PYOUS Inc.

YATS CORP
Employee Retention

Andre Faddoul
Dóra Linda Kocsis
Noémie Quéré
Kilian Tep

MISSION

MINIMIZE THE RETENTION RATE

24%

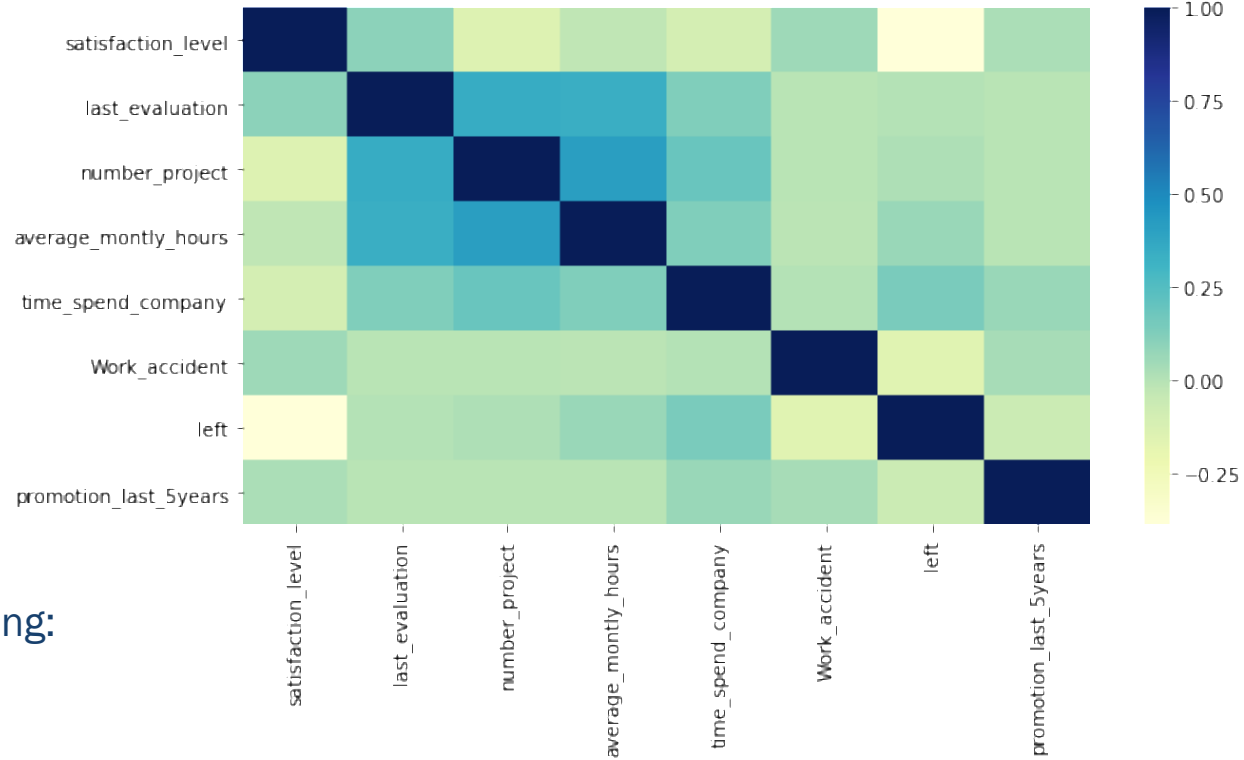
15%

BY Determining the significant factors leading to employee's resignations

TO Create or improve retention strategies on different employees

DATA PREPROCESSING: DATASET

- 14999 observations, 10 features



- Data preprocessing:
  - *missing values*
  - *outliers*
  - *normalizing data*
  - *turning categorical variables into numerical (salary, sales)*

DATA PREPROCESSING: INITIAL CONCLUSIONS (I/III)

The distribution of **satisfaction level** follows **similar trends** across the different departments.

