HR Analytics Logistic Regression Report

Github link: https://github.com/hilman1998/HR-Analytics

Executive Summary

Over the years, employee attrition has been a massive problem for companies the world over. One paper (Chen 2023) notes that the overall employee turnover rate in 2021 was as high as 53.7%, with many industries experiencing rates near 19%, significantly above the 10% basic standard.

The goal of this study is to model the probability of attrition for employees in a company and also find the features that are most important for influencing employee attrition. This can then help HR managers the world over to change up their policies in order to keep employees loyal and happy. The data was obtained from Kaggle.

The metrics used were “attrition” to see how many employees left the company. Looking at several features describing the characteristics of each employee, such as their education level and number of years since last promotion, the results found that some factors were more important than others for influencing employee attrition.

One risk with this study is the rather low amount of data points as there were only 4410 rows of employee data.

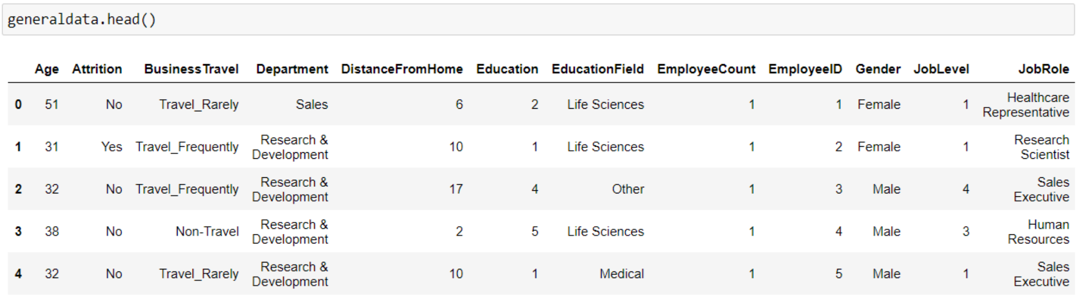
Data cleaning

Some columns are not needed in this logistic regression model such as EmployeeID and Over18 (which all employees in the dataset are). These were removed to help increase the accuracy.

Some columns had missing numbers. The mean of the numbers were used for numerical columns and the mode was used for categorical columns (if any).

Summary of statistical analysis

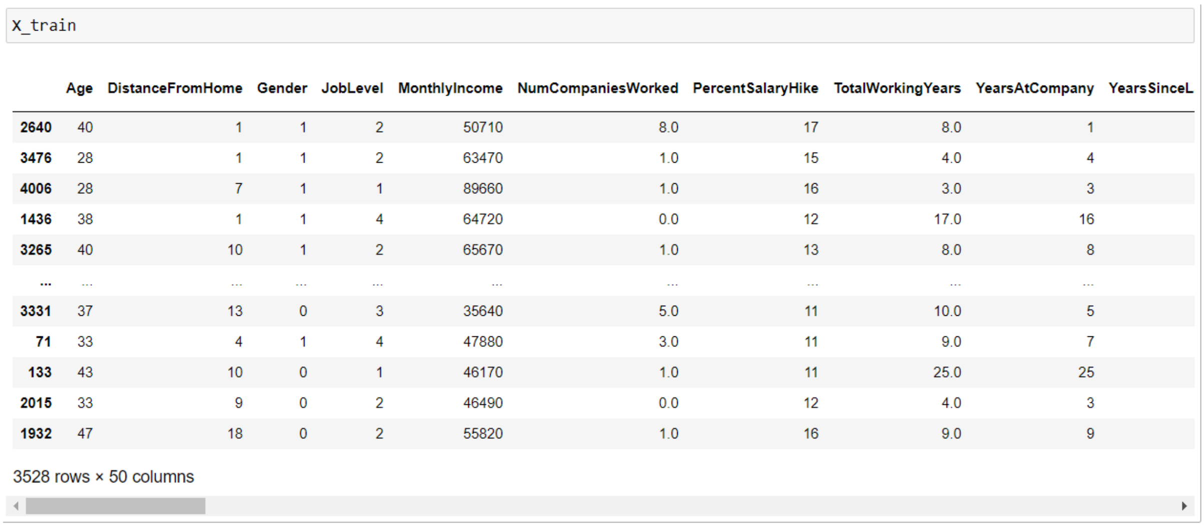
The model used was logistics regression after data cleaning was done. The reason for using this model was that the model is easy to interpret and is well-suited to model the binary outcomes of the study. One hot encoding was used on the categorical features to find their precise effects on the outcome. The features that were one hot encoded were 'BusinessTravel', 'Education', 'EducationField', 'JobRole', 'MaritalStatus', 'StockOptionLevel', 'TrainingTimesLastYear' and 'Department'.



```
generaldata.head()
```

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	JobLevel	JobRole
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	1	Healthcare Representative
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	1	Research Scientist
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	4	Sales Executive
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	3	Human Resources
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	1	Sales Executive

Figure 1: The original dataframe before any preprocessing was done.



