Cornell University
Operations Research and
Information Engineering

Human Resources Analytics: Predicting Attrition

Majda Essadiqi, Silvia Ruiz, Alejandro Ruiz

Abstract: Increasing turnover rates represent higher hiring and training expenses. Understanding the demands of new generations of employees can help companies retain their employees, boost productivity and build a stronger public image. The goal of our project is to predict based on the available data of a Human Resources department whether an employee is likely to leave the company in the next year. To do so four different models are implemented: Logistic Regression, k Nearest Neighbors, Random Forest and Support Vector Machines. The best model in terms of overall accuracy and recall score was Random Forest. The project also identifies the most significant features that the company should focus on.

1. Project Application

The results of this project could significantly help the Human Resources department of this company detect protentional attritioners. The insights obtained could also help them understand the work motivations of these employees and design strategies to meet their job expectations in order retain their talent.

2. Data Description

2.1 Understanding and merging data

The data we used comes from three different sources. The first one comes from an employee survey and contains the following predictor variables: environment satisfaction, work satisfaction and work life balance. These are measured as: 1-low, 2-medium, 3-high and 4-very high. The second one includes the response variable attrition (measured as "yes" or "no") and general information of the employees such as gender, job level, marital status, etc. The third one contains the results of a manager survey evaluation: job performance and job involvement (measured as 1-low, 2-good, 3-excellent, 4-outstanding). We merged the data by employee ID, which resulted in a data set containing 4,410 observations of 28 response variables. The information we are going to analyze comes from 3 different departments within the same company. Through the report we will use the color orange to represent the attritioners and the color blue the non attritioners.

2.2 Data Encoding

After merging the data, we omitted the feature employee ID and those variables that had the same value for each observation, namely: over 18, employee count, standard hours.

Then, since some of the categorical and numerical predictors were stored as strings, we converted them first to floats and assigned the data type missing to the missing values. The variables converted were: environment satisfaction, job satisfaction, work life balance, number of companies worked and total working years.

We converted the following variables to boolean: attrition (yes=1, no=-1) and gender (female=1, male=0).

Next we applied the one hot encoding to the following categorical variables: departments, education field, marital status and job role.

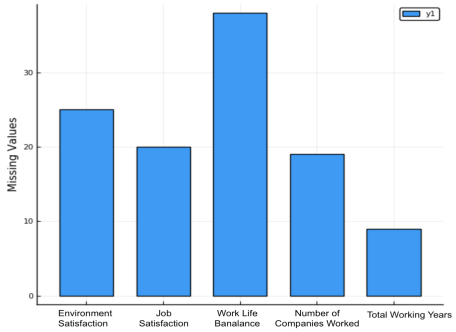
For the ordinary predictors we decided to use rank hot encoding. The ordinal variables converted were: education level, travel, job involvement, job satisfaction, performance rating, work life balance, job level, stock option and environment satisfaction.

Example of hot rank encoding:

Education Level	College	Bachelor	Master	Doctor
Bellow College	0	0	0	0
College	1	0	0	0
Bachelor	1	1	0	0
Master	1	1	1	0
Doctor	1	1	1	1

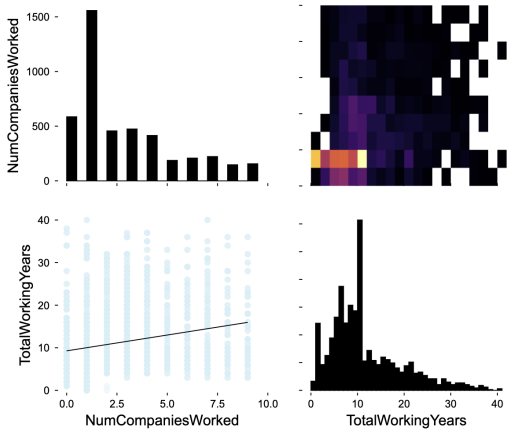
2.3 Missing Values

The columns of our data set that contained missing values were: environment satisfaction, job satisfaction, work life balance, number of companies worked and total working years. The first three variables come from the employee survey data. We made the assumption that the employees that did not answer these questions had a low rating for this ordinal variable but did not feel comfortable answering it. Therefore, we replaced these missing values by the column minimum. For the last two variables (companies worked and total working years) we decided to replace the missing values with the median of the columns, as these variables come from the company records.



