# Improving Employee Retention at YATS Corp.

Technical Report - Py-ous Inc.

André Faddoul Dóra Linda Kocsis

Noémie Quéré Kilian Tep

## 1 INTRODUCTION | MOTIVATION

Employee turnover is part of the regular business cycle, with employees leaving a company for reasons as diverse as retirement, career opportunities or other life circumstances. However, there exists excessive employee turnover, where the amount of people leaving is abnormally high and should raise a red flag for a company to review its management style.

The reason why companies should aim to retain their employees is, in addition to the talent loss, the cost of hiring new workers. Indeed, whenever an employee leaves, the incoming replacement must be hired, trained, and it usually takes up a few weeks or months for his performance to be optimal. This creates a costly process for the company, especially in the case of high-level employees.

At YATS Corporation, the average retention rate of employees is 76%, whereas the standard in the industry is around 85%. Hence, there is a discrepancy between the number of YATS' employees leaving the company and the number that should actually be leaving. Our main business problem is therefore to find the underlying causes of departure, as well as the type of employees leaving (e.g. if they are good or bad workers). Ultimately, the goal will be to retain the 'good' employees to raise the retention rate while maintaining a good performing workforce.

As YATS counts around 15,000 employees, it can lose 2,250 workers within a calendar year without it becoming a problem. While several causes can be at the origin of the departure of good workers, the ones we are focusing on this analysis are mainly linked to the amount of work per employee (number of projects, number of hours worked), their history at the company (number of years working for the company, promotions received, work accidents) and their benefits (salary, satisfaction level).

We then chose to tackle this problem by creating a binary classification model to predict if an employee will leave YATS Corp. Several models were tested, including Random Forest, Logistic Regression, Support Vector Machines, Gradient Boosting and K Nearest Neighbors. Ultimately, we managed to achieve an over 90% accuracy, which allows us to predict confidently the list of employees leaving YATS Corp. to advise on a better employee retention strategy.

## 2 PROBLEM DEFINITION

The problem we are trying to deal with is to build a model capable of predicting whether an employee is going to leave a certain company, based on certain criteria collected. From these criteria, some are more important than others and will have a bigger impact on employees' resignations.

Determining the most important features will lead to better understanding of the employees' behavior and can be very effective in terms of finding solutions to improve employees' retention.

Hence, the issue presented will be treated by building a binary classification model for which the output is whether an employee will leave. The importance of each feature will be determined using the Random Forest Classifier.

To get the best prediction accuracy, different classification models will be tested such as Decision Tree, Random Forest, Gradient Boosting, Support Vector Classification (SVC), Logistic Regression, and K-Nearest Neighbor.

While visualizing data, the density plots of the employees that left the company indicated a certain pattern similar to a bimodal distribution with respect to the employees' evaluation and average numbers of hours spent at the company each month. The shapes of these curves divided the employees that left into two groups: employees with high evaluation and spending a lot of time at work and others that have low evaluation and not spending much time at the company.

From this point of view, some additional experiments have been done on a certain group of employees that can be considered the best employees in terms of evaluation, number of projects they were working on and the number of years spent at the company. This group of employees can be considered the one that has a priority when it comes to its retention. The point of the additional modeling was to determine the main reasons behind the resignation of the best employees and to fit a model that can have a higher accuracy on this type of employees.

## 3 RELATED WORK

The employee retention problem is certainly one of the most important issues in the human resources industry and a lot of work has been done in this field.

A lot of books discuss the reasons behind employees' turnover and the strategies to ameliorate their retention. From these books, we can cite: "Love 'Em Or Lose 'Em" by Beverly L. Kaye and Sharon Jordan-Evans, "Managing Employee Retention by Adele O'Connell and Jack J. Phillips, "Managing Employee Turnover: Dispelling Myths and Fostering Evidence" by David Allen, and many more. The mentioned books are based on theories, researches and surveys done on employees.

In addition, there exist some HR tools and softwares that calculate employees' performances and can be used to predict their turnover. Apps like "CompanyMood", "Culture Amp", "STAYview", "TINYpulse" and "WeThrive" are used for the same reason. They are mainly handled by psychologists and data scientists to create surveys with the right questions and analyze the results after these questions are answered by a company's employees.

The way our work relates to the above-mentioned books and tools is by the way the problem was addressed. The data from YATS Corp. was analyzed based on the key features affecting their employees' turnover. The same predictive analytics techniques were used when it comes to modeling and results comparisons.

