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

Title: Employee Attrition Analysis and Predictive Modeling for Salifort Motors

Overview

This project aims to analyze employee data from Salifort Motors to understand the factors that contribute to employee attrition and to develop a predictive model that can identify employees who are at high risk of leaving the company. The ultimate goal is to provide actionable insights and recommendations to Salifort Motors to improve employee retention, reduce turnover costs, and enhance overall workforce stability. This project will involve data exploration, visualization, statistical analysis, and machine learning techniques using Python.

Milestones	Tasks	PACE stages
Project Initiation	Define project scope, identify data sources, gather initial requirements.	Plan
Data Exploration	Perform exploratory data analysis (EDA) to understand data characteristics and identify patterns.	Analyze
Data Preparation	Clean, transform, and prepare the data for modeling, including handling missing values and outliers.	Analyze
Feature Engineering	Create new features and select the most relevant variables for predicting attrition.	Analyze / Construct
Model	Build and train a predictive model (e.g., logistic regression,	Construct

Development	**random forest) to predict attrition.**	
Model Evalution	**Evaluate model performance using appropriate metrics and techniques.**	**Execute**
Result Interpretation	**Interpret the model results and identify key drivers of attrition.**	**Execute**
Recommendations	**Develop actionable recommendations for Salifort Motors based on the analysis.**	**Execute**
Final Report	**Document the entire project process, findings, and recommendations in a final report.**	**Execute**
Present	**Present the project findings and recommendations to stakeholders.**	**Execute**

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of Data Science

- Who is your audience for this project?
    - My primary audience is the HR department and senior management of Salifort Motors. They are the stakeholders who need to understand employee attrition and implement strategies to address it.
- 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?
    - I am trying to solve the problem of employee attrition at Salifort Motors. I aim to identify the key factors that contribute to employees leaving the company and develop a predictive model to identify at-risk employees. The anticipated impact is a reduction in employee turnover, which will lower hiring and training costs, improve productivity, and maintain organizational knowledge.
- What questions need to be asked or answered?
    - Key questions include:
        - What are the primary drivers of employee attrition?
        - Can we predict which employees are likely to leave?
        - Which departments or employee groups have the highest attrition rates?
        - What are the costs associated with employee turnover?
        - What interventions can be implemented to improve employee retention?
- What resources are required to complete this project?
    - Resources required include:
        - Employee data from Salifort Motors' HR systems.
        - Python programming environment (Jupyter Notebook).
        - Python libraries (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, Statsmodels).
        - Computational resources (computer with sufficient processing power).
        - Time and expertise for data analysis and model development.
- What are the deliverables that will need to be created over the course of this project?
    - Deliverables include:
        - Project proposal.
        - Exploratory data analysis (EDA) notebook.
        - Predictive model(s).
        - Model evaluation report.

- Final report with findings and recommendations.
- Presentation for stakeholders.

Get Started with Python

- How can you best prepare to understand and organize the provided information?
  - I can best prepare by:
    - Reviewing the data dictionary to understand each variable.
    - Identifying the data types of each column.
    - Planning the data cleaning and preprocessing steps.
    - Outlining the analysis and modeling approach.
- What follow-along and self-review codebooks will help you perform this work?
  - Follow-along and self-review codebooks on:
    - Pandas for data manipulation.
    - NumPy for numerical operations.
    - Matplotlib and Seaborn for data visualization.
    - Scikit-learn for machine learning.[1]
- What are a couple of additional activities a resourceful learner would perform before starting to code?
  - Additional activities:
    - Researching best practices for employee attrition analysis.
    - Exploring relevant documentation for Python libraries.
    - Looking for examples of similar projects or analyses.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
  - Data columns include:
    - 'satisfaction_level', 'last_evaluation', 'number_project', 'average_monthly_hours', 'time_spend_company', 'Work_accident', 'left', 'promotion_last_5years', 'Department', 'salary'.[2]
    - Relevant variables for the deliverable (predicting attrition) are likely to be all of these, but EDA will help confirm their importance. 'left' is the target variable.
- What units are your variables in?
  - Units vary:
    - 'satisfaction_level' and 'last_evaluation' are likely on a scale (e.g., 0 to 1).
    - 'number_project' is a count.
    - 'average_monthly_hours' is in hours.
    - 'time_spend_company' is in years.
    - 'salary' is categorical (e.g., 'low', 'medium', 'high').
    - Other variables are binary or 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?
    - Initial presumptions:
        - Lower satisfaction and higher working hours will be associated with higher attrition.
        - Salary level will be a significant factor.
        - Employees with more experience (time spent at the company) may be less likely to leave.
- Is there any missing or incomplete data?
    - This needs to be investigated during EDA.
- Are all pieces of this dataset in the same format?
    - This also needs to be checked during data exploration.
- Which EDA practices will be required to begin this project?
    - EDA practices:
        - Descriptive statistics.
        - Data visualization (histograms, box plots, scatter plots, etc.).
        - Correlation analysis.
        - Missing data analysis.
        - Data profiling.

