Salifort Motors

Project Report

Providing data-driven suggestions to HR

1. Project overview

This project aims to provide insights for the Salifort Motors HR department. They want to take some initiatives to improve employee satisfaction levels at the company. They collected data from employees, and the goal is to analyze it and build a model that predicts whether an employee will leave the company.

If employees who are likely to quit can be predicted, it might be possible to identify factors that contribute to their leaving. Because it is time-consuming and expensive to find, interview, and hire new employees, increasing employee retention will benefit the company.

Key research question: what's likely to make the employee leave the company?

2. Data

2.1 Data description

The dataset has 15,000 rows and ten columns (listed below).

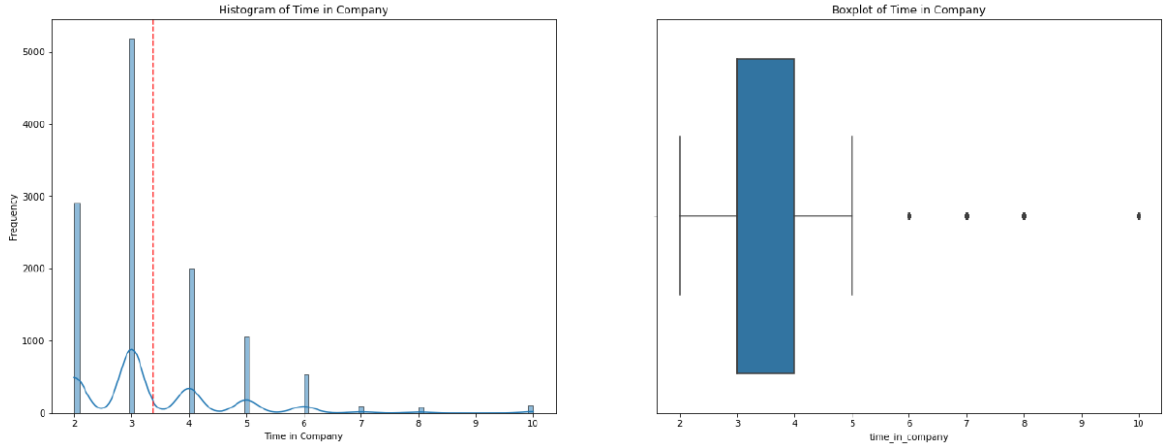
Variable	Description
satisfaction_level	Employee-reported job satisfaction level [0–1]
last_evaluation	Score of employee's last performance review [0–1]
number_project	Number of projects employee contributes to
average_monthly_hours	Average number of hours the employee worked per month
time_spend_company	How long the employee has been with the company (years)
Work_accident	Whether or not the employee experienced an accident while at work
left	Whether or not the employee left the company
promotion_last_5years	Whether or not the employee was promoted in the last 5 years
Department	The employee's department
salary	The employee's salary (U.S. dollars)

2.2 Data cleaning and structuring

- All columns have 14,999 no-null values
- We performed descriptive statistics on all eight numeric columns
- Seven out of the ten column names were changed
- We checked for duplicates: 3008 rows
- We dropped the duplicates; the dataset now is 11991x10
- The histogram of time_in _company is right-skewed, the mean is 3.36
- Checked for outiliers in time_in_company, there are 824 rows
- The histogram of satisfaction_level is left-skewed, the mean is 0.63
- There are no outliers in satisfaction_level