The main difference is that our predictive analytics are based more on objective criteria that have less to do with an employee's psychology or subjectivity since all the features except the employee satisfaction are based on facts and not emotions.

## 4 METHODOLOGY

### 4.1 DATASET

The dataset [1]at hand aggregates records of 15,000 employees at YATS Corp. Without information on reasons for leaving the company the following attributes were known to us:

- `satisfaction level`: *numerical*
- `latest evaluation (yearly)`: *numerical,*
- `number of projects worked on`: *numerical*
- `average monthly hours`: *numerical*
- `time spend in the company (in years)`: *numerical*
- `work accident (within the past 5 years)`: *numerical*
- `promotion within the past 5 years`: *numerical (binary)*
- `sales (i.e.: department)`: *categorical*
- `salary`: *categorical*
- `and whether the employee has left the company or not`: *numerical (binary)*

The latter constitutes as the target of our classification model. During the modeling phase we decided to retain all of these features.

### 4.2 DATA PRE-PROCESSING

During the data cleaning and feature engineering processes, we have carried out the following steps:

- **missing values analysis:** there were no missing values found during the analysis
- **outlier detection:** preliminary data and boxplot analyses returned no reason to exclude or impute any of the variables at hand at this stage.

As a next step, we carried out two types of variable transformation. The `salary` variable was codified into numerical. The feature, `sales` was encoded into dummy variables creating the following additional attributes: `'sales_IT'`, `'sales_RandD'`, `'sales_accounting'`, `'sales_hr'`, `'sales_management'`, `'sales_marketing'`, `'sales_product_mng'`, `'sales_sales'`, `'sales_support'`, `'sales_technical'`.

Variable transformation is an indispensable step as most of the algorithms implemented and further described in Section 4.4 require numerical input and output variables.

### 4.3 PRELIMINARY DATA ANALYSIS

Given the scope of this project is first to understand reasons why the most valuable employees leave and given this information predict who will leave next, a preliminary data analysis can help in identifying these reasons.

We begin by analysing the correlation plot of the features at hand. Fig. 1 suggests, that `satsfaction_level`, `last_evaluation`, `number_projects`, `average_monthly_hours` are the most significant factors in an employee's departure from YATS Corp.

[1] Source: https://www.kaggle.com/ludobenistant/hr-analytics

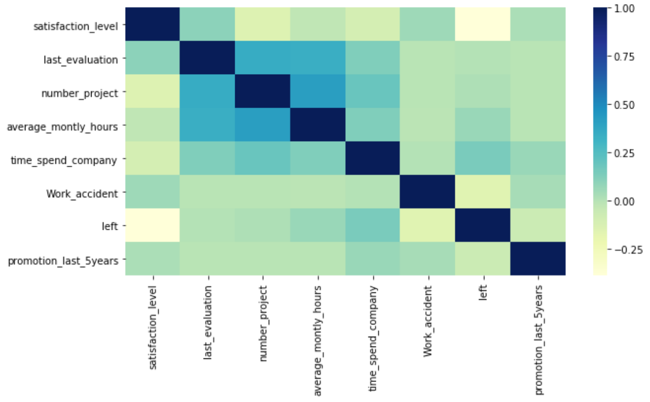


**Figure 1. Correlation Matrix**

Analyzing the identified features described above one by one we managed to gather further insight for setting our modeling strategy.

### Satisfaction Level

With regards to satisfaction level, we were primarily curious to know, the general feeling of YATS Crop.'s employees across the different departments.

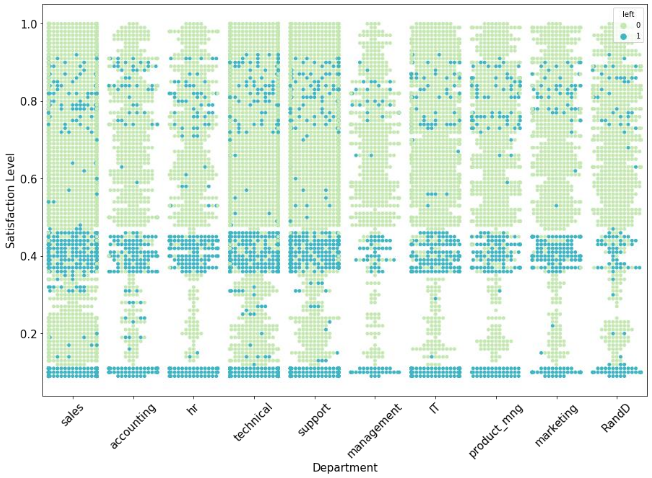


**Figure 2. Department & Satisfaction Level vs Left**

As Fig. 2 clearly indicates, the department does not play a role in the distribution of employee departure as a function of satisfaction level.