2.4 Correlated Variables

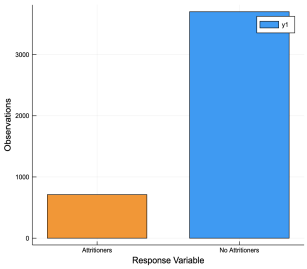
After doing a correlation analysis for all the variables in our data set, the only two variables that seemed to be highly correlated were: number of companies worked and total working years. We decided to keep the later, as it had fewer missing values.



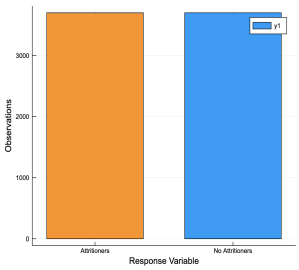
2.5 Imbalanced Data

Before running our models, we realized that our data set was imbalanced. In other words, it contained approximately 80% more observations for non attritioners, than attritioners. Here is a graph of our training data set:

Oversampling and Undersampling



Original Training Set



Oversampled Training Set

Oversampling and Undersampling are techniques to work with imbalanced data sets. The first one works by repeatedly sampling the minority class data points with replacement. The second one consists of randomly removing observations of the majority class. While the first one can introduce bias, the second one can lead to underfitting and poor generalization of the test.

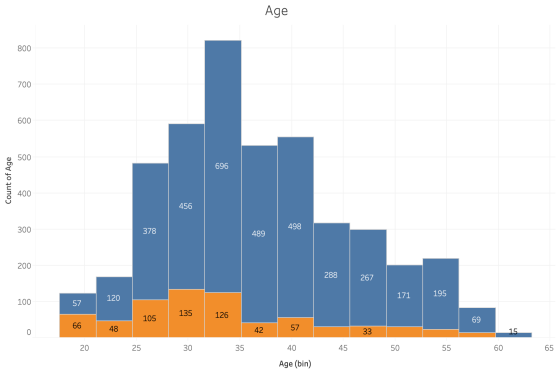
Since we were trying to predict the attritioners, it was important to balance our data set and we used

both techniques. Later in the report we compare the results of the models trained on the original, the oversampled and undersampled data.

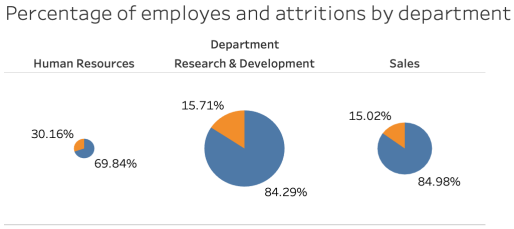
3. Initial Analysis

3.1 Exploratory Analysis

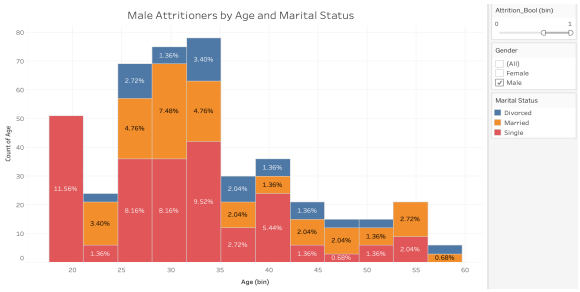
Before running the models, we decided to visually explore our predictors to understand the distribution of our data.