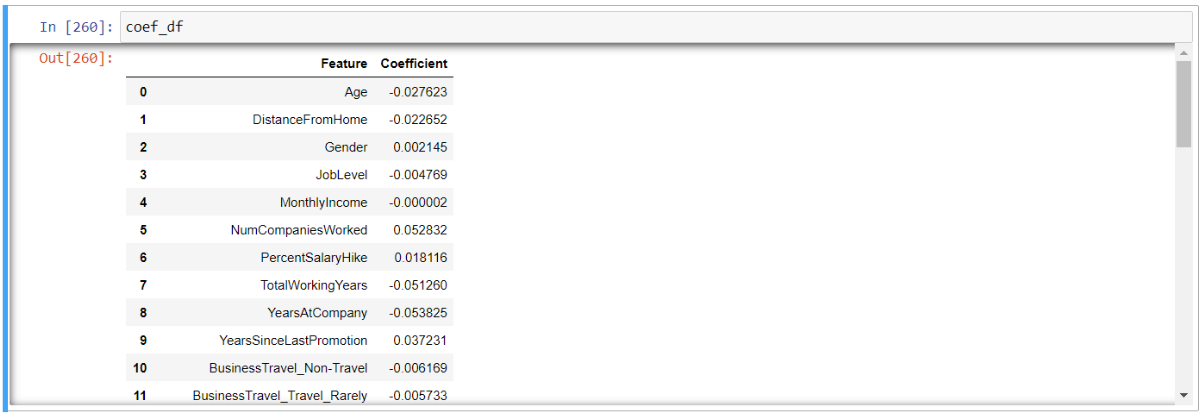
```
X_train
```

	Age	DistanceFromHome	Gender	JobLevel	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	YearsAtCompany	YearsSinceL
2640	40	1	1	2	50710	8.0	17	8.0	1	
3476	28	1	1	2	63470	1.0	15	4.0	4	
4006	28	7	1	1	89660	1.0	16	3.0	3	
1436	38	1	1	4	64720	0.0	12	17.0	16	
3265	40	10	1	2	65670	1.0	13	8.0	8	
...	...	...	...	...	...	...	...	...	...	
3331	37	13	0	3	35640	5.0	11	10.0	5	
71	33	4	1	4	47880	3.0	11	9.0	7	
133	43	10	0	1	46170	1.0	11	25.0	25	
2015	33	9	0	2	46490	0.0	12	4.0	3	
1932	47	18	0	2	55820	1.0	16	9.0	9	

3528 rows × 50 columns

Figure 2: The X_train data used to train the model. This dataframe is after doing all the data cleaning, performing one hot encoding on the aforementioned features and splitting the data into training sets and testing sets.

Looking at the coefficients of the model in figure 3 below, it was found that number of companies worked (NumOfCompaniesWorked) and years at the company (YearsAtCompany) are the two strongest factors that determine how likely someone will leave a company. The former had a coefficient of 0.05 (indicating that a 1 unit increase in this will make the employee 1.05 times more likely to leave) while the latter had a coefficient of -0.05 (indicating that a 1 unit increase in this will make the employee 1.05 times less likely to leave).

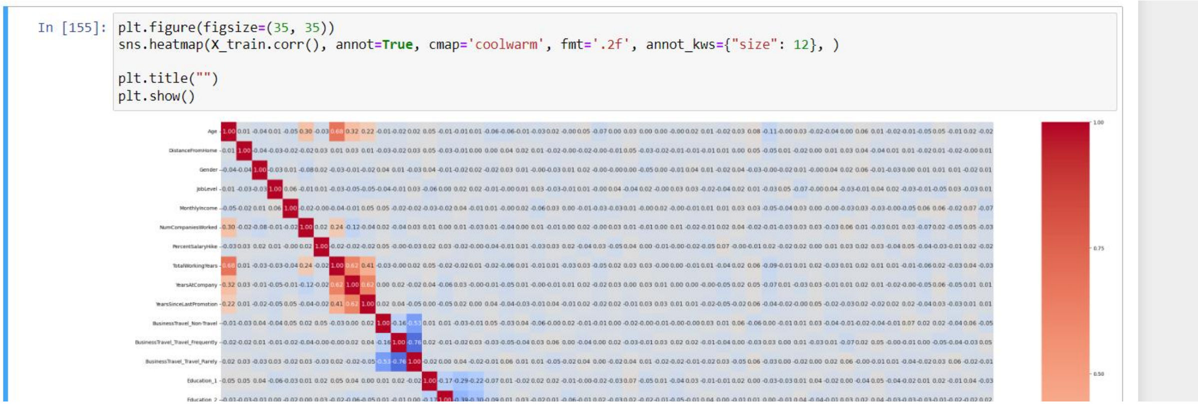


```
In [260]: coef_df
```

Out[260]:

