Attrition Analysis

Benjamin Edwards (bje43), Artina Maloki (am2495), Lillyan Pan (ldp54)

Abstract
This analysis identifies the strongest correlating features that relate to employee attrition and explores models to identify the most potent warning features at an employee level. To solve this problem we trained and iterated over several models that would best predict attrition based of non-trivial features. The best model in terms of overall accuracy and recall score was Logistic Regression with $l2$ regularization and $\lambda = 1$. Over all models, whether or not an employee works overtime regularly was one of the strongest predictors of attrition. For companies this will reduce incurred expenses to find replacements, recruiting, advertising, hiring and training, etc. and increase morale to boost employee productivity and the company's public image.

Contents

1. Exploratory Data Analysis

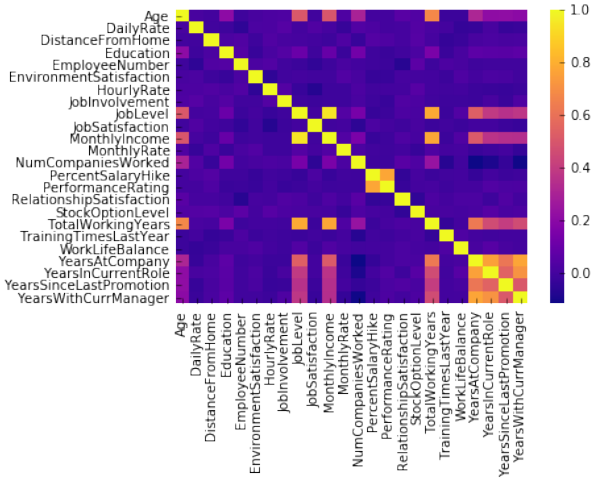

Figure 1. Correlation Matrix between Continuous and Ordinal Features

1.1 Data Characteristics

The *IBM HR Analytics Employee Attrition and Performance* dataset provides comprehensive information about employees, including demographic and financial data. The data set contains employee attrition data in the form of a variable matrix with 36 features and 1470 rows of employees. Each employee has a field specifying whether they left the company and how long they worked there. It also quantifies job satisfaction, work life balance and other company related metrics on a scale from 1 to 4, with 1 being low and 4 being very high. This dataset gives details into many possible causes for an employee to become unsatisfied with their current work situation. It also provides us with the granularity to see how factors contributing to employee attrition differ by age, gender and education.

The dataset is comprehensive and complete. We performed a data validation test to assure that there are no null or missing values. The features in the data set are broken

down into the following data types, 17 continuous variables, including age, monthly income, and years at company.

- 8 categorial variables, including education level, job level, and department.
- 7 ordinal variables, including job satisfaction and work life balance.
- 4 binary variables, including employee attrition.

Note, as this dataset was generated by IBM (due to privacy reasons) this dataset has been explicitly okayed by Prof. Udell.

1.2 Data Visualization

To get an overview of how the features are related to one another we created a correlation matrix on the continuous and ordinal variables. After visualizing results on a heatmap Figure 1 we can observe that the majority of the data is uncorrelated, an advantageous attribute in creating our predictive model.

To analyze the distributions of the features in the dataset and their relationship to attrition we created several histogram and density plots. We plotted the distribution for three continuous features that we thought were interesting using density plots: *Age*, *MonthlyIncome*, and *DistanceFromHome* (Figure 2). From these plots we realized that these selected features will be important for our classifications since the distributions between the employees who left and those who stayed are dissimilar. For this specific company we discovered that on average the employees lost are younger, have lower monthly incomes than those who stay, and have longer commute distances.

We were also interested in the following categorical and ordinal features: *JobLevel, JobInvolvement,* and *DistanceFromHome*. By plotting their distributions (Figure 3) we discovered that employees lost in attrition work in lower job levels, are less involved in their jobs and are less satisfied with their work environment than those who stayed.

