Evaluation of Prediction Models

In Understanding Employee Attrition Rate

By:

John Arvee Flores

Business Understanding

Every year, there are a handful of employees moving from one company to another. Various reasons arise on each and every person who are moving to another company. As employees leave every now and then, companies would need to hire again for new employees for the positions being left by the old employees which goes to show that having high attrition rate causes issues in the company which this study would be able to help by understanding the dataset available with regards to the employees' profile. The dataset presented on this study provides background information on the status of the employees last year which includes the profile of the employees who already left the company.

Research Question

Due to high employee attrition rate of the company, the researchers decided to use the dataset comprising around 4000 observations to understand the behavior of the employees for this study.

The researcher wants to understand and answer the question:

What are the key determining factors of employees leaving the company?

This study would be beneficial for future companies who are on the same problem of having a high employee attrition rate as they would be able to have reference on the factors which could contribute on having high employee attrition rate.

This study is limited only on identifying the key determining factors of having high employee attrition rate for this company. This will not directly give information on how these key determining factors affect employees but would give understanding on which among the attributes are contributing to the employee attrition rate.

Objectives

The goal of this study is to build prediction models by performing the learnings taken from the course “Predictive Analytics and Machine Learning”. Another objective of this study is to evaluate the performance of the models created to identify the best fit model for the dataset provided and eventually use by the company for future business decisions.

Data Understanding

Data Dictionary

The table below presents the variables to be used for this study.

Variable	Description	Categories
Age	Age of the employee	
Attrition	Whether the employee left in the previous year or not	
BusinessTravel	How frequently the employees travelled for business purposes in the last year	
Department	Department in company	
DistanceFromHome	Distance from home in kms	
Education	Education Level	1 'Below College'
		2 'College'
		3 'Bachelor'
		4 'Master'
		5 'Doctor'
EducationField	Field of education	
EmployeeCount	Employee count	
EmployeeNumber	Employee number/id	
EnvironmentSatisfaction	Work Environment Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
Gender	Gender of employee	
JobInvolvement	Job Involvement Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
JobLevel	Job level at company on a scale of 1 to 5	
JobRole	Name of job role in company	
JobSatisfaction	Job Satisfaction Level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
MaritalStatus	Marital status of the employee	
MonthlyIncome	Monthly income in rupees per month	
NumCompaniesWorked	Total number of companies the employee has worked for	

Over18	Whether the employee is above 18 years of age or not	
PercentSalaryHike	Percent salary hike for last year	
PerformanceRating	Performance rating for last year	1 'Low'
		2 'Good'
		3 'Excellent'
		4 'Outstanding'
RelationshipSatisfaction	Relationship satisfaction level	1 'Low'
		2 'Medium'
		3 'High'
		4 'Very High'
StandardHours	Standard hours of work for the employee	
StockOptionLevel	Stock option level of the employee	
TotalWorkingYears	Total number of years the employee has worked so far	
TrainingTimesLastYear	Number of times training was conducted for this employee last year	
WorkLifeBalance	Work life balance level	1 'Bad'
		2 'Good'
		3 'Better'
		4 'Best'
YearsAtCompany	Total number of years spent at the company by the employee	
YearsSinceLastPromotion	Number of years since last promotion	
YearsWithCurrManager	Number of years under current manager	

Timeframe of Data Gathering

The dataset presented is a snapshot of the status of the employees' profile last year. This mean that employee attrition rate presented on this study was the employee attrition rate based on last year's data.

Data Preparation

Data Source and Processing

The data was pulled from Kaggle which is a popular website for data scientists which provides a lot of dataset available for analysis. A dataset of employees with its attrition status was picked by the researcher and has become the source of this study. The dataset is downloaded and stored in a CSV file having over 4000 observations and around 20 fields included.

The tool used for preprocessing the data is R in which the data issues were treated to come up with a clean data before the analysis.

Data Issues and Remedies

Among all the data issues encountered, one of the most common data issue was the incompleteness of dataset. There are columns in which the data are missing. In some columns, there are NA values which can be considered as 0 like in 'TotalWorkingYears' as well as in 'YearsWithCurrManager'.

Columns like 'Over18' and 'EmployeeCount' were removed as well as these does not have any other values as all the records have the same values for these columns.

For the target variable, 'Attrition' was originally marked with 'Yes' or 'No' but changed this into 1 and 0 for easier manipulation.

New Features

Categorical variables were transformed into numeric values by associating each category into a numeric value for better analysis with these fields.

