Salifort Motors: Employee Turnover Prediction Project

PACE Strategy Document - Execute stage

Google Advanced Data Analytics Capstone

Introduction

I will use this PACE strategy document to record my decisions and reflections as I move into the Execute stage of this capstone project. This document will guide my efforts to finalize model selection, interpret results, and communicate insights to stakeholders through clear visualizations and summaries. It will help ensure that my findings are actionable and aligned with business objectives. Additionally, this document will serve as a valuable reference for evaluating the overall impact of my analysis and for supporting my continued growth as a data professional.

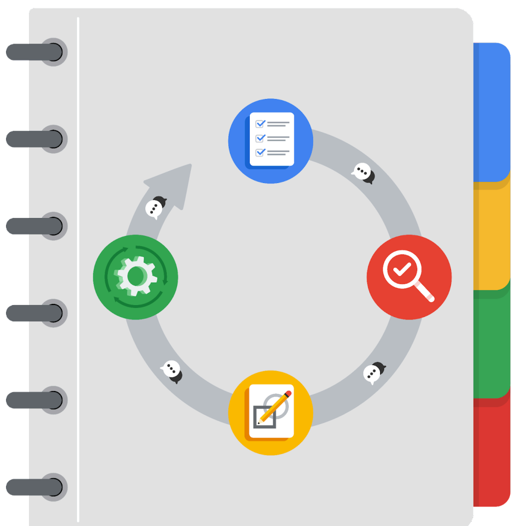
Portfolio Project Recap

Many of the goals I accomplished in my individual course portfolio projects are integrated into this Advanced Data Analytics capstone project. These include:

- Creating a clear and structured project proposal
- Demonstrating my understanding of Python's form and function
- Using Python to load, explore, extract, and organize information through custom functions
- Organizing and analyzing a dataset to uncover meaningful insights and tell the underlying "story"
- Developing a Jupyter notebook for exploratory data analysis (EDA)
- Computing descriptive statistics.
- Evaluating the model to assess its performance
- Applying machine learning techniques in a notebook environment to solve a defined problem
- Communicating results effectively by summarizing findings in an executive summary for external stakeholders

This capstone brings together all the skills I've developed across the program, allowing me to apply them in a cohesive, real-world project.

THE PACE WORKFLOW



[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

I will demonstrate to the company's HR team how I would apply the PACE workflow to the upcoming Salifort Motors project. For each question presented in this PACE strategy document, I will provide a structured and thoughtful response aligned with the corresponding stage of the PACE framework — Plan, Analyze, Construct, and Execute. This approach ensures clarity in my methodology, highlights my strategic planning skills, and illustrates a well-organized path from problem understanding to actionable insights.

Project Tasks

The following questions have been identified as essential to the Employee Turnover Prediction Project. These questions are primarily situated within the Execute stage of the PACE workflow, focusing on the interpretation of results, communication of insights, and evaluation of the model's impact. I will address each question by aligning it with the most appropriate phase of the PACE framework — Plan, Analyze, Construct, or Execute — based on the specific project stage it reflects. To ensure each response is informed and well-reasoned, I will draw on relevant materials from the project notebook, the PACE framework, and best practices in communicating data-driven findings to stakeholders.

Data Project Questions & Considerations

PACE: Execute Stage

Get Started with Python

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing an exploratory data analysis?

Given my initial understanding of the dataset, I recommend investigating two key areas before conducting a full exploratory data analysis:

- Organizational Context:

I would ask the manager to provide recent context about the company or their specific team — for example, whether employees have appeared engaged, stressed, or disengaged during recent interactions. While this qualitative insight may not offer a comprehensive or unbiased view of the entire workforce, it can serve as a useful starting point to shape initial perspectives. Although my analysis will be grounded in the data, understanding the manager's observations may help guide more thoughtful and targeted hypotheses early in the process.

- Data Timing and Potential Leakage:

I also recommend confirming when key variables — such as satisfaction_level, last_evaluation, and average_monthly_hours — were last updated. Variables like satisfaction_level and last_evaluation could introduce data leakage if they reflect information collected after an employee had already decided to leave (e.g., after submitting their notice) or had been flagged for termination. Verifying the timing of these fields will help ensure the model is trained only on data that would have been available at the time of prediction.

Together, these investigations will help validate the data's integrity and support a more strategic, informed approach to the exploratory analysis.

- What data initially presents as containing anomalies?

