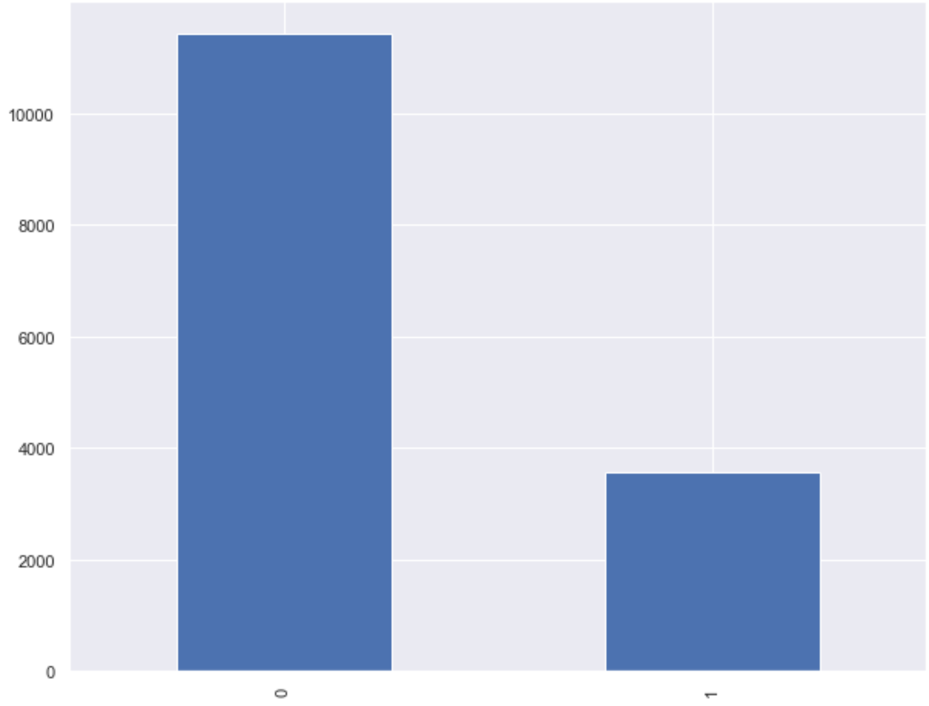
- More time spent at a company, the less likely an employee will leave