However, one insight we could deduce is, that there exist at least 3 clusters of reasons to leave for the 15,000 employees:

- Low Satisfaction (<0.1)
- Medium Satisfaction (ca. 0.4)
- The departure rate is also higher for Very High Satisfaction level (around 0.8)

### Last Evaluation

By observing the density plot of last evaluation as a function of whether the employees left or not, we observed a bimodal distribution for those who leave.

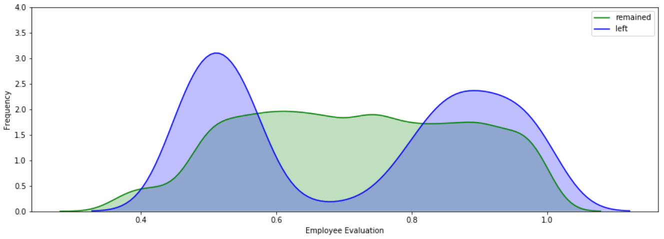


**Figure 3. Last Evaluation vs Left**

Such picture, clearly signaled a red flag during our analysis. As the next variable in our analysis was average monthly hours, we decided to plot the same graph with this variable to find any similarity in distribution and to conclude whether the amount of work affects employee turnover for the highly evaluated or not. Indeed, by looking at Fig. 4, there is a similarity between the two density plots.

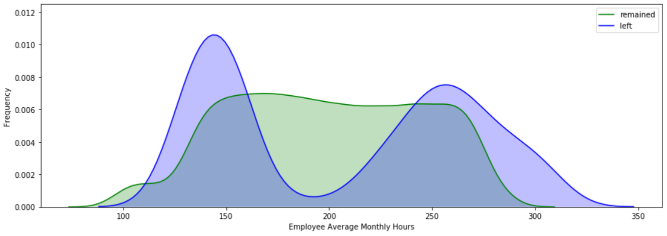


**Figure 4. Avg. Monthly Hours vs Left**

Clearly, overtime affects the decision-making process of employees evaluated both high or low.

### Number of Projects & Avg. Monthly Hours

The correlation between these two features seem obvious, however after projecting it against employee departure, we could extract actionable insights for our modeling strategy. The following boxplot summarizes our findings.

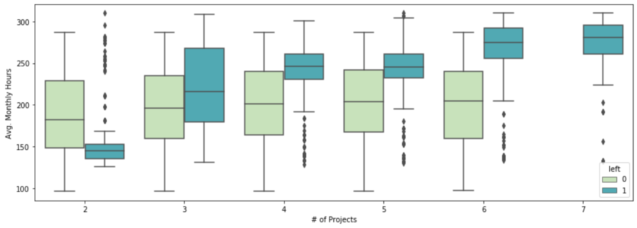


**Figure 5. Avg. Monthly Hours vs # of Projects**

The # of projects is positively correlated with avg. monthly hours. Employees who stay seem to work less in terms of hours, despite the increase in projects. Those who leave are the ones who experience sufficient increase in working hours. The threshold for an employee in terms of project count seems to be around 3-4 projects.

### Further Insights

Clearly the two main variables (being the most prone to subjectivity) are satisfaction and evaluation levels. When further investigating these variables we found, that:

- Highly evaluated employees are more likely to leave after 4+ projects are assigned to them.
- In turn, satisfaction level also drops and employee turnover increases after being assigned to 4+ projects
- Optimum # of projects should be 3-4 for good employees to keep them satisfied and increase chances of retention.
- Employees receiving the highest evaluation tend to leave after spending more than 4 years at YATS Corp.
- Satisfaction & evaluation levels seem to follow similar trends when projected against # of Years