The Power of Statistics

- What is the main purpose of this project?
    - The main purpose is to understand and predict employee attrition.
- What is your research question for this project?
    - The primary research question is: What factors significantly influence employee attrition at Salifort Motors, and can we build a model to predict which employees are most likely to leave?
- 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 is important to ensure that the sample used for analysis is representative of the entire population of employees. This helps to avoid bias and ensures that the findings can be generalized.
    - Example of sampling bias: If we only analyzed data from employees who voluntarily participated in a survey, we might miss the perspectives of those who are dissatisfied and chose not to participate, leading to an underestimation of attrition risk.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
    - Stakeholders include:
        - HR department.

        - Senior management.
        - Department managers.
- What are you trying to solve or accomplish?
    - I'm trying to identify the relationships between various employee factors and the likelihood of them leaving the company. This will help in understanding the drivers of attrition.
- What are your initial observations when you explore the data?
    - Initial observations will be made during EDA, but I expect to see relationships between variables like satisfaction, salary, and attrition.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
    - Resources:
        - Pandas documentation: https://pandas.pydata.org/
        - Scikit-learn documentation: https://scikit-learn.org/
        - Statsmodels documentation: https://www.statsmodels.org/
        - Seaborn documentation: https://seaborn.pydata.org/
- Do you have any ethical considerations in this stage?
    - Ethical considerations:
        - Ensuring data privacy and confidentiality.
        - Avoiding biased analysis.
        - Using the model responsibly and not for discriminatory purposes.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
    - I'm trying to build a predictive model that can accurately identify employees who are at high risk of leaving the company.
- What resources do you find yourself using as you complete this stage?
    - Resources:
        - Scikit-learn documentation: https://scikit-learn.org/
        - Machine learning tutorials and documentation.
- Is my data reliable?
    - Data reliability needs to be assessed through data quality checks during EDA.
- Do you have any additional ethical considerations in this stage?
    - Additional ethical considerations:
        - Model fairness and potential for bias.
        - Explainability of the model's predictions.
        - Impact of predictions on employees' lives.
- What data do I need/would I like to see in a perfect world to answer this question?
    - Ideal data:

        - More detailed employee feedback (e.g., from surveys).
        - Data on employee engagement and morale.
        - Information on external factors (e.g., job market conditions).
- What data do I have/can I get?
    - This depends on the data provided by Salifort Motors.
- What metric should I use to evaluate success of my business objective? Why?
    - Metrics:
        - AUC-ROC: Measures the model's ability to distinguish between employees who leave and stay.
        - Recall (for employees who leave): Measures the model's ability to identify employees who will actually leave (minimizing false negatives is crucial in this scenario).
        - Precision (for employees who leave): Measures how often the model is correct when it predicts an employee will leave (important to avoid unnecessary interventions).
        - F1-score: Balances precision and recall.

Data Project Questions & Considerations

PACE: Analyze Stage

PACE Stage: Analyze

Get Started with Python

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?
    - It's difficult to say with complete certainty before conducting the analysis, but based on my initial intuition and the variables provided, I believe the information will be sufficient to achieve the goal. The dataset includes variables that are commonly associated with employee attrition, such as satisfaction level, salary, working hours, and time spent at the company. However, the actual sufficiency will depend on the strength of the relationships between these variables and attrition, and the quality of the data.

