Midterm Report: Human Resources Analytics

Majda Essadiqui, Silvia Ruiz, Alejandro Ruiz
Cornell University, College of Engineering
Course: ORIE 4741 - Learning with Big Messy Data

I. Project Goal and Applications

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. The results of our project could significantly help Human Resources understand the work motivations of this group of people, design strategies to meet their job expectations and lower employee turnover within this company. The decrease in employee turnover could help companies lessen their expenses of hiring, training and build a stronger public image.

II. Data Description

A. Understanding and merging the data set

The data we are working on comes from three different sources: an employee survey, an employee data base and a direct manager survey. The first one 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 (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 (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.

B. Missing and cleaning data

After merging the data, we omitted the feature employee ID and those 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
- Total working years

We converted the following variables to boolean:

- Attrition (attrition: yes=1, no=-1)
- Gender (female=1, male=0)

Next we applied the one hot encoding to the following categorical variables:

- Departments
- Education field
- Marital status
- Job Role

For the ordinal predictors we decided to use rank hot encoding. ”Using rank hot encoding changes the question from ”is the level achieved x” to ”is the level achieved at least x”. With this representation of the data, the linear model can explain the effect of a high rank as additive composition of the effect of each rank in turn.” (Lowe, 2016)
The ordinal variables converted were:

- Education Level
- Travel
- Job Involvement
- Job Satisfaction
- Performance Rating
- Work Life Balance
- Job Level
- Stock Option
- Environment Satisfaction

Example of rank hot 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

Finally, since just less than 5% of the values were missing, we decided to replace the missing values by the median of each column. Then, we standardized the numerical columns.

III. Initial Analysis

A. *Exploratory Analysis*

The color orange will represent the attritioners and the color blue non attritioners.

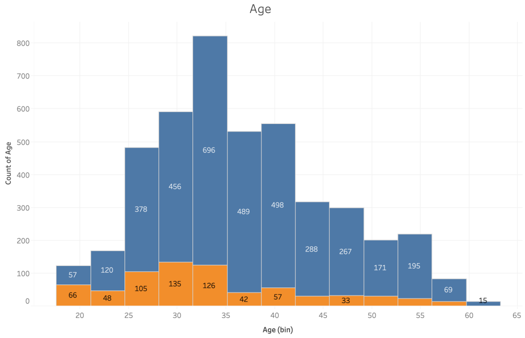

Fig. 1. Distribution of ages of employees

As we can see in Figure 1 the distribution of the ages of 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. We infer the age will be a significant predictor for our model.

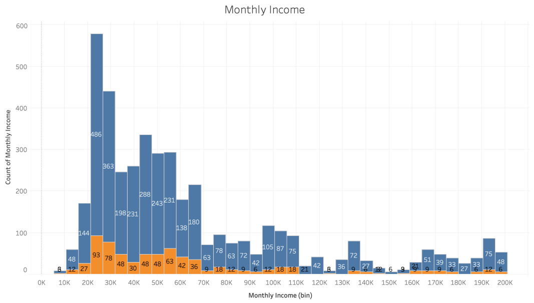

Fig. 2. Distribution of Monthly Income of employees

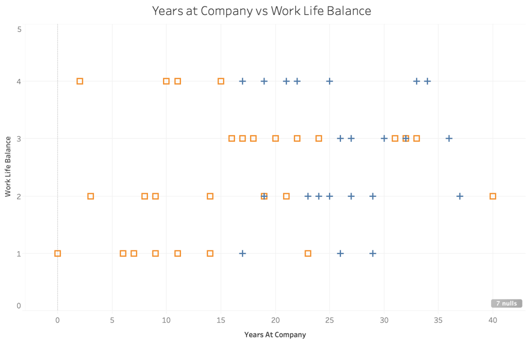

Fig. 3. Years at company vs Work Life Balance

The distribution of the monthly income of the attritioners seems to behave accordingly to the distribution of all employees.

In this graph we wanted to visually classify the two kinds of employees using two predictors: work life balance and years at the company. As we can see, the years at the company (loyalty to the company) seem to have a higher impact than the work life balance.

