Kristie Nguyen
DATA 240

HR Employee Attrition Report

Introduction

In a competitive business environment, attrition rate, or employee turnover, is a huge problem that affects businesses because it causes significant costs to the businesses, which includes costs of business productivity, hiring and training new employees. Also, employee attrition impacts both the businesses and customers because businesses can lose customers if the customers are dissatisfied with the service from the new hires. This results in a loss of revenue for the businesses. Usually, the causes of employee attrition are career advancements and higher job salary, but many other factors affect the attrition rate. It is crucial for businesses to prevent profit loss, so there is a great business interest to reduce employee attrition rate. To do this, businesses have required the Human Resources Department to intervene and fix the situation if an employee is likely to leave. By using the data provided by HR, a model can be developed to predict if an employee is likely to leave based on a number of factors.

For this project, the dataset, IBM HR Analytics Employee Attrition and Performance, used is a fictional dataset created from IBM data scientists to indicate if there is an attrition or not. The dataset contains approximately 1500 entries. There are 35 columns, which consists of 34 features and 1 target variable. The 34 features include: Age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number of Companies Worked, Over 18, Over Time, Percentage of Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, and Years with Current Manager. The target variable is Attrition, which is labeled with either ‘yes’ or ‘no’. It is also important to note that there is an imbalance data with 84% of employees who stayed and 16% of employees who left. This class imbalance is taken into consideration when splitting the data into training and testing data by using *stratify*.

Methodology

a. Decision Tree

A decision tree is one of the predictive modelling approaches that uses a tree structured model where the data is continuously split according to a certain parameter. The decision tree consists of 3 components: nodes, edges/branches, and leaf nodes. The decision node has two or more branches that correspond to the outcome of the nodes and connects to the next leaf nodes. The leaf nodes represent a classification or decision. Lastly, the root node corresponds to the best predictor of the topmost decision node in the tree. A decision tree uses Information Gain (Entropy) or Gini Impurity to construct the model. In this case, Information Gain (Entropy) is used since it is the most common impurity measure used in classification problems. Also, there are two main types of decision trees, classification and regression. Since this is a classification problem, using a decision tree is a great baseline model to see how well the algorithm performs on the dataset. Based on the many features in the dataset, the decision tree model learns a series of questions with their corresponding class labels. Also, decision trees are able to work with categorical and numerical data, which contains both in the dataset. Although a decision tree already detects the most important variables by identifying the top few nodes, using

Random Forest's built-in feature selection is more accurately since decision trees are poorer predictors.

b. Logistic Regression

A logistic regression is a model that is used for predicting classes using the probability of the target variable. Compared to linear regression where it uses expected values of the response model, logistic regression uses the probability or odds of the response variable to model based on the combination of values taken be the predictors [1]. Also, unlike linear regression, logistic regression uses the sigmoid function that maps predicted values to probabilities. In addition, logistic regression work well on linearly separable class with its easy implementation, making it a popular choice for classification problems. There are two types of logistic regression model for classification, binary and multinomial. Binary logistic regression requires a dependent variable with only two possible outcomes whereas a multinomial requires 3 or more outcomes. In this case, this dataset is working with the binary logistic regression since the target variable is binary (Yes, No). Logistic regression is applicable to this problem since it wants to predict the probabilities and classify the employees into two categories (Yes, No) based on the explanatory variables. For the solver in the logistic regression model, liblinear is picked since it supports both L1 and L2 regularization.

c. Data Preprocessing

First action for data cleaning is to check for nulls. Fortunately, this data was very clean, so there was not a need to remove or replace a missing value. Nothing else was needed to be done in the data cleaning process, so I moved onto the next step, data preprocessing. Although decision trees can handle both categorical values, logistic regression cannot. To convert all the categorical values into numerical values, *LabelEncoder* is used. For example, the variables in 'BusinessTravel' ('Travel_Rarely', 'Travel_Frequently', 'Never') would convert to 1,2,3, respectively. After encoding the categorical values, the entire dataset must be standardized using *StandardScaler*. This is especially useful since the data has varying scales and it prevents the algorithms like linear regression, logistic regression, and linear discriminate analysis from making assumptions that the data has a Gaussian distribution.