2. Data Cleaning

2.1 Data Corruption Analysis

By running statistics on the dataset, we found that no data values were missing or corrupted as NULL or N/A values were checked for and summary statistics were run on the dataset to find any invalid data entries (ie. outliers). However, some columns were found to have the same value over all data points- hence in training, models assigned those features a weight of 0. Specifically, feature *StandardHours* only had value of 80, *Over18* only had value of "Yes", *EmployeeCount* only had value of 1.

2.2 Data Transformation

The dataset contains several categorical and binary features, such as *JobRole* and *OverTime*. We used one-hot encoding to change the categorical variables into numeric values to allow for regression (expanding the number of features to 55). Similarly, we used indicator encoding to assign the binary variables to integer values.

2.3 SMOTE

Synthetic Minority Over-sampling Technique (SMOTE)[1] was employed as originally, the data exhibited a large imbalance in classes with "No" labels far outweighing "Yes" labels in both the training and test data partitions as shown in Figure 4. Unbalanced datasets are a problem in that when one classes is highly represented in comparison to other classes, the final model may be a poor predictor of the underrepresented class. This is specifically shown in the recall score of the different models. Note, recall the number of correctly predicted "positives" ("Yes" to Attrition) divided by the total number of "positives". Because it is most beneficial to companies to understand what drives attrition, recall scores are important to optimize. In order to balance the class values, SMOTE was employed. SMOTE is technique that oversampling the minority class observations to improve the quality of predictive modeling. By oversampling, models can better learn patterns to differentiated classes. The balanced label dataset class counts can be seen in Figure 5.

3. Support Vector Machine

3.1 Model Description

We decided to implement a Support Vector Machine (SVM) classification since it is widely recognized as a good tool for high dimensional data. SVM performs classification by constructing an optimal hyperplane that maximizes the margin between two classes. For our algorithm we explored different Support Vector Clustering methods that use a "one-against-one" approach for multi-class classification. It was necessary to linearly scale each attribute since greater numeric ranges may dominate those in smaller numeric ranges. For example in linear and polynomial kernels case, the kernel values depend on the inner products of feature vectors. Thus large values may cause numerical problems.

To determine which model most accurately captured our feature space we analyzed the following combinations of parameters using 5-fold cross validation and grid search,

1. **Kernel Type**: Functions that return the inner product between two points in feature spaces.
   - Linear $\langle x, x' \rangle$
   - Polynomial $(\gamma \langle x, x' \rangle + r)^d$
   - RBF $\exp(-\gamma ||x - x'||^2)$
2. **C**: Penalty parameter applied to all kernels $\in$ [0.1, 1, 10, 100].
3. γ: Kernel Coefficient for Polynomial and RBF kernels $\in$ [0.1, 0.01, 0.001, 0.0001].
4. **d** Degree of Polynomial kernel $\in$ [1,2,3]

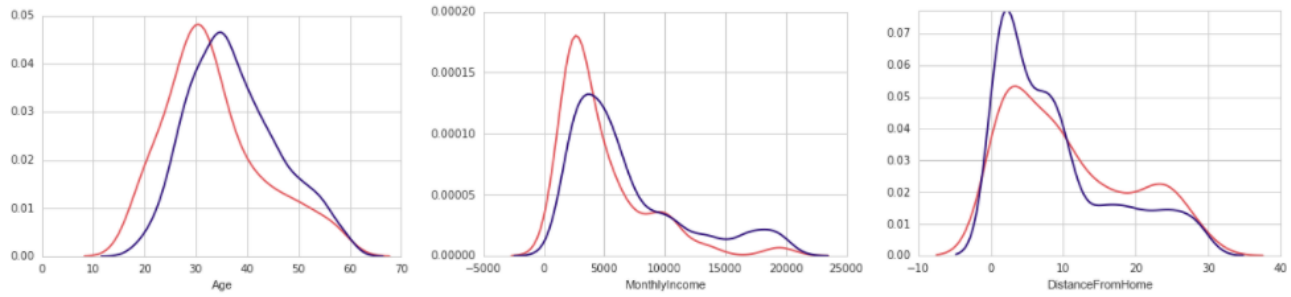

Figure 2. Distribution of Age, Monthly Income, and Distance from Home.