	Feature	Coefficient
0	Age	-0.027623
1	DistanceFromHome	-0.022652
2	Gender	0.002145
3	JobLevel	-0.004769
4	MonthlyIncome	-0.000002
5	NumCompaniesWorked	0.052832
6	PercentSalaryHike	0.018116
7	TotalWorkingYears	-0.051260
8	YearsAtCompany	-0.053825
9	YearsSinceLastPromotion	0.037231
10	BusinessTravel_Non-Travel	-0.006169
11	BusinessTravel_Travel_Rarely	-0.005733

Figure 3: The coefficients of the model after being fitted into the logistic regression model.



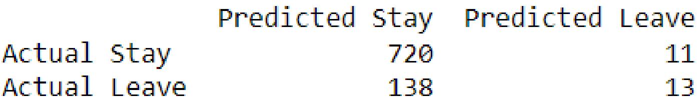
```
In [155]: plt.figure(figsize=(35, 35))
          sns.heatmap(X_train.corr(), annot=True, cmap='coolwarm', fmt='.2f', annot_kws={"size": 12}, )

          plt.title("")
          plt.show()
```

Figure 4: The heatmap showing the correlations of the X_train data. Some data features were correlated with each other so they had to be taken out. These were BusinessTravel_Travel_Frequently and Department_Research & Development

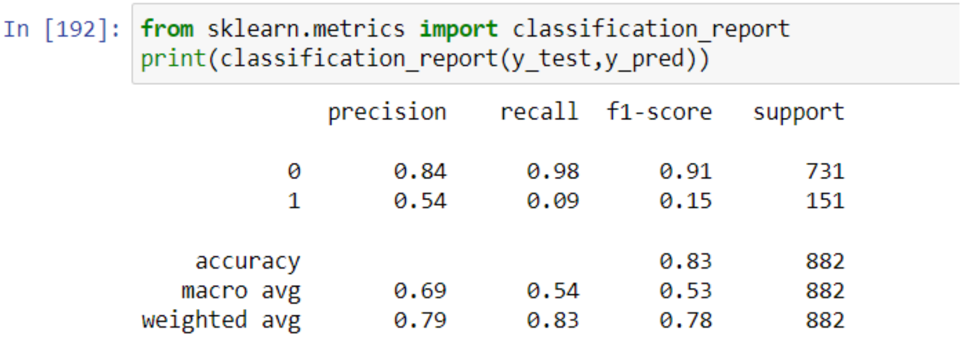
The evaluation phase showed that the accuracy score of 84.9% was obtained for the testing set and 83.1% was obtained for the training set. This may show slight overfitting.

The following are the confusion matrix and the classification report.



```
              Predicted Stay  Predicted Leave
Actual Stay              720               11
Actual Leave             138               13
```

Figure 5: The confusion matrix



```
In [192]: from sklearn.metrics import classification_report
          print(classification_report(y_test,y_pred))

                        precision    recall  f1-score   support

                   0       0.84      0.98      0.91       731
                   1       0.54      0.09      0.15       151

            accuracy                           0.83       882
           macro avg       0.69      0.54      0.53       882
        weighted avg       0.79      0.83      0.78       882
```

Figure 6: The classification report.

Discussion of classification report

Class 0 (Employees Who Stay):

Precision: 0.84 - When the model predicts an employee will stay, it is correct 84% of the time.

Recall: 0.98 - The model correctly identifies 98% of the employees who actually stay.

F1-Score: 0.91 - A high F1-score indicates a good balance between precision and recall for this class.

Class 1 (Employees Who Leave):

Precision: 0.54 - When the model predicts an employee will leave, it is correct 54% of the time.

Recall: 0.09 - The model correctly identifies only 9% of the employees who actually leave.

F1-Score: 0.15 - A low F1-score indicates that the model is not performing well in predicting this class.

Overall Model Performance:

Accuracy: 0.83 - Overall, the model correctly predicts the status (stay or leave) of 83% of the employees.

Macro Average: Averages for precision, recall, and F1-score are 0.69, 0.54, and 0.53 respectively, indicating moderate performance.

Weighted Average: Averages for precision, recall, and F1-score are 0.79, 0.83, and 0.78 respectively, weighted for class imbalance.

References

1. Chen, B., 2023. Factors of Employee Attrition: A Logistic Regression Approach. [online] Available at: https://www.researchgate.net/publication/373896134_Factors_of_Employee_Attrition_A_Logistic_Regression_Approach