Data source:

https://www.kaggle.com/anupammajhi/hr-analytics-predictive-analysis

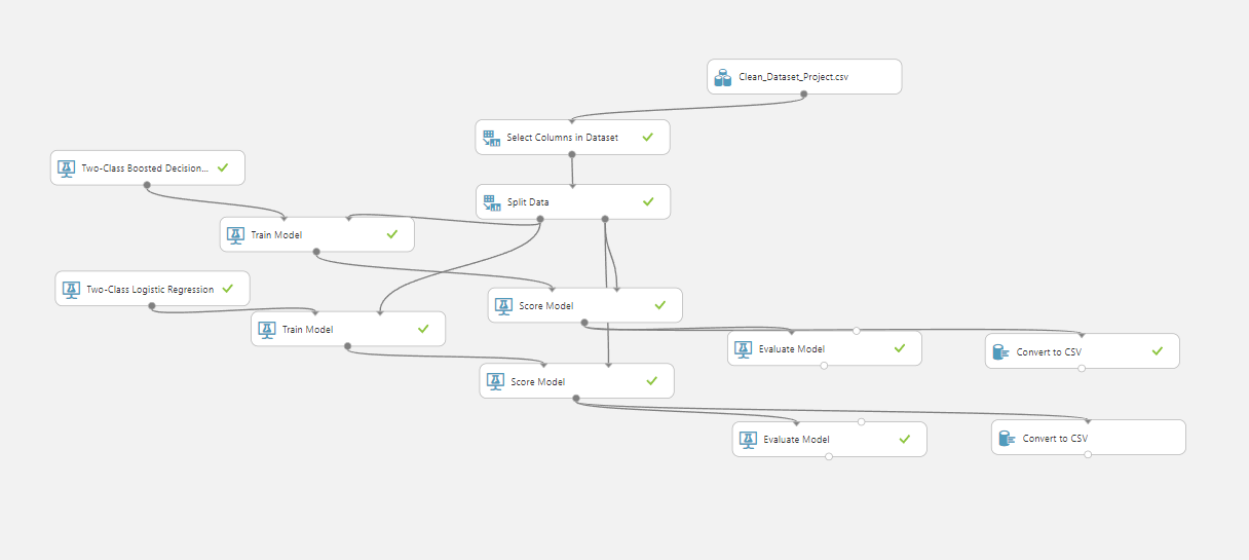
Modelling and Evaluation

The tools used for this study are as follows:

- R - Data Cleaning & Data Manipulation
- Azure Machine Learning – Modelling and Evaluation

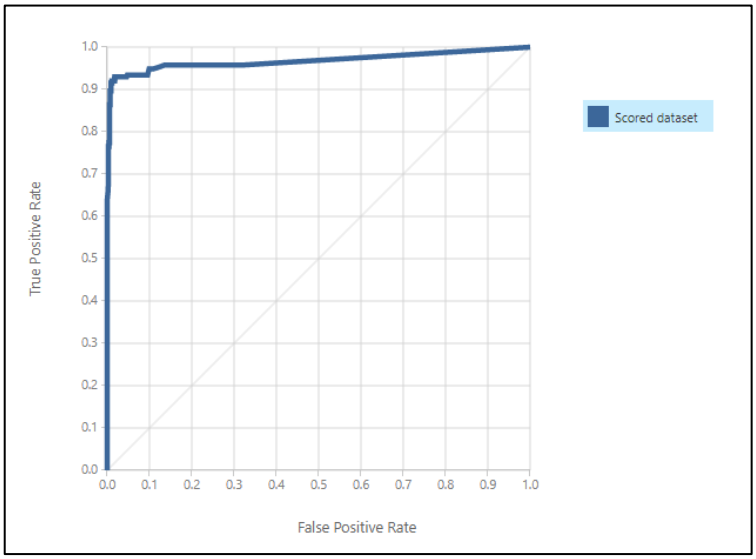
The classification algorithms used in this study are Boosted Decision Tree and Logistic Regression.

