Predictive HR Analytics for Employee Engagement

July 6, 2024

Jeff Nguyen, JT Herren, Sreya Kumar

Contents

1 Introduction, Background and Data Description

Our project centers around addressing questions concerning employee retention, performance, and satisfaction by utilizing an HR dataset. High attrition and dissatisfied employees negatively affect productivity, organizational knowledge, and the company's financial health. Therefore, we will use data-driven strategies to understand and predict employee behavior based on numerous factors.

Our dataset was retrieved from Kaggle, which can be found here. The dataset contains a variety of employee data. There is a mix of categorical and numerical features – the categorical features will have to be one hot encoded. Some redundant features were removed to clean the dataset like duplicate columns or irrelevant variables, such as employee number, before starting any analysis. Our data set initially consisted of 39 features including different variables related to forms of job satisfaction, employment details, different forms of income, and various demographic data such as gender and age.

2 Exploratory Data Analysis

Many of the variables in the initial data set were redundant or unimportant for our analysis. For example, the binary column, 'Attrition', with class labels, 1 and 0, was represented in another column 'CF_attrition label' with the class label "Yes" and "No". We noticed similar patterns throughout the listed features and removed 6 features.

Some of the categorical variables in our dataset were labeled with words instead of numbers, which is not very model friendly. For example, "Education" had the class labels "High school", "Bachelor's", "Master's" and "PhD". The following columns – 'Business Travel', 'CF_age band', 'Department', 'Education Field', 'Job Role', 'Marital Status', 'Education' – followed the same principle. We applied one hot encoding through the pandas get_dummies function, which converted the features to multiple one hot encoded columns. The final data set had 59 columns

3 Methodology

We used various statistical learning techniques to accomplish our objectives. Python will be the main language used to conduct the following analyses. Packages used include pandas, numpy, sklearn, and statsmodels. Our methodology encompassed both supervised and unsupervised methods, with our second objective being to compare both types.

The main predictor was Attrition, meaning whether an individual left the company. This was encoded with binary labels, hence the primary supervised model used was logistic regression. Different subsets of data were tested using information gained from forward/backward selection and Principal Component Analysis (PCA). Forward/backward selection are regression techniques that choose k number of relevant and significant features. PCA aids in dimensionality reduction.

The main metrics used to evaluate logistic regression performance will be accuracy and the ROC AUC score. Accuracy provides information on how often the model's predictions are correct compared to the actual label. ROC AUC is a measure of the true positive and false positive rates – values extracted from a model's confusion matrix. The higher the ROC AUC, the better the model's ability to classify. If the ROC AUC is closer to 0.5, the model is performing at a chance level, and not classifying based on the specified parameters.

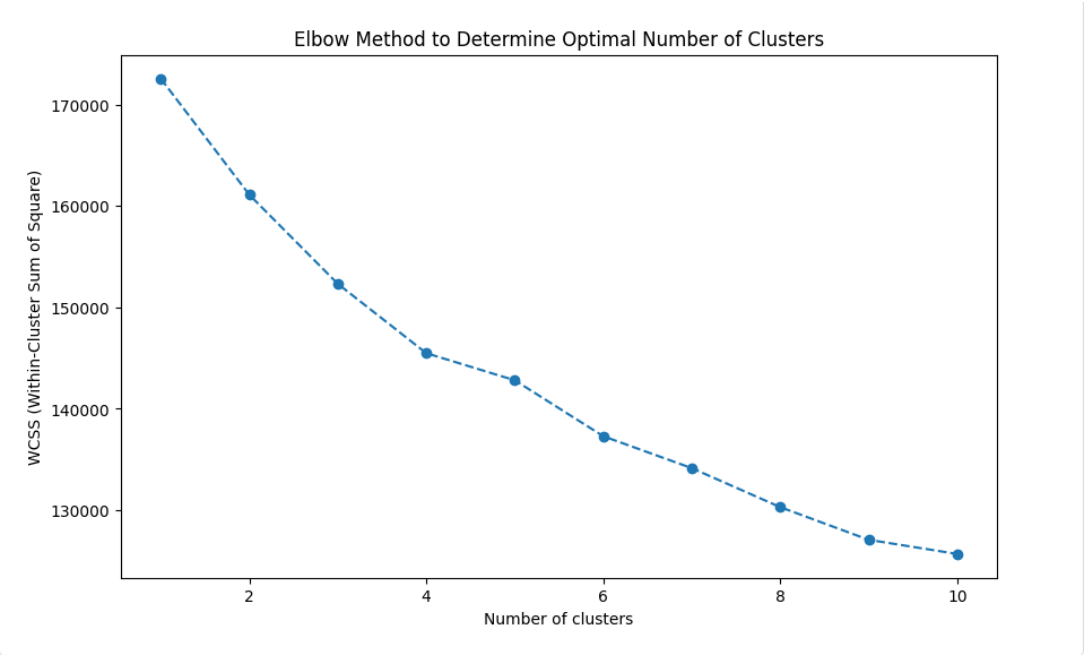
After testing the preliminary models, the job satisfaction and income variables appeared to have a significant impact. Therefore, we decided to focus on variables related to these categories. To start our analysis we created a subset of the data with these variables. The Satisfaction variables included Job Satisfaction, Relationship (Work Relationships) Satisfaction, Environmental (Work Environment) Satisfaction, and work life balance. While the income related variables were composed of Stock option level (Available Stock options), Percent Salary hike, Monthly income, and Hourly rate. Initially, we attempted a linear regression predictive model to try to get a baseline understanding of the relationship. While the training and test errors were low the R-squared values were rather small.

