Course Seven

Google Advanced Data Analytics Capstone

Instructions

Use this PACE strategy document to record your decisions and reflections as a data professional as you work through the capstone project. As a reminder, this document is a resource guide that you can reference in the future and a space to help guide your responses and reflections posed at various points throughout the project.

Portfolio Project Recap

Many of the goals you accomplished in your individual course portfolio projects are incorporated into the Advanced Data Analytics capstone project including:

- Create a project proposal
- Demonstrate understanding of the form and function of Python
- Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
- Demonstrate understanding of how to organize and analyze a dataset to find the “story”
- Create a Jupyter notebook for exploratory data analysis (EDA)
- Create visualization(s) using Tableau
- Use Python to compute descriptive statistics and conduct a hypothesis test
- Build a multiple linear regression model with ANOVA testing
- Evaluate the model
- Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to solve a problem
- Articulate findings in an executive summary for external stakeholders

Project proposal

Salifort Motors Employee Turnover Prediction

Overview

This project aims to develop machine learning models to predict employee turnover at Salifort Motors, analyzing survey data to identify key factors influencing employee departures. The goal is to provide actionable insights for the leadership team to implement targeted retention strategies, thereby reducing turnover costs and enhancing employee satisfaction.

Milestones	Tasks	PACE stages
1	Establish project goals, identified stakeholder needs, confirmed data availability, defined project scope and deliverables, selected analysis methods and tools, and addressed ethical concerns.	**Plan**
2	**Data Preprocessing and Cleaning:** Loading the dataset, handling missing values, standardizing data formats, encoding categorical variables (one-hot encoding), scaling numerical variables (if necessary).	**Analyze**
3	**Exploratory Data Analysis (EDA):** Calculating descriptive statistics, creating visualizations (bar charts, heatmaps, etc.), analyzing correlations between variables, feature importance analysis.	**Analyze**
4	**Feature Engineering and Selection:** Creating new features (if necessary), selecting relevant features for modeling, addressing multicollinearity if present.	**Construct**
5	**Model Development and Training (Decision Tree, Random Forest, XGBoost):** Splitting the dataset into training and testing sets, training Decision Tree, Random Forest, and XGBoost models.	**Construct**
6	**Model Evaluation and Comparison:** Evaluating model performance using metrics (AUC-ROC, F1-score, Cohen's Kappa, Balanced Accuracy), comparing the performance of different models.	**Construct**
7	**Hyperparameter Tuning (XGBoost):** Using RandomizedSearchCV or other techniques to optimize XGBoost model parameters, evaluating the tuned model.	**Construct**
8	**Results Interpretation and Visualization:** Interpreting model results and feature importance, creating visualizations to communicate findings.	**Execute**
9	**Executive Summary and Recommendations for Salifort Motors:** Summarizing the project findings, providing actionable recommendations for Salifort Motors, creating a report for the stakeholders.	**Execute**

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?

  Salifort Motors' senior leadership team and HR department.

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger business need?

  The project aims to predict employee turnover and identify key factors driving it, enabling Salifort Motors to implement targeted retention strategies, reduce costs, and improve employee satisfaction.

- What questions need to be asked or answered?

  What are the primary drivers of employee turnover?
  Which employees are at high risk?
  What effective retention strategies can be implemented?

- What resources are required to complete this project?

  Python (pandas, scikit-learn, xgboost), Jupyter Notebook, relevant documentation, online tutorials.

- What are the deliverables that will need to be created over the course of this project?

  Jupyter Notebook, executive summary, model evaluation report, actionable recommendations.

Get Started with Python

- How can you best prepare to understand and organize the provided information?

  Reviewing data analysis and machine learning libraries, and understanding the data structure, is essential.

- What follow-along and self-review codebooks will help you perform this work?

Scikit-learn, pandas, XGBoost documentation are useful.

- What are a couple additional activities a resourceful learner would perform before starting to code?

Researching best practices for handling imbalanced datasets, and reviewing feature importance techniques are helpful.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?