Modelling Workflow:



Discussion

Boosted Decision Tree

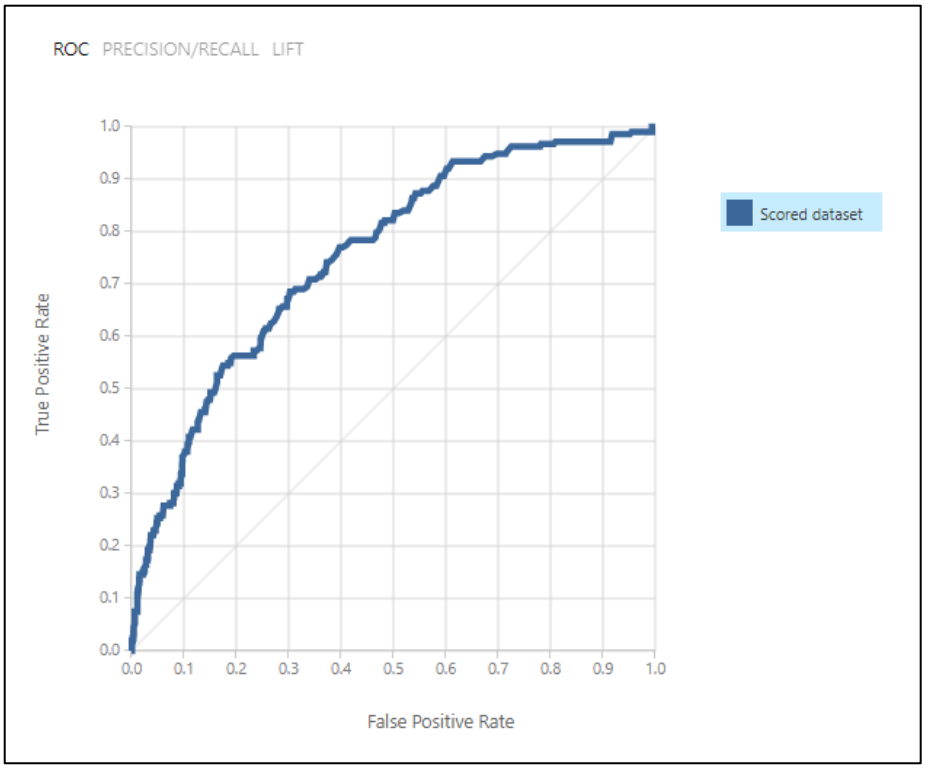


One of the evaluation measures we look into our model results is the AUC in which by using the Boosted Decision Tree it garnered 96.7% which explains that our model almost completely predicted true positives.

True Positive	False Negative	Accuracy	Precision	Threshold
196	17	0.978	0.942	0.7
False Positive	**True Negative**	**Recall**	**F1 Score**	
12	1098	0.920	0.931	
Positive Label	**Negative Label**			
1	0			

Confusion Matrix result using boosted decision tree model. All the evaluation metrics yielded above 90% results. Seem to be an overfitted model but could still be remedied through different overfitting treatments available.

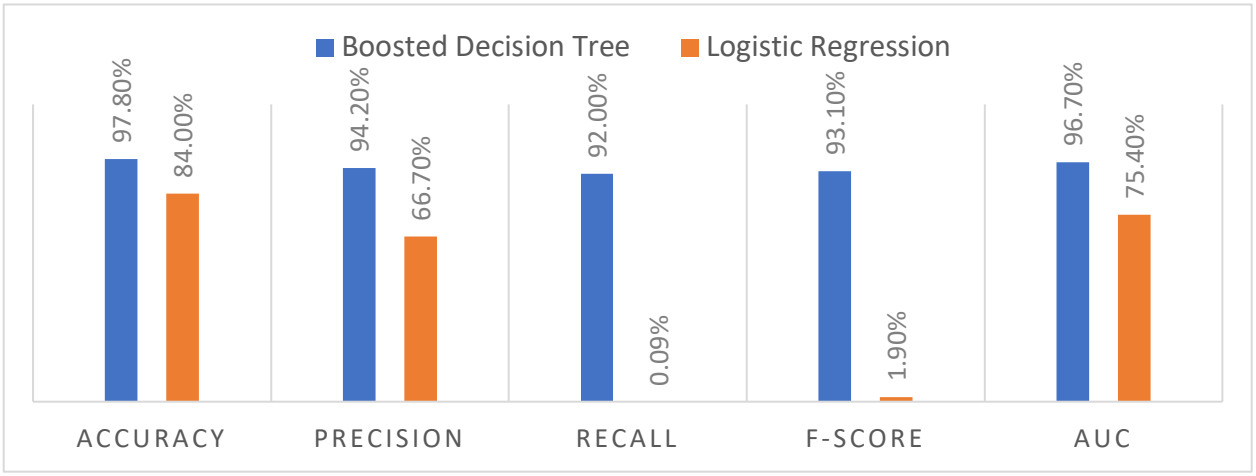
Logistic Regression



The result for this model had a lot less AUC result compared with Boosted Decision Tree. With the value of 75.4% this shows to be a nice fitting model to be used for employee attrition rate prediction.