Initial data exploration revealed two key areas of potential anomalies:

- Outliers in tenure:

The tenure variable (time spent at the company) exhibited a right-skewed distribution, with a long tail of employees who had unusually high tenure values. To assess these outliers, both conservative (median + 1.5 × IQR) and standard (Q3 + 1.5 × IQR) methods were applied.

- The conservative method identified 1,886 outliers, reflecting a stricter threshold suitable for skewed data.

- The standard IQR method identified 824 outliers, offering a more lenient cutoff.

Ultimately, the standard method was retained for modeling, as it preserved more data while still flagging extreme cases.

- Duplicate Records:

A total of 3,008 duplicate rows were detected — identical across all columns. After careful inspection, these were determined to be unlikely to represent valid, repeated employee records due to the number of continuous variables involved. Given the low probability of all values matching across multiple continuous fields by chance, these duplicates were removed to maintain data integrity.

These anomalies were addressed to ensure cleaner, more reliable data for the modeling phase and to prevent misleading signals during analysis.

- What additional types of data could strengthen this dataset?

While the current HR dataset provides a strong foundation for analyzing employee turnover and satisfaction, several additional data types could further enhance the analysis:

- Demographic Information:

In an ideal scenario, adding employee demographics — such as age, gender, education level, and marital status — could provide deeper insights into trends across different employee segments. However, these variables should be used cautiously to avoid introducing bias into the model. It's important to assess them critically and remove any that may lead to unintended or unethical patterns during prediction.

- Workplace Experience and Engagement:

Data reflecting the employee experience, such as exit interview feedback, participation in social or engagement activities, or employee conduct records, would offer a more holistic view of job satisfaction and retention risks. These insights could help explain behavioral drivers behind attrition beyond what quantitative metrics reveal.

- Behavioral or Performance Indicators:

Information on peer or manager evaluations, internal surveys, or training participation might help highlight trends related to performance, growth opportunities, or burnout — all of which can impact turnover.

If such data were available, it would need to be properly validated and joined with the existing dataset. That includes ensuring structural consistency and applying ongoing validation throughout the analysis process.

As for the current dataset:

- No missing values were found.

- 3,008 duplicate records were identified and removed.

- Outlier detection on tenure was performed using both conservative and standard IQR methods. The lenient method was ultimately chosen to retain more data.

- Field names were reviewed and updated for clarity.

With these steps completed, the dataset is well-prepared for analysis — and would be further strengthened by integrating these additional data types in a thoughtful and ethical way.

<u>Go Beyond the Numbers: Translate Data into Insights</u>

- What key insights emerged from your EDA and visualizations(s)?

Key Insights from EDA and Visualizations

- Class Imbalance:

The target variable left exhibits a significant class imbalance, with approximately 83% of employees staying and only 17% leaving. This imbalance could affect the performance of predictive models, especially in terms of precision and recall for the minority class (leavers).

- Workload and Attrition:

Attrition was observed at both extremes of workload: employees who were either underworked or overworked were more likely to leave. This suggests that workload balance plays a critical role in employee retention.

Employees working 7 projects and around 255–295 hours/month tended to fall within the interquartile range for monthly hours, indicating a more optimal balance between workload and retention.

A spike in attrition was observed around the 3-year tenure mark, indicating that mid-tenure employees might be at a higher risk of leaving, possibly due to stagnation or unmet expectations.

- Satisfaction and Attrition:

Satisfaction level is a strong indicator of employee retention. Employees with low satisfaction and low hours worked were most likely to leave, as were those with medium tenure but high satisfaction, indicating a possible mismatch between expectations and career growth.

A scatterplot of satisfaction vs. monthly hours revealed three distinct clusters:

- Low satisfaction, low hours — leading to attrition.
- High hours, mixed satisfaction — indicating burnout and potential disengagement.
- High satisfaction, moderate hours — associated with lower attrition.

- Burnout Risk:

Employees with high monthly hours and high evaluation scores may indicate burnout rather than strong performance, suggesting that long hours without corresponding rewards or promotions lead to disengagement and turnover.

Monthly hours vs. promotions showed that employees with high monthly hours often did not receive promotions, indicating potential lack of recognition as a contributing factor to attrition.

- Department-Specific Attrition:

HR had the highest proportion of employees leaving, while management had the lowest. This department-based analysis can inform targeted retention strategies and intervention programs.

- Salary vs. Tenure:

Short-tenured employees were more likely to have low or medium salaries, while longer-tenured employees tended to maintain medium salaries throughout. This suggests that salary growth may be limited over time, which could contribute to attrition as employees seek better financial opportunities.