Relevant Columns: satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company (tenure), work_accident, promotion_last_5years, department, salary. **Target Variable:** left (employee departure).

- What units are your variables in?

satisfaction_level, **last_evaluation**: float64 (0-1 score) **number_project**, **average_monthly_hours**, **time_spend_company**: int64 (count, hours, years) **work_accident**, **left**, **promotion_last_5years**: int64 (0/1 binary) **department**, **salary**: object (categorical)

- What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

High workload, low satisfaction, and specific departments or job titles may correlate with higher turnover.

- Is there any missing or incomplete data?

Missing Data: None. **Duplicate Data:** 3008 rows. **Incomplete Data:** Potentially, due to high duplication. **Action:** Duplicates removed.

- Are all pieces of this dataset in the same format?

No. The dataset contains a mix of numerical (**int64**, **float64**) and categorical (**object**) data types.

- Which EDA practices will be required to begin this project?

  - Descriptive statistics (mean, median, etc.)
  - Data visualization (histograms, boxplots, scatter plots, heatmaps)
  - Correlation analysis
  - Categorical variable analysis (value counts, bar charts)
  - Missing value checks
  - Duplicate checks
  - Outlier detection

The Power of Statistics

- What is the main purpose of this project?

  To understand the statistical relationships between employee attributes and turnover, and to build a predictive model based on these relationships.

- What is your research question for this project?

  What factors (satisfaction, evaluation, tenure, etc.) significantly influence employee turnover at Salifort Motors?

- What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn't use random sampling?

  Random sampling ensures the model is trained on a representative subset of the population, preventing bias. If, for instance, the HR survey only included employees from one department or those who recently experienced a work accident, the model would be biased and not generalize well to the entire company.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?

  Salifort Motors' senior leadership team and the HR department.

- What are you trying to solve or accomplish?

  To identify and quantify the relationships between employee attributes and turnover using regression analysis, providing insights for retention strategies.

- What are your initial observations when you explore the data?

  There are no missing values.
  There are 3008 duplicate rows, which have been removed.
  There are 824 rows with outliers in the "tenure" column.
  The data includes numerical and categorical variables.
  The "left" column is the target variable (binary, 0 or 1).

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

  Coursera: Specifically, Google Advanced Data Analytics Professional Certificate projects for practical application and advanced techniques. coursera.org

- Do you have any ethical considerations in this stage?

  Ensuring employee data is handled confidentially.
  Avoiding the use of sensitive data for discriminatory purposes.
  Communicating the limitations of the model to stakeholders.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?

  To build a predictive model that accurately predicts employee turnover.

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

  Coursera: Specifically, Google Advanced Data Analytics Professional Certificate projects for practical application and advanced techniques. coursera.org

- Is my data reliable?

  The data has no missing values, but it did have many duplicate rows. The tenure column also had many outliers. The data needs to be preprocessed carefully.

- Do you have any additional ethical considerations in this stage?

  - Avoiding biased models that could unfairly target specific employee groups.
  - Ensuring the model is used to improve employee retention, not to justify layoffs.

- What data do I need/would I like to see in a perfect world to answer this question?

  Employee feedback, performance reviews, detailed career development data, and reasons for leaving.

- What data do I have/can I get?

  The provided dataset, containing employee attributes and turnover information.

- What metric should I use to evaluate success of my business objective? Why?

  AUC-ROC: This metric represents the area under the Receiver Operating Characteristic curve, illustrating the model's ability to distinguish between employees who left and those who stayed across various thresholds. It's particularly valuable because it remains reliable even with imbalanced datasets, which are common in employee turnover scenarios.

  F1-score: This score calculates the harmonic mean of precision (correctly predicted departures) and recall (identifying all actual departures), providing a balanced measure of the model's accuracy. It's crucial when the cost of failing to identify employees who will leave (false negatives) is high, as it minimizes such errors.