We then moved on to the logistic regression predictive model with an $20 - 80$ training and test data split to attempt to accurately predict the binary attrition variable with the variables of interest. After creating the model we calculated the odds values for each of the variables and confidence intervals to determine the effect of each of the variables on attrition and if the variables were significant to the model.

The key unsupervised method used was K-means clustering, which aimed to segment employees into distinct groups based on their characteristics to ultimately predict which groups of factors contributed most to a higher or lower risk of attrition.

To determine the optimal number of clusters, we used the Elbow Method, which involves running the K-means algorithm for a range of cluster numbers and calculating the Within-Cluster Sum of Squares (WCSS) for each. The optimal number of clusters is identified at the "elbow point" where the WCSS starts to decrease at a slower rate. Since WCSS measures the compactness of clusters, the elbow method points to the idea that increasing k past a certain point does not significantly improve the compactness of clusters, and therefore the "elbow point" is chosen. We then applied the K-means algorithm with the optimal number of clusters and visualized the results using PCA to reduce the dimensions and plot the clusters. Below is the plot where k = 4 was interpreted to be the "elbow point".

Figure 1: Elbow Method Plot



4 Main results

4.1 General Logistic Regression Analysis

The goal of our supervised learning technique was to create a classifier for Attrition and analyze the factors that most strongly impact the target variable. A baseline logistic regression model was created with the entire dataset to assess the performance and fit of the data. We predicted that models with fewer features would classify better and employed various techniques to test this.

Forward and backward selection were also performed to extract the top 4 features selected by either model. The forward selection model chose Gender, Overtime hours worked, Job involvement, and Younger than 25 as significant predictors while backward selection chose Overtime hours worked, Distance from home, Job level, and Stock option level.

To understand the size of our data, we applied PCA. We have 59 features in the dataset, and PCA allows us to extract a lower number of components – aiding the runtime efficiency of a model.

Figure 2: Explained Variance Bar Chart

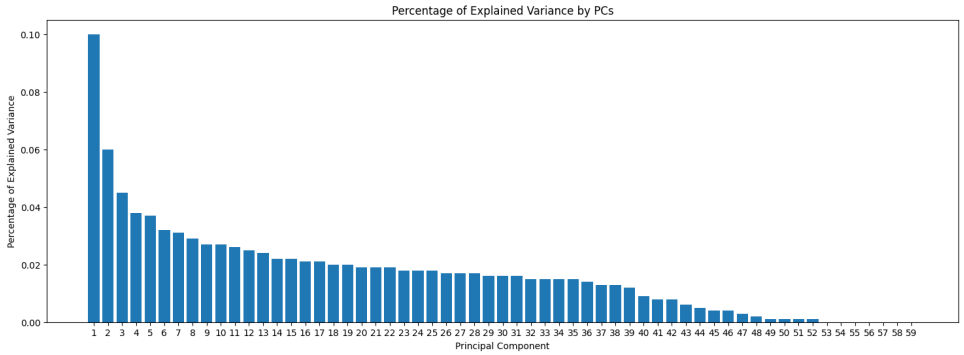
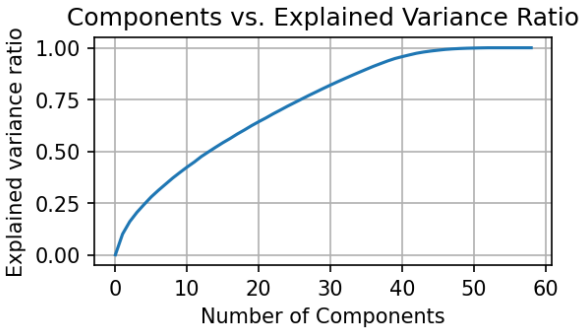