- Correlation Insights:

There was a positive correlation among variables like number of projects, monthly hours, and last evaluation, indicating that these factors often increase together. However, a negative association was found between satisfaction level and attrition, reinforcing the idea that satisfied employees are less likely to leave.

- Employee Satisfaction and Performance:

The combination of low satisfaction and high workload emerges as a strong indicator of burnout and potential attrition, underscoring the importance of balancing employee workload with adequate rewards and recognition.

- Monthly Hours Anomaly:

The average monthly hours of 201.05 hours per month was notably higher than the expected 166.67 hours, suggesting that many employees are working beyond the typical full-time schedule, which likely contributes to attrition risk.

These insights, drawn from the EDA and various visualizations, emphasize the critical role of workload balance, employee satisfaction, and recognition in determining turnover. They also highlight areas for potential intervention, such as reducing overwork, improving satisfaction, and ensuring that high performers are properly recognized and rewarded.

- What business recommendations do you propose based on the visualization(s) built?

Business Recommendations Based on Visualizations

- Cap the Number of Projects per Employee:

The visualizations clearly indicate that employees who are overworked (particularly those managing a higher number of projects and working excessive monthly hours) are at a significantly higher risk of attrition. Limiting the number of projects employees can handle — to a maximum of four projects — would help balance their workload, reduce burnout, and likely lower turnover rates.

- Reevaluate Employee Promotion Criteria:

The analysis of tenure and satisfaction levels reveals that employees who stay with the company for at least four years experience notable dissatisfaction. To improve retention, HR should either consider promoting employees after four years of service or conduct further investigations to understand why this group is dissatisfied. By offering more growth opportunities, employees may feel more valued and motivated to stay.

- Reassess Work Hours and Overtime Policies:

Many employees are working well beyond the expected full-time hours, contributing to burnout. To address this, the company should either reward employees for working longer hours or reduce the expectation for overtime. Additionally, it's crucial that employees are aware of the company's overtime pay policies if they aren't already. Clear communication around workload expectations, time off, and overtime pay can alleviate stress and contribute to a healthier work environment.

- Clarify Workload Expectations and Time Off:

To ensure alignment between management and employees, the company should make workload expectations and time-off policies explicit. This includes ensuring that employees understand the company's overtime policies and feel comfortable discussing workload issues without fear of reprisal. This transparency will reduce misunderstandings and promote a healthier work-life balance.

- Promote Open Discussions Around Company Culture:

The visualizations suggest that there may be an underlying issue with the company's work culture contributing to employee dissatisfaction and turnover. It's essential to hold company-wide and within-team discussions to better understand and address specific cultural or workload-related issues. Regular dialogues can help foster a more inclusive and supportive environment, improving employee morale and retention.

- Rewarding High Performance Beyond Workload:

The analysis shows that high evaluation scores are often tied to employees who work excessively long hours, yet this may not always indicate true performance. To incentivize sustainable performance, the company should consider a more proportionate scale for rewarding employees based on performance, rather than just the number of hours worked. This will prevent burnout and promote long-term engagement and loyalty.

- Incorporate Satisfaction Levels into Performance Reviews:

Satisfaction levels were identified as a strong indicator of employee attrition. HR should incorporate satisfaction surveys into regular performance reviews to identify early signs of disengagement, particularly among high-performing employees. Addressing dissatisfaction early allows the company to take proactive steps to improve morale before employees decide to leave.

- Balance Workload and Motivation:

The visualizations suggest that employees with high satisfaction and moderate workloads tend to stay longer, while those with imbalanced workloads (either too low or too high) are more likely to leave.

By regularly monitoring both workload and employee motivation, the company can uncover hidden retention risks and take corrective actions before they escalate.

- Track Burnout Indicators:

Burnout is a critical factor contributing to attrition, as revealed by the analysis of satisfaction vs. monthly hours. HR teams should incorporate burnout indicators (such as high monthly hours and low satisfaction) into performance reviews and employee feedback systems. This proactive approach will allow HR to intervene early and provide necessary support, reducing the likelihood of burnout and associated turnover.

- Model-Based Recommendations:

Based on the performance of the Random Forest model, I recommend deploying this top-performing model to enhance proactive retention strategies. This model has proven to be effective in predicting employee attrition and identifying at-risk employees.