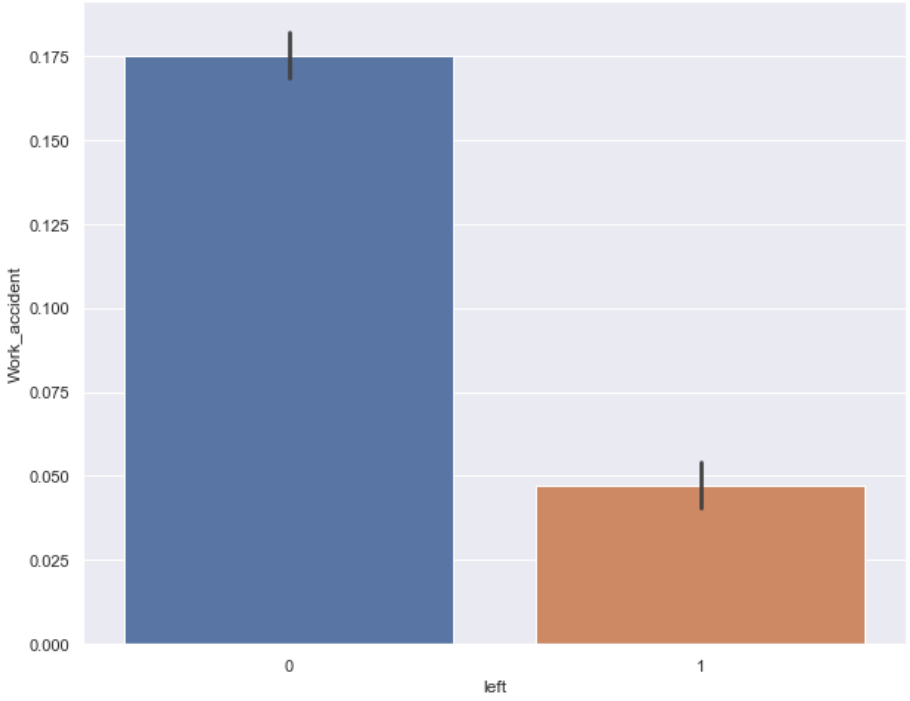


Bar Plot of Left Frequency and Left vs Work Accident

- More people staying at a company than leaving

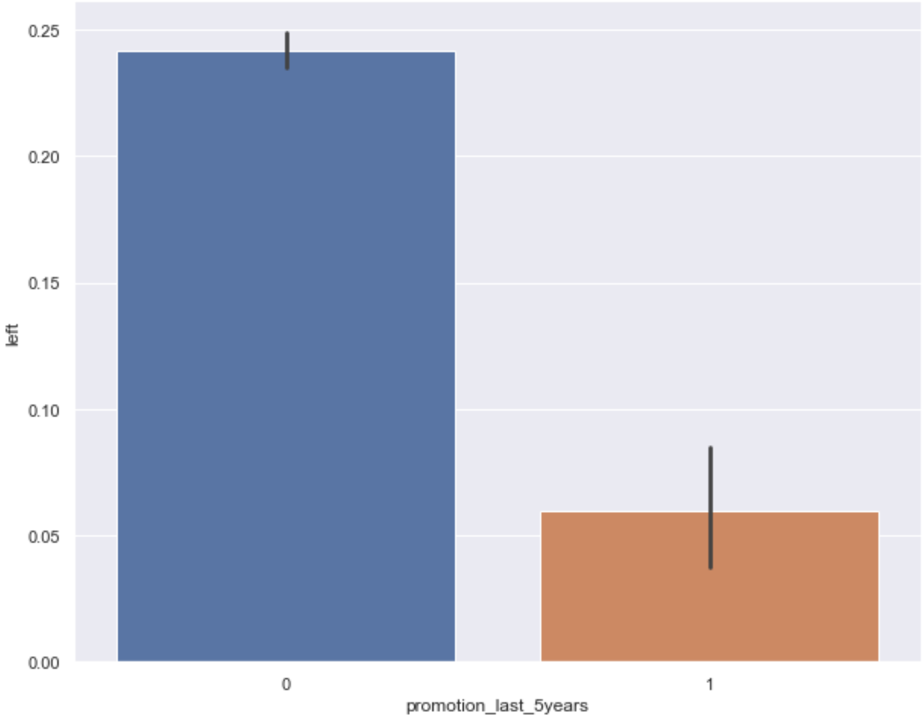


- Contradicts with Correlation Matrix since more employees with Work accidents are staying instead of leaving