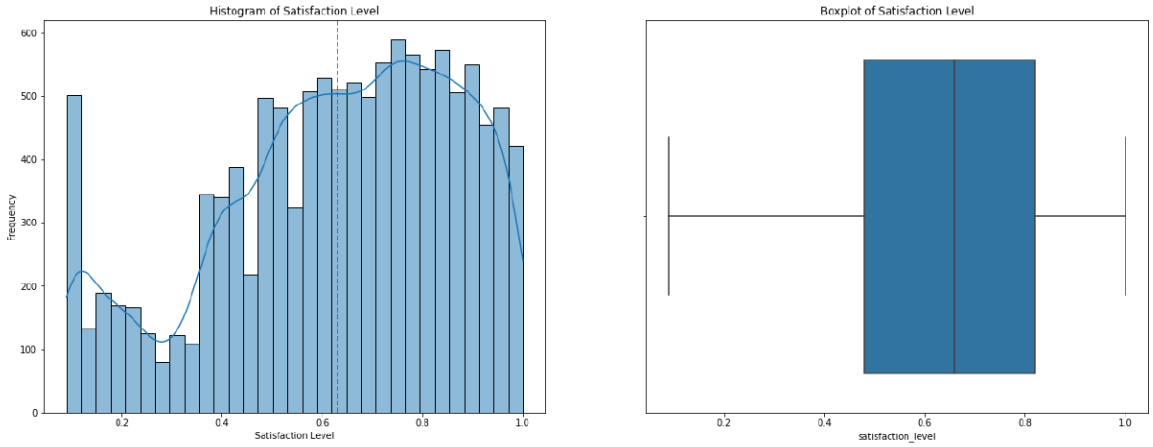
Based on the histograms and boxplots:

For time_in_company, the shape of the distribution suggests that a larger proportion of the data is concentrated at lower values, with fewer observations at higher values. This tells us most of the employees are located on the lower side of time spent working for the company.



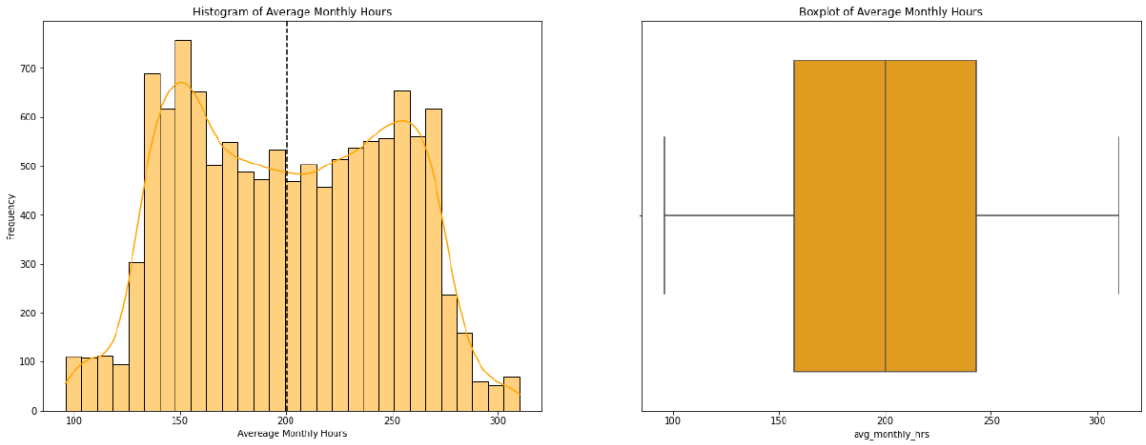
Histogram and boxplot for time_in_company.

The distribution of satisfaction_level works to the contrary of time_in_company, a more significant proportion of the data is concentrated at higher values, with fewer observations at lower values. From this, we can take away that the satisfaction is mostly positive.



Histogram and boxplot for satisfaction_level.

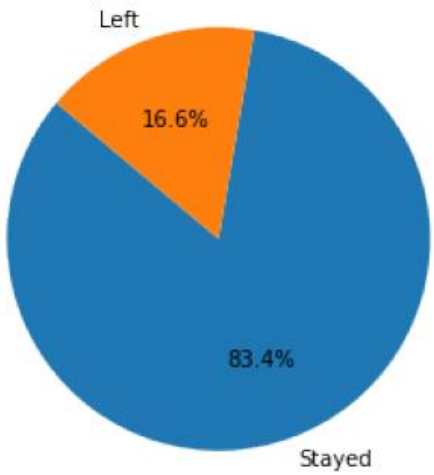
The histogram for avg_monthly_hrs appears to be approximately normally distributed. This means the data is clustered around a central value, with the frequency gradually decreasing as you move away from the center in both directions.