Key Model Metrics:

- Recall: 90.36% — ensuring that most employees at risk of leaving are identified, minimizing false negatives.
- Precision: 87.04% — balancing the identification of leavers while avoiding false positives.
- Accuracy: 96.16% — overall effectiveness in classifying employees accurately.

Deploying the Random Forest model will allow HR to prioritize interventions more efficiently. The model can be integrated into the existing HR workflow to support targeted outreach and root cause analysis of potential resignations. Over time, this approach can significantly improve retention rates and support more informed workforce planning.

By leveraging this data-driven approach, the company can make more informed, proactive decisions that benefit both employees and the business as a whole.

These recommendations are designed to improve employee satisfaction, reduce burnout, and ultimately lower turnover by addressing workload imbalances, enhancing communication, and using data-driven insights to guide HR decisions.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

Additional Questions for Further Research

- Is salary allocation fair relative to experience and tenure?

Exploring whether compensation is equitably distributed across different levels of experience and length of service can help identify potential disparities and inform more transparent salary policies.

- Are performance evaluations being conducted fairly?

Analyzing how evaluation scores vary across departments, managers, or employee demographics can reveal whether the evaluation process is consistent and unbiased.

- Is there a relationship between employee satisfaction and evaluation scores?

Investigating whether more satisfied employees tend to receive higher evaluations could help understand if satisfaction impacts perceived performance or if performance drives satisfaction.

These questions could guide further analysis and support strategic decision-making aimed at improving employee engagement, retention, and fairness.

- How might you share these visualizations with different audiences?

To effectively communicate the visualizations to different audiences, I would adjust the complexity of the visuals based on the technical knowledge of each group. Here's how I would approach it:

For Technical Stakeholders:

- Boxplots: Useful for analyzing distributions, skewness, and identifying outliers in variables like satisfaction level and evaluation scores.

- Correlation Heatmaps: Visualize relationships between variables and identify multicollinearity, aiding in understanding interdependencies.

- Bar Charts for Feature Importance: Display the ranking of features based on their importance from decision trees and random forests.

- Line Plots: Analyze the linearity of the logit in logistic regression models, helping to evaluate model assumptions and AUC-PR.

- Decision Tree Diagrams: Show how decision trees split data, providing interpretability of model logic.

- Confusion Matrices: Assess classification model performance, highlighting true positives, false positives, true negatives, and false negatives.

For Non-Technical Stakeholders:

- Histograms: Show the distribution of key variables like satisfaction level, evaluation scores, and monthly hours, making data easy to interpret.

- Grouped Histograms: Compare attrition rates across salary levels, departments, or tenure groups to spot trends.

- Stacked Column Charts: Visualize proportions of employees who stayed or left within departments or job roles, simplifying turnover patterns.

- Grouped Column Charts (Department vs. Employee Count): Display the distribution of employees who stayed versus those who left across departments for clear, comparative insights.

For Both Audiences:

- Scatterplots: These show relationships between continuous variables (e.g., satisfaction level vs. monthly hours). For technical stakeholders, scatterplots can be used for deeper analysis, while for non-technical audiences, they should be explained more intuitively. When presented with context and a narrative, scatterplots can be a helpful tool for both groups to understand correlations and patterns.

By tailoring the visualizations to the specific audience, I can ensure effective communication of the findings while making the insights accessible to both technical and non-technical stakeholders.

The Power of Statistics

- Did any part of the project involve or require A/B testing, or would it align with the overall goals of this project? If so, how would I design it, and what key business insights could potentially be gained from it?

No part of my project involved A/B testing, nor was it required to meet the goals of this analysis. A/B testing wasn't part of any current or previous phase of the project, and it doesn't align with the core objective — which was to build a predictive model for employee attrition.

That said, A/B testing could be considered in the future if the company decides to test specific interventions inspired by the model's insights. For example, if HR wants to assess the impact of capping employees at 4 projects, an A/B test could be set up:

- Group A (Control): No changes — employees continue under current conditions.

- Group B (Treatment): Limit employees to a maximum of 4 projects.

Metric: Compare attrition rates between the two groups over a 6-month period.

This kind of testing would fall outside the scope of my current project but could provide valuable insights into which actions are most effective at reducing turnover.

- What business recommendations do you propose based on your results?

Based on the visualizations and analysis, I propose the following business recommendations to improve employee retention and satisfaction:

- Cap the Number of Projects per Employee: Limiting the number of projects to four will help balance workloads, reduce burnout, and lower turnover rates.

- Reevaluate Employee Promotion Criteria: Consider promoting employees after four years of service or investigate why this group is dissatisfied to improve retention.