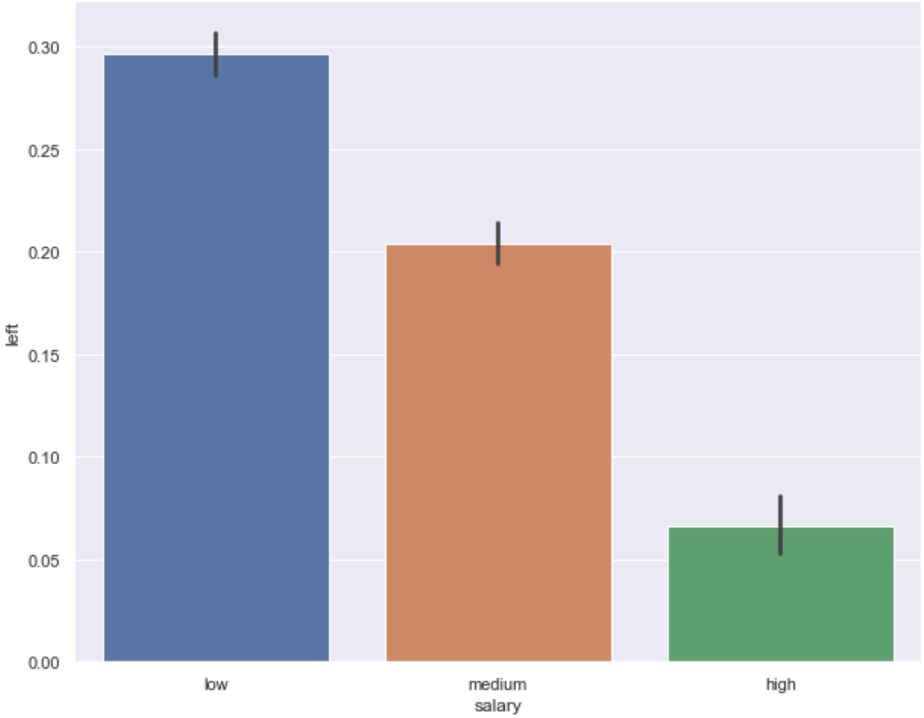


Bar Plot of Promotion in 5 Years and Salary vs Left

- Employees without a promotion in last 5 years are more likely to leave



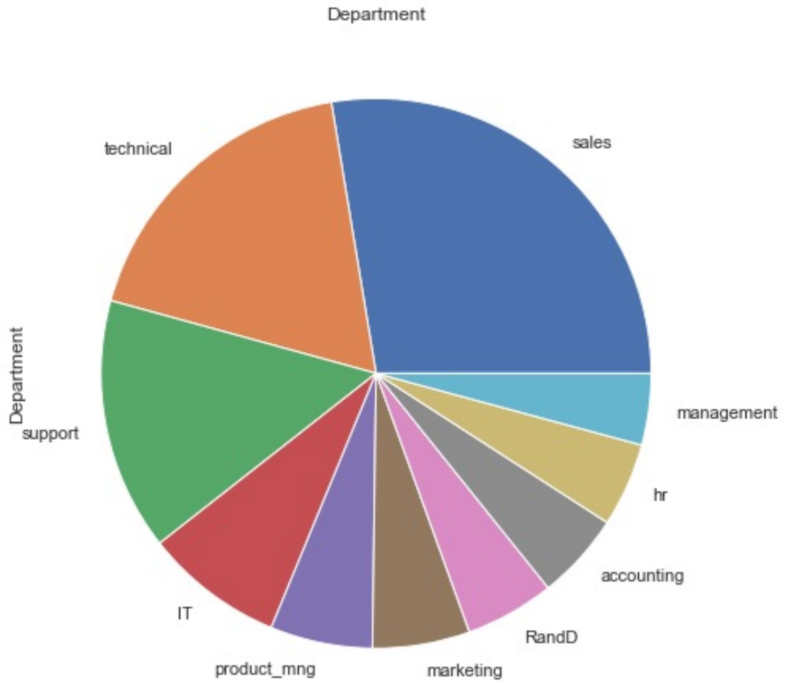
- Shows salary does have an impact on leaving since employees with a low salary are likely to leave the company