Feature Selection

Feature selection is important in many machine learning projects with predictive modeling because it allows the algorithms to train faster and reduces the complexity of the models. For this project, feature selection is applied using Random Forest under the category of Embedded methods. Embedded methods combine the qualities of filter and wrapper methods [2]. These embedded methods are advantageous because they are highly accurate, have better generalizations, and are interpretable. With these advantages, I thought I would get the best accuracy using feature selection with Random Forest. To do this, a Random Forest ensemble must be created and the split training data are fitted into the ensemble. From there *feature_importances_* is implemented to extract the most important features of the data. Since there are 34 features, I wanted to extract the top 25 most impactful features as shown in Figure 1.

Out of the 9 columns that had to be dropped from the feature selection result, the two most unexpected features to be dropped were 'Performance Rating' and 'Relationship Satisfaction'. The bar graph is organized from the feature with the highest to lowest probability. As expected, monthly income is the most important feature because many employees leave businesses due to a higher salary offer at another business. One of the most surprising things I noticed that the feature, 'Years Since Last Promotion', had one of the lowest probabilities in the top 25. Since many people leave their current jobs for career advancement, it would be expected for that feature to have a higher probability.

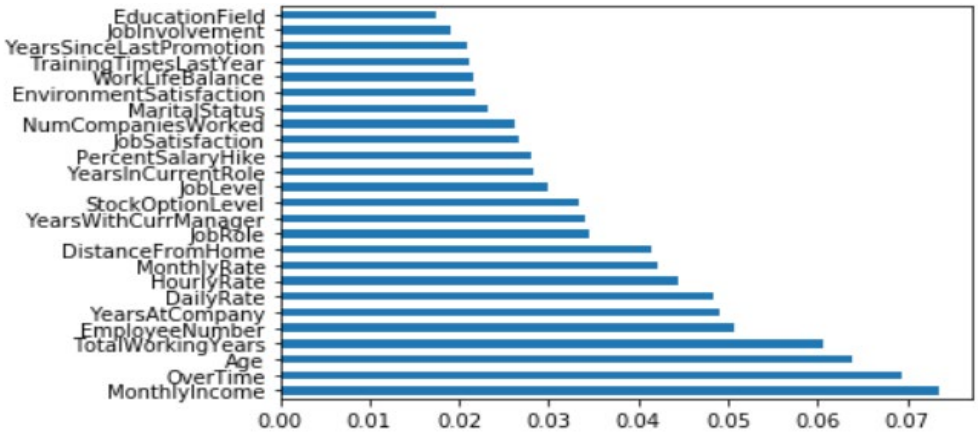

Figure 1. Bar Graph of Top 25 Features

Results & Discussion

The objective of this section is to evaluate the two algorithms by the accuracy on all and the selected features. For decision tree and logistic regression, both models are evaluated by the following metrics: accuracy, precision, recall, F1, and Receiver Operating Characteristic (ROC) curve. In Figure 2, it shows a bar chart comparison of the metrics for the two algorithms with selected features, and it is obvious logistic regression performed the best in all metrics. One of the more important metrics to look out for is the precision. It is the optimal to have a higher precision because it determines how precise the algorithm is out the predicted positive. In this case, a false positive means that an employee that is actually not leaving (actual negative) is detected as leaving (predicted attrition). This allows HR to reach out to those who are detected as false positive and check up on the employees to see how happy they are in the company. The false positive employees may not be actually leaving the company, but they can think about it. This is where HR can intervene and discuss potential problems that may arise in the future to prevent the attrition rate to go higher. Another important metric is recall, which is basically follows the same logic as precision. For instance, employees who are predicted to not leave are actually leaving the business, which is costly for businesses. For the decision tree, the average percentage came out to be 79% while the percentage for logistic regression is 86%. This is considered an okay score, but it is important to build a better predictor model to prevent a false negative, otherwise it can result in inefficiencies in resource allocation.