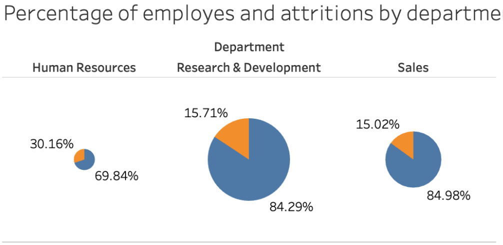

Fig. 4. Percentage of attritioners by department

We can see that although Human Resources is the smallest department, it has the highest attrition rate (double compared to the other departments). We infer that the department within the company will be a significant predictor.

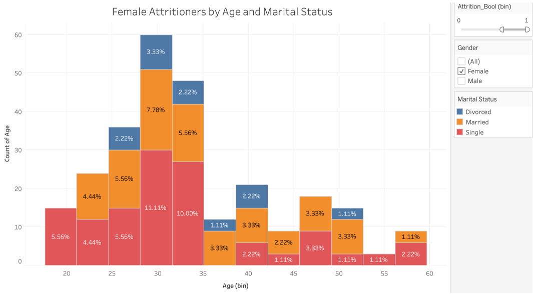

Fig. 5. Distribution of female attritioners by age and marital status

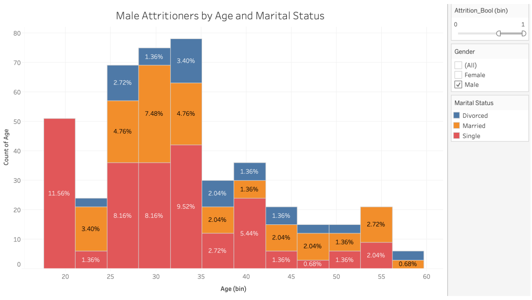

Fig. 6. Distribution of male attritioners by age and marital status

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 persons. Therefore, we infer that gender and marital status will be a significant predictors.

	Attrition rate by travel frequency (in %)
Non Travel	8.00
Travel Rarely	14.96
Travel Frequently	24.91

We also noted that travel frequency could be a key factor in our prediction, since attrition rate reaches 25% for employees that travel frequently whereas this rate is at 10% for employees rarely or not traveling.

IV. Model Selection

A. *Logistic Regression*

The first model implemented was regularized 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. Given the amount of features included in the data set, regularization is necessary in order to avoid over fitting. The regularization term added to the normal logistic regression loss function was $||w||_A$. A regularization parameter of $\lambda = 1$ was chosen both for simplicity and to reduce coefficient sparsity.

Some features that came out as relevant were consistent with our reasoning performing the exploratory analysis, but we have to keep digging and improving our model because our error rates are high and we think that some other features might have a greater impact.

B. *Under fitting and Over fitting*

With over 3,200 employees data in our training set, we believe that removing columns with redundant information will still leave us with a large and diverse enough training set. We plan on building our initial model upon the variables we looked at in our preliminary analysis, since we prefer a simpler model that helps us avoid over fitting. Furthermore, cross validation will also be used to train and test the model on different subsets of training data and further estimate the performance. This will lead to a better regularization parameter for the test model.

C. *Forward Steps*

A next step would be implementing Leave-One-Out Cross Validation (LOO) in order to compute the average error and used it to evaluate the accuracy of our model. Even if LOO is computationally expensive, the amount of data we have is not large enough for this to be a problem. On the other hand, we want to explore some other classification models like Support Vector Machines (SVM) and Random Forest. This way we can have a broader spectrum of results and our conclusion can be more accurate. The scope of this project is just predicting if an employee stays or leaves the company. But, if we conclude that we have the correct data, we can try to predict the years before an employee leaves the company. This is subject to completing the previous steps.

References

[1] Lowe, S. (2016) Rank-Hot Encoder for Ordinal Features. Academic blog: https://scottclowe.com/2016-03-05-rank-hot-encoder/

[2] Yaser, S. (2012) ”Learning from Data: A short course”. AML books.