HR Analytics

Predicting Employee Attrition
Machine Learning II
Individual Assignment
17 February 2019
Federico C. Loguercio

INTRODUCTION

Loosing current employees and hiring new ones can be an expensive process. Apart from needing to find a suitable candidate, they then will need potentially costly and time-consuming training in order to be able appropriately replace the previous worker. And it obviously is most painful when a good employee is lost - perhaps to a competing business. The mechanisms underlying **an employee's decision to leave** can be classified into two buckets: **Push** and **Pull**. Either something was bothering the employee at the current workplace, making them want to leave, or the received a better offer from a different company. Often, the reason is a combination of the both. The more important it is to **recognise in a timely manner when an employee is at risk of leaving** the company in order to be able to **take action**. It turns out that this can indeed be **predicted with satisfyingly high accuracy**.

The aim of this paper is to create a model using logistic regression which can optimally predict who is likely to leave. It shall be noted that **this paper does not attempt to infer causal relationships** between variables and an employee's likelihood to leave the company, or at least not primarily. **The estimated model is optimised for prediction** instead of for the isolation of causal effects. It will still yield an explanation as to which characteristics of an employee influence the chance that they will leave, and in which direction this influence points, but the magnitude of the effect will not be accurately estimated due to high correlation between certain variables. The next section will take a look at the dataset and analyse first evident patterns.

DATA PREPARATION & EXPLORATORY DATA ANALYSIS

The dataset comprises information on **14999 employees** who either left or did not leave, specifying **7 variables** which describe different attributes of each employee, ranging from the average monthly hours they worked to their reported satisfaction level. This preliminary analysis will focus on identifying differences in the distribution of each variable between the pool of employees who left and those who did not, in order to later leverage these insights to **extract, construct and select the ideal variables for the prediction** of the target variable, indicating who left the company. Particular focus will be given to non-linear patterns in the data, as logistic regression tends to struggle to fit these patterns without external help.

Before proceeding with any analysis, **a holdout dataset will be set apart.** The entire validation strategy will be outlined in more detail later on; this holdout set will serve to estimate the true accuracy achievable through the model. Cross-Validation (CV) will be performed on the non-holdout dataset throughout the process in order to determine which features add predictive power to the model and avoid overfitting to that data. More on that later. **10%** of the data were randomly selected into the holdout set, minimising the loss of data to be used for fitting while still reserving a portion of 1500 observations for true accuracy estimation. The following sections will be based on the **remaining 13499 observations.**

	Overall Variable Means	Variable Means for Leavers
Satisfaction Level	0.61	0.44
Last Evaluation	0.72	0.71
# of Projects	3.80	3.86

	Overall Variable Means	Variable Means for Leavers
Avg. Monthly Hours	201.11	207.26
Years at Company	3.50	3.87
Work Accident	0.14	0.04
Promotion in last 5 Years	0.02	0.01
Left	0.24	1

Table 1: Variable Means

The above table shows the overall mean for each variable on the left and the mean for each variable when only employees who have left are considered, on the right. The overall attrition rate amounts to 23.8%, 2.1% of employees have received a promotion within the last 5 years, and an impressive 14.5% have suffered a work accident. On average, employees have spent 3.5 (presumably) years at the company, working 201.1 hours per month and completing 3.8 projects (presumably) per year.

The dataset is very clean, there are **no missing values**. None of the correlations between these unprocessed variables are very strong, apart from monthly hours, number of projects and last evaluation all being fairly correlated with each other. One perhaps unexpected finding is that whether someone had a work accident has a negative correlation with their propensity to leave. It also is obvious from the table above that employees who left very rarely had received a promotion within the last 5 years.

The variables indicating the **department** and the **salary** of an employee are **unfolded into dummies** as they represent a categorical variable.

Skewness

While logistic regression does not assume normality, it is still explored. None of the non-dichotomous variables exhibit skewness in excess of 0.5. No correction is performed.

Scale

Again, scaling is not a necessity for logistic regression. In order to avoid worsening the model fit through scaling and in the light of ease of exploratory data analysis, scaling was moved from being a step of data preparation to being an iteration suggested within the feature engineering pipeline.

Analysis of Variable Distributions

Powerful insights about **key differences between employees who left and others who did not** can be inferred from a comparison of the distribution of each variable for the two classes of employee.

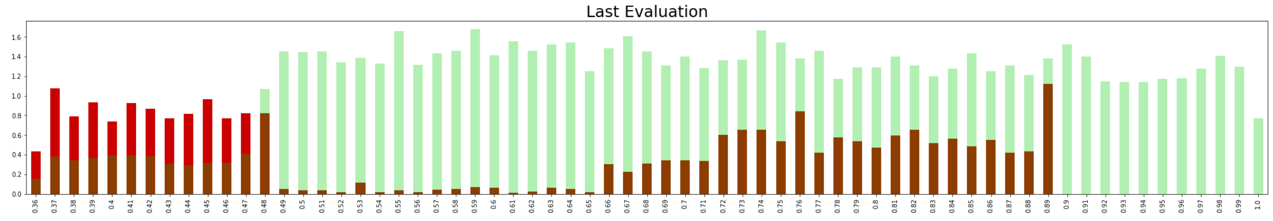


Fig 1: Histogram of Last Evaluation by Left

A clear separation can be identified in the score leaving employees received in the last evaluation. **None of the leavers obtained a score of 0.9 or higher**, and a significant portion of leavers scored below 0.49.