  Balanced Accuracy: This metric calculates the average of recall obtained on each class. Therefore it corrects for the data imbalance that exists in the target variable. It is valuable because it gives a more accurate representation of the model's performance.

  Cohen's Kappa: This metric measures the agreement between predicted and actual outcomes, accounting for the possibility of agreement occurring by chance, thus indicating the model's reliability. It is valuable because it shows how much better the model performs compared to random guessing.

Data Project Questions & Considerations

PACE: Analyze Stage

Get Started with Python

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, the available information appears sufficient. The variables cover key aspects of employee satisfaction, workload, compensation, and tenure, all of which are strongly linked to turnover. The EDA has revealed clear relationships and patterns that can be used to build a predictive model and derive actionable insights.

Go Beyond the Numbers: Translate Data into Insights

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

Key steps taken include: • Calculating descriptive statistics. • Visualizing variable distributions and relationships (heatmaps, countplots, boxplots, pairplots). • Analyzing correlations between variables. • Examining the impact of categorical variables (department, salary) on turnover. • Addressing duplicate data. Further steps could include: • Creating interaction features. • Performing more in depth analysis of the tenure variable, to find why long tenured employees are leaving.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

Based on the current analysis, additional data joining is not immediately necessary. However, if available, external data on industry benchmarks or economic indicators could provide valuable context. Structuring done included: • Dropping Duplicates. • Renaming columns. Further structuring could include: • Filtering the dataset to examine specific departments or salary levels. • Sorting the data by satisfaction level or tenure to identify at-risk employee segments.

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

For HR and senior leadership, clear and concise visualizations are essential:
- Bar charts showing turnover rates by department and salary.
- Box plots illustrating the impact of workload (hours, projects) and tenure on turnover.
- Correlation heatmaps to highlight key relationships.
- Countplots to show the distribution of the target variable, and categorical variables.
- Simple histograms to show the distribution of numerical variables.

The Power of Statistics

- Why are descriptive statistics useful?

Descriptive statistics provide a summary of the data's central tendency (mean, median), dispersion (standard deviation), and range, enabling a quick understanding of variable distributions and potential outliers.

- What is the difference between the null hypothesis and the alternative hypothesis?

The null hypothesis assumes no significant relationship or effect, while the alternative hypothesis posits that a significant relationship or effect exists.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?

- Checking for linearity between independent and dependent variables.
- Detecting multicollinearity among independent variables.
- Identifying outliers that could skew the model.
- Assessing the distribution of residuals.

- Do you have any ethical considerations in this stage?

Yes:
- Ensuring data privacy and confidentiality.
- Avoiding biased interpretations that could lead to discriminatory practices.
- Being transparent about data limitations and potential biases.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?

  The goal is to predict employee turnover and identify key contributing factors. The plan is still valid, but may require revising based on the results of the model.

- Does the data break the assumptions of the model? Is that ok, or unacceptable?

  The data exhibits class imbalance, which can affect some models. This needs to be addressed. Outliers are also present, and depending on the model chosen, may need to be addressed.

- Why did you select the X variables you did?

  The selected variables (satisfaction, salary, hours, tenure, etc.) are known to have a strong influence on employee turnover, based on both intuition and previous research.

- What are some purposes of EDA before constructing a model?

  - Understanding variable relationships.
  - Identifying data quality issues.
  - Informing feature selection and engineering.
  - Guiding model selection.

- What has the EDA told you?

  EDA has revealed key drivers of turnover (low satisfaction, low salary, high workload), identified data quality issues (duplicates, imbalance), and highlighted important variable relationships.

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

  Python libraries (pandas, matplotlib, seaborn, scikit-learn, NumPy), course documentation, Python documentation, and online resources (YouTube) and Coursera

- Do you have any ethical considerations in this stage?

Yes:
- Data privacy and security.
- Bias awareness and mitigation.
- Responsible use of model predictions.
- Transparency with stakeholders.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?

  From the EDA, the average monthly hours and tenure for employees who left were slightly higher than those who stayed. This could indicate potential issues with workload and long-term retention. Additionally, the low average satisfaction levels indicate a widespread issue.