Histogram and boxplot foravg_monthly_hrs.

2.3 Data Exploration

From the 11,991 employees left in the dataset (after removing duplicates): 10,000 (83.4%) stayed and 1,991(16.6%) left.



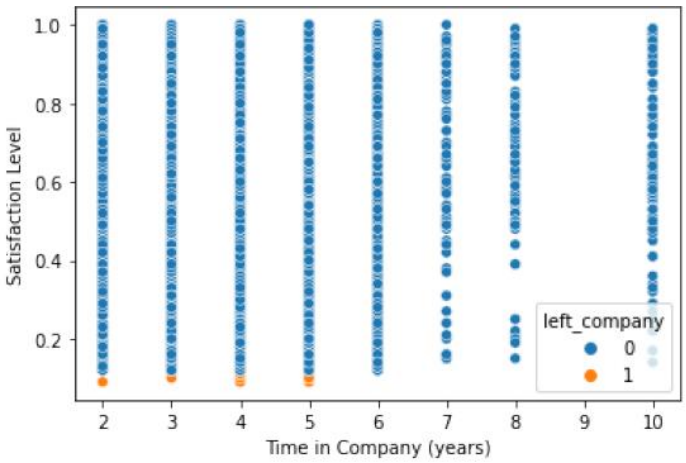
Employee turnover.

The initial intuition was that satisfaction_level could give us some insights into the company turnover, so we compared this variable against some others in various charts, always making the distinction between employees who stayed or left.

- satisfaction_level versus time_in company

  – The scatter plot shows that employees who left were on the 2–5-year mark and had the lowest satisfaction levels of said years.

  – Apart from that, the employees who stayed are present on all satisfaction levels, it looks constant from year 2-6, from year 7 and 8 there are fewer values, none for 9-year employees and again constant satisfaction levels for 10-year employees.

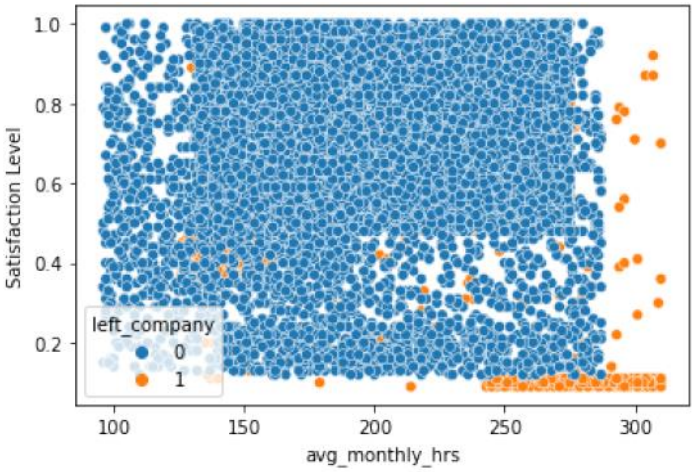


Scatterplot for satisfaction vs time_in_company.

- satisfaction_level versus avg_monthly_hrs

  –The scatter plot shows that the most employees who left had lower satisfaction levels and the most monthly hours (between ~250 and ~300).

  –Also, even though some of the employees who left had the most hours, there are some higher satisfaction levels present.