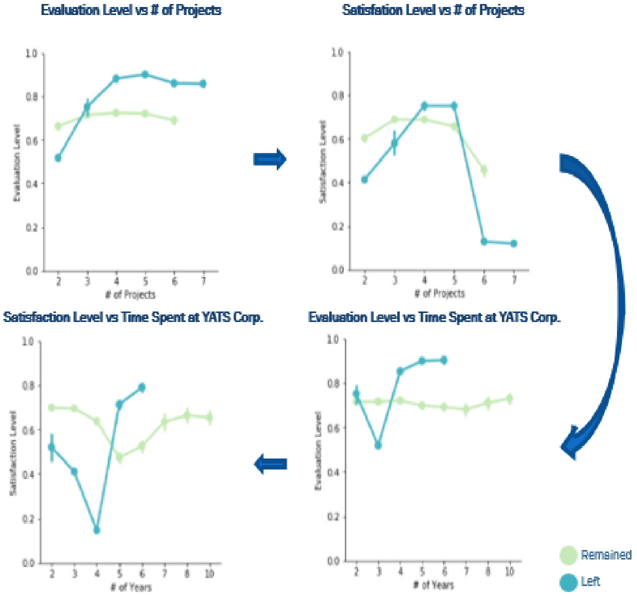
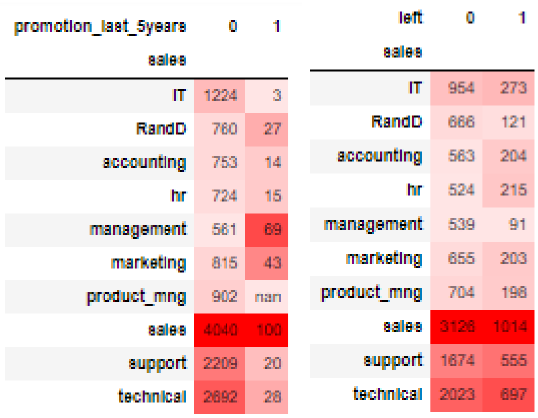


**Figure 6. Evaluation & Satisfaction Level vs # of Project & Time Spent at YATS Corp.**

Furthermore, it is interesting to highlight the high concentration of employees in the lowest salary range. Although such concentration does not necessarily signal a red flag in a company's strategy, further analysis is required regarding the proportion of employees receiving a promotion in the last 5 years.



| promotion_last_5years | 0 | 1 |
|---|---|---|
| sales | | |
| IT | 1224 | 3 |
| RandD | 760 | 27 |
| accounting | 753 | 14 |
| hr | 724 | 15 |
| management | 561 | 69 |
| marketing | 815 | 43 |
| product_mng | 902 | nan |
| sales | 4040 | 100 |
| support | 2209 | 20 |
| technical | 2692 | 28 |

| left | 0 | 1 |
|---|---|---|
| sales | | |
| IT | 954 | 273 |
| RandD | 666 | 121 |
| accounting | 563 | 204 |
| hr | 524 | 215 |
| management | 539 | 91 |
| marketing | 655 | 203 |
| product_mng | 704 | 198 |
| sales | 3126 | 1014 |
| support | 1674 | 555 |
| technical | 2023 | 697 |

**Figure 7. Department vs Promotion & Left**

Indeed, observing the relative distribution of those leaving the company and not receiving a promotion over the last 5 years, the results suggest a high correlation.

### Employee Clustering

As suggested by the insights described above, we suspected three main clusters of employees working at the Company:

- Good but Sad employees: evaluation high, satisfaction low
- Bad and Sad employees: evaluation low, satisfaction low or medium
- Good and Happy employees: both evaluation and satisfaction high

After performing K-means Clustering, we obtained this hypothesis was further supported by Fig. 8. This technique is often used in classification problems. Essentially, we treat a set of observations as a d-dimensional real vector (where d is the number of features). K-means clustering segregates these observations into a pre-specified number of sets, k (≤ n) with the aim to minimize the within-cluster sum of squares (i.e. within-cluster variance).

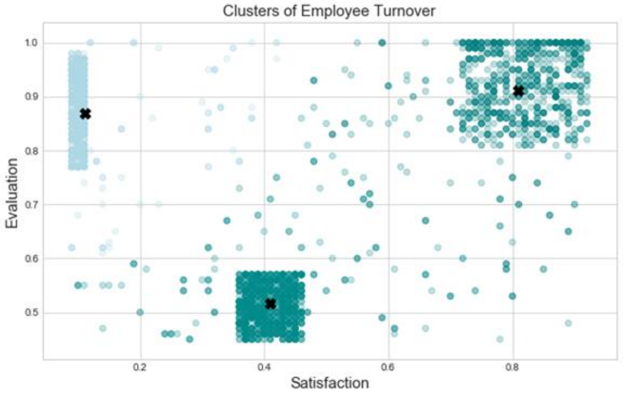


**Figure 8. K-means Cluster Analysis**

As our primary aim is to find out, why good employees leave the Company, these results help us in setting the thresholds for satisfaction and evaluation levels when subsetting the data on the cluster of Good but Sad employees.