- Reassess Work Hours and Overtime Policies: Either reward employees for overtime or reduce expectations to avoid burnout, while ensuring employees understand the company's overtime policies.

- Clarify Workload Expectations and Time Off: Ensure clear communication about workload and time-off policies to promote a healthier work-life balance.

- Promote Open Discussions Around Company Culture: Regular team discussions to address cultural or workload issues can improve morale and retention.

- Reward High Performance Beyond Workload: Incentivize sustainable performance by rewarding results, not just hours worked.

- Incorporate Satisfaction Levels into Performance Reviews: Regularly measure satisfaction to identify disengagement early and take proactive steps.

- Balance Workload and Motivation: Monitor workload and motivation to prevent imbalance and hidden turnover risks.

- Track Burnout Indicators: Implement burnout indicators into performance reviews to intervene early and reduce turnover.

- Deploy the Random Forest Model: Integrating the top-performing Random Forest model into HR processes will enable targeted interventions and improve retention rates by identifying at-risk employees.

These recommendations can help the company create a more supportive, balanced, and proactive work environment, ultimately reducing attrition and improving retention.

Regression Analysis: Simplify Complex Data Relationships

- To interpret model results, why is it important to interpret the beta coefficients?

Interpreting beta coefficients is crucial because they provide insights into the relationships between the independent variables and the dependent variable in a regression model. Here's why:

Intercept and Slope Interpretation:

- The intercept (constant term) represents the predicted value of the dependent variable when all independent variables are zero.

- The slope (beta coefficient) represents the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant. This tells us the strength and direction of the relationship.

Model Insights:

- Beta coefficients allow us to interpret how each predictor variable influences the outcome. For instance, in a logistic regression model, a positive beta coefficient indicates that as the independent variable increases, the likelihood of the outcome increases. Conversely, a negative coefficient means the outcome is less likely as the predictor increases.

Significance:

- When paired with statistical tests, the beta coefficients help us understand whether a relationship is statistically significant. A non-zero coefficient typically indicates that the predictor has a meaningful impact on the dependent variable, assuming the p-value is below a certain threshold (e.g., 0.05).

Comparison Across Variables:

- By interpreting beta coefficients, we can compare the relative importance of different variables. A larger magnitude of a coefficient typically suggests a stronger effect on the dependent variable, providing a guide for prioritizing interventions or decisions.

In summary, beta coefficients are essential for understanding the direction, magnitude, and significance of relationships in the model, offering actionable insights beyond just model performance metrics like accuracy or R-squared.

- What potential recommendations would you make to your manager/company?

Here are the key recommendations to improve employee retention and reduce burnout: - Cap Projects per Employee: Limit employees to four projects to prevent overwork and burnout. - Promote After Four Years: Consider promoting employees after four years or investigate their dissatisfaction to improve retention. - Review Overtime Policies: Reward or reduce excessive overtime, and ensure employees are aware of overtime pay policies. - Clarify Workload Expectations: Make workload and time-off policies transparent and encourage open discussions. - Foster Open Culture Conversations: Hold regular discussions to identify and address cultural or workload-related issues. - Reward True Performance: Tie high evaluation scores to actual performance, not just hours worked. - Consider Satisfaction in Attrition Predictions: Include satisfaction levels in performance reviews to spot disengagement early. - Monitor Workload and Motivation: Regularly track employee workload and motivation to prevent burnout. - Include Burnout Indicators in Reviews: Use burnout and satisfaction data in performance reviews to enable early intervention. These recommendations aim to balance workloads, improve satisfaction, and reduce turnover.

- Do you think your model could be improved? Why or why not? How?

Yes, I believe the model could still be improved. While it performs well, there are several potential areas to enhance its performance further: - Increase Recall: Since recall is critical for identifying at-risk employees, improving this metric would help capture more leavers and reduce the likelihood of false negatives. - Feature Engineering: Additional or modified features could help better capture turnover patterns. By creating new features or refining existing ones, the model might find stronger signals and improve its predictive power.

- Investigate Data Leakage: While previous checks were made, revisiting potential data leakage is essential. Certain features like last_evaluation and satisfaction_level may reflect outcomes rather than causal factors. Evaluating model performance with and without these features could help understand their true influence.

- Alternative Modeling Focus: If features like last_evaluation are strong predictors of attrition, it may be worth reframing the problem. For instance, predicting evaluation or satisfaction scores themselves could be useful as leading indicators of turnover.

