Employee Attrition Prediction with Machine Learning

Introduction

A classification problem is typically considered supervised learning in machine learning, with the goal of determining whether a new sample belongs to a known sample class based on certain characteristics of known samples. The project can be considered as binary classification problem of employee attrition of the IBM with various predictive features. ML is helping to narrow the key features and predict attrition among a lot of data.

Methodology

A. Data Preprocessing

A data quality report is generated and correlation matrix is used to have a overview of the data. No missing value is detected. Using interquartile range, 12 outliers are detected and deleted. Adopting description function, 3 categorical features Over18, EmployeeCount and StandardHours are with the same value. In addition, and EmployeeNumber is irrelevant to attribution. Therefore, the above features are deleted. Visualizing data, monthly income,monthly rate and daily rate are detected as larger range. To standardize the measurements, these features' range are scaled between 0 and 1. Next step is to use a supervised machine learning model to judge the importance of each feature based on model-based feature selection. Adopting a random forest classifier with 100 trees to compute the feature importance and using median as a threshold, half of 30 features are selected. Then, the numerical and categorical features are selected and the latter is encoded. Before training, the data set is split for training and testing. (for feature importance see figure1)

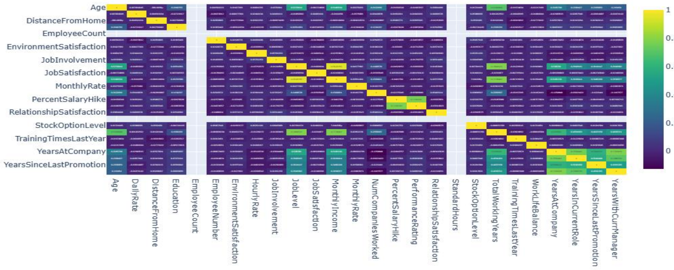

Figure1. correlation matrix for feature importance

B. Classification Algorithm

1. The Logistic Regression: The objective or dependent variable in binary logistic regression can only be one of two possible forms. It enables us to mimic the relationship between a number of predictor factors and a binomial target variables, such as the "Yes" and "No" values in the "Attrition" feature.
2. The Decision Tree: It is a tree-structured classifier, with internal nodes denoting dataset features, branches representing decision rules, and each leaf node indicating the classification results. It begins with the root node and expands on additional branches to create a structure similar to a tree. (figure2)
3. The Random Forest: Random Forest classifier uses multiple decision trees on different subsets of the data. Rather than relying on a single decision tree, the random forest collects the results of each tree and predicts the final output based on the majority of predictions. The optimal estimator is 100 in the project.

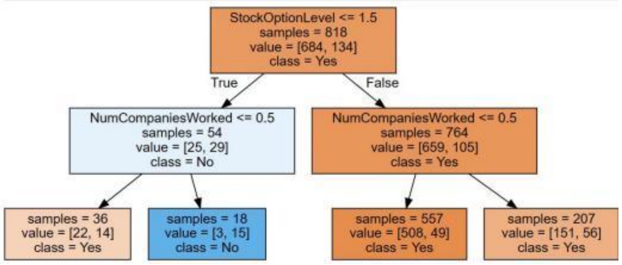

Figure2 decision tree structure

Results

The performances of the models are compared based on accuracy tested in test dataset Kaggle and train dataset. Considering interpretability and accuracy, the decision model is the best model (with max depth is 2). Besides, the decision tree is not overfitted since the test score is lower than the train score. (see analysis in the table below)

Models	Dataset	Accuracy	Pros	Cons
Logistic Regression	Test	0.8333	Easy to explain Faster pace	Easier to be over-fitting due to the multicollinearity among features
	Train	1		
Decision Tree (best model used)	Test	0.85294 (kaggle)	Easy to explain and interpret	The depth of the tree can be easily look complex and therefore over-fitting
	Train	0.86		
Random Forest	Test	0.85	Deal with imbalance data with less errors	Harder to interpret than Decision tree and define estimator
	Train	O.87		

Discussion

There are some challenges for the machine learning process. After encoding there are high dimensions in the data set. The limited data size also restricts the model training process. Therefore, data preprocessing and feaure selections are essential before data training. Besides, improper parameters affect the model's efficiency. The decision tree performs worse than the random tree with a max depth equal to 4. Hence, grid research is conducted to optimize the model.

Conclusion

1. Data understanding and preprocessing are crucial to modeling. 2. In the process of model choosing, a company should take interpretability into great consideration instead of depending merely on the accuracy, since the reasons for leaving are highly personal and various.