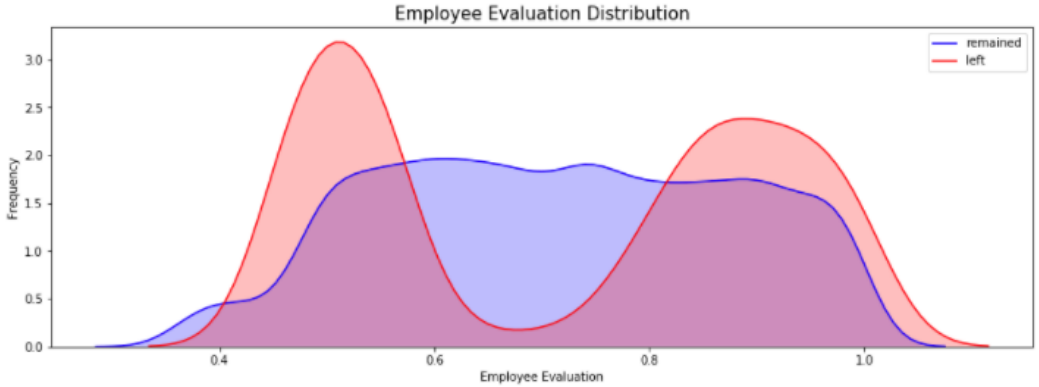
In each department, **3 observable clusters** tend to leave the company:

- **Low Satisfaction** (<0.1)
- **Medium Satisfaction** (around 0.4)
- The departure rate is also higher for **Very High Satisfaction** level (around 8)

WHAT CAUSES THE DEPARTURE OF VERY SATISFIED EMPLOYEES?

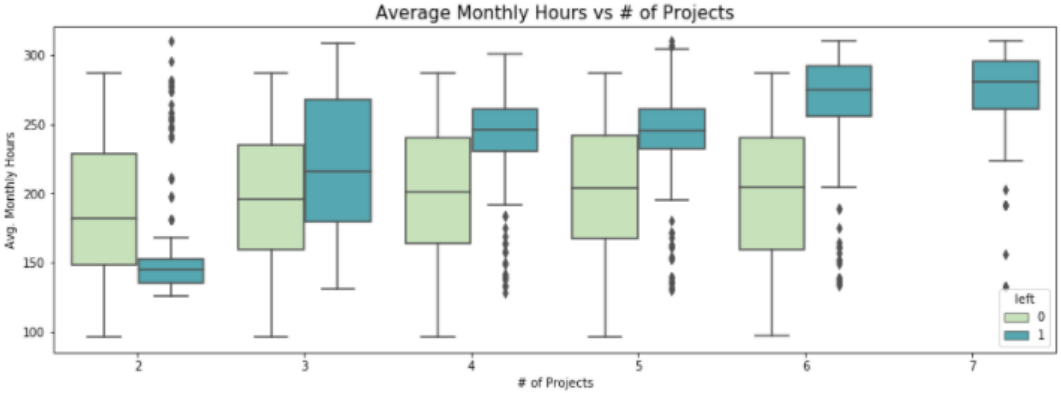
DATA PREPROCESSING: INITIAL CONCLUSIONS (II/III)

EMPLOYEE EVALUATION DISTRIBUTION



- Employees with **low (<0.6)** and **high (>0.8)** evaluation have a higher tendency to leave.
- The **optimal rating** seems to be **between 0.6 and 0.8.**

AVG MONTHLY HOURS DISTRIBUTION/ # OF PROJECTS

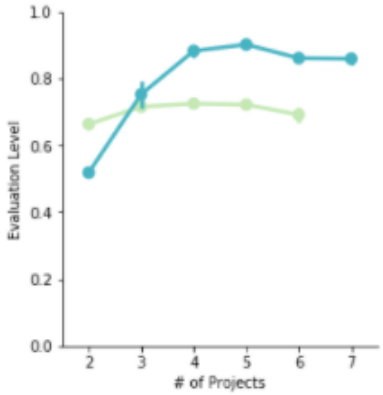


DOES THE AMOUNT OF WORK AFFECT THE EMPLOYEE TURNOVER FOR THE HIGHLY EVALUATED?

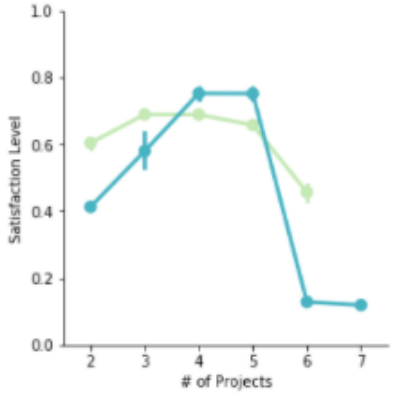
- The # of projects is **positively correlated** with avg. monthly hours
- Employees who **stay seem to work less** in terms of hours, despite the increase in projects
- Those who **leave** are the ones who experience **sufficient increase**

DATA PREPROCESSING: INITIAL CONCLUSIONS (III/III)