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?
    - To perform EDA effectively, I will take these steps:
        - **Data Cleaning:** Handle missing values, correct data types, and address any inconsistencies.
        - **Descriptive Statistics:** Calculate summary statistics (mean, median, standard deviation, etc.) for numerical variables and frequency counts for categorical variables.
        - **Data Visualization:** Create visualizations to explore relationships between variables and the target variable ('left'). This will include:
            - Histograms and box plots for numerical variables.
            - Bar charts for categorical variables.
            - Correlation matrices to understand relationships between variables.
        - **Outlier Detection:** Identify and address any outliers that may skew the results.
        - **Feature Analysis:** Examine the distribution of each variable and its relationship to attrition.
- 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.?
    - At this point, I do not anticipate needing to join additional data. The provided dataset seems to contain the necessary information. However, I will perform the following structuring steps:
        - **Filtering:** Filter the data, if necessary, to focus on specific subsets of employees (e.g., by department).
        - **Sorting:** Sort the data to facilitate analysis and visualization.
        - **Data Type Conversion:** Ensure all variables are in the correct data type.

        - **Data Transformation:** Transform variables, if needed, to make them suitable for modeling.
- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
    - I assume the following visualizations will be best suited for the intended audience (HR and senior management):
        - Clear and concise charts that highlight key relationships and trends.
        - Visualizations that show the impact of different factors on attrition (e.g., how satisfaction level or salary affects the likelihood of leaving).
        - Executive-friendly summaries that avoid technical jargon and focus on actionable insights.
        - I will use tools like Seaborn and Matplotlib to create these visualizations.

The Power of Statistics

- Why are descriptive statistics useful?
    - Descriptive statistics are useful because they provide a concise summary of the main features of a dataset. They help to understand the distribution, central tendency, and variability of the data, which is essential for identifying patterns, anomalies, and potential relationships between variables.
- What is the difference between the null hypothesis and the alternative hypothesis?
    - The null hypothesis (H0) is a statement that there is no significant effect or relationship in the population. It's what I will try to disprove. The alternative hypothesis (Ha or H1) is a statement that there is a significant effect or relationship. In this project, the null hypothesis might be that there is no difference in key metrics between employees who leave and those who stay, while the alternative hypothesis is that there *is* a difference.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
    - Purposes of EDA before constructing a multiple linear regression model:
        - **Check Assumptions:** Verify if the data meets the assumptions of linear regression (linearity, independence, normality, homoscedasticity).
        - **Identify Relationships:** Explore the relationships between the independent variables and the dependent variable to inform model specification.
        - **Detect Multicollinearity:** Identify if there are high correlations between independent variables, which can affect the stability and interpretability of the model.
        - **Identify Outliers:** Detect and handle outliers that may disproportionately influence the regression results.
        - **Feature Selection:** Help in selecting the most relevant independent variables for the model.
- Do you have any ethical considerations in this stage?
    - Yes, ethical considerations in this stage include:
        - Ensuring that the data is used responsibly and not for discriminatory purposes.
        - Being transparent about the limitations of the analysis.
        - Protecting the privacy of employees.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
    - I am trying to solve the problem of predicting employee attrition. The plan is still fundamentally sound, but it may need revising based on the EDA. For example, if I find that the data violates the assumptions of Logistic Regression, I may need to consider alternative modeling techniques.
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
    - This needs to be thoroughly checked during the modeling phase. If the data severely violates the assumptions of the chosen model (e.g., Logistic Regression), the results may be unreliable, and it would be unacceptable. I would then need to consider data transformations or a different model.
- Why did you select the X variables you did?
    - I selected the X variables (independent variables) based on my understanding of the factors that typically influence employee attrition (e.g., satisfaction, salary, working hours, etc.) and the variables available in the dataset.
- What are some purposes of EDA before constructing a model?
    - Purposes of EDA before constructing a model:
        - (Same as in Regression Analysis) Check assumptions, Identify Relationships, Detect Multicollinearity, Identify Outliers, Feature Selection.
        - Additionally, understand feature scaling requirements and identify potential issues.
- What has the EDA told you?
    - The EDA will tell me about:
        - The distribution of the variables.
        - The relationships between the variables.
        - The presence of missing data or outliers.
        - The potential need for data transformations.
        - Which variables are most likely to be predictive of attrition.
- What resources do you find yourself using as you complete this stage?
    - Resources:
        - Pandas, NumPy, Matplotlib, Seaborn documentation.
        - Scikit-learn documentation and tutorials.
        - Statistical and machine learning textbooks.
        - Online resources and forums (e.g., Stack Overflow).
- Do you have any ethical considerations in this stage?
    - Yes, ethical considerations include:
        - Ensuring data privacy and confidentiality.
        - Avoiding bias in model development and evaluation.
        - Using the model responsibly and transparently.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
    - This will require a close look at the descriptive statistics of the numerical variables (e.g., 'satisfaction_level', 'average_monthly_hours', 'time_spend_company'). I'll be looking for averages that seem significantly different from what I'd expect in a typical employee dataset, which could indicate data errors or specific patterns in this company.
- How many vendors, organizations, or groupings are included in this total data?
    - This question relates to the categorical variables, primarily 'department' and 'salary'. I'll need to determine the number of unique values within each of these columns to understand the diversity of departments and salary levels represented in the data.