The next section aims at giving a brief overview on the implemented models, followed by the description of the strategies used and conclusion for the project at hand.

## 4.4 MODEL TUNING & RESULTS

Given that we face a binary classification problem, we should explore the capabilities of classification algorithms. We chose to implement the following models: Decision Tree, Random Forest, Logistic Regression, Support Vector Machines, Gradient Boosting, and K Nearest Neighbors. We decided to run these models both with and without parameter tuning where it was possible.

### 4.4.1 Overview of deployed models

### Decision Tree

The Decision Tree algorithm belongs to the supervised learning algorithms. When used for classification problems, the model leaves serve as the class labels and its internal nodes represent the association of attributes leading to those labels. The main challenge of this algorithm, is to identify the right attributes are to be considered as the root of the "tree" and each level. Primarily, there are two fundamental measures for attribute selection: information gain and the Gini index. During this study, the latter has been implemented. This metric measures the number of times a randomly chosen data point has been incorrectly identified. The lower the metric, the more preferable is the attribute selection for the algorithm.

**Random Forests (RF)**

Random Forest is an ensemble machine learning method, based on the development of a multitude of decision trees. The aim is to make decision trees more independent by adding randomness in features choice. For each tree, the algorithm draws first a bootstrap sample, obtained by repeatedly sampling observations from the original sample with replacement. Then, the tree is estimated on that sample thanks to feature randomization, which implies that searching for the optimal node is preceded by a random sampling of a subset of predictors. The final result can either be an average or weighted average of all of the terminal nodes that are reached, or, in the case of categorical variables, a voting majority.

**Logistic Regression**

Logistic regression belongs to the same family as linear regression, e.g. the generalized linear models. In both cases the goal is to link an event to a linear combination of features, except that for logistic regression, we assume that the target variable is following a Binomial distribution. For the first part of this project that concerns the prediction of the message substance, we will use a multinomial logistic regression since our problem is a multi-class one.

**Support Vector Machines (SVM)**

Support Vector Machines are based on the research of the optimal margin hyperplanes, which goal is to correctly classify or separate data to be as far away as possible from all other observations, if possible. This objective can be achieved by:

1. Defining the hyperplane as a solution of a constrained optimization problem which objective function is only expressed through vector scalar products. The number of active constraints or support vectors controls the model's complexity.
2. Introducing a kernel function in the scalar product inducing implicitly a nonlinear transformation of the data into an intermediate space of higher dimension. These kernels (specifically adapted to a given problem) give more flexibility to the model to adapt to diverse situations.

**Gradient Boosting**

The principle of Boosting is to construct a family of models which are then aggregated together thanks to a weighted average of estimations, or a vote. However, the family is constructed differently from other ensemble methods. In this case, each model is an adaptive version of the previous one: it gives more weight to wrongly predicted observations. Intuitively, this algorithm focusses on observations which are more difficult to adjust. In other words, it improves capacities of weak learners. In this project we implemented Gradient Boosting (GBM), an iterative algorithm based on the gradient descent.

**K Nearest Neighbors (KNN)**

KNN is a simple algorithm that computes a similarity measure across observations, such as Euclidean distance, and makes predictions by searching in the dataset for the K most similar cases called neighbors. The final output is obtained by taking the most common value among these K instances.

### 4.4.2 Classification Results

### Approach 1: Using the Entire Dataset

Before setting any specific strategy, we decided to check overall model performance on the entirety of the data also without any parameter tuning.

### Feature Importance

Using the `RandomForestClassifier` we deduced the feature importance for our attributes.