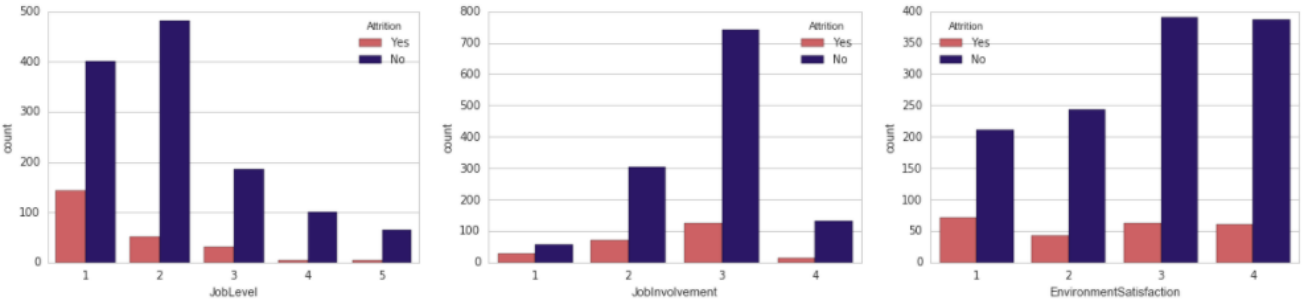

Figure 3. Distribution of Job Levels, Job Involvement and Environment Satisfaction

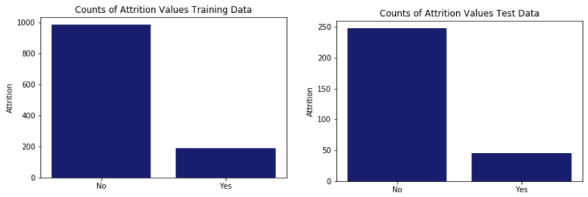

Figure 4. Imbalanced Class Counts. *Left*: Training Data Imbalance. *Right*: Test Data Imbalance

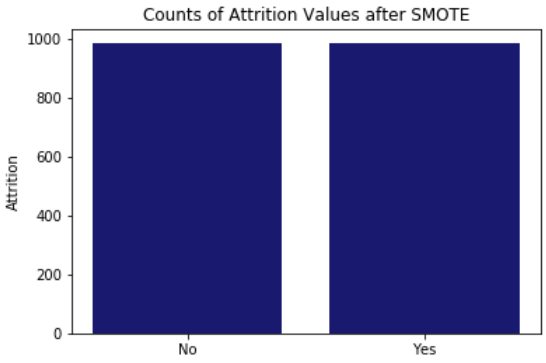

Figure 5. Balanced Class Counts Over Training and Test employing SMOTE

3.2 Results

A 5-fold cross validation revealed that the best performing model was the linear kernel with C = 1. We implemented a SVC algorithm with our optimal parameters on the original standardized data and on standardized data after applying the SMOTE method.

Data Treatment	Training	Test	Recall Score
Regular	0.8937	0.8877	0.500
SMOTE	0.8284	0.6598	0.8283

Table 1. Accuracy statistics for the best SVC model with and without SMOTE

Our results reveal that although the Linear SVC model achieved a 89 % accuracy score on the training and test sets, the classification achieved a poor recall score. The recall score measures how well our model predicts "Yes" for attrition. For employers it is beneficial to achieve higher recall score while maintaining a high test accuracy score. As a result we tested our model on standardized data with the SMOTE method. This model achieved a higher recall score of 82% however our test error fell to 65.9%.