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?
    - To complete the project goals, I anticipate needing to build the following:
        - Data Visualizations:
            - Histograms and box plots to visualize the distribution of numerical variables.
            - Bar charts to visualize the distribution of categorical variables.
            - Scatter plots to explore relationships between numerical variables.
            - Correlation matrices to visualize relationships between all relevant variables.
            - ROC curves to evaluate the performance of classification models.
        - Machine Learning Algorithms:
            - I'll start with Logistic Regression as a baseline model for predicting employee attrition.
            - Depending on its performance and the data characteristics, I may also explore other algorithms like Random Forest or Gradient Boosting Machines.
        - Other Data Outputs:
            - Tables of descriptive statistics.
            - Model evaluation metrics (e.g., accuracy, precision, recall, F1-score, AUC-ROC).
            - Feature importance rankings from the machine learning model.
- What processes need to be performed in order to build the necessary data visualizations?
    - The processes involved in building the data visualizations include:
        - Data Cleaning and Preprocessing: Ensuring the data is in the correct format and handling any missing values or inconsistencies.
        - Data Transformation: Transforming variables as needed (e.g., scaling numerical variables, encoding categorical variables).
        - Visualization Design: Selecting the appropriate type of visualization for each variable and relationship.

        - Implementation: Using Python libraries like Matplotlib and Seaborn to create the visualizations.
        - Refinement: Iteratively refining the visualizations to improve clarity and effectiveness.
- Which variables are most applicable for the visualizations in this data project?
    - The most applicable variables for visualizations in this project are:
        - 'left' (the target variable, indicating attrition).
        - 'satisfaction_level'.
        - 'last_evaluation'.
        - 'number_project'.
        - 'average_monthly_hours'.
        - 'time_spend_company'.
        - 'salary'.
        - 'department'.
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
    - In the Plan stage, I outlined that missing data would be addressed during the Analyze stage (EDA). The specific approach will depend on the nature and extent of the missing data, but common strategies include:
        - Removing rows with missing values (if the amount of missing data is small).
        - Imputing missing values using the mean, median, or mode (for numerical variables).
        - Imputing missing values using more advanced techniques like k-nearest neighbors or regression imputation.
        - Treating missing data as a separate category (for categorical variables, if appropriate).

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
    - For this project, I will formulate the null and alternative hypotheses to investigate the relationship between employee attributes and attrition. For example:
        - Null Hypothesis (H0): There is no significant difference in satisfaction levels between employees who leave and those who stay.
        - Alternative Hypothesis (Ha): There is a significant difference in satisfaction levels between employees who leave and those who stay.
- What conclusion can be drawn from the hypothesis test?
    - The conclusion drawn from the hypothesis test will depend on the p-value obtained.
        - If the p-value is less than the significance level (alpha, typically 0.05), I will reject the null hypothesis and conclude that there is evidence to support the alternative hypothesis.
        - If the p-value is greater than alpha, I will fail to reject the null hypothesis, meaning there is not enough evidence to support the alternative hypothesis.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?
    - During the model building phase, I will be looking for:
        - High multicollinearity among the independent variables.
        - Non-linear relationships between the independent and dependent variables.
        - Outliers that may disproportionately influence the model.

