Predicting Employee Attrition

Abstract

This project aims to predict attrition rates of employees using a Linear Discriminant Analysis (LDA). Attrition is a critical issue for companies, leading to a loss of talent, increased recruitment costs, and reduced productivity. Using a dataset of employee information, including demographic, job-related, and performance-related features, we trained a LDA to predict whether an employee is likely to leave the company.

Our approach involved data preprocessing, including data cleaning, handling missing values, and removing irrelevant features, feature engineering, model training and model testing. Specifically, we split the final dataset into training and testing sets, iteratively trained various models on the training data, and evaluated their performance on the testing data.

Upon trying various models, we decided on our final model, a linear discriminant analysis classifier, which is able to predict whether an employee will leave the company or not to quit or not with an AUC of 80%. The top five most important features affecting attrition are 'Marital Status', 'Sales Representative', 'Sales Executive' , '3 years or less at the company' and 'being under 31 years old'.

We can conclude that the 5 features above should be monitored to determine attrition rate of employees, the company can go a step further and reduce unwanted attrition through proper talent management systems.

1. Introduction

The dataset we are using was fictionally created by IBM data scientists. The goal of the dataset is to gain insight into the attrition rates and how to prevent unwanted attrition. It is estimated that attrition can cost a company 1.5 to 2 times the employee's annual salary. This is a significant financial burden that a company can prevent if they know the root cause of what can lead to employee attrition. When an employee leaves their workload needs to go elsewhere and this typically falls to the colleagues of the employees who have pre-existing roles. The current employees are then spread over multiple roles which results in less than optimal quality of work. In addition, these projects that no longer have leaders or project owners result in delayed releases and lost revenue.

Additionally, if the attrition rate of a company is high this can cause damage to an employer brand, which means that the company will have a revolving door of employees and will have a hard time attracting top tier team members. This revolving door of employees means that the company has a higher turnover rate and will result in unengaged employees, burnout, increased cost and poor company culture.

We will discuss what data we have, the approach we took in feature engineering, algorithms tried and challenges we faced. Then we will wrap up with the key takeaways and recommendations for future work.

2. Data Preparation

2.1. Dataset Description

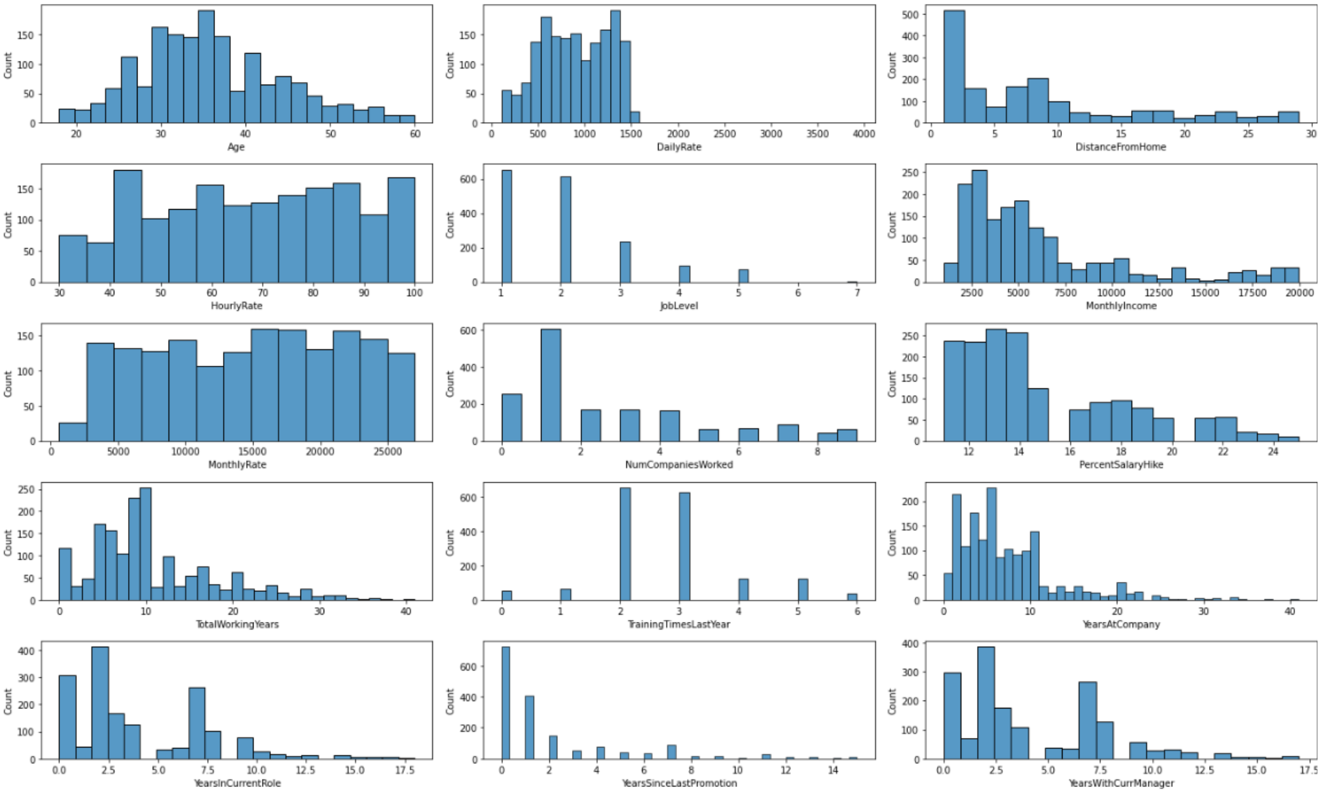
Our data set has a total of 35 variables including our target variable. There is a combination of categorical and numerical variables. Our data has 3 types of features: 7 categorical, 9 ordinal, and 17 numerical features.