3.3 Analysis

In the case of the linear kernel, our model outputs a set of weights assigned to features that can be interpreted as the primal coefficients in the optimization problem. The size of the coefficients show us the amount of influence the feature has on the prediction and the sign corresponds to if it's positively or negatively correlated to attrition.

After running our algorithm we found that *Overtime*, *TotalWorkingYears*, *JobSatisfaction* and *YearsSinceLastPromo-*

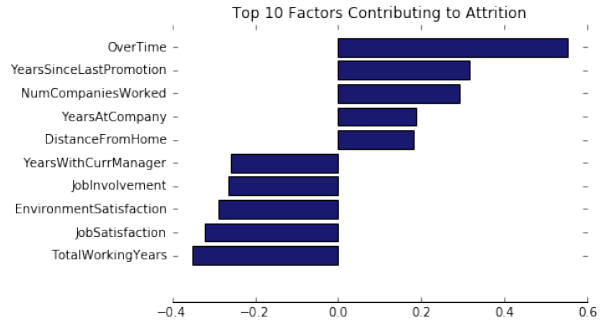

Figure 6. Top 10 Factors Contributing to Attrition

tion have the largest coefficients. See Figure 6 for the list of the top 10 factors contributing to attrition. The features with the smallest coefficients, approximately 0, were *EducationField_Life Sciences*, *MonthlyIncome*, *MonthlyRate*, and *Department_Human Resources*

To assess which areas the company can make improvements to according to our classification results given by SVM, we categorized the features into three groups,

1. **Satisfaction** Figure 7 We can observe that all the satisfaction metrics are negatively correlated with attrition rate (as expected). A lower satisfaction rate corresponds to a higher attrition rate. The two most influential factors are Job Satisfaction and Environmental Satisfaction.

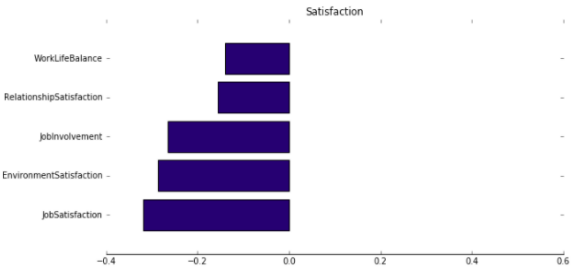

Figure 7. SVM Coefficient for Ordinal Satisfaction Features

2. **Financial** Figure 8 Here we can observe a variety of positive and negatively correlated features. We learn that employees working overtime correspond to high attrition rates and a low stock option level corresponds to high attrition rates.

3. **Demographic** Figure 9 From these features we learn that younger employees and employees with fewer working years contribute to higher attrition rates. Employees who live farther from work and are single also contribute to higher attrition rates.

4. Random Forest

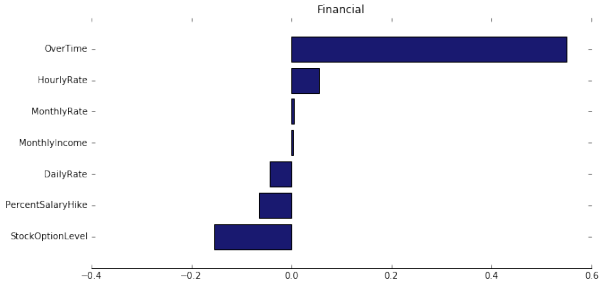

Figure 8. SVM Coefficient for Financial Related Features

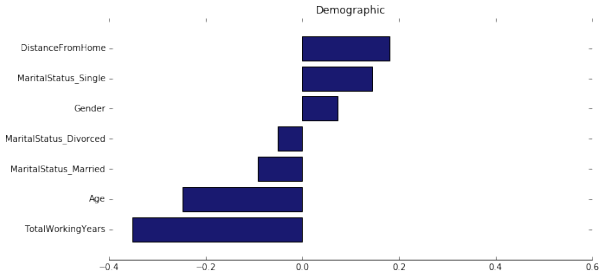

Figure 9. SVM Coefficient for Demographic Features

4.1 Model Description

A popular bagged algorithm is Random Forest, which is a collection of bagged decision trees with an altered splitting criteria.