Figure 3: Finding the ideal number of principal components



80% of our data is explained by 29 components while 90% of our data is explained by 36 components. The table below shows the accuracy and ROC AUC metrics for the Logistic Regression model tested with different predictors. Accuracy and ROC AUC were both used as a measure of reliability, and higher scores in each would indicate better model performance. Since our dataset is imbalanced with attrition having a $80-20$ split, it is likely that when the accuracy is high, our model might be predicting the majority class more often.

Table 1: Performance metrics of Logistic Regression models with various subsets of data

Predictors used	Accuracy	ROC AUC
All	0.88	0.85
Forward Selection	0.84	0.71
Backward Selection	0.85	0.75
PCA (29 components)	0.85	0.78

From the results of the table above, it is evident that the model that used all 58 predictors to classify Attrition has the best performance, with the highest accuracy and ROC AUC score. This proves the hypothesis for this objective wrong. Forward selection performed the worst, with a ROC AUC of 0.71, making the subset of data used the poorest to aid in classification. PCA had relatively good performance, with an accuracy of 0.85 and ROC AUC of 0.78. However, we theorize that the nature of the imbalance in the dataset plays a part in the poor classification performance with less data.

To find out the features that most significantly impacted the PCA components, PCA loadings were extracted from each feature. Loadings hold the value of how much a feature impacts a principal component (PC). The loadings across all PCs were aggregated with the absolute value taken. The top 10 most impactful features after examining the loadings were education level, an individual's age group, and the department they worked in. Using this knowledge we proceeded to explore other aspects of our data with supervised learning.

4.2 Focused Logistic Regression Analysis

From the confidence intervals, it is notable that the confidence intervals of Percent Salary Hike and Hourly Rate contain zero, which would suggest that they are not statistically significant to this model. The most standout results for the satisfaction variables are Environmental and Job Satisfaction with a 26% decrease in the odds of attrition. In the other group, Monthly Income had the most noteworthy results with a 48% decrease in odds of attrition. When it comes to model performance the accuracy score was 83.76% and the ROC was 0.67. The most likely cause of this disparity is the model being biased towards the majority outcome in this case no attrition which makes up 80% of the data.

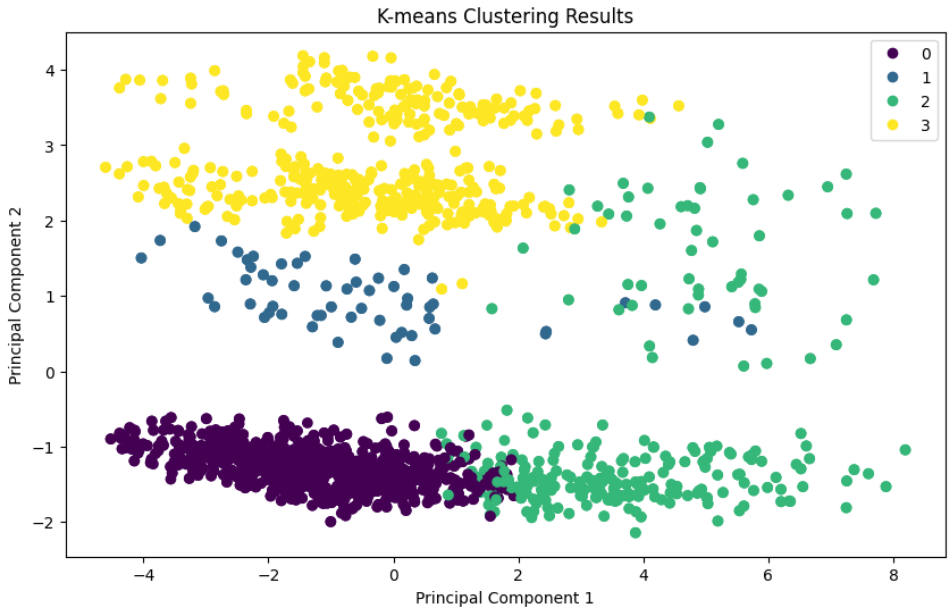
Table 2: Job Satisfaction and Salary Analysis