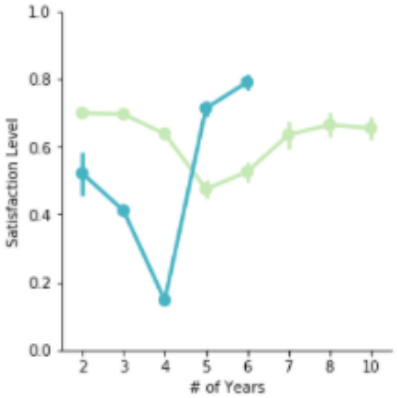
Evaluation Level vs # of Projects



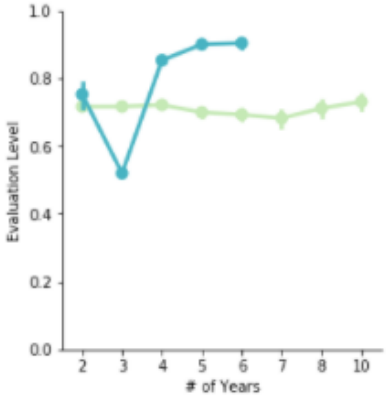
Satisfation Level vs # of Projects



Satisfaction Level vs Time Spent at YATS Corp.



Evaluation Level vs Time Spent at YATS Corp.



Remained

Left

- People **who stay are less impacted** by those variables
- It seems, that **highly evaluated employees** are more likely to leave after **4+ projects** are assigned to them.
- **Satisfaction drops and employee trunover increases** after being assigned to 4+ projects
- **Optimum # of projects should be 3-4** for good employees to be satisfied.
- Employees receiving the **highest evaluation** tend to **leave after >4 years**
- **Satisfaction & evaluation levels** seem to follow **similar trends** when projected against # of Years

HOW CAN YATS INC. RETAIN GOOD EMPLOYEES AFTER 4 YEARS?

DATA PREPROCESSING: K-MEANS CLUSTERING

- **Good but Sad employees:** evaluation high, satisfaction low
- **Bad and Sad employees:** evaluation low, satisfaction low/medium
- **Good and Happy employees:** evaluation high, satisfaction high

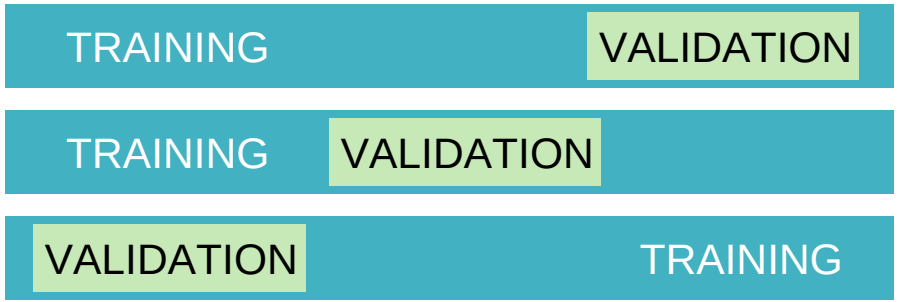
ARE THE GOOD EMPLOYEES LEAVING?

DATA PREPROCESSING: DATASET SPLIT / CROSS-VALIDATION

DATASET SPLIT

- The dataset was split int two: a **training (80%) and testing (20%) sample**
- The testing sample allows to calculate the **prediction error** from which the prediction capacity of the applied models can be calculated