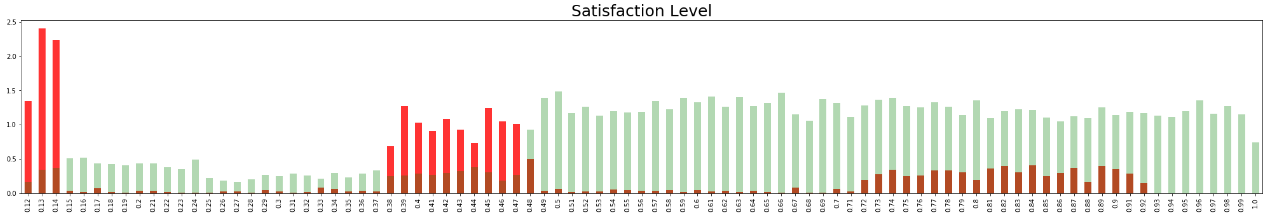


Fig 2: Histogram of Satisfaction Level by Left

A similar pattern can be identified in the satisfaction level: **No employee leaves when they are very highly satisfied**, and there are three clusters of values where many leaving employees can be found.

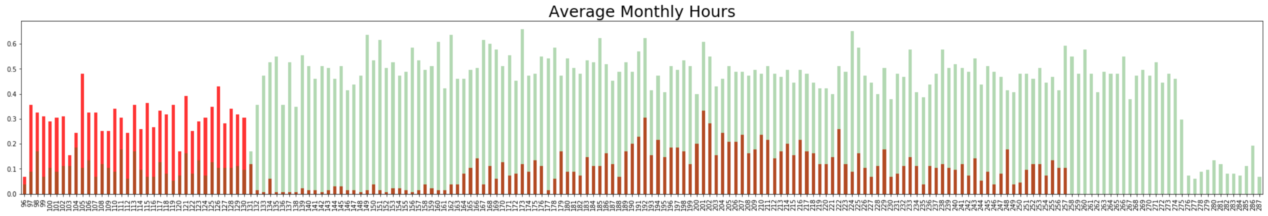


Fig 3: Histogram of Avg. Monthly Hours by Left

The structure continues when the average monthly hours are analysed. Short working hours can be an indicator for attrition, whereas **someone working 258 monthly hours or more should not be expected to leave anytime soon**.

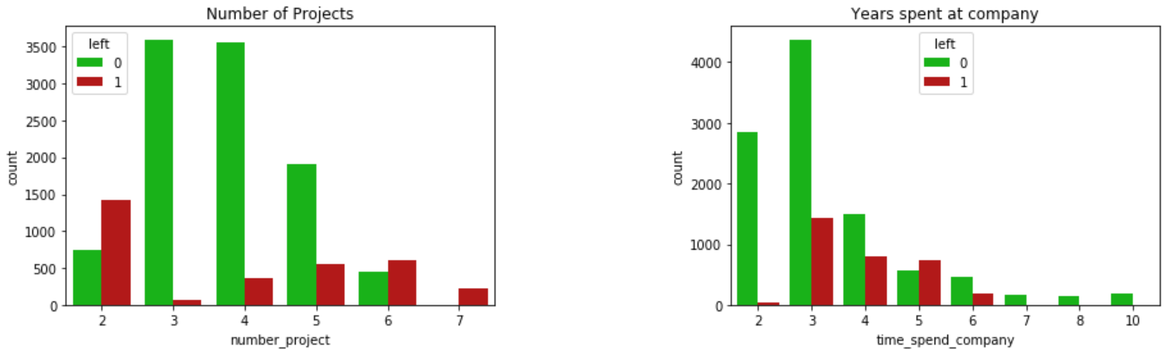
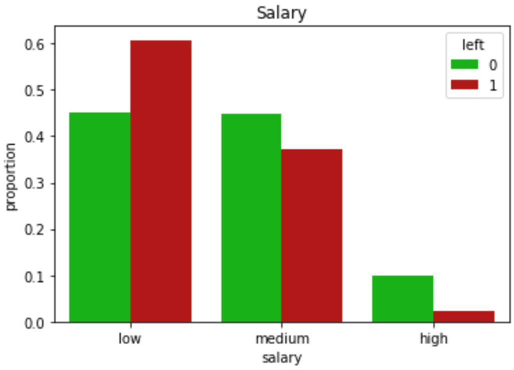


Fig 4 & 5: Histogram of Number of Projects and of Years spent at Company by Left

Regarding the number of yearly projects an employee was on, those on the lower bound display a significantly increased likelihood to leave the company; however, the same is true for employees on the higher bound. Not many have had to complete 7 projects per year, but those who did are, based on past frequencies, certain to leave the company. This is interesting as it show a somewhat different picture from the analysis of monthly work-hours: The employees with the highest number of projects are not necessarily the ones who work the longest hours.