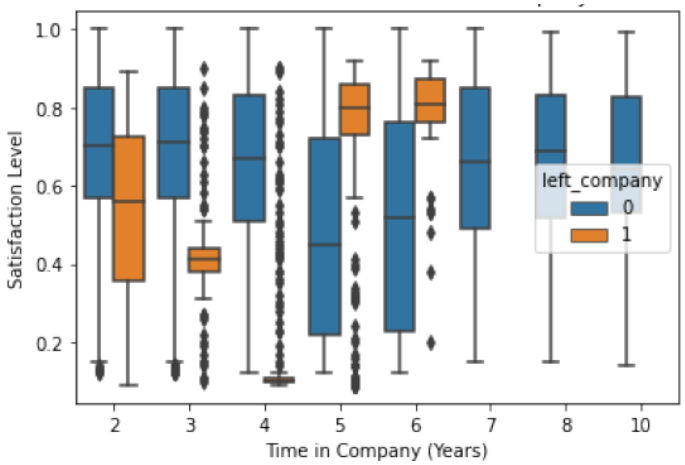


Scatterplot for satisfaction_level versus avg_monthly_hrs.

- satisfaction_level versus salary
  - Of the three salary categories, the employees who left were the ones with the lowest satisfaction scores.
- satisfaction_level versus num_projects
  - For 2-6 projects, the employees that left were the ones with the lowest satisfaction levels.
  - For the employees with 7 projects, all left, no matter the satisfaction level.
- satisfaction_level versus promotion
  - For the employees that didn't get promoted, only the ones with the lowest satisfaction levels left.
  - For the employees that did get promoted, the ones with the lowest satisfaction levels also left, but it appears some of them with a higher satisfaction level left, too.
- satisfaction_level and time_in_company

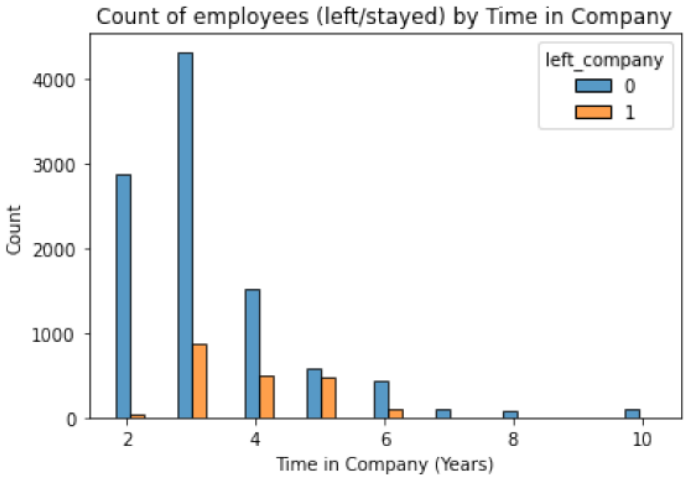
  – The most people who left stayed at the company for a little over 2 years, with a satisfaction level almost reaching the mean of 0.63.

- There are some employees who left that had 5 and 6 years at the company with a higher satisfaction level (~0.8)
- On the four-year mark there is a lower satisfaction level for the employees who left.
- From the employees that stayed: for the first 3 years, the boxes don't change much, from years 5 and 6, the values drop. Years 7, 8, and 9 show an increase in satisfaction.
- From employees who left: year 4 has 50% of the values with less than 0.2 in satisfaction. Employees in the 5- and 6-year mark have higher satisfaction levels.



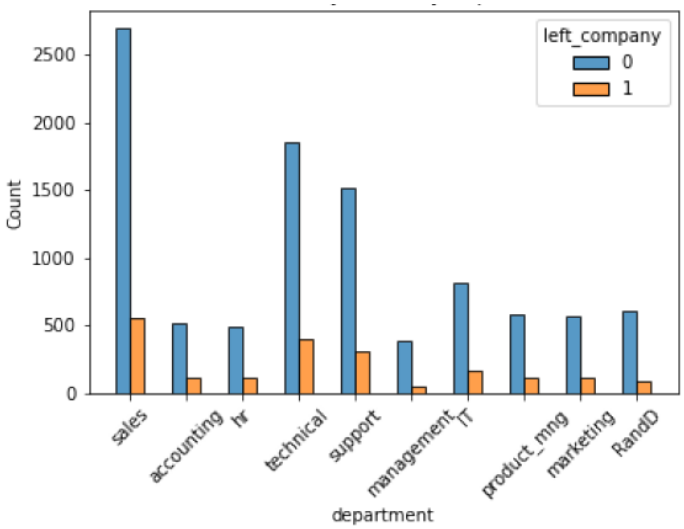
Satisfaction_level versus time_in_company.

