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

Employee Turnover Prediction and Retention Strategies at Salifort Motors project proposal

Overview

Salifort Motors is experiencing a high employee turnover rate, leading to significant financial costs and impacts on corporate culture. This project aims to analyze employee survey data, build predictive models to identify employees at risk of leaving, and provide actionable recommendations to improve employee retention and job satisfaction.

Milestones	Tasks	PACE stages
Data Collection	**- Gather employee survey data from HR.**	**Plan**
Date Preprocessing	**- Handle missing values and convert categorical variables to numerical.**	**Analyze**
Exploratory Analysis	**- Conduct initial data analysis to identify patterns and correlations.**	**Analyze**
Model Development	**- Develop predictive models: logistic regression, decision tree, random forest, XGBoost.**	**Construct**
Model Evaluation	**- Evaluate model performance using accuracy, precision, recall, and**	**Construct**

	F1-score metrics.	
Feature Importance	- Analyze feature importance to identify key drivers of employee turnover.	Analyze
Recommendations	- Provide actionable recommendations based on model insights.	Execute
Model Deployment	- Implement the best-performing model for continuous turnover risk assessment.	Execute
Reporting	- Create an executive summary and detailed report for stakeholders.	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?

The audience for this project includes stakeholders such as data analysts, business managers, and decision-makers within the organization who rely on data-driven insights to guide strategic planning and operational decisions.

- 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 primary goal is to leverage data science techniques to extract valuable insights from the dataset, enabling better decision-making, optimizing processes, and identifying new opportunities for growth. The impact of this work will be to improve business performance through informed decisions, increase operational efficiency, and uncover new market opportunities. The insights derived could lead to cost savings, revenue growth, and enhanced customer satisfaction.

- What questions need to be asked or answered?

What are the trends and patterns within the data?

How can we predict future outcomes based on historical data?

What are the key factors influencing business performance?

Are there any anomalies or outliers in the data that need investigation?

- What resources are required to complete this project?

Data scientists and analysts

Access to relevant datasets

Data processing and analysis tools (e.g., Python, Jupyter Notebook)

Visualization tools (e.g., Matplotlib, Seaborn)

Domain experts for contextual understanding

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

Cleaned and preprocessed datasets

Exploratory Data Analysis (EDA) reports

Predictive models and their evaluations

Visualization dashboards

Final report summarizing insights and recommendations

Get Started with Python

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

To best understand and organize the provided information, I will start by familiarizing myself with the dataset and its structure. Documenting the data types, column names, and any initial observations also helps.

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

Python tutorials and documentation

Data science and machine learning textbooks

Online courses

Jupyter Notebook

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

Review basic Python programming concepts.

Study examples of data analysis and visualization in Python.

Familiarize self with libraries like Pandas, NumPy, and Matplotlib.

Go Beyond the Numbers: Translate Data into Insights

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

satisfaction_level

last_evaluation

number_project

average_monthly_hours

time_spend_company

left

salary

- What units are your variables in?

satisfaction_level: Scale (0 to 1)

last_evaluation: Scale (0 to 1)

number_project: Count

average_monthly_hours: Hours

time_spend_company: Years

Work_accident: Binary (0 or 1)

left: Binary (0 or 1)

promotion_last_5years: Binary (0 or 1)

Department: Categorical

salary: Categorical (low, medium, high)

- 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?

Higher satisfaction levels might correlate with lower attrition rates.

Employees with higher last evaluation scores may have higher satisfaction and lower attrition rates.

Employees with a very high or very low number of projects might have different attrition rates.

Higher average monthly hours might correlate with lower satisfaction and higher attrition.

Longer tenure in the company might correlate with lower attrition.

Employees who have experienced work accidents might have different satisfaction and attrition rates.

Employees who have received promotions might have higher satisfaction and lower attrition.

Departments may have varying attrition rates.

Salary level might correlate with satisfaction and attrition rates.

- Is there any missing or incomplete data?

I did not find any.

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

Some are numerical, some are categorical.

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

Descriptive Statistics: Summary statistics for numerical variables.

Data Visualization: Histograms, bar plots, and box plots for distribution and comparison.

Correlation Analysis: Heatmaps to identify relationships between numerical variables.

Categorical Data Analysis: Count plots and comparison of categorical variables.

Handling Missing Data: Identifying and deciding on strategies to handle missing values.

Outlier Detection: Identifying and analyzing outliers.

The Power of Statistics

- What is the main purpose of this project?

The main purpose of this project is to analyze employee data to understand the factors influencing employee satisfaction and attrition, and to develop predictive models that can help the organization reduce turnover and improve overall employee satisfaction.

- What is your research question for this project?

What are the key factors that influence employee attrition, and how can we predict which employees are at risk of leaving the company?