Kristie Nguyen
DATA 240

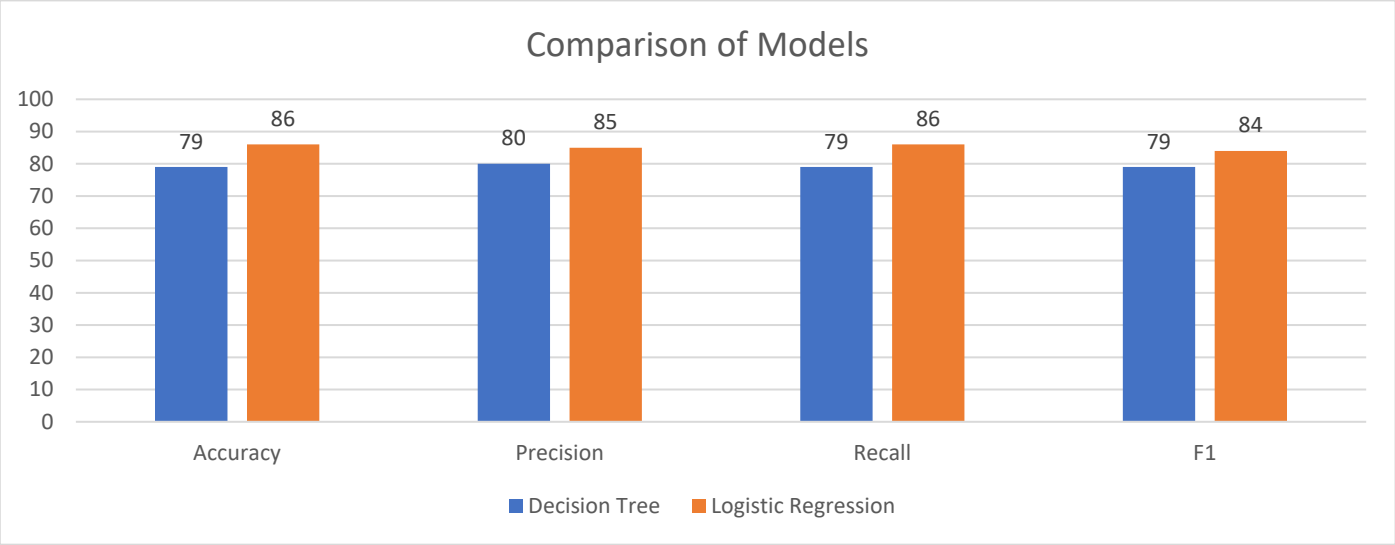

Figure 2. Comparison of Decision Tree and Logistic Regression Results

Usually, decision trees are supposed to be better predictor models than logistic regression, but the results show that logistic regression has outperformed the decision tree in every aspect. In Figure 3, it displays that the ROC curve for the decision tree is pretty close to the baseline compared to the ROC curve of the logistic regression, which means the decision tree is a bad model for this problem. I believe the reason why decision tree did not perform as well as expected because the continuous variables were not transformed into categorical values. Also, there were many categories, so the large number of categories might have caused cardinality problems and overfitted the model. What helped the logistic model perform better was the standardization of the data, allowing the model to optimize efficiently. In addition, logistic regression is robust to noise whereas decision trees are significantly affected by noise, which contributes to the performance of each model. Next time, working on an ensemble of extreme gradient boosted trees (XGBoost) might have a better chance at performing instead of the decision tree.

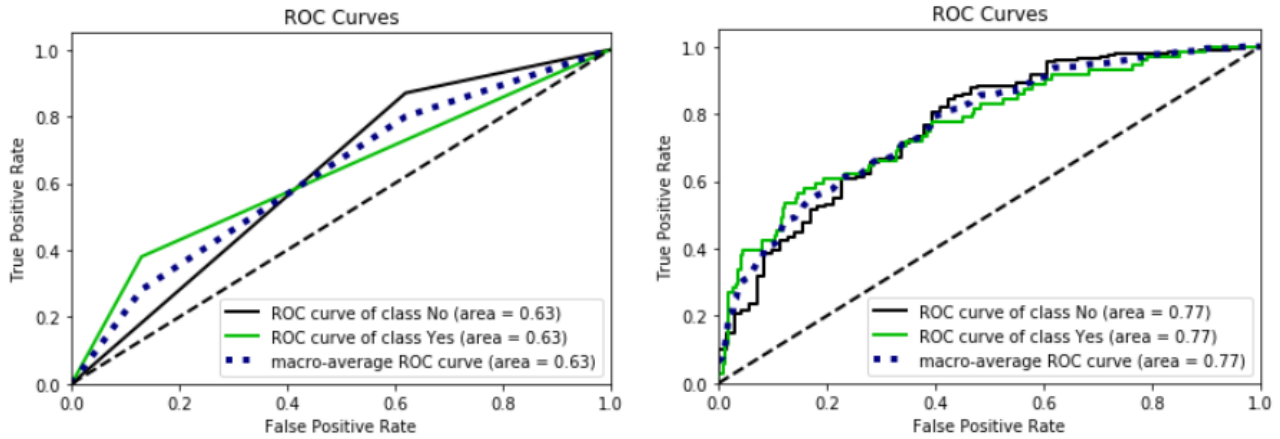

Figure 3. ROC Curves for Decision Tree (Left) and Logistic Regression (Right)