- Employees (left/stayed) by Time in Company
  - Year 3 is the one with the largest number of employees who left.



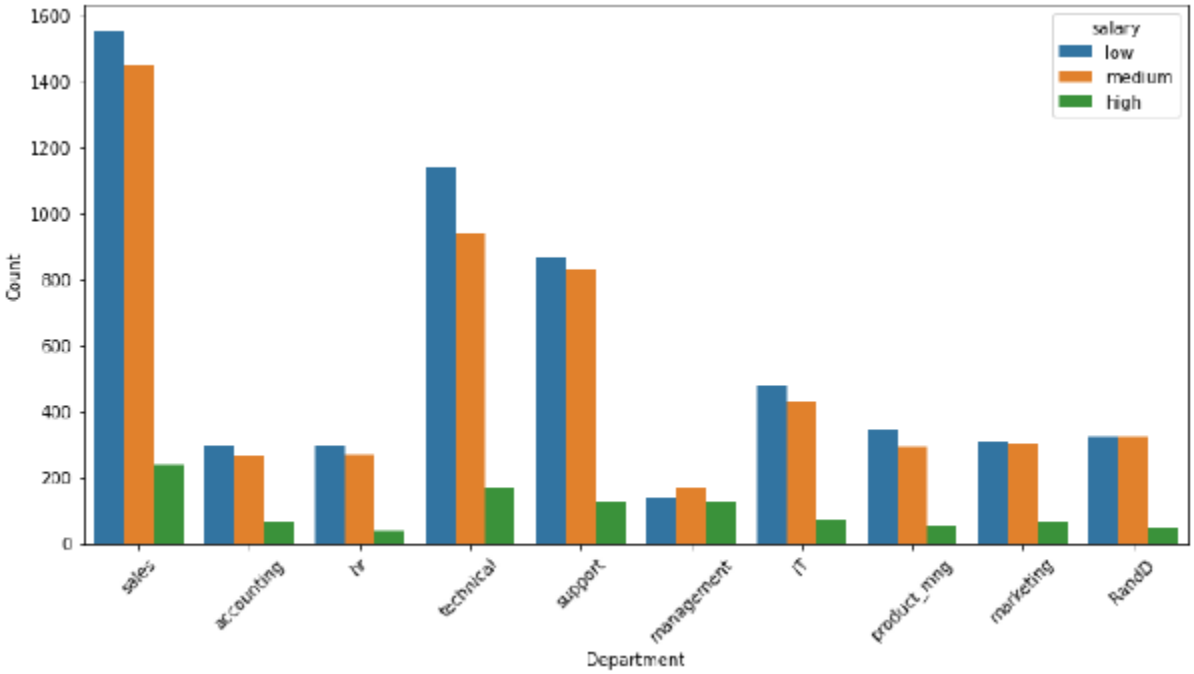
Count of employees (left/stayed) by time_in_company.

- monthly_hours vs salary
  - No matter the salary category, the employees who left had, on average, more monthly hours.
- num_projects vs salary
  - No matter the salary, the employees who left had more projects.
- Employees that stayed/left by salary
  - More employees with lower salaries left.
  - 20% of people with low salaries left.
  - 15% of people with medium salaries left.
  - 5% of people with high salaries left.
- Employees that stayed/left by department
  - The departments with more employees who left were sales, technical, and support.