Predictor	Odds Ratio	Lower CI	Upper CI
const	0.158143	-1.978435	-1.710071
Job Satisfaction	0.735804	-0.419534	-0.194049
Relationship Satisfaction	0.865428	-0.256089	-0.032972
Environment Satisfaction	0.738466	-0.415738	-0.190621
Work Life Balance	0.850452	-0.271671	-0.052305
Stock Option Level	0.650045	-0.561957	-0.299470
Percent Salary Hike	0.940857	-0.176333	0.054405
Monthly Income	0.511778	-0.832948	-0.506780
Hourly Rate	0.974691	-0.139703	0.088433

4.3 K-means clustering analysis

Applying the K-means algorithm resulted in 4 distinct clusters, which can be seen in the plot below.

Figure 4: Clustering Plot



The characteristics can be summarized in the following table:

Feature	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Age	Younger	Varied	Older	Younger
Job Level	Lower	Varied	Higher	Varied
Monthly Income	Lower	Lower	Higher	Lower
Total Working Years	Fewer	Varied	More	Fewer
Attrition Risk	Slightly Higher	Higher	Lower	Slightly Higher
Department	R&D	HR	Varied	Sales

Table 3: Cluster Characteristics

Overall trends pointed to a likely inverse relationship between age, job level, monthly income, total working years, and attrition risk. That is, younger employees with lower job levels, income, and experience were at higher risk of attrition whilst older employees with higher job levels, salaries, and tenures correlated with a lower risk of attrition. In addition, the R&D, HR, and Sales departments were especially prone to attrition.

5 Discussion

One of the main limitations of our supervised methods was the logistic model's bias toward the majority outcome. To combat the imbalanced classes, we tested oversampling the training set and noting the effects it had on model performance. Employing the oversampler, SMOTE, on the baseline logistic regression model (all features against attrition) resulted in a ROC AUC of

0.83 and an accuracy of 0.76. While these are high values, the baseline model without oversampling performed better with an ROC AUC of 0.85 and an accuracy of 0.88. Some possible future directions include improving the logistic regression model during training by using advanced data augmentation techniques to create a more accurate classifier.

On the other hand the unsupervised method is less likely to be affected by the imbalance, as it doesn't train on the Attrition labels. Hence, k-means clustering may be a more reliable indicator of patterns in our data compared to the supervised learning methods. However, that does not mean that it is free from possible future improvements. In future applications, more advanced clustering techniques such as hierarchical clustering or DBSCAN can be employed. In addition, weighting clustering may prove useful as the 4 clusters previously generated had widely varying employee counts, as well as evaluation techniques such as the silhouette score, may be utilized to check model performance.

6 Conclusion

Our goals were to explore supervised and unsupervised techniques with real world data. We met this goal with the implementation of logistic/linear regression models to classify and the k-means algorithm for clustering.

Through our supervised methods we were able to obtain some substantial results, which showed that improvements in job/environmental satisfaction and monthly income result in a lower rate of attrition. However, these results were relatively one-dimensional because while they do highlight key factors affecting attrition, they didn't capture the complexity of interactions between multiple features of the data.

The unsupervised k-means clustering allowed for a better understanding of the complex characteristics of different groups of employees and their relationship with attrition. The most noteworthy discoveries from this method were that the main contributors of higher attrition found within 3 clusters were younger age, fewer years with the company, and lower monthly income. While the contributors to lower chances of attrition were found within 1 cluster, which is composed of older employees, higher job levels and monthly income, and more years with the company.

While these methods gave different levels of results it is evident that we were able to meet our goal of identifying the primary characteristics that lead to attrition. With this knowledge, companies could attempt to create more incentives to keep younger people working in their positions. These results would suggest that younger people tend to feel less connected to where they are working resulting in them leaving if the characteristics mentioned are met somewhere else.

7 File Appendix

To view and/or download resources used in this project, including the source code and ipynb file, click here.