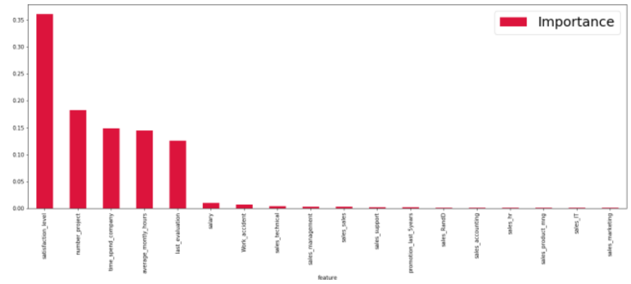


**Figure 9. Feature Importance – Entire Dataset**

The function returned `number project`, `satisfaction level`, `average monthly hours`, `time spend company` and `last evaluation` as the most significant variables for building this model. As mentioned earlier in the report, we decided to retain all features during the modeling phase.

To compare all the models presented before, we implemented a Monte Carlo cross-validation. 50 different training and testing sub-samples were randomly sampled from the training set. We decided on sampling 25% of the dataset with each iteration. The accuracy scores for each method and each validation sample were computed and plotted in the boxplot below.

During this step, we also applied the StandardScaler utility class. This is a common requirement for many algorithms we decided to deploy. This class takes care of the normalization of the dataset, to approximate a standard normal distribution. This is done by computing the mean and standard deviation on the training set to be able to later re-apply the same transformation on the testing set.

The figure indicates that the best performing and most stable models are: Decision Trees, Random Forest, and Gradient Boosting.

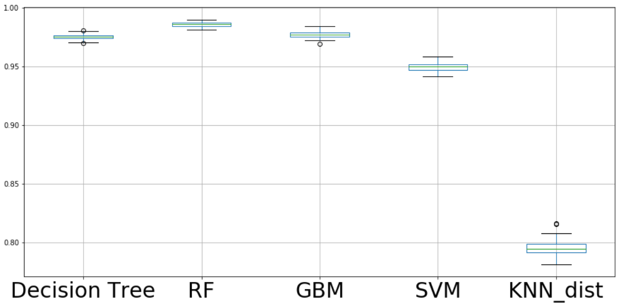


**Figure 10. Model Comparison – Entire Dataset**

### Approach 2: Dataset Split Between 'Good' and 'Bad' Employees

Additional modeling was carried out on a specific group of employees. These employees represent the cluster described above as Good and Sad employees. They are evaluated highly, assigned to the highest number of projects spending more than the average number of years at the Company. This group of employees can be considered as the one that has a priority when it comes to retention. The key objective here was to determine the main reasons behind the resignation of this group and to fit a model that can result in higher accuracy.

The dataset split was carried out by subsetting employees, who:

- were evaluated with score of over 0.75
- were assigned to more than 5 projects

- have been part of the Company for over 4 years

This subset resulted in over a thousand observations.

The necessity of this step is further underlined by the refitted `RandomForestClassifier`, finding, that: `satisfaction level`, `average monthly hours`, `last evaluation` and `time spend at the company` are the most significant factors for this group.

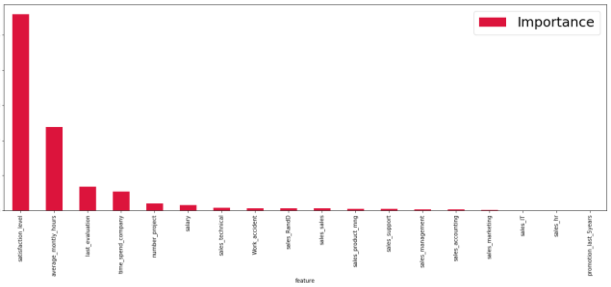


**Figure 11. Feature Importance – Good Employees**

The same models were fitted as before, with Random Forest and Boosting performing the best, but each with an accuracy score well above 90%.

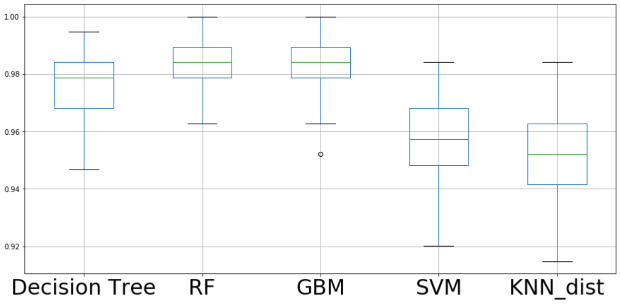


**Figure 12. Model Comparison – Good Employees**

To improve model performance, we decided to apply parameter tuning on Random Forest and Decision Trees. Finally, using the Voting Classifier that combines these two models into a new one. This model results in a more powerful version with an expected higher performance.