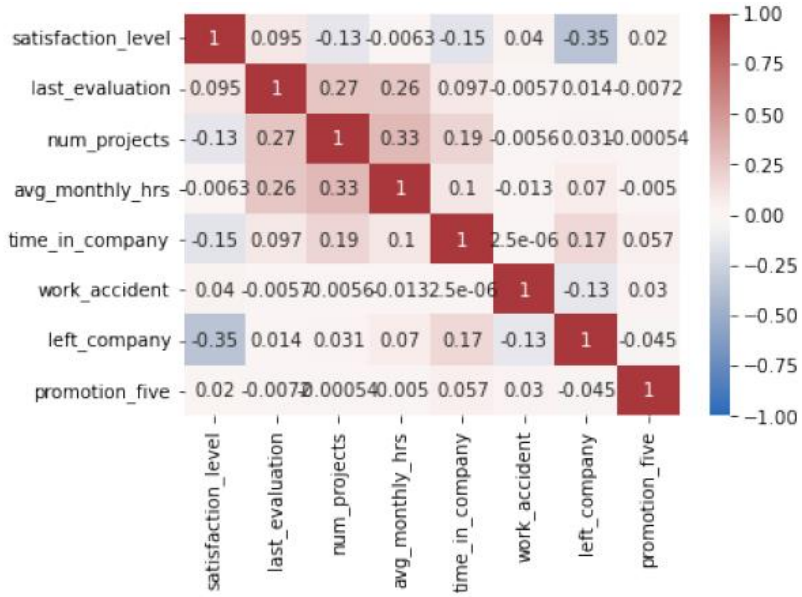
Counts of stayed/left by department.

- Salary vs department
  - The biggest departments are sales, technical, and support (that might be the reason those departments show more people who left than the others).
  - Most of their salaries are low and medium.



Employee count by department and salary.

- Correlation heatmap
  Higher correlation between:
  - last_evaluation and num_projects.
    The people with more projects left the most, even with a higher evaluation.
  - last_evaluation and avg_monthly_hrs
    For people who left: less average monthly hours, less evaluation score or more hours, higher evaluation score.
  - num_projects and avg_montly_hrs
    The more projects, the more employees left.
    Higher hours and a larger number of projects meant that more employees left.



Correlation heatmap.

3. Logistic Regression Model

What's likely to make the employee leave the company? The model should predict whether an employee will leave the company based on: job title, department, number of projects, and average monthly hours. A good model will help the company increase retention and job satisfaction for current employees and save money and time training new employees.

After re-checking variable correlations and dropping outliers, we encoded the categorical variables department (one-hot encoding) and salary (label encoding).

The outcome variable was left_company and the rest of the variables were the independent variables.

3.1 Model performance evaluation and metrics

- Precision

Predicted would not leave: 86% of the instances predicted as "would not leave" were correct.

Predicted would leave: 45% of the instances predicted as "would leave" were correct.

- Recall

Predicted would not leave: 93% of the actual "would not leave" instances were correctly predicted.

Predicted would leave: 26% of the actual "would leave" instances were correctly predicted.

- Accuracy

Overall, the model correctly predicted 82% of the instances.