- Model Tuning:

  - Adjusting cross-validation parameters can ensure more robust hyperparameter optimization.

  - Resampling techniques (such as SMOTE or undersampling) could address class imbalance and improve model performance.

  - Class weights or threshold tuning can help optimize the model for specific business needs, such as minimizing false negatives and prioritizing retention efforts.

Despite these possible improvements, it's also possible that this model is close to achieving the best performance with the current dataset, given the thorough preprocessing and quality checks already performed. However, exploring these areas could still unlock valuable insights and performance improvements.

- What business recommendations do you propose based on the models built?

Based on the models I built, especially the top-performing Random Forest model, I recommend the following:

- Deploy the Random Forest Model: With high recall (90.36%), precision (87.04%), and accuracy (96.16%), this model is effective at identifying employees at risk of leaving. It’s reliable for supporting proactive retention strategies.

- Integrate the Model into HR Workflow: Use the model to flag at-risk employees early, allowing HR to prioritize interventions and investigate root causes of dissatisfaction before resignations occur.

- Use Predictions to Guide Actions: Model outputs can help inform more targeted and timely outreach efforts, improve retention strategies, and support better workforce planning.

Overall, the model enables data-driven decision-making and allows the company to act before attrition becomes a larger issue.

- What key insights emerged from your model(s)?

The key insights from the models are as follows:

- Logistic Regression Insights:

  - Good Performance for Stayers: The logistic regression model accurately identified employees who were likely to stay, with a strong performance for true negatives (stayers). However, it struggled with the minority class (leavers).

  - Challenges with Attrition Prediction: The model misclassified a significant number of leavers, with low recall (26%) and precision (44%) for predicting employees who would leave. This is problematic for the business, as the primary goal is to identify at-risk employees to prevent attrition.

- Decision Tree Insights (Round 1):

  - Better Performance on Attrition: The decision tree model outperformed logistic regression, especially in recall and F1-score for predicting leavers. This is crucial for detecting employee turnover.

  - Strong Model for Imbalanced Data: With a high ROC AUC (0.9698), the decision tree showed good handling of class imbalance. It also demonstrated robustness without significant overfitting.

  - Clear Interpretability: Decision trees provide easy-to-understand, interpretable results, making it easier for stakeholders to grasp how the model makes predictions.

- Random Forest Insights (Round 1):

  - Best Overall Performance: The random forest model was the highest performing across all evaluation metrics (ROC AUC: 0.9804), with great recall for both leavers and stayers. This makes it ideal for dealing with the class imbalance.

  - Effective Handling of Imbalance: By using ensembling techniques, random forest reduced overfitting and improved generalization. It handled the class imbalance better than logistic regression and decision trees, making it particularly valuable for attrition prediction.

- Post-Data Leakage Model Insights (Round 2):

  - Improved Model Performance: After addressing data leakage and refining features, the random forest model continued to outperform the decision tree model, with a higher precision-recall balance and reduced false negatives.

- Critical Role of False Negatives: The random forest model achieved excellent results in minimizing false negatives, correctly identifying 450 out of 498 leavers. This is crucial, as the business goal is to proactively identify employees at risk of leaving.

- Acceptable Trade-Offs in False Positives: While the model misclassified some stayers as leavers, this trade-off was acceptable for the business, as it allows for proactive intervention in at-risk employees.

Final Insight:

The Random Forest Round 2 model emerged as the champion model. It outperformed both logistic regression and decision trees, particularly in predicting leavers (attrition) while maintaining high accuracy and recall. This model will serve as the basis for deploying targeted retention interventions, supporting the business goal of improving employee retention by identifying at-risk employees early.

- Do you have any ethical considerations at this stage?

Yes, at this stage, I will focus on minimizing potential bias to ensure that the model produces fair and equitable predictions. My goal is to develop a baseline model that is technically robust and also aligned with the overall problem-solving objective. I'll pay close attention to selecting and evaluating the most relevant performance metrics for the scenario to minimize the risk of misclassification and unintended consequences. Throughout this phase, fairness, transparency, and reliability will remain core priorities.

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?

The key insights from the models revealed that while logistic regression performed well in identifying employees who would stay, it struggled with predicting attrition. Specifically, it had low recall and precision for the "leaver" class, misclassifying a significant number of employees who were likely to leave. Decision trees provided better performance for attrition prediction, with improved recall and F1 scores, but showed slight signs of overfitting. The model also handled class imbalance well, offering strong interpretability, making it easier to understand the decision-making process.