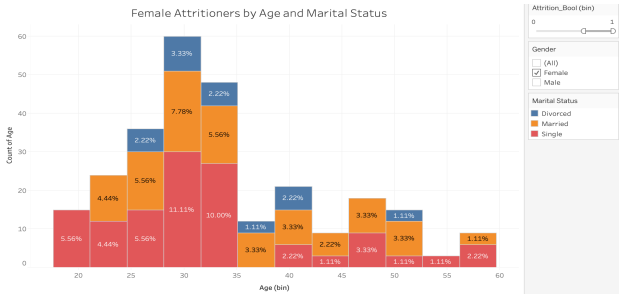


As it can be seen, the distribution of the ages of the attritioners is slightly skewed to the left compared to the distribution of total of the employees. Also, approximately 70% of the attritioners are younger than 35.



In this graph it can be seen that although Human Resources is the smallest department, it has the highest attrition rate (double compared to the other departments).





In these graphs we were trying to observe whether the age, gender and marital status affected the distribution of only the attritioners. For men, we can observe that the distribution is slightly skewed to the left. It can be possible that young men tend to change work more frequently in their twenties, whereas for women that happens in their thirties. Although there does not seem to be any difference in the distribution of the marital status between women and men, we do observe that the majority of attritioners are single.

4. Model Selection

4.1 Logistic Regression

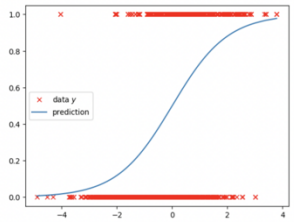
The first model we ran was logistic regression. This classification method was used as opposed to other regression models due to the binary outcome of the project: determining attrition in employees. A quadratic regularized logistic model was also implemented to reduce overfitting. Both models were run first on the original data set and then in the balanced one.

Remember that the logistic model is defined as:

$$logit(p(\boldsymbol{x})) = log\left(\frac{p(\boldsymbol{x})}{1\text{-}p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

The regularized model modifies to:

$$\text{minimize} \sum_{i=1}^{n} l(x_i, y_i; w) + \lambda \parallel w \parallel_2^2$$



Logistic Regression for the original data set

4.1.1 Results:

Rates	Normal	Undersampling	Oversampling	Oversampling with Quad Reg.
Correct Prediction	49.60%	49.50%	84.80%	85.38%
Prediction Error	50.40%	50.50%	15.19%	14.62%
False Positive	0.05%	0.10%	0.00%	0.57%
False Negative	50.35%	50.40%	15.19%	14.06%

4.1.2 Significant Coefficients

Variable	Coefficient	Standard Error	Odds Ratio
Marital Status Married	0.0310	0.025	1.15469
Marital Status Single	0.9455	0.012	2.900605
Marital Status Divorced	-0.2228	0.015	0.878589
Years at Company	-1.1407	0.946	4.804671
Years since Last Promotion	2.5524	0.386	10.56305
Travel Frequently	0.8853	0.127	2.209835
Female	-0.1070	0.112	0.977827
Age	-1.4923	0.366	0.267454
Environment Satisfaction at least 1	-0.1346	0.165	0.533318
Environment Satisfaction at least 2	-0.2014	0.162	0.832041
Environment Satisfaction at least 3	-0.7035	0.147	0.823489
Stock Option Level at least 3	-0.2518	0.280	0.55632

4.1.3 Interpretation

First, we ran logistic regression on the imbalanced data set. As it can be seen in the first table, we got an error of 50.40%. This error was in majority attributed to a large false negative rate. Since our data train consisted almost only of non-attritioners observations, our model was predicting "non attrition" for almost all observations of the test set. Using the undersampled data set did not help solving this problem, as the training set became too small and our model was overfitted. However, the oversampled training data significantly reduced our error rate to 15.19%. Although the error rate is still attributed to the false negative rate, we were able to reduce it from 50.35% to 15.19%. We also ran a regularized version of the logistic regression ($\lambda = 0.05$) in order to reduce overfitting. The quadratic regularization reduced our total prediction error to 14.62%. The false negative rate also reduced to 14.06% and the false positive rate increased to 0.57%.

