DSC 680

Submitted to: Frank Neugebauer

HR Analytics

Using Machine Learning techniques to understand why employees leave

SUBMITTED BY| MARIA COUTO

Business Problem

This Project focuses on the use of Machine Learning techniques to determine employee turnover. In every business, Human Resources (HR) manages the people within the workforce. Data science and analytics allows us to use behaviors and patterns from people are keys to maximizing HR performance. We see the application of Machine Learning algorithms in HR functions. Perhaps, the most critical function of talent management is keeping and developing talent to support organizational goals.

Datasets

IBM HR analytics Employee Attrition & Performance dataset With 35 columns, 1500 rows, and the attrition column (True/False) as the target variable. This dataset contains different features that could potentially influence employee turnover such as travel, pay rate, satisfaction level, etc

Data Explanation and Prep

With 35 features in the dataset, I would like to narrow down the variables that are most likely to affect attrition. I used exploratory data analysis to help understand how the information relates to each other and which ones are the most important. For the integer type variables, I used acorrelation matrix, along with a heatmap to better identify features that have a high correlation. For the categorical/object datatypes, I used chi-square test to determine the features'relationship with each other.

Methods

This project uses Logistic Regression and Random Forest Classification to predict features that affect employee turnover. The target variable, Attrition, was treated with Label encoder to fit both models. For the categorical data, the pandas get_dummies function was used to transform the variables into numerical values that can be applied in our algorithm. The dataset was split into an 80/20 train-test split and was treated with a scaler to normalize the values so that metrics that have a high value do not affect the performance of the models. Finally, even though only 2 models were used, the Random Forest classification was repurposed with hyperparameter tuning to see if finding the right parameters would improve the performance of the model.

Analysis - Logistic Regression

ACCURACY SCORE:
0.8997
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.91	0.98	0.94	986
1	0.82	0.48	0.61	190
accuracy			0.90	1176
macro avg	0.86	0.73	0.78	1176
weighted avg	0.89	0.90	0.89	1176

Using the Logistic Regression model, we see that both the training and test data had a high accuracy score (90% and 86%, respectively) which means that it is able to distinguish patterns within the dataset which is able to identify what belongs to the attrition and no attrition class. However, the recall for 1 (attrition_yes) is at 48%. We see this in the correlation matrix where the models predicted a high negative, attrition_no, which mirrors the characteristics of our dataset. However, we see a lot of false negatives where the model predicted attrition_no and the actual value is an attrition_yes. This means that while our model performs well at predicting attr_yes, it struggles to identify attr_no, which is the key value we want to identify in this project

Methodology - Random Forest Classification

ACCURACY SCORE:
0.8639
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.91	0.98	0.94	986
1	0.82	0.48	0.61	190
accuracy			0.90	1176
macro avg	0.86	0.73	0.78	1176
weighted avg	0.89	0.90	0.89	1176

Random Forest Classifier

ACCURACY SCORE:
0.8469
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	986
1	1.00	0.05	0.10	190
accuracy			0.85	1176
macro avg	0.92	0.53	0.51	1176
weighted avg	0.87	0.85	0.78	1176

Random Forest Classifier

GridSearchCV

ACCURACY SCORE:
0.8520
CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0	0.85	1.00	0.92	986
1	1.00	0.08	0.16	190
accuracy			0.85	1176
macro avg	0.93	0.54	0.54	1176
weighted avg	0.87	0.85	0.80	1176

Random Forest Classifier

RandomizedSearchCV

Conclusion

Feature Importance

Overtime - YES

Coefficient - .53

Years Since Last Promotion

Coeffiicient - .52

Job Role - Laboratory Technician

Coefficient - .48

Ethical Consideration

In my experience working with HR data, people analytics contain sensitive information protected under privacy laws. We need to be careful in handling any personal information that could identify employees. Some HR institutions perform de-identification so the data cannot be linked to a specific person. Doing so will also protect employees from discrimination especially if they show that they are highly likely to leave the company. Finally, access to employee data must be limited so only qualified people can gain information about employees.