        - Patterns in the residuals that violate the assumptions of linear regression.
- Can you improve it? Is there anything you would change about the model?
    - Yes, the model can potentially be improved. I will explore techniques such as:
        - Feature engineering to create new, more informative variables.
        - Regularization to prevent overfitting.
        - Transforming variables to meet the assumptions of linear regression.
        - Trying different regression models.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?
    - Potential problems in the machine learning phase include:
        - Low model performance (poor accuracy, precision, recall, etc.).
        - Overfitting (the model performs well on the training data but poorly on unseen data).
        - Class imbalance (if there are significantly more employees who stayed than left).
        - Violations of model assumptions.
    - These problems can be addressed through techniques such as:
        - Hyperparameter tuning.
        - Regularization.
        - Ensemble methods (e.g., Random Forest, Gradient Boosting).
        - Resampling techniques (e.g., SMOTE).
        - Feature selection.
- Which independent variables did you choose for the model, and why?
    - I will choose independent variables based on their relevance to employee attrition, as suggested by domain knowledge and the EDA. These will likely include: 'satisfaction_level', 'last_evaluation', 'number_project', 'average_monthly_hours', 'time_spend_company', 'salary', and 'department'.
- How well does your model fit the data? (What is my model's validation score?)
    - Model fit will be evaluated using appropriate metrics, such as:
        - Accuracy.
        - Precision and recall.
        - F1-score.
        - AUC-ROC.
    - I will use cross-validation to obtain a robust estimate of the model's performance on unseen data.
- Can you improve it? Is there anything you would change about the model?
    - Yes, the model can likely be improved. I will explore:
        - Different machine learning algorithms.
        - Hyperparameter tuning.
        - Feature engineering.
        - Ensemble methods.
- Do you have any ethical considerations in this stage?
    - Yes, ethical considerations in this stage include:
        - Ensuring that the model is fair and does not discriminate against any group of employees.
        - Being transparent about the model's limitations and potential biases.
        - Protecting the privacy of employee data.

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 that my manager investigate the data collection process in more detail. Specifically, I'd want to understand:
        - How the data was collected and from what systems.
        - Whether there were any changes in data collection methods during the period covered.
        - The accuracy and completeness of the data.
        - How the categorical variables (e.g., department, salary) were coded and if there are inconsistencies.
- What data initially presents as containing anomalies?
    - The VIF results indicate that the one-hot encoded 'department' and 'salary' variables show extremely high multicollinearity, which is a statistical anomaly. This needs to be addressed before the model can be reliably interpreted.
- What additional types of data could strengthen this dataset?
    - Additional data that could strengthen the dataset include:
        - Employee performance reviews.
        - Employee satisfaction survey results (more detailed than the single 'satisfaction_level' variable).
        - Data on employee demographics.
        - Information on any employee turnover initiatives or programs.
        - Data on external factors, such as local job market conditions.

Go Beyond the Numbers: Translate Data into Insights

- What key insights emerged from your EDA and visualizations(s)?
    - Key insights include:
        - The Logistic Regression model, while having a reasonable AUC-ROC (0.8420), has limited practical value.
        - Severe multicollinearity exists in 'department' and 'salary,' making it hard to interpret their individual effects.
        - The model is better at predicting employees who stay (class 0) but struggles with those who leave (class 1).
        - Factors like satisfaction level, evaluation, working hours, salary, promotion, and time at the company are generally related to attrition.
- What business recommendations do you propose based on the visualization(s) built?
    - Recommendations:

        - Address multicollinearity.
        - Improve the model's ability to predict which employees will leave.
        - Conduct further analysis into the root causes of attrition (department issues, low satisfaction, etc.).
        - Use the current model's predictions cautiously, especially for individual employees.
- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
    - Additional research questions:
        - What are the underlying reasons for the high attrition rate in certain departments?
        - What are the specific factors that contribute to low satisfaction among employees?
        - What is the cost of employee attrition to the company?
        - What is the potential ROI of implementing different retention strategies?
        - Can we identify specific "at-risk" employee profiles based on a combination of factors?
- How might you share these visualizations with different audiences?
    - When sharing visualizations:
        - Tailor the presentation to the audience's technical level.
        - For non-technical audiences, focus on high-level insights and business implications, using clear and concise language.
        - For technical audiences, provide more detail on the data, methodology, and model performance.
        - Use interactive visualizations where possible to allow users to explore the data themselves.