- 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 crucial to ensure that the sample represents the entire population accurately, minimizing biases and allowing for generalizable conclusions. If the sample only includes data from a particular department or time period, it might not represent the entire workforce. For instance, using only recent hires or a specific department could lead to biased conclusions that do not reflect the overall attrition patterns in the company.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?

HR Managers

Senior Executives

Department Heads

Data Science Team

- What are you trying to solve or accomplish?

To identify and quantify the relationships between employee characteristics (e.g., satisfaction level, number of projects, salary) and their likelihood of leaving the company.

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

Variations in satisfaction levels and last evaluation scores

Differences in average monthly hours and time spent at the company

Potential correlation between salary levels and employee attrition

- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)

Python libraries (Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib)

Online tutorials and documentation

- Do you have any ethical considerations in this stage?

Ensuring employee privacy and confidentiality in the data

Avoiding any bias in the analysis and ensuring fair treatment of all employees

The Nuts and Bolts of Machine Learning

- What am I trying to solve?

Develop a predictive model to identify employees at risk of leaving the company, helping HR to implement targeted retention strategies.

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

Machine learning libraries (Scikit-learn)

Tutorials on machine learning algorithms

- Is my data reliable?

It appears so.

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

Avoiding discrimination based on sensitive attributes (e.g., gender, race)

Ensuring the model does not reinforce existing biases in the data

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

Comprehensive data on all employee characteristics (e.g., demographics, job performance)

Updated and complete records of employee satisfaction and performance evaluations

- What data do I have/can I get?

Employee satisfaction level

Performance evaluations

Number of projects, average monthly hours

Tenure, work accidents, promotions

Department and salary information

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

Accuracy: To measure the overall correctness of the model's predictions.

Precision and Recall: To evaluate the model’s ability to correctly identify employees at risk of leaving (especially important for imbalanced datasets).

F1 Score: To balance precision and recall, providing a single metric for model performance.

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?

Based on the variables available (satisfaction level, last evaluation, number of projects, average monthly hours, time spent at the company, work accidents, promotions, department, and salary), the dataset appears to have sufficient information to achieve the goal of understanding employee attrition and satisfaction. These variables cover key aspects of an employee's experience and performance, which are critical to analyzing attrition.

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?

Data Cleaning:

Handle missing values.

Ensure data consistency and correct data types.

Descriptive Statistics:

Calculate mean, median, mode, standard deviation for numerical variables.

Frequency counts for categorical variables.

Data Visualization:

Histograms and box plots for distribution analysis.

Scatter plots for relationships between variables.

Heatmaps for correlation analysis.

Outlier Detection:

Identify and analyze outliers that may affect the analysis.

- 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.?

If additional relevant data is available (e.g., demographic information, additional performance metrics), it could be beneficial to join this data to enhance the analysis. Structuring may involve:

Filtering: Removing irrelevant or redundant data.

Sorting: Organizing data to facilitate analysis (e.g., by department, tenure).

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

Bar charts for categorical variable distributions (e.g., department, salary).

Box plots to show variations in satisfaction and average monthly hours across different departments.

Heatmaps to show correlations between numerical variables.

Scatter plots to illustrate relationships (e.g., satisfaction vs. last evaluation).

The Power of Statistics

- Why are descriptive statistics useful?

Descriptive statistics summarize and describe the main features of a dataset, providing a simple overview of the sample and measures. They are useful for:

Understanding the basic characteristics of the data.

Identifying trends and patterns.

Highlighting anomalies.

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

Null Hypothesis (H0): A statement that there is no effect or no difference, and any observed effect is due to sampling variability. Example: "There is no difference in attrition rates across different departments."

Alternative Hypothesis (H1): A statement that there is an effect or a difference. Example: "There is a significant difference in attrition rates across different departments."

Regression Analysis: Simplify Complex Data Relationships

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

Understanding Data Structure: Identify relationships and interactions between variables.

Data Cleaning: Ensure data is free from errors and inconsistencies.

Feature Selection: Determine which variables are most relevant for the model.

Checking Assumptions: Ensure data meets the assumptions of regression analysis (e.g., linearity, independence, homoscedasticity).

- Do you have any ethical considerations in this stage?

Avoiding bias in model training and interpretation.

Ensuring fairness and avoiding discrimination based on sensitive attributes.

Maintaining data privacy and confidentiality.

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 attrition and identify key factors influencing it. The plan may need revising based on EDA findings or model performance.

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

Ensure data meets model assumptions (e.g., linearity, normality, independence).
If assumptions are violated, consider alternative models or data transformations.

- Why did you select the X variables you did?

Variables were selected based on their relevance to employee attrition and satisfaction, and their potential predictive power (e.g., satisfaction level, last evaluation, number of projects).

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

Feature Engineering: Create new features that might improve model performance.

Data Transformation: Normalize or standardize data if necessary.

Identify Patterns: Understand data distributions and relationships.

- What has the EDA told you?

EDA might reveal which factors have the strongest relationships