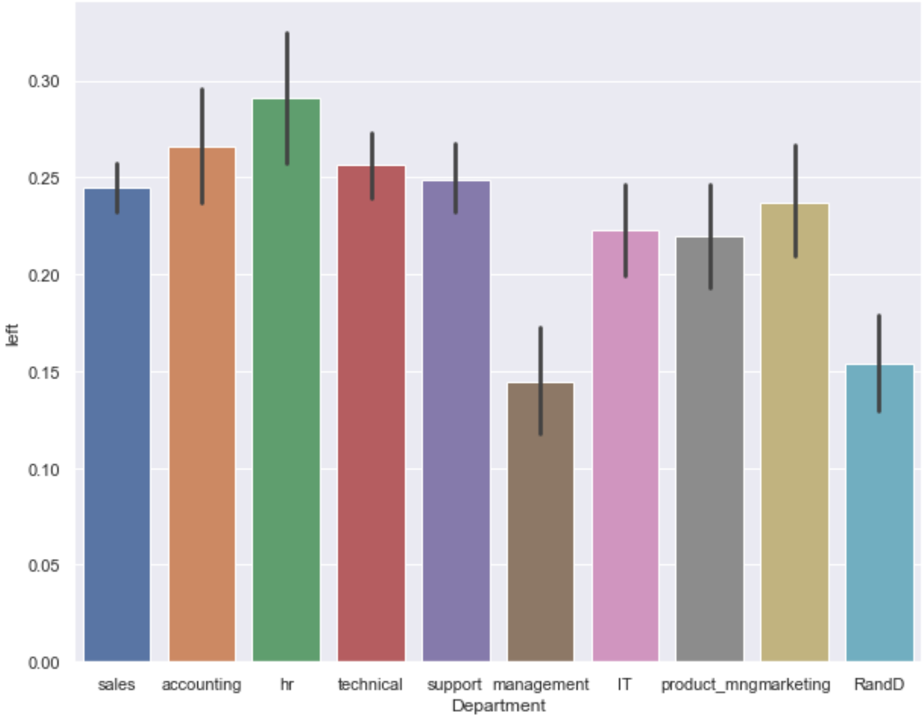


Plots of Department and Department vs Left

- Biggest departments are sales, technical, and support

- Highest turnover rate is in HR, Accounting, and Technical departments

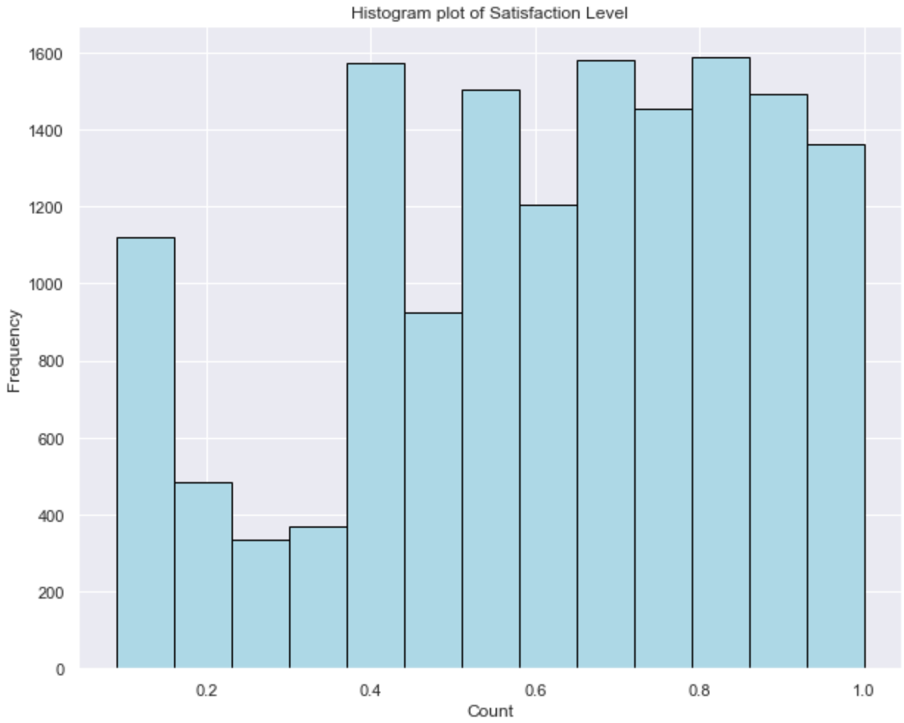


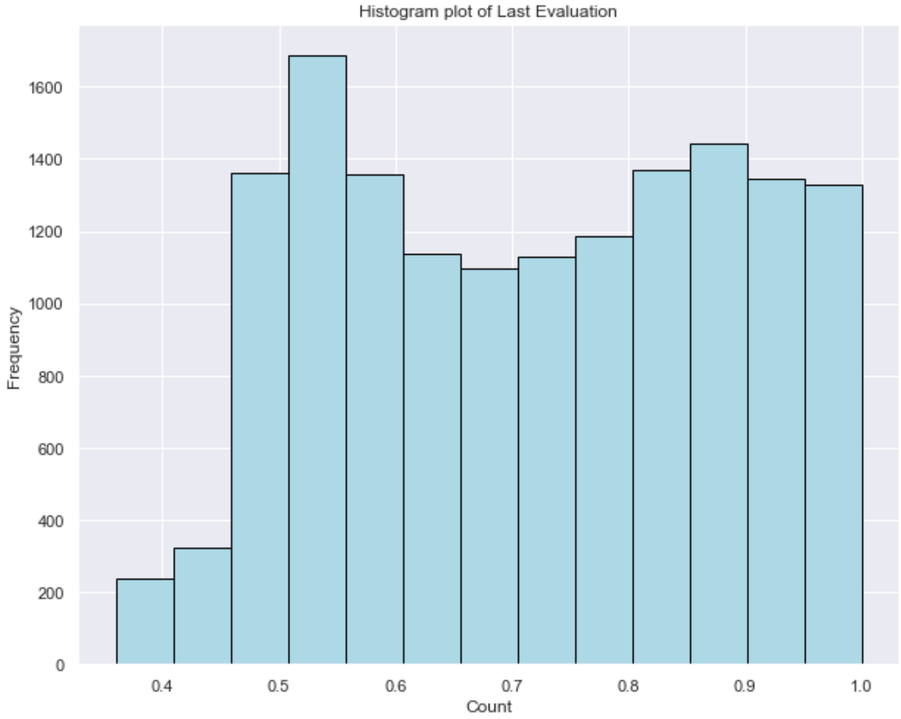


Plots of Satisfaction Level and Last Evaluation

- Left skewed where most employees have good satisfaction levels

- Also left skewed where last evaluations were mostly on the high end



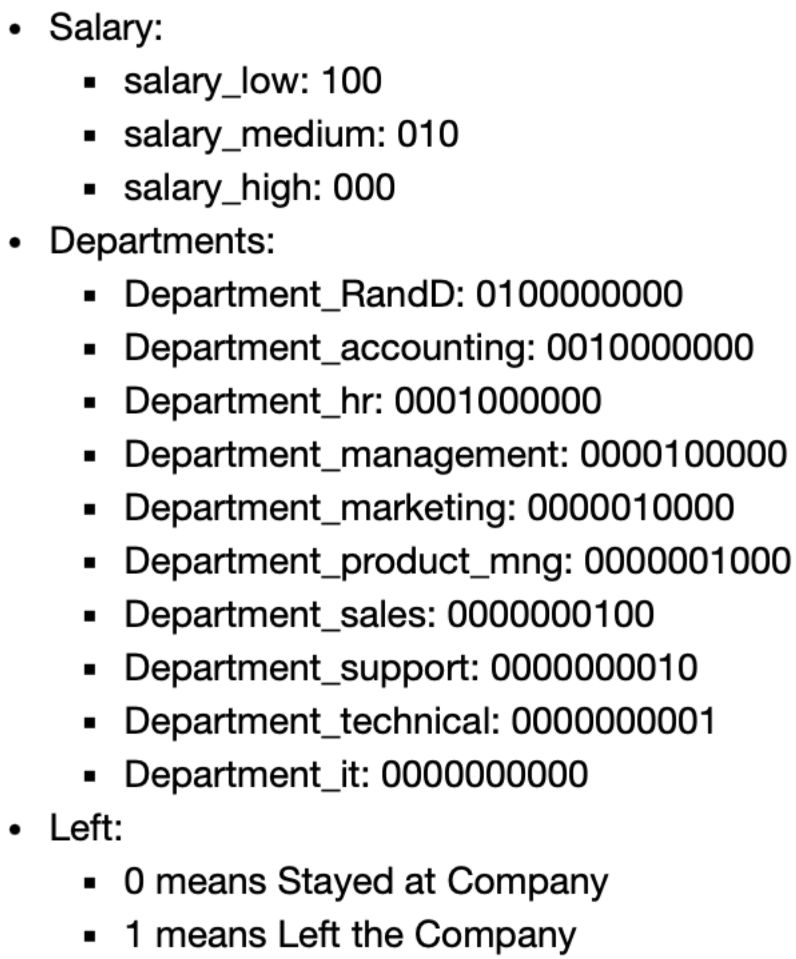


Results of Exploration

- Correlation matrix was correct about satisfaction levels and promotion in the last 5 years impacted if an employee left
- Satisfaction levels also had a positive correlation with last evaluation showing that this correlates with an employee leaving
- Found that the longer an employee is at a company, the more likely they will stay
  - Possibly due to higher salary
- Correlation matrix was not accurate about work accidents making an employee leave a company
  - Possibly due to workers compensation
- Employees without a promotion in the last 5 years or who have a low salary are likely to leave
- HR, Accounting, and Technical departments have the highest turnover rate
  - Possibly due to low salary, high work hours, and stress
- Most columns in the dataset had a negative correlation with an employee leaving a company

Logistic Regression

- Split into x and y variables with x being independent and y being dependent (‘left’ column)
- Used get dummies to handle the categorical data
- Used 3 different split sizes and 2 random states:
  - 85/15 split with random state of 42
  - 85/15 split with random state of 20
  - 80/20 split with random state of 42
  - 80/20 split with random state of 20
  - 70/30 split with random state of 42
  - 70/30 split with random state of 20