We also decided to interpret the coefficients of our predictors to understand which variables increase the probabilities that an employee will leave the company. Since our transformed data consisted on 62 predictors, the most interesting results are summarized in the second table. The odds ratios of the coefficients are included in order to understand how a unit increase or decrease in the predictor variable will most likely affect the odds of attrition. To calculate the odds ratios, we took the exponential of each one of the coefficients.

5. Random Forest

Random forest is an ensemble learning method for classification. It trains multiple decision trees and outputs the class that is the mode of the individual trees. Random decision forests correct for the overfitting resulting from decision trees.

Training a random forest applies the general technique of bootstrap. The algorithm works as follows:

1. m datasets are sampled from the original dataset D with replacement
2. For each generated dataset Di, a decision tree is trained to a maximum 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

After training, predictions for test samples x' can be made by averaging the predictions from all the individual regression trees on x' or by taking the majority vote in the case of classification trees.

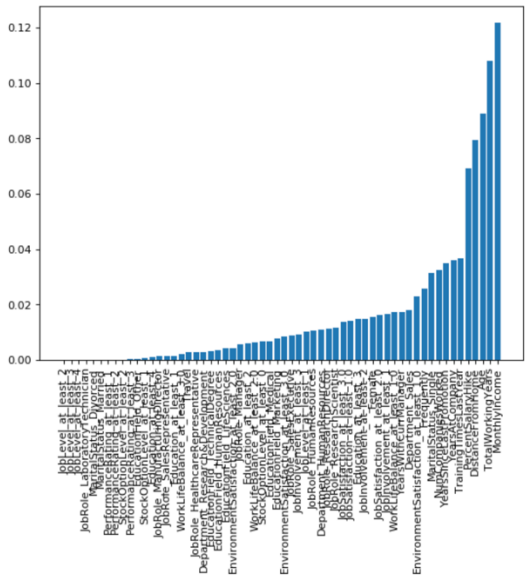
This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its

training set, the average of many trees is not, as long as the trees are not correlated.

5.1 Results

Model	Accuracy training data	Accuracy on test data	Cross Validation score
RF	0.921	0.902	0.923

Feature importance graph:



5.2 Interpretation

This graph does not show the specific correlation of every feature, but rather the predictive power each feature has. Feature importance is determined as the mean decrease impurity (how well the tree splits the data). We used the Gini index (expected error rate of the system) to measure the weighted impurity in a tree. Overall, monthly income, total working years, age, distance from home and percent salary hike were the highest in predictive power.

6. Support vector machine

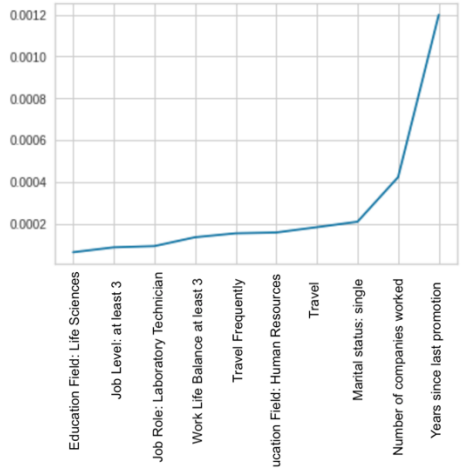
This model is a supervised machine learning tool that classifies the data through a separating hyperplane. Given an input labeled data, the algorithm provides a hyperplane that separates between 2 classes. The particularity of SVM is that it allows some mistakes by trading off the severity of mistakes with the safety margin.

The model depends on several parameters, we tested different values of each through a 5-fold cross-validation and grid search.