The algorithm works as follows:

1. m datasets are sampled from the original dataset D with replacement
2. For each generated dataset D_i, a decision tree is trained to a max depth p and before each split a subsample of $k \leq d$ features are only considered from this split.
3. Final classification is chosen using a majority vote among the trees.

The model took in as input all non-trivial features from the dataset to predict the *Attrition* column.

4.1.1 Telescopic Search and Cross Validation

5-Fold cross validation was implemented in order to better tune and compare our models. The number of folds was chosen due to computational constraints of higher folds and Leave One Out Cross Validation, as well as diminishing accuracy returns versus time complexity as the number of folds increased. By using cross validation scores, telescopic parameter search was employed to determine optimal parameters of the model. Telescopic search is performed using two searches: 1^{st}, finding the best order of magnitude for λ; 2^{nd}, performing a more fine-grained search around the best λ found so far. Parameters for Random Forest include the number of trees in the forest and the maximum depth of the tree, and function used to calculate the number of features to subsample. Initial parameters values in the first round of the telescopic search were set as follows:

- `n_estimators` $= [10, 20, 40, 80, 160]$
- `max_depth` : $[10, 20, 30, 40, 50, 55]$
- `max_features` : $['sqrt', 'log_2', 10]$

with an optimal parameter finding of {`max_depth`: 55, `max_features`: $'log_2'$, `n_estimators`: 160}. Note, `n_estimators` is the number of trees in the forest, `max_depth` is maximum depth of the tree, and `max_features` is number of features to consider when looking for the best split– specifically if log_2, then `max_features` $= log_2$(`max_features`). In the second, fine-grain round of telescopic search, the parameters values were set as follows:

- `n_estimators` $= [10, 20, 40, 80, 160]$
- `max_depth` : $[51, 52, 53, 54, 55]$
- `max_features` : $['sqrt', 'log_2', 10]$

with an optimal parameter finding of {`max_depth`: 51, `max_features`: $'log2'$, `n_estimators`: 160}. Note, there was only marginal improvement in accuracy between what the course-grain parameter search found and the fine-grain parameter search.

4.2 Results

Data Treatment	Training	Test	Recall Score
Original Dataset	0.9761	0.8639	0.1521
SMOTE Augmented (Large-Grain Search)	1.0	0.8707	0.2608
SMOTE Augmented (Fine-Grain Search)	1.0	0.8776	0.2608

Table 2. Summary of Accuracy Statistics for Random Forest Model

The optimal Random Forest model used SMOTE augmented data with 160 trees for the forest, each tree was restricted to a depth of 51 (out of 55 features), and a maximum feature selection function of log_2. Gini impurity was selected as the impurity function (as opposed to Entropy) as computationally intensive logarithmic functions are not required. A table of training, test accuracy, and recall score can be found in Table 2. Note, Random Forest feature importance does not return the specific correlation, but rather the predictive power each feature has. Feature importance is determined as the mean decrease impurity. Hence, feature importance can be interpreted as how much each feature decreases the weighted impurity in a tree. A chart of feature importance can be found in Figure 10. Overall, *Overtime* and *Income* were the highest in predictive power, while *JobRole_ResearchDirector* and *EducationField_HumanResources* had the least.