True Positive	False Negative	Accuracy	Precision
2	211	0.840	0.667
False Positive	**True Negative**	**Recall**	**F1 Score**
1	1109	0.009	0.019
Positive Label	**Negative Label**		
1	0		

Accuracy has a decent results of 84% using logistic regression. Precision got 66.7% which would be a considerable metric result. The recall and f1 score obtained from this model are less than 5%. These results were affected mainly because of the dataset provided as there is a small amount of actual true values in the dataset.



Model	Accuracy	Precision	Recall	F-Score	AUC
Boosted Decision Tree	97.80%	94.20%	92.00%	93.10%	96.70%
Logistic Regression	84.00%	66.70%	0.09%	1.90%	75.40%

Comparing the two models under the 70% threshold, boosted decision tree result to have the higher metric results in all evaluation. Although this almost predicted the target variable completely, overfitting issues arise due to this concern. This can be remedied through different techniques and is subject for further analysis. Looking at their AUC, both models had decent results which can be considered to be used for this prediction. In my opinion, having some tweaks updated with the boosted decision tree model would be the best model in predicting the employee attrition

Correlation:

	Age	Attrition	DistanceFromHome	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion
Age	1.000000000	-0.159205007	0.006963332	-0.0443139217	0.29929657	-0.033136611	-0.031752826	0.678436347	-0.027307864	0.3113087697	0.216513368
Attrition	-0.159205007	1.000000000	-0.009730141	-0.0311762817	0.04150333	0.032532595	-0.006838852	-0.170237940	-0.049430576	-0.1343922140	-0.033018775
DistanceFromHome	0.006963332	-0.009730141	1.000000000	-0.0216070230	-0.01261699	0.038124615	0.011168676	0.008925272	-0.009001456	0.0316840455	0.002289598
MonthlyIncome	-0.044313922	-0.031176282	-0.021607023	1.0000000000	-0.01991534	0.004324699	0.026929826	-0.033693886	0.050112340	0.0009949458	0.065219286
NumCompaniesWorked	0.299296571	0.041503330	-0.012616990	-0.0199153444	1.00000000	0.031682605	0.017685456	0.237472333	-0.032123167	-0.1163223293	-0.035420909
PercentSalaryHike	-0.033136611	0.032532595	0.038124615	0.0043246991	0.03168260	1.000000000	0.012548325	-0.019495342	-0.037392066	-0.0297069085	-0.029542382
StockOptionLevel	-0.031752826	-0.006838852	0.011168676	0.0269298256	0.01768546	0.012548325	1.000000000	0.003128139	-0.069901761	0.0078858806	0.019062713
TotalWorkingYears	0.678436347	-0.170237940	0.008925272	-0.0336938863	0.23747233	-0.019495342	0.003128139	1.000000000	-0.041842121	0.6244676130	0.403405452
TrainingTimesLastYear	-0.027307864	-0.049430576	-0.009001456	0.0501123403	-0.03212317	-0.037392066	-0.069901761	-0.041842121	1.000000000	-0.0078936282	0.016120987
YearsAtCompany	0.311308770	-0.134392214	0.031684046	0.0009949458	-0.11632233	-0.029706909	0.007885881	0.624467613	-0.007893628	1.0000000000	0.618408865
YearsSinceLastPromotion	0.216513368	-0.033018775	0.002289598	0.0652192855	-0.03542091	-0.029542382	0.019062713	0.403405452	0.016120987	0.6184088652	1.000000000

Conclusion and Recommendation

Conclusion

In conclusion, boosted decision tree is the best fit model among the two models created. This is considered to be the best fit as all the model evaluation metrics resulted to above 90%.

Looking at the employee attrition, the evaluation metric which should be prioritized would be the precision as the business would like to make sure that all actual positives should be taken care of

Recommendation

For the model to become not overly fitted to the dataset, remedies are required for building better model.

Among the presented predictor variables, we can say that age and total working years are the closest variables correlated with the employees attrition.

As a recommendation to the business, it would be best to drill down more data by conducting surveys to understand better the factors in which majority are considering upon leaving their company.