However, the Random Forest model emerged as the most effective model, outperforming both logistic regression and decision trees across all metrics. It showed the best overall performance in identifying leavers, with high recall, precision, and ROC AUC, while effectively managing class imbalance. After addressing potential data leakage in Round 2, the Random Forest model continued to perform exceptionally well, minimizing false negatives and correctly identifying most leavers, which aligns with the business goal of improving employee retention. This model is now the recommended choice for predicting attrition and supporting targeted employee retention interventions.

- What are the criteria for model selection?

The criteria for model selection in this analysis are primarily based on performance metrics and the business goal of predicting employee attrition. Several factors were considered to determine the best model:

- ROC AUC: This metric is important for assessing the model's ability to discriminate between employees likely to leave (leavers) and those likely to stay (stayers). Models were optimized for ROC AUC through hyperparameter tuning using GridSearchCV with a focus on maximizing this metric, as it reflects overall model performance in imbalanced datasets.

- Recall: Since the business objective is to predict employee attrition effectively, recall is the most crucial metric. It indicates the model's ability to correctly identify employees who will leave, minimizing false negatives (leavers misclassified as stayers). Given the importance of not missing potential leavers, high recall was prioritized.

- Precision, F1 Score, and Accuracy: While precision and F1 score were also considered, they were secondary to recall in this case. A balance between precision and recall ensures that the model not only correctly identifies leavers but also minimizes the number of false positives. Accuracy was considered but was not the primary selection criterion due to the class imbalance in the dataset.

- Overfitting Risk: Models were evaluated for overfitting, especially with the decision tree model. Random Forest, being an ensemble method, showed a stronger ability to generalize, which led to its selection over the decision tree model.

- Class Imbalance Handling: All models were evaluated for their effectiveness in managing the class imbalance between stayers and leavers. Random Forests handled this well, maintaining strong performance across both classes, making it the top choice.

The Random Forest model (Round 2) emerged as the champion due to its superior performance across these criteria, especially in identifying potential leavers with high recall and robust performance metrics.

- Does my model make sense? Are my final results acceptable?

Yes, the model makes sense, and the final results are acceptable, especially given the context of employee attrition prediction. The Random Forest model, after addressing potential data leakage and performing rigorous hyperparameter tuning, outperformed both logistic regression and decision tree models in identifying employees likely to leave. This aligns well with the primary business goal of improving employee retention by focusing on attrition prediction.

The Random Forest model demonstrated high accuracy, strong recall, and a good balance between precision and recall, particularly for the "leaver" class. The low number of false negatives (employees misclassified as stayers) is crucial, as missing potential leavers can have significant consequences for retention efforts. Furthermore, the ensemble nature of Random Forest helped reduce overfitting while effectively handling the class imbalance in the dataset. The results from both the training and test sets suggest the model is robust and generalizable, making it suitable for deployment in real-world scenarios.

Therefore, the final results are acceptable, and the model can be confidently used for supporting targeted retention interventions.

- Were there any features that were not important at all? What if you take them out?

Based on the analysis above, it seems that tenure is one feature that appears to have very weak correlations with most other variables. Specifically, its correlation with other features like satisfaction level (-0.15) and number of projects (-0.13) is quite low, suggesting that tenure may not be strongly predictive of employee attrition, at least not in relation to the other features.

If we were to remove tenure from the model:

- The impact would likely be minimal in terms of performance, given its weak correlations with other features.

- Model simplicity could improve as it might reduce noise, especially if tenure is not contributing significantly to predictions.

- However, tenure could still have some explanatory power that is not fully captured by the correlation coefficients alone, so we'd need to validate through model performance (e.g., cross-validation or feature importance) before definitively removing it.

It would be useful to evaluate the model performance with and without tenure and observe if it has a noticeable impact on accuracy, precision, recall, or other relevant metrics. In any case, tenure seems to be a weaker predictor compared to other features like satisfaction level or average monthly hours.

- Given what you know about the data and the models you were using, what other questions could you address for the team?

Given the data and models, I could address several additional questions for the team, including:

- Is salary allocation fair relative to experience and tenure?

By analyzing salary distribution across different levels of experience and tenure, I can uncover any disparities that might exist, helping to inform more transparent and equitable compensation policies.

- Are performance evaluations being conducted fairly?

Investigating whether evaluation scores differ across departments, managers, or demographics will reveal if the evaluation process is consistent and unbiased.