As expected, the models that used the selected features outperformed the models that used all the features in the provided dataset. By using the selected features that the Random Forest produced, both models achieved a greater accuracy with a small margin of difference because it removed unnecessary, redundant, and irrelevant features from the dataset. The dataset reduced from 34 to 25 features. The features removed were the following: Employee Count, Standard Hours, Business Travel, Department, Performance Rating, Education, Over18, Gender, and Relationship Satisfaction. Some of the features were redundant such as Employee Count and Employee Number, so it made sense that one of them were removed. Compared to benefits given by the businesses like Stock Option and Work Life Balance, features like Business Travel and Relationship Satisfaction were not as significant. Also, it seemed that the employee's education and department did not influence the attrition rate, which was a little interesting to see because the field sometimes determine how rigorous the work is.

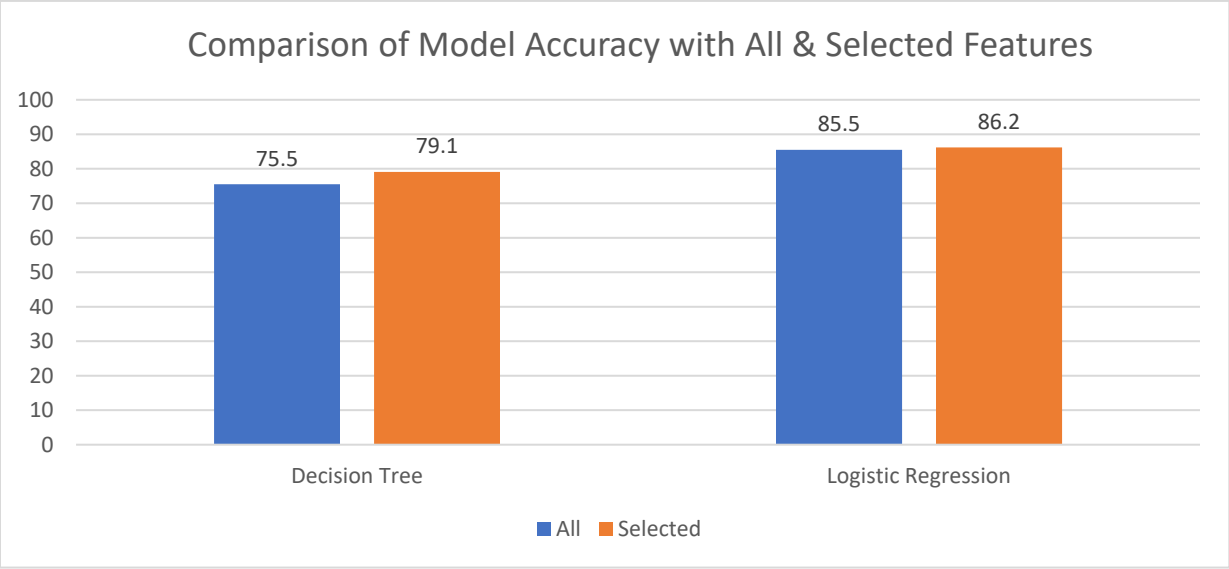

Figure 4. Comparison Chart of Model Accuracy with All & Selected Features

After seeing the results, it shows that HR should pay attention to the salary, age, overtime and total working years since they are important factors responsible for the attrition in the company. Since salary, age, and total working years have some correlation, HR should consult with the managers of the departments and discuss the possibility of a raise or even a job promotion to increase the salary as the employee gains more working experience in the business. If an employee has been working 3-5 years at a company and has been performing well, then a job promotion is definitely an option that should be discussed with the managers. Also, since the employees are salary-based, working overtime does not provide extra income. Working overtime affect mental and physical health, can cut into personal time, and can lead to job dissatisfaction. This is especially impactful to employees who are married and have children than employees who are single. To reduce overtime, HR can establish an overtime policy and allow flexible work schedules. As long as the employee does their work, attends meetings, and reach the deadlines, then a flexible work schedule can be openly discussed with HR and the managers.

Kristie Nguyen
DATA 240

Reference

[1] “Lesson 6: Logistic Regression.” Lesson 6: Logistic Regression | STAT 504, online.stat.psu.edu/stat504/node/149/.

[2] Dubey, Akash. “Feature Selection Using Random Forest.” Medium, Towards Data Science, 15 Dec. 2018, towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f.