- How many vendors, organizations or groupings are included in this total data?

  The primary groupings are departments (IT, RandD, accounting, etc.) and salary levels (low, medium, high). There are 10 departments, and 3 salary levels.

Go Beyond the Numbers: Translate Data into Insights

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

  - **Data visualizations:** Confusion matrices, ROC curves, and visualizations of feature importance from tree-based models.
  - **Machine learning algorithms:** Logistic Regression, Decision Tree, XGBoost, Random Forest, and SVM.
  - **Outputs:** Classification reports, AUC-ROC scores, Cohen's Kappa scores, and feature importance analyses.

- What processes need to be performed in order to build the necessary data visualizations?

  - Data preprocessing (one-hot encoding, scaling).
  - Model training and prediction.
  - Calculation of evaluation metrics.
  - Generation of confusion matrices and ROC curves using matplotlib and seaborn.

- Which variables are most applicable for the visualizations in this data project?

  The target variable ('left'), satisfaction level, salary, department, tenure, and average monthly hours are most applicable.

- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

The provided data was cleaned of duplicates, and no missing data was noted. If missing data were found, methods like imputation (mean, median, or mode) or removal of rows/columns would be used.

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?

• **Null hypothesis:** There is no significant relationship between the independent variables (satisfaction, salary, etc.) and employee turnover ('left'). • **Alternative hypothesis:** There is a significant relationship between the independent variables and employee turnover.

- What conclusion can be drawn from the hypothesis test?

The high performance of the tree-based models, particularly XGBoost, indicates that the alternative hypothesis is supported. There is a significant relationship between the independent variables and employee turnover.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?

Yes, high Variance Inflation Factor (VIF) values indicate significant multicollinearity among the one-hot encoded 'department' features. Multicollinearity is when independent variables in a model are highly correlated, providing redundant information and potentially destabilizing model results.

- Can you improve it? Is there anything you would change about the model?

Yes, multicollinearity should be addressed using feature selection or dimensionality reduction. Additionally, hyperparameter tuning and class imbalance handling are crucial for model improvement.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?

Yes, multicollinearity and class imbalance are problems. They can be fixed using feature selection/PCA and resampling techniques (SMOTE).

- Which independent variables did you choose for the model, and why?

  All available variables were used to capture all potential relationships. This led to the discovery of multicollinearity, which needs to be addressed.

- How well does your model fit the data? (What is my model's validation score?)

  XGBoost (tuned) had the highest validation scores, with an AUC-ROC of 0.9804 and a Cohen's Kappa of 0.9248.

- Can you improve it? Is there anything you would change about the model?

  Yes, addressing multicollinearity, tuning hyperparameters, and handling class imbalance are crucial for improvement.

- Do you have any ethical considerations in this stage?

  Yes:
  - Bias in the data.
  - Privacy of employee data.
  - Fairness and transparency in model application.
  - The potential negative impact on employee morale.

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?

I would recommend investigating the high number of duplicate entries, and why they were present. Also, the high correlation between the department variables, this points to underlying issues that need to be understood.

- What data initially presents as containing anomalies?

The duplicate entries, the high multicollinearity in the "department" variables, and the imbalanced target variable ("left") initially presented as anomalies.

- What additional types of data could strengthen this dataset?

Additional data that could strengthen the dataset includes: • Employee performance metrics (beyond last evaluation). • Employee feedback from exit interviews. • Industry benchmarks for turnover rates. • Economic indicators affecting employee behavior. • Employee interaction data, and team dynamics. • Data on the effectiveness of previous retention strategies.

Go Beyond the Numbers: Translate Data into Insights

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

Key insights include: • Low employee satisfaction is a strong predictor of turnover. • Salary and workload significantly impact retention. • Certain departments have disproportionately high turnover rates. • Tenure has a positive correlation with leaving the company.