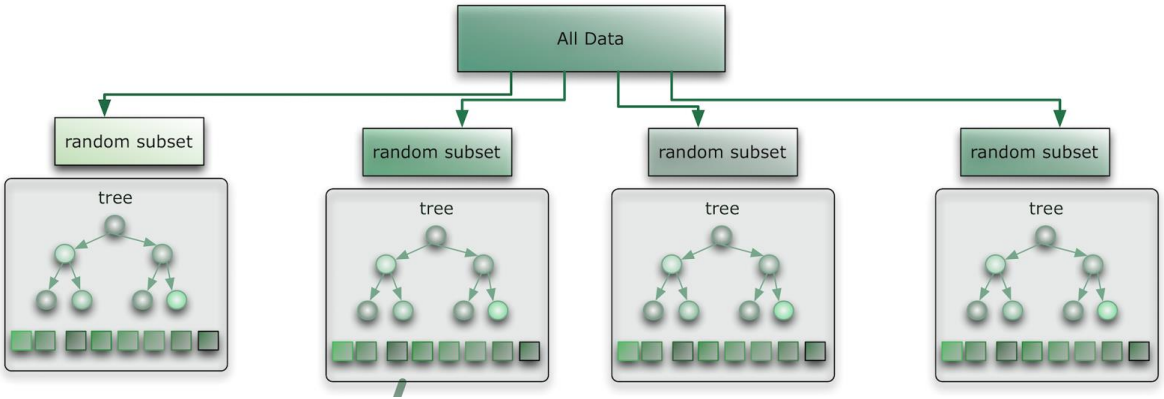
K-FOLD CROSS-VALIDATION



- Next is to **test the number of K folds** by testing different values and looking at the variance of the prediction error
- We will then use **K with the least number of misclassified observations** since this is a qualitative classification problem

MODEL SELECTION

RANDOM FOREST



- Classification algorithm
- Selecting the mode among all predictions
- Need to discretize our continuous features by turning them into dummy features

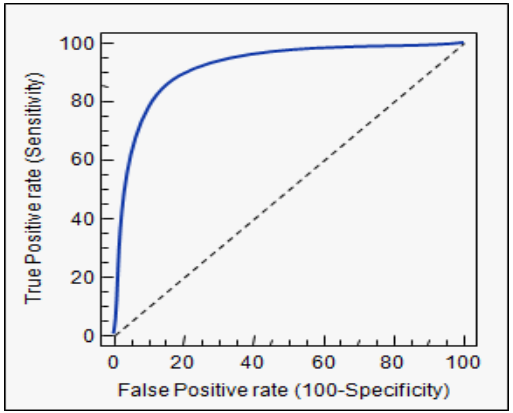
LOGISTIC REGRESSION

$$P('Left' = 1|(\Theta_n) = \Theta_1 x_1 + \Theta_2 x_2 + \ \ldots + \Theta_n x_n$$

- Dependent variable “left” is discrete
- We broke down the categorical feature “sales” into several dummy variables

MODEL EVALUATION

ROC CURVE

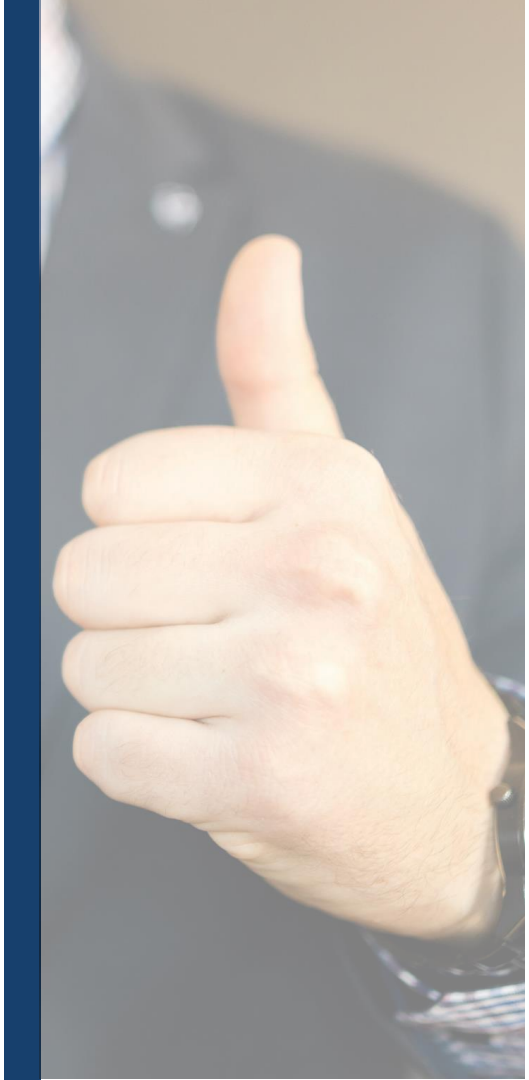


CONFUSION MATRIX

	Predicted Left	**Predicted Stayed**
Left	TP	FN
Stayed	FP	TN

TESTS APPLIED:

- **Random Forest:** The decision tree algorithm will perform feature selection by itself based on an entropy-based purity measure. The least pure features will be removed by the classifier.
- **Logistic Regression:** We will evaluate the overall performance of the model using the pseudo r-squared. We will also perform feature selection using the Wald test. Any feature with a weighted p-value of more than 5% will be discarded from the model.



Any questions?

You can find us at:

- datateam@pyous.com