- Is there a relationship between employee satisfaction and evaluation scores?

Exploring the correlation between satisfaction and evaluation scores can help determine whether higher satisfaction influences performance evaluations or if the reverse is true.

These questions could lead to actionable insights that promote fairness and enhance employee engagement and retention.

- What resources do you find yourself using as you complete this stage?

As I complete this stage, I find myself relying on the following resources:

- Pandas: For data import/export, cleaning, transformation, and exploratory data analysis (EDA). It's my primary tool for data manipulation in Python.

- Seaborn and Matplotlib: For creating various visualizations, including distribution plots, box plots, heatmaps, and decision trees to support both EDA and model evaluation.

- Matplotlib: Used for additional visualizations, particularly for model evaluation tools like ROC curves and confusion matrices.

- Scikit-learn: This library is central for training and evaluating models, including logistic regression, decision trees, and random forests. It helps with dataset splitting, hyperparameter tuning, and performance metrics like accuracy, precision, recall, F1 score, and AUC-PR.

- XGBoost: I use this framework for gradient boosting and to visualize feature importance in tree-based models, which helps improve predictive performance.

- Pickle: For saving and loading trained models, which ensures reproducibility and avoids retraining from scratch.

- GitHub: Essential for tracking version history, documenting changes, and maintaining a structured and reproducible workflow.

These resources are crucial for building, tuning, and evaluating machine learning models effectively at this stage.

- Is my model ethical?

Yes, my approach to model development and evaluation is ethical. I've taken several important steps to ensure that the model is both fair and transparent. Here are the key aspects that demonstrate the ethical considerations I've incorporated:

- Transparency in Communication:

I am committed to communicating the model's results with full transparency. By providing both strengths and limitations of the model, I avoid overstating its effectiveness and acknowledge potential weaknesses. This helps ensure that stakeholders have a clear, realistic understanding of what the model can and cannot do.

- Minimizing Bias:

I am actively working to minimize bias in my model. This includes careful selection of features and attention to potential sources of bias that could skew the predictions, particularly when it comes to sensitive factors like employee attrition. Ensuring fairness by addressing and reducing bias is a core ethical responsibility.

- Balanced Evaluation:

My focus on constructive criticism of the model's performance, rather than presenting an overly optimistic view, is a sign of ethical integrity. This approach helps prevent the model from being used inappropriately or irresponsibly.

- Consideration of Unintended Consequences:

By emphasizing the importance of avoiding harmful or misleading decisions, I am considering the broader impact of my model. Recognizing that misclassification of employee attrition could lead to ineffective interventions or potentially harmful HR decisions shows a commitment to responsible AI use.

- Focus on Fairness and Equity:

I'm mindful of ensuring the model produces fair and equitable predictions. This includes considering the ethical implications of using features such as demographic data and making sure that the model's recommendations don't inadvertently discriminate against certain employee groups.

In summary, my approach aligns with ethical best practices in data science. By prioritizing fairness, transparency, and accountability, I'm ensuring that the model serves the broader goal of responsible decision-making and minimizes the risk of harm or bias.

- When my model makes a mistake, what is happening? How does that translate to my use case?

When the model makes a mistake, it can either misclassify leavers as stayers or stayers as leavers, both of which have different implications for your use case:

- False Negatives (Leavers Misclassified as Stayers):

  - What's Happening: The model predicts an employee is likely to stay when, in fact, they are at risk of leaving. This means the model fails to identify some employees who are at high risk of attrition.

  - Implication for Use Case: False negatives are particularly critical because the business goal is to proactively address attrition risk. Failing to identify these employees could lead to missed opportunities for retention strategies, such as targeted interventions or engagement activities, which might prevent their departure.

- False Positives (Stayers Misclassified as Leavers):

  - What's Happening: The model predicts an employee is at risk of leaving when, in fact, they are likely to stay. This means the model incorrectly identifies some employees as at-risk leavers.

  - Implication for Use Case: While false positives lead to unnecessary interventions or engagement efforts for employees who aren't at risk, this is generally less problematic than false negatives. The business can still reach out to these employees to confirm satisfaction and perhaps prevent any potential dissatisfaction before it becomes a larger issue. This approach is less costly than failing to act on potential leavers.

In summary, the primary risk of the model's mistakes lies in false negatives — failing to identify employees likely to leave, which would result in missed opportunities to intervene and retain those individuals. However, false positives, while requiring some additional effort, are more manageable because they allow the business to engage with employees preemptively.