The Power of Statistics

- What key business insight(s) emerged from your A/B test?
    - This section is not applicable, as the provided document focuses on a classification problem (employee attrition), not A/B testing.
- What business recommendations do you propose based on your results?
    - This section is not applicable.

Regression Analysis: Simplify Complex Data Relationships

- To interpret model results, why is it important to interpret the beta coefficients?
    - In Logistic Regression, the coefficients represent the change in the log-odds of the outcome (employee leaving) for a one-unit change in the predictor variable, holding other variables constant. They indicate the direction and strength of the relationship between each predictor and the likelihood of attrition.
- What potential recommendations would you make to your manager/company?
    - Recommendations to the manager/company:
        - Acknowledge the model's limitations.
        - Prioritize data quality.
        - Conduct further analysis.
        - Focus on understanding the root causes of attrition.
        - Use the model ethically.
- Do you think your model could be improved? Why or why not? How?

    - Yes, the model can be improved by:
        - Addressing multicollinearity.
        - Improving the prediction of employees who leave.
        - Trying alternative models.
        - Gathering more data.
        - Feature engineering
- What business recommendations do you propose based on the models built?
    - Business recommendations:
        - Address multicollinearity.
        - Improve model performance.
        - Focus on high-level insights related to factors influencing attrition.
- What key insights emerged from your model(s)?
    - Key insights:
        - Multicollinearity is a major issue.
        - The model is better at predicting who stays than who leaves.
        - Several factors are related to attrition (satisfaction, etc.).
- Do you have any ethical considerations at this stage?
    - Yes, the ethical considerations are critical:
        - Responsible use of predictions.
        - Transparency and explainability.
        - Data privacy.
        - Fairness and bias.
        - Continuous monitoring.

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?
    - Key insights:
        - (Same as in Regression Analysis) Multicollinearity is a major issue.
        - The model is better at predicting who stays than who leaves.
        - Several factors are related to attrition (satisfaction, etc.).
- What are the criteria for model selection?
    - The document doesn't go into detail about criteria for *selecting* a model, but it does evaluate the Logistic Regression model based on:
        - AUC-ROC score.
        - Accuracy.
        - Precision and recall.
        - Confusion matrix.
- Does my model make sense? Are my final results acceptable?
    - The model's results are not entirely acceptable in its current state due to the multicollinearity and poor performance in predicting employees who leave.
- Were there any features that were not important at all? What if you take them out?
    - The document doesn't identify unimportant features. VIF analysis suggests that the encoded 'department' and 'salary' features are problematic due to multicollinearity, not lack of importance.
- Given what you know about the data and the models you were using, what other questions could you address for the team?

    - (Same as in Regression Analysis)
        - What are the underlying reasons for the high attrition rate in certain departments?
        - What are the specific factors that contribute to low satisfaction among employees?
        - What is the cost of employee attrition to the company?
        - What is the potential ROI of implementing different retention strategies?
        - Can we identify specific "at-risk" employee profiles based on a combination of factors?
- What resources do you find yourself using as you complete this stage?
    - Resources:
        - Scikit-learn.
        - Matplotlib and Seaborn.
        - Business analytics resources.
        - HR best practices resources.
        - Statsmodels.
- Is my model ethical?
    - The document emphasizes the importance of ethical considerations, including:
        - Responsible use of predictions.
        - Transparency.
        - Data privacy.
        - Fairness.
        - Continuous monitoring.
- When my model makes a mistake, what is happening? How does that translate to my use case?
    - When the model makes a mistake, it is either:
        - Incorrectly predicting that an employee will leave (false positive). In this use case, this might lead to unnecessary intervention or wasted resources.
        - Incorrectly predicting that an employee will stay (false negative). This is more problematic, as it means the company misses an opportunity to address the issues causing that employee to leave. Given the low recall for class 1, this is a significant concern with the current model.