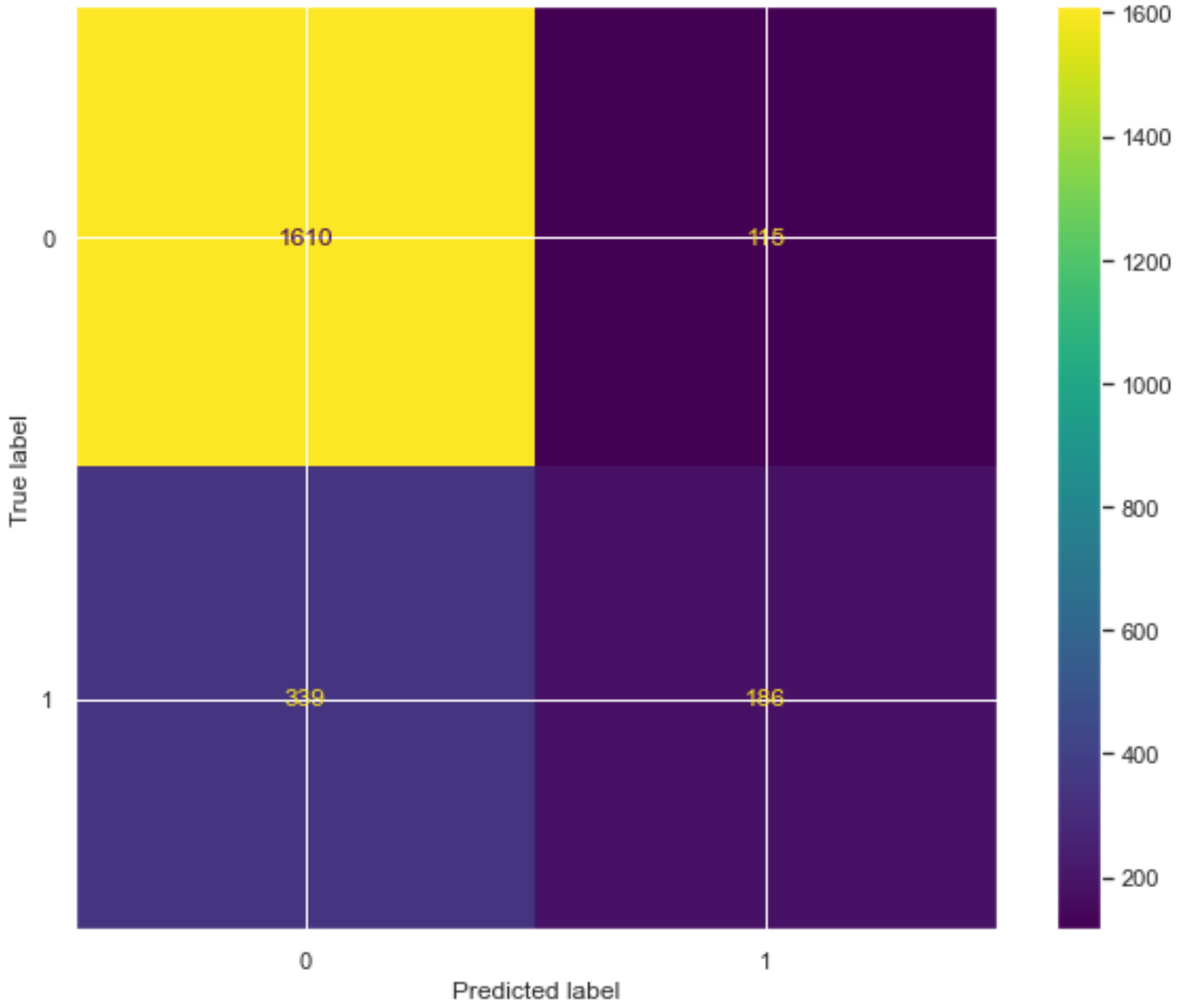
- Salary:
  - salary_low: 100
  - salary_medium: 010
  - salary_high: 000
- Departments:
  - Department_RandD: 0100000000
  - Department_accounting: 0010000000
  - Department_hr: 0001000000
  - Department_management: 0000100000
  - Department_marketing: 000010000
  - Department_product_mng: 0000001000
  - Department_sales: 0000000100
  - Department_support: 0000000010
  - Department_technical: 0000000001
  - Department_it: 0000000000
- Left:
  - 0 means Stayed at Company
  - 1 means Left the Company