2.2. Exploratory Data Analysis

Removing irrelevant features: We dropped irrelevant features such as "Employee Count", "Standard Hours" and "Over the age of 18", as these features contained the same value for each observation and thus did not provide any relevant information.

Numerical Features: For the numerical features, we observed that none of them are normally distributed, in fact the majority are skewed. This distribution is important when determining the type of scaling to use and needs to be taken into consideration for models that assume a gaussian distribution, as is the case with our final model. To address these issues we use the Power Transformer to make the data more gaussian-like and then we applied Standard Scaler to scale the numerical features.

Figure 1: Numerical Feature Plot



Ordinal & Categorical Features: For the ordinal and categorical features we observed that many of them are highly imbalanced which needs to be taken into consideration when analyzing model results. We did not scale either the ordinal or the categorical features but we one hot encoded the categorical ones.

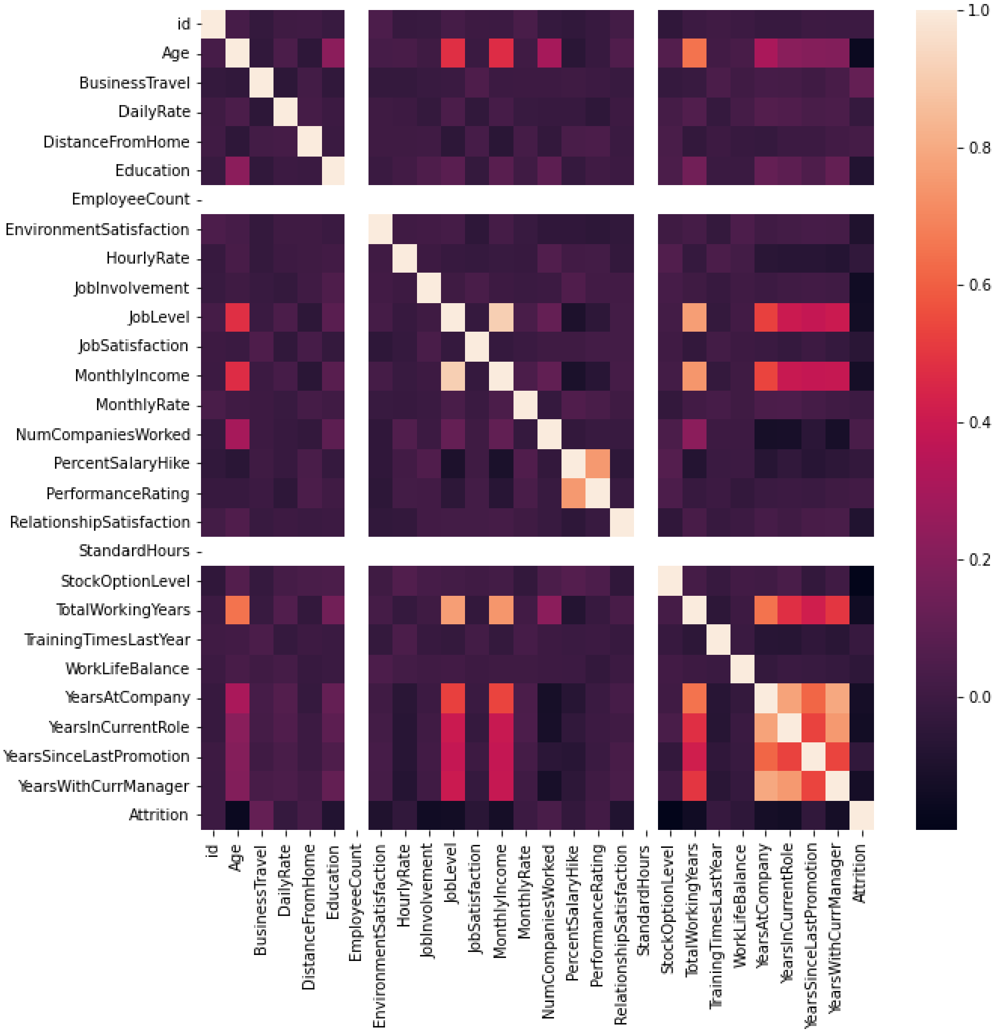
Feature engineering: We created 7 new features based on the profile of employees that quit. The goal is to capture the risk profile of those employees and add them as an additional column to all employees. The first feature added was a ‘Monthly Income/Age’ ratio, that would help us identify any relationships between the salary and age of each employee. ‘Average working time’ which outputs the average number of years worked at a company for each employee. ‘Attrition Rate’ which captures the median age of those employees that have already quit. ‘Attrition Distance’ calculates the distance between home and work. ‘Attrition Hourly Rate’ provides the hourly salary per employee that quit. ‘Attrition Years’ captures if an employee had been working for less than 3 years.