During this step, both soft and hard voting were applied. The underlying difference is that the prior uses majority rule voting for predicting class labels, while the latter uses the argmax of the sums of the predicted probabilities. The boxplot did not indicate significant difference of in the performance of the two-different type of Voting methods. However, it suggests a more stable performance for the tuned Random Forest model.

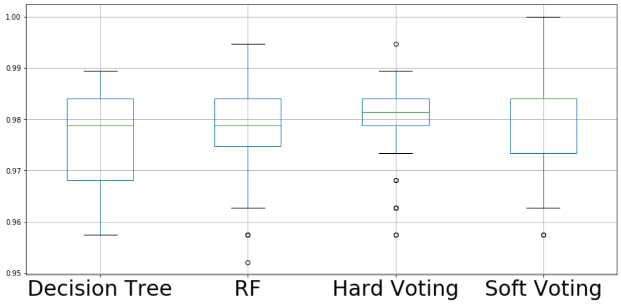


**Figure 13. Model Comparison – Tuned Models**

### Final Model Comparison

Finally, running our models on the true test set resulted in the highest performance for the Voting classifier with no difference in Accuracy between the soft or hard methods. As the confusion matrix shows below, the model misclassified only one observation. This misclassification is an equivalent of a Type I error, however, as the employee was classified as someone staying at YATS Corp. whereas, he/she decided to leave.

Clearly, for the model to increase reliability, we need to work with a bigger sample of employees we can analyze.

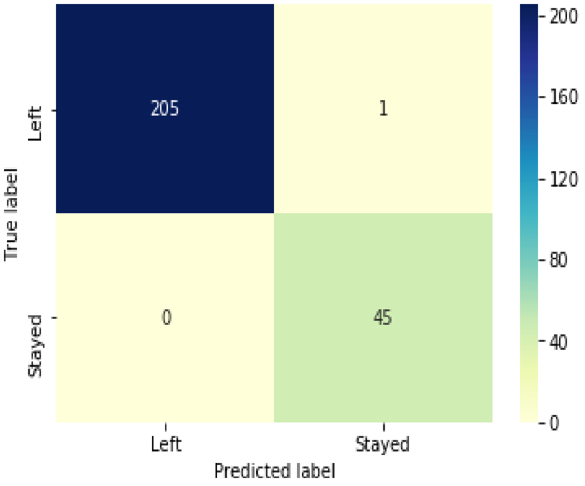


**Figure 14. Confusion Matrix – Voting Model**

## 5 LIMITS AND CONCLUSIONS

Our analysis led us to believe that the primary factors affecting employee retention rate were:

- Satisfaction level: how fulfilled the employee is at work. Fulfilled employees are less likely to leave the company.
- Number of projects the employees work on, which leads to more employees leaving if the number of projects is too high.
- The average monthly hours, which also leads to more departures if employees work too much time.
- Time spent at the company: fewer employees leave as they spend more years working for the company.
- Finally, a promotion in the last 5 years also decreases the chances of employees leaving.

Moreover, a key takeaway from this report is that there exists a divide between 'Good' (last evaluation greater than or equal to 0.75) and 'Bad' (evaluation less than 0.75) employees when it comes to employee retention. While having an overall understanding of employee retention is important, it is more economically impacting to focus on retaining the employees who are best evaluated as they are the most productive. Our main recommendation is to adapt the aforementioned levers to minimize the likelihood of best employees leaving. For instance, YATS could deploy initiatives to improve the employees' comfort and well-being, which would aim to increase employee satisfaction.

On the technical side, the best classification model eventually turned out to be the Voting Classifier, with F1, Precision, Recall and Accuracy hovering around a score of 1 on the test set. We are thus confident our model can accurately predict whether the employee will depart from the company.

However, given the nature of the data, our report presents some limitations. First and foremost, we do not know whether employees who left were laid off, transferred, or resigned. Resignations would be the one we would seek to most minimize; yet we have no indication for differentiating them. Furthermore, while it is the determining criterion impacting the decision to leave, employee satisfaction is a subjective feature, which makes it hard to know how exactly act upon this lever. We also do not know how the economic context potentially affects the employees' decisions to leave.

Further leads for improving our comprehension can be to add economic variables (i.e. the country's unemployment at the time when the employee left the company, etc.). We will also need to refine the distinction among layoffs, transfers and resignation for future studies.