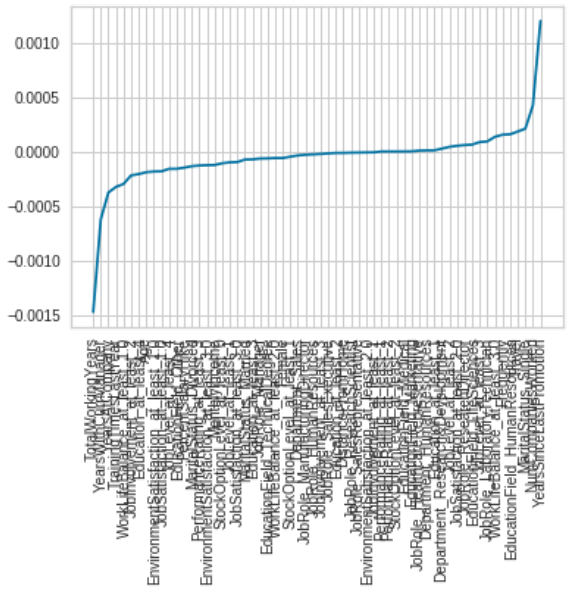
- Kernel function: defines the inner product in the transformed space and reduces the complexity of finding the mapping function.
  - Linear:
    $K(x, x_i) = \sum x \cdot x_i$
  - Polynomial:
    $K(x, x_i) = 1 + \sum (x \cdot x_i)^d$
  - Radial Basis function (rbf)
    $K(x, x_i) = e^{-\gamma \sum (x - x_i)^2}$
- Softening parameter C: trades off correct classification of training examples against maximization of the decision function's margin. We used:
  C =[0.001, 0.01, 0.1, 1, 10]
- Gamma: defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. We used:
  Gamma= [0.001, 0.01, 0.1, 1]

6.1 Results:

Model	Accuracy on training data	Accuracy on test data	Cross Validation score
SVM	0.836	0.844	0.838



The 9 most important features



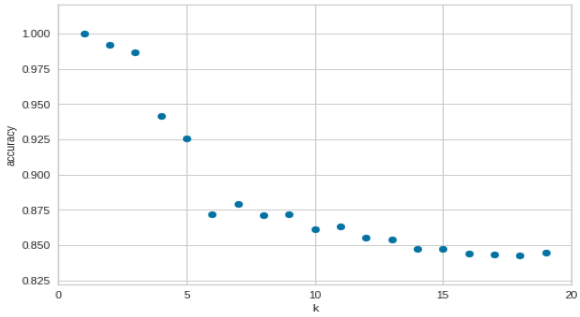
Features' weights

Interpretation:
The cross-validation method shows that the best model considering our data is a linear kernel with C=0.001.

When constructing the model, the features that contribute the most to attrition are the years since last promotion, single marital status, education in human resources, number of companies they have worked in and travel frequency. These results should be interpreted as the higher the coefficient, the higher is the likelihood of the employee to leave. Which is consistent with previous results.

7. K-nearest neighbors

K-nearest neighbor is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space.
It works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbors. Once clustered (we gather K nearest neighbors), we simply take the majority of these K-nearest neighbors to perfect the query instance.



Accuracy of k-NN classifier on the training set with k neighbors:

The accuracy on the training set decreases with the number of closest neighbors, therefore we picked k=1.

7.1 Results:

Model	Accuracy on training data	Accuracy on test data	Cross-Validation score
K-nearest neighbors (1 neighbor)	0.92	0.87	0.91
K-nearest neighbors (3 neighbors)	0.88	0.86	0.89

7.2 Interpretation:

k-NN is a simple yet effective classifier for our dataset. Another advantage is that there's no need to build a model, tune several parameters, or make additional assumptions. Nonetheless, the results obtained do not allow a clear interpretation. Although the accuracy of this model is high, the application is limited given that it is limited for one neighbor.

8. Recommendations and Conclusion

It is important for companies' success not only to attract talented and experienced employees, but also to retain them. A decreasing company retention rate can have a negative impact in many aspects. For example: loss in productivity, loss in talent and increasing training and hiring costs.