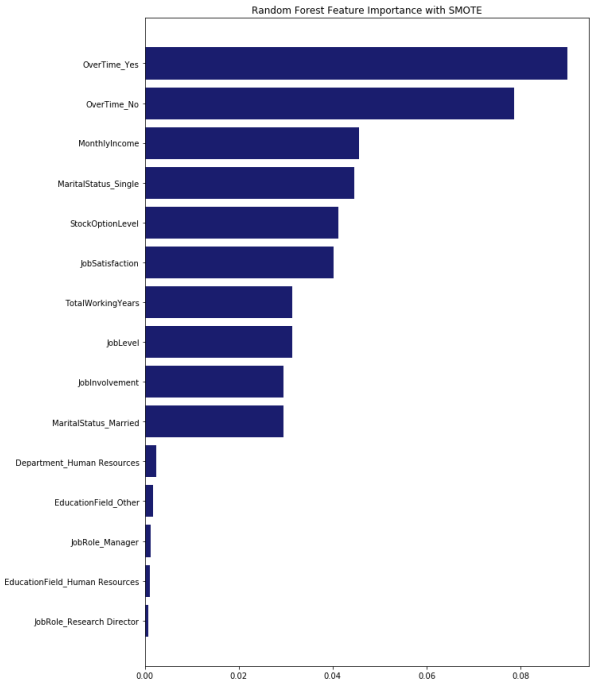

Figure 10. Random Forest with SMOTE Top 10 and Bottom 5 Important Features

4.3 Analysis

4.3.1 Model

Random Forest gives great flexibility to account for all characteristics of an employee with minimal preprocessing. Random Forest decreases variance by before each split a random subsample of features are only considered for the split– in the final Random Forest model we consider the log_2 of the total number of features at each split.

4.3.2 Feature Importance

OverTime_No represents an one-hot encoding to the question: *Do you regularly work overtime*? This vector is 1 if the employee answered "No", and 0 if the employee answered "Yes". This is consistent with general understanding of employee happiness. If an employee has to work over time hours often, the employee may be under a large amount of stress and is less likely to stay at the job- with similar logic applying to *OverTime_Yes*. *MaritalStatus_Signle* was also in the top 10 features of predictive power of attrition as as single people are more likely to shift jobs, while married couples prefer stability. On the other hand, *EducationField_HumanResources* had one of the lowest feature importance in predicting attrition. However Educational background in highly technical/scientific fields (Medical) were given much higher predictive power. Hence, as HR often does not emphasize formal educational background and instead practical experience for functioning

within an HR role, educational background in HR likely does not affect why an employee would want to leave. As these feature rankings are in line with general social behavior, we are sufficiently confident in this model. However, the low recall score– signaling poor performance in correctly predicting "Yes" Attribution cases– motivated us to look to other models.

5. Logistic Regression

5.1 Model Description

A regularized logistic regression model was implemented as another classification method. Logistic regression was used as opposed to other regression models because determining attrition in employees is a binary classification problem and other types of regression are meant to predict real valued outcome variables. A regularized version of logistic regression was used in order to reduce overfitting and to prevent the model from depending on every feature. 5-fold cross validation was used to evaluate models with all combinations of

- Regularization parameter $\lambda \in \{1, 10, 100, 1000\}$
- Loss functions: $l_1 - norm, l_2 - norm$

The best model selected from cross validation was then trained and tested with data treated with the SMOTE method.

5.2 Results

Cross validation revealed that the most accurate model had a regularization parameter of $\lambda = 1$ and a $l_2 - norm$ penalization. The training accuracy, test accuracy and recall score, for this model are shown in Table 3 with and without SMOTE.

Data Treatment	Training	Test	Recall Score
Original Dataset	0.8869	0.8843	0.4317
SMOTE Augmented	0.8247	0.8155	0.6739

Table 3. Summary of Accuracy Statistics for Logistic Regression Model

As seen in Table 3, SMOTE increased the recall score of the model significantly with a small penalty to training and test accuracy. For the purposes of predicting attrition and analyzing the factors causing it, the model using SMOTE has better performance.

5.3 Analysis

5.3.1 Model

Logistic regression is a reasonable algorithm to employ as the outcome variable (prediction) is binary and the predictor variables are continuous and/or categorical– a requirement of the algorithm. Analyzing the cross validation results, the $l_2 - norm$ penalization favors sparse coefficients and may have performed better than $l_1 - norm$ penalization because some of the features like *EmployeeNumber* are not good predictors of attrition.