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

Recommendations include:
- Implement targeted retention strategies for high-turnover departments.
- Review and adjust salary structures to ensure competitiveness.
- Implement workload management strategies to prevent burnout.
- Improve employee satisfaction through targeted programs.

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

Other questions include:
- What are the underlying causes of low satisfaction in specific departments?
- How do changes in company policies affect turnover rates?
- What is the ROI of different retention strategies?
- How do team dynamics affect turnover?

- How might you share these visualizations with different audiences?

For senior management, focus on high-level summaries and key performance indicators (KPIs). For HR, provide detailed visualizations and actionable insights. Use interactive dashboards for ongoing monitoring.

The Power of Statistics

- What key business insight(s) emerged from your A/B test?

Simulated A/B tests showed:
- Mentorship programs significantly reduced turnover rates.
- Flexible work hours significantly increased employee satisfaction.
- Salary increases significantly improved performance ratings.
- Enhanced training significantly increased employee skill levels.

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

Implement and rigorously evaluate HR interventions like mentorship, flexible work, salary adjustments, and training programs using A/B testing to measure their impact.

Regression Analysis: Simplify Complex Data Relationships

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

This project focused on classification, not regression. Therefore, beta coefficients were not used.

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

I would recommend that the company:

- **Invest in Proactive Retention Programs:** Implement programs that address the key factors contributing to employee turnover, such as improving employee satisfaction and workload management.
- **Enhance Data Collection and Management:** Improve the quality and completeness of employee data to enhance the models' accuracy and reliability.
- **Establish Ethical Guidelines:** Develop clear ethical guidelines for using predictive models in HR, ensuring transparency and fairness.
- **Provide Training and Support:** Train HR staff on how to interpret and use the model's predictions effectively.
- **Conduct Regular Audits:** Perform periodic audits to ensure that the models are not producing biased or discriminatory outcomes.
- **Integrate A/B Testing into HR Strategy:** Use A/B testing as a standard tool to evaluate the effectiveness of HR initiatives and make data-driven decisions.
- **Prioritize Employee Feedback:** Establish mechanisms for regular employee feedback and integrate this feedback into HR strategy and model development.

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

Yes, the models could be improved.

- **Feature Engineering:** Creating more relevant features, or combining existing ones, can increase model accuracy.
- **Further Hyperparameter Tuning:** More extensive hyperparameter tuning, perhaps using Bayesian optimization, may enhance performance.
- Class Imbalance Mitigation: Implementing advanced resampling techniques, or cost-sensitive learning.
- **External Validation:** Testing the model on datasets from other time periods or companies, to test the model's generalizability.
- **Outlier Analysis:** Implement isolation forest or other methods to remove outliers.
- **Assumption Validation:** Validating the assumptions of the chosen models. Especially in the case of logistic regression.
- **Real-World A/B Testing:** Validating the simulated A/B test findings with real-world A/B tests to measure the actual impact of HR interventions.
- **Continuous Model Monitoring:** Implementing a system for continuous model monitoring to detect and address any performance degradation or bias over time.

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

Based on the model results, I recommend the following:

- **Implement XGBoost (tuned) or Random Forest:** Deploy these models for practical employee turnover prediction due to their high accuracy and reliability.
- **Focus on Key Predictors:** Analyze the feature importance from XGBoost and Random Forest to identify the most significant factors contributing to employee turnover (e.g., satisfaction, project load, working hours).
- **Develop Targeted Retention Strategies:** Use the model's insights to develop targeted retention strategies for employees identified as high-risk.
- **Address Departmental Issues:** Investigate the high multicollinearity within the 'department' features, as this implies that there are underlying issues related to departments that influence employee turnover.
- **Monitor and Update Models:** Regularly monitor the models' performance and update them as needed to reflect changes in employee behavior and company policies.
- **Implement and Evaluate HR Interventions:** Based on the insights from the A/B test simulations, implement and rigorously evaluate HR interventions like mentorship programs, flexible work hours, salary adjustments, and enhanced training programs. A/B testing can be used to measure the impact of these initiatives on key employee metrics.