This project modeled the probability of attrition using several models (Logistic Regression, k-NN, Random Forest, SVM). The accuracies obtained in all models are high (above 85%).

Therefore, we would trust the results obtained to make changes and decisions only in this company. Before detailing our recommendations, it is important to highlight that this model does not generalize for other companies, as the reasons for attritions in each company can drastically vary. However, it can serve for different employers as a baseline to create similar models for their companies to determine which factors are driving people out of them and make internal changes to lower that rates.

As we can see in section 4.1.2 having a degree in Human Resources or belonging to the Human Resources Department, increases the chances of attrition almost three times. This is an interesting result, not only because it is consistent with our exploratory analysis, but also because it can highlight a problem in this specific department of the company. This high attrition rate could be due to a toxic work environment in that department, a bad department leader or an unfair compensation. We would recommend the company to interview the employees from this department to find out what is causing the problem.

The results of 4.1.2 and 5.1 are consistent with our exploratory analysis as it can be seen that being single affects the chances of attrition by almost 3 times, as opposed to married and divorced individuals. Being female, older and having more years in the workforce decreases the chances of attrition.

Also, some of the variables that seem to increase the probability of attrition the most across all models are: years since last promotion and travel frequently. The first one can be related to the feeling of not being recognized enough and the latter due to increased fatigue that comes with frequent travel. We would recommend the employers to design a strategy or policy to reward hard work and dedication. The rewards do not have to necessarily be material and they can be adapted to the individual preferences of the employees. Also, employers of this company should revise whether they are pressuring their traveler employees too much and either reward them or consider lowering their travel frequency.

It is also interesting that increasing work-life balance does not seem to impact the chances of attrition, as much as increasing environment satisfaction does. Which also evidences that a revise in company culture and environment might be needed.

One of the variables that decrease the probability of attrition the most is years at the company. This can be due to loyalty to the company and identifying with it. This can also be seen in the "option level at least 3" coefficient. Stock options level is a way of compensation in which employees own a stake in the company. Having a greater stock options level might represent that the employees do believe in the success of the company and therefore are not planning to leave the company soon.

9. Fairness

The results obtained in our project could be used by this specific company in the context of understanding their employees in more depth and modifying their culture in order to meet the employees' expectations.

An attrition model like this should not be used when considering hiring or promotions, as it can introduce an unfair bias against certain group of people. In our specific case, given the results of section 4, 5 and 6 we conclude that there would be an unfair bias against young, single, male employees.

As Cathy O'Neil says: "As Data Scientist we should be willing to sacrifice a bit of efficiency in the interest of fairness".

It is also important to mention that if our model had a high false-positive rate, we could target an employee (without present intentions of leaving the company) as potential attritioner and if the actions of the company are targeted as to not promote the growth of potential attritioners, using a model like this would reduce their success within the company and actually increase their chances of leaving (although that was not their initial intention). Since our model has a higher false-negative rate, than false-positive rate, this would not be the case. At least this metric can help us measure that we are in fact not building a "weapon of math destruction" in this case.

10. References:

1. Shalizi, C. (2018) "Logistic Regression", Undergraduate Advanced Data Analysis, Carnegie Mellon University. Consulted: https://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf
2. Lowe, S. (2016) Rank-Hot Encoder for Ordinal Features. Academic blog: https://scottclowe.com/2016-03-05-rank-hot- encoder/
3. Cathy O'Neil (2016) Weapons of Math Destruction. Crown Editorial.
4. Badr, W. (2018) "Different Ways to Handle Imbalanced Datasets". Towards Data Science: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb
5. Gilles Louppe (2015) "Understanding Random Forest: from theory to practice". Department of Electrical Engineering & Computer Science, University of Liège: https://arxiv.org/pdf/1407.7502.pdf
6. Udell, Madeleine (2019). ORIE 4741: Learning with Big Messy Data