Lastly 'Unstable Employee' which marks as a risk (with 1) those employees that had worked at over 2 companies and for less than 2 year in average.

Outlier detection: We found that the outliers for each feature vary between 0% to as high as 23%, we took no action on the outliers as we determined they could be important.

Correlation Analysis: The correlation matrix showed a high correlation between several features such as 'Job Level', 'Total Working Years', 'Years at Company', 'Years in Current Role' and 'Performance Rating'. Correlated variables were not removed when running decision tree based models as these are robust to multicollinearity but were removed when running our final model, LDA, as it is not robust to multicollinearity.

Figure 2: Correlation Matrix



Balancing Data: Lastly, we noticed that the target variable was highly unbalanced, which can lead to a biased model that favors the majority class, and thus results in lower model performance. To address this issue we used SMOTE, a technique for oversampling the minority class, which in this case was oversampling the attrition cases. We tried various sampling methods and proportions and finally created a 1:1 balanced sample.

3. Methods and Algorithms:

We ran various classification models iteratively until reaching our final model. We used a grid search to find the optimal hyperparameters for each model, which allowed us to test different combinations of various hyperparameters and return the best ones. The models attempted are described below.

3.1. Decision Tree

We first tried a decision tree model, which is a supervised machine learning algorithm, but due to a low performance on the test set, the model was abandoned. In addition, the model was overfitting on our training set, meaning that the model was fitting the training data too closely and was not able to generalize to the test set.

3.2. Random Forest

We ran a random forest classifier which decreased overfitting. Random forest algorithms are less prone to overfitting as they create multiple decision trees and use a different subset of features for each decision tree, thus decorrelating the trees. However, the AUC score was lower compared to other models later pursued.

3.3. XGBoost

We ran an XGBoost model, which is similar to a random forest but instead of building each decision tree using a subset of features from the dataset, it builds decision trees based on the errors that it learns from previously built decision trees. XGBoost is a commonly used model for classification tasks but once again, this caused overfitting on the training set.