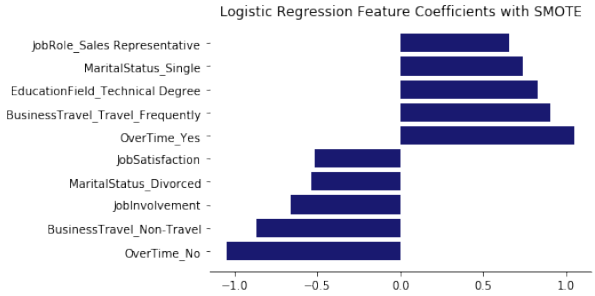

Figure 11. Logistic Regression with SMOTE Top 5 and Bottom 5 Coefficients

5.3.2 Feature Importance

Figure 11 shows the top five positive and negative coefficients in the best logistic regression model using SMOTE. *Overtime_Yes* and *Overtime_No* are the two most significant coefficients suggesting that employees are more likely to leave the company if they are forced to work overtime. Similarly the models shows that employees who leave the company often travel for business frequently. The most important coefficients

As these feature weights are in line with general social behavior, we are confident in this model. Additionally, the high recall score– signaling good performance in correctly predicting "Yes" Attribution cases– further boosts confidence with this model.

6. Conclusion

6.1 Model Selection

Over the course of this analysis three models, SVM, Random Forest and Logistic Regression were used to predict attrition and analyze its attributing factors. Logistic Regression performed the best out of the three models by considering both test accuracy and recall score. Because companies value accuracy prediction on whether an employee is likely the leave the company ('Yes' for attrition), high recall score is crucial. While the Logistic Regression model has 6.21% lower test accuracy over the Random Forest model, the 41.31% higher recall score, lead us to chose the Logistic Regression model as our final model.

The Logistic Regression model heavily weighs overtime and frequency of business travel as factors that contribute to attrition. All three models put importance on *Overtime*, *JobSatisfaction*, *JobInvolvement*, and *MaritalStatus*. This suggests that employers should expect higher attrition rates from single employees who aren't satisfied or invested in their work. Marital status is important in predicting attrition as single people are more likely to have the flexibility to shift jobs, while married couples prefer stability. If employers want to encourage their employees to stay, they should make efforts to ensure their employees are emotionally invested into his/her work, profession, and company.

6.2 Applications

Companies seek to retain and attract top talent and experience in order to grow their company's success and influence, especially as the average cost-per-hire for companies is $4,129[2]. A decreasing company retention rate can severely impact a company in the following social and economic ways:

- Finding new recruits: Including hiring recruiters/recruitment company, spending money on forming applications, application review, etc.
- Loss in productivity: Due to an employee leaving a company, there may be a significant knowledge gap, hindering progress until the organization can find a replacement
- Lost opportunity/talent: The loss of a highly talented employee can be a financial loss
- Decreasing company morale and company public image: High attrition rates may lead to lowered confidence with both the company employees as well as the public with assumptions of a poor workplace
- Training new employees: Training processes will add additional financial burden to the company

With such strong incentive for companies to identify and reverse high employee turnover, an answer to what causes talented and/or experienced employees to leave prematurely is highly sought after.

Employers seeking to minimize their attrition rate can utilize these models to determine what factors are driving people out of their company and hopefully make internal changes to lower their rates. Since our model is not company specific, employers in different industries and companies of different sizes can run our model on data specific to their company and obtained catered results. The output of our model is constructed by predicting employee attrition and the model's coefficients and weights are used to determine which factors are most (and least) influencing attrition rates. These will be unique for every company dataset given. Hence, we are only willing to use our best model in production after being trained on the dataset specific to the company that will make hiring/retention decisions based off the feature weights from the model.

References

[1] Bowyer K. Hall L. Kegelmeyer W. Chawla, N. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[2] Society for Human Resource Management. Shrm human capital benchmarking survey. *Human Capital Benchmarking Report.*