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

The key insight from the predictive models is the significant superiority of tree-based models, particularly XGBoost (tuned), in predicting employee turnover. XGBoost (tuned) demonstrated exceptional performance across all metrics, including a high AUC-ROC, Cohen's Kappa, and balanced accuracy, indicating its ability to accurately distinguish between employees who leave and stay, even in the presence of class imbalance. Conversely, Logistic Regression exhibited substantial limitations, struggling with class imbalance and multicollinearity, resulting in poor recall for employees who left. This highlights the importance of using models capable of capturing non-linear relationships and handling complex data patterns in employee turnover prediction. Additionally, the tuned XGBoost model's performance improvement demonstrated the importance of hyperparameter tuning.

Furthermore, the simulated A/B tests indicated that interventions like mentorship programs, flexible work hours, salary increases, and enhanced training could significantly impact employee metrics. For instance, the mentorship program simulation showed a statistically significant decrease in turnover rates, and the flexible work hours simulation showed a significant increase in employee satisfaction. These results suggest the potential effectiveness of targeted HR initiatives and the value of A/B testing in evaluating their impact.

- Do you have any ethical considerations at this stage?

- Yes, ethical considerations are crucial, including fairness, transparency, and data privacy.

The Nuts and Bolts of Machine Learning

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

  - Tree-based models, particularly XGBoost (tuned), significantly outperform Logistic Regression in predicting employee turnover.
  - Hyperparameter tuning is essential for model optimization.
  - Class imbalance and multicollinearity must be addressed.

- What are the criteria for model selection?

  Criteria include:
  - AUC-ROC, F1-score, Cohen's Kappa, and balanced accuracy.
  - Model robustness to class imbalance and multicollinearity.
  - Training time and computational efficiency.
  - Interpretability.

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

  Yes, the models make sense, and the results are acceptable. XGBoost (tuned) provides high predictive accuracy.

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

  Feature importance analysis from XGBoost and Random Forest can identify less important features. Removing them may improve model efficiency but could also impact accuracy.

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

  We could investigate the long term effects of implemented retention strategies.

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

  Scikit-learn documentation, academic research papers, online articles, business intelligence tools, and Coursera.

- Is my model ethical?

Ethical considerations were a primary focus throughout this project. I took several steps to ensure the model's responsible use. I rigorously tested for biases across departments and salary levels and prioritized transparency by using interpretable models and clearly communicating feature importance. I recommend that the company create clear guidelines for model use, and that they ensure that human oversight is always involved in the decision making process. I have also recommended that the company perform regular audits to ensure fairness. I recognize that ongoing vigilance is essential, and I recommend continuous monitoring and feedback mechanisms to ensure ethical and equitable outcomes.

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

False Positives (Predicting an employee will leave when they stay):

- **What's happening:** The model identifies factors that resemble those of employees who left, but the employee remains with the company.
- **Impact:**
  - Wasted resources on unnecessary retention efforts (e.g., offering unnecessary salary increases or promotions).
  - Potential damage to employee morale if they feel unfairly targeted.
  - HR staff will spend time and resources on employees that did not need intervention.

False Negatives (Predicting an employee will stay when they leave):

- **What's happening:** The model fails to identify employees who are at high risk of leaving, missing crucial indicators.
- **Impact:**
  - Loss of valuable employees and institutional knowledge.
  - Increased recruitment and training costs.
  - Disruption to team productivity and project timelines.
  - This type of error is generally more costly to the company.

Translation to My Use Case:

- Given the potential costs of losing employees, minimizing false negatives is crucial.
- However, too many false positives can also strain resources and damage morale.
- The optimal balance depends on the company's risk tolerance and priorities.
- Since the data is imbalanced, the model has a higher chance of predicting that an employee stays, rather than leaves. This needs to be taken into account.
- HR should use the model's predictions as a starting point for further investigation and intervention, not as a definitive judgment.
- The model should be used to support, and not replace human decision making.