3.4. Linear Discriminant Analysis (LDA)

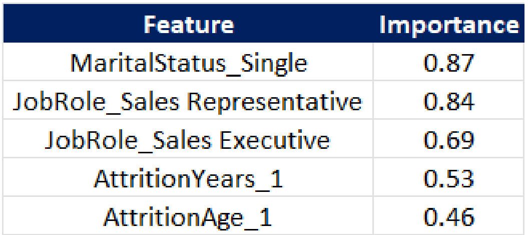
Lastly, we decided to try a statistical based model and for this we chose to use a Linear Discriminant Analysis (LDA) classifier. LDA is a supervised learning algorithm that can be used for classification tasks such as the one here. It is a statistical method that aims to find a linear combination of features that aims to maximize the distance between classes while minimizing the distance within classes. It is a method that reduces dimensionality of data while still capturing important information. This was the final model we decided to use as it provided the best model performance on the test set and did not show signs of overfitting, as exemplified by a similar score on the train and test set.

4. Final Results

We achieved our best model performance with the LDA model, such that it is able to predict whether an employee will leave the company or not with an 80% AUC score. When testing the model on our test set we get a 78% AUC score and 76% recall.

According to our LDA model, the most important features determining Attrition are Marital Status: Single, Job Role: Sales Representative, and Job Role: Sales Executive, 3 years or less at the company (Attrition_Years) and being under 31 years old (Attrition_Age).

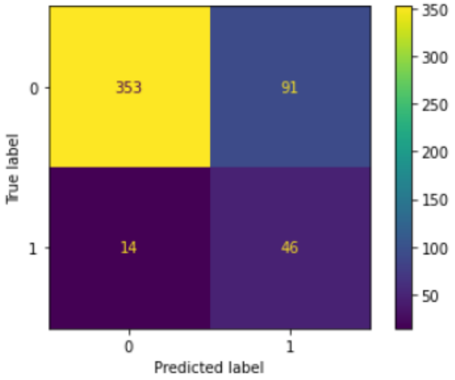
Figure 3: Table of Feature Importance according to LDA Model



Feature	Importance
MaritalStatus_Single	0.87
JobRole_Sales Representative	0.84
JobRole_Sales Executive	0.69
AttritionYears_1	0.53
AttritionAge_1	0.46

As seen on the Confusion Matrix below, our model minimizes False Negatives (FN) 14 observations and False Positives (FP) 91 observations, therefore reducing the costs related to unforeseen attrition cases.

Figure 4: Confusion Matrix on Testing Dataset



5. Advantages & Limitations

5.1. Advantages

LDA is an algorithm that can tolerate many features, thus the model not only can handle the features currently used but could be used if additional features were to be added in the future. It is also an algorithm that is able to generate predictions in a short amount of time and is not prone to overfitting. Lastly, LDA has good interpretability, as looking at the most important features provides useful insight into what factors contribute most to employee attrition.

5.2. Limitations

It is important to note that our solution has some limitations. Firstly, our chosen model assumes that the employees who are likely to leave the company and employees that are likely to stay are linearly separable, which may not be the case. If this is true, a different algorithm, such as quadratic discriminant analysis may be better suited to the data.

Additionally, we know that our model will not make accurate predictions in 22% of the cases, which is important to take into consideration when using the model to make business decisions. Model performance could be further improved by improving the quality of the data or the size of the dataset, thus by conducting further collection.

6. Conclusion and Next Steps

We have reached an acceptable AUC score which means that our model can be put into production and we can predict the attrition of future employees. In addition the model highlights which variables can make employees more likely to resign.

The most important features shown in the previous section can be used to take corrective action. The first strategy the human resource department can take for recent hires prior to their 3 year work anniversaries is to do weekly check ups with their managers to better understand their mental health and how they have been coping with their role. This can actively help prevent attrition as managers can adjust their working schedules and check in on their well being. It is also shown that sales roles tend to have a higher attrition which is correlated to travel within the role. The sales executives can dig further to understand how much travel is really needed for the role, and how the role can evolve to reduce travel.

We also found 2 interesting features that are inversely related to attrition, these are stock option levels and overtime. The company can do a cost benefit analysis of awarding more stock options vs cost of attrition to those roles that are likely to resign such as the sales team. This could result in savings since an employee will want to stay longer if they are more vested in the company. Overtime can also further be analyzed to determine which roles have the most over time and how this can be reduced operationally.

Currently we have roughly a 12% attrition rate which is slightly higher than industry standard. A good attrition rate to target is roughly 10%. In reality most human resources are strained with time, so an active strategy that the human resources team can do is to target the attrition with the highest salaries, since unwanted attrition of higher salaried employees will affect the business more. It will always hurt a company more when a C suite resigns as opposed to a manager.