Accuracy Scores, R2, MSE, and RMSE

- Accuracy scores for experiments with different splits and random states:
  - 85/15 split with random state of 42: 78%
  - **85/15 split with random state of 20: 80%**
  - 80/20 split with random state of 42: 78%
  - 80/20 split with random state of 20: 79%
  - 70/30 split with random state of 42: 78%
  - 70/30 split with random state of 20: 79%
- R2 Score: -0.1279503105590063
  - Model is not a good fit for the dataset since it's negative
  - Possibly from the line of best fit being skewed from there being more 0's than 1's in the dataset
- MSE: 0.20177777777777778
  - Low score which means model is fit well and accuracy will be high
- RMSE: 0.4491968140779471
  - Low score which means model is fit well and makes accurate predictions

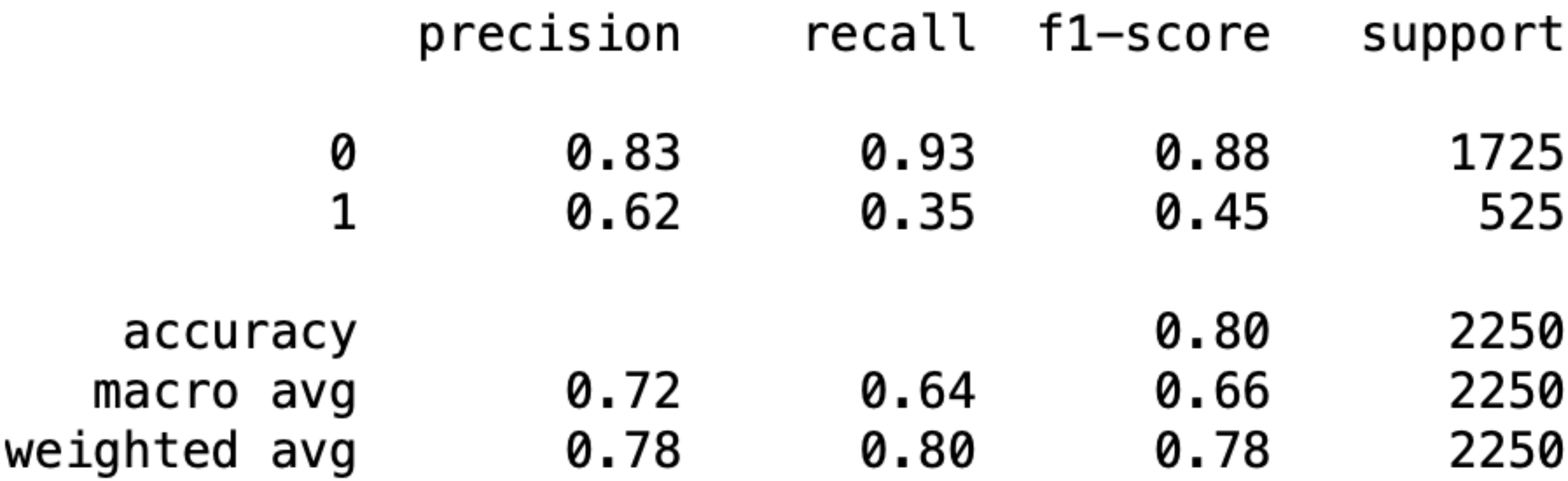
Confusion Matrix

- Predicted 1,610 true negatives, 339 false negatives, 115 false positives, and 186 true positives
- Did a better job of predicting 0’s than 1’s meaning the model did a better job of predicting if an employee would stay than if an employee would leave
- Due to the skewed distribution of 0’s and 1’s in the dataset



Classification Report

- About 3 times as many 0’s than 1’s
- Model did better at predicting 0’s than 1’s
- Could have higher accuracy score to raise the average if there was a more even distribution



```
              precision    recall  f1-score   support

           0       0.83      0.93      0.88      1725
           1       0.62      0.35      0.45       525

    accuracy                           0.80      2250
   macro avg       0.72      0.64      0.66      2250
weighted avg       0.78      0.80      0.78      2250
```

Conclusion

- Employee’s satisfaction level, if they received a promotion, salary, and time spent at the company impacted if they left the company the most
- All the independent variables played a role in determining if an employee would stay or leave
- Most employee’s who left the company had only been there between 2 and 5 years, so they would likely not have had a promotion, and their salary and satisfaction levels would be low
- Model was skewed in distribution of 0’s and 1’s
- Better at predicting if an employee would stay
- Good accuracy of 80%
- Could improve with a better distribution of 1’s
- Decent dataset to work with but might work better under a different model
- Overall, a decent accuracy score with good MSE and RMSE values