Lastly, observing the permanence at the company, whoever has endure for 7 years or longer is very unlikely to leave, and rather **new employees with less than 3 years at the company generally stay as well**.

Differences in the likelihood to leave across departments are minimal.

Salary-wise, as expected, **the portion of leavers is highest in the low-earning bucket**.

Two-way graphs, coloured by leave-status, convey some further information. They allow to detect which combinations of characteristics are particularly strong predictors for whether someone left. For example, the combination of a high evaluation together with high monthly hours set leavers apart even more than the two characteristics by themselves. However, when the point of feature engineering will be reached, the fit achieved by a decision tree will be analysed. Since that, as we will see, yields not only clear combined thresholds that can be used to differentiate leavers, but also indicates their importance in terms of predictive power, it is deemed a superior method.

Validation Procedure

It was already mentioned at the beginning that a holdout set was put apart for overall model validation. The various variables' distributions in the holdout set do not display any major differences from the non-holdout set. This holdout will be used for true accuracy estimation.

Within the non-holdout set, 5-fold cross validation (CV) was performed. In order to verify the outcome's stability, 10-fold CV was also attempted, trading lower bias in overestimating accuracy for lower variance. The best model is chosen through a pipeline; first, a baseline score is obtained by performing 5-fold CV on the non-holdout set, fitting a logistic regression and returning the average accuracy obtained from prediction on the left out fold. Thereafter, **iteratively, features are added to the model / to the data**, the same accuracy-estimation is performed, and if the accuracy is improved, the feature is added to the model / data. Finally, the selected features are applied to the holdout set. The accuracy is estimated for each of the k estimated models when predicting on the holdout set, and the best one is returned.

A problem would arise if the previously detected patters would be used directly for feature engineering; for example if dummy variables were created which subset the continuous variables into bins that exactly separate the distinct distribution patterns between leavers and stayers. These would likely be accepted by the pipeline but they would be overfitting the non-holdout set. Therefore, only the overall structures found will be taken into consideration.

Baseline

The best **CV-accuracy** obtained through logistic regression including all variables is **0.80**. The same maximum accuracy is obtained by the best of the k models predicting on the holdout set.

FEATURE ENGINEERING

A **first set of features** was **created using business understanding.** These are **hours per project**, which is a simple division of monthly hours by the number of projects (the hours were not rescaled to be yearly as it was deemed unnecessary), and **long hours many projects**, which is a multiplication of the two. Both were accepted in the pipeline as improving the model. On the other hand, excluding one level (one dummy) of the two categorical variables, salary and department, was rejected.

Next, a series of steps was suggested based on the exploratory data analysis' findings. The histograms suggested that there were clear, non-linear boundaries between leavers and non-leavers. In order to facilitate the creation of these boundaries to the estimator, the **continuous variables are binned into 10 equal width bins** as a suggestion in the pipeline. Similarly, the time spent at the company and the projects per year are turned into categorical variables. All of these indeed improve the model, at least based on cross validation.

Further, a **decision tree is fit to the data**. In an attempt to take advantage of the tree's capability to capture non-linear relationships while at the same time not overfitting (since the tree is fit to the entire non-holdout set, outside of the cross validation), the first two decisions of the tree are mimicked through dummies. In other words, one dummy is created indicating whether an employee whether an employee reported **high satisfaction and** has been at the company for a **long time**, and another one indicating the employees with a **low satisfaction** level who completed **few projects**. In order to further reduce the risk of overfitting, the satisfaction level, which is used for both splits, is not being binned. These steps are clearly accepted in the pipeline.

Scaling the variables is rejected. Similarly, a function **removing outliers** in the truly continuous variables (avg, hours, satisfaction level and last evaluation) based on a z-score exceeding 3 is also rejected; it appears not to detect outliers in any of the training-folds.

In an attempt to perform feature reduction, **recursive feature elimination** is applied, automatically tuning the number of features selected. Thereby, features are recursively left out, a model is fit, and it is scored based on the cv-accuracy reached (5-fold). Features whose reduction improve model accuracy are thereafter excluded. In our case, the algorithm suggests to drop on base category for each of the real categorical variables (salary and department), as that increases cv-accuracy. However, holdout performance is lowered by doing so.

Given the non-linear boundaries observed in the exploratory analysis, fitting **polynomial features** might help to explain them. During that process, all possible combinations of features with degree less than or equal to 2 are created and a model including them is fit. Indeed, a maximum cv-accuracy of 0.97 is reached. However, the accuracy on the holdout set is lower than without polynomial features; the features are leading to significant overfitting. Higher degrees than 2 are not explored as they would lead to even stronger overfitting.

At last, the **optimal parameter for regularisation** is detected through a grid search with 5 fold CV. From a set of 10 values logarithmically distributed between 1 and 10000, 21.54 is found to result in the best cv-accuracy.

Finally, model accuracy is estimated on the holdout dataset by using the best generalising model, as in the model of the 5 CV iterations performing best on the holdout set, including the features selected throughout the pipeline and performing regularisation using the optimised threshold. **An accuracy score of 0.958 is reached**.