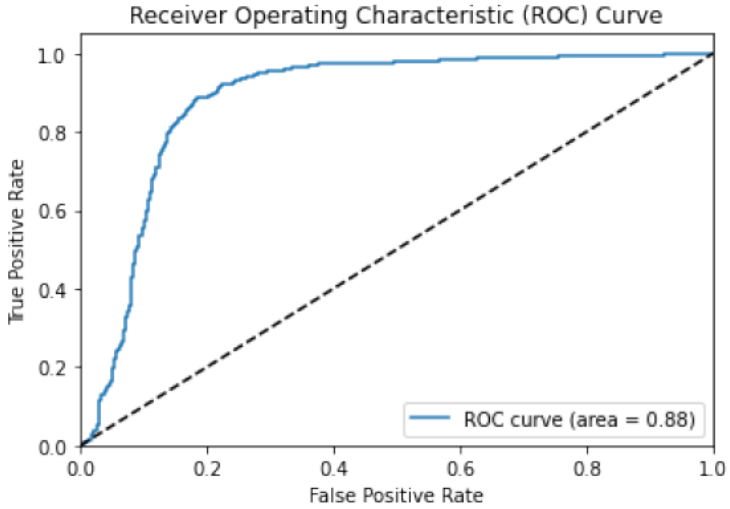
- F1-Score

Predicted would not leave: 90% Predicted would leave: 33%

- Beta coefficients
  - More satisfaction, less turnover (one-unit increase in *satisfaction_level*, the odds of an employee leaving decrease by approximately 98.7%)

- Higher evaluation, less turnover (not as much as satisfaction, but for a one-unit increase in *last_evaluation*, the odds of an employee leaving decrease by approximately 0.154%
- More projects, less turnover (one-unit increase in num_projects, the odds of an employee leaving decrease by approximately 38.527%)
- More hours in a month, more turnover (for each additional hour worked per month, the odds of an employee leaving increase by approximately 0.36%)
- More years in the company, more turnover (for each additional year of tenure in the company, the odds of an employee leaving increase by approximately 301%)
- Getting a promotion, less turover (employees who have received a promotion in the past five years are approximately 69% less likely to leave the company compared to those who haven't)
- The departments with higher odds of turnover are: sales (3%), support (4%), and technical (4%)
- More salary, less turnover (a one-unit increase in the salary level means the odds of an employee leaving the company decrease by approximately 41%)

- ROC curve



Receiver Operating Characteristic Curve

4. Conclusions, recommendations, next steps

4.1 Conclusions

The model has a precision of 86% for "predicted would not leave" and 45% for "predicted would leave". Recall had a 93% for "predicted would not leave" and 26% for "predicted would leave". For accuracy, the model correctly predicted 82% of the instances. For the F1-Score, the model predicted 90% for "would not leave" and 33% for "predicted would leave".

Class Imbalance: the model seems to be biased towards the "would not leave" class, as evidenced by the higher precision and recall for this class. This could be due to an imbalance in the training data. Consider techniques like oversampling, undersampling, or class weighting to address this issue.

Low Recall for "Would Leave": The low recall for the "would leave" class indicates that the model struggles to identify actual "would leave" instances. This might be due to a lack of informative features or insufficient training data for this class. Consider feature engineering or collecting more data to improve performance.

The ROC curve shows that this is a good model, given that the curve is closer to the top-left corner (the model is better at classifying the data).

- More hours/month, more projects, and having a promotion in the last five years meant more people leaving.
- On the four-year mark there seems to be a lower satisfaction level for the employees who left.
- No matter the salary category, the employees who left had, on average, more monthly hours.
- The people with more projects left the most, even with a higher evaluation.
- For people who left: less average monthly hours = less evaluation score or more hours = higher evaluation score.
- The more projects, the more employees left.
- Higher hours and a larger number of projects meant more employees left.

4.2 Recommendations

- We suggest having a limited amount of projects an employee can have. To reduce burnout and increase productivity
- There should also be discussions about working extra time. Either with already enrolled employees (so that no one works more than they physically/mentally could) and consider the legal amount of hours a person can work in France (research shows is a max of 10 hours/day, some sources say 35 hours/week). This information should also be provided to applicants as they engage in the interview process.
- There should also be fair compensation for working more hours.
- Investigate the reason for having a so-low satisfaction level of employees that left at the four-year mark.
- The higher evaluation scores should go to the employees who deserve them and not only to employees who worked more hours.
- Make sure the employees know that the company is set on having a good work environment, schedule meetings so the topic can be discussed, and find the issues the employees have with the company's culture, so the satisfaction can increase.

4.3 Next steps

- Share the results of the model to see if it worked as expected.
- As for the model, there are some adjustments that should be made (as mentioned in the conclusion section).
- There should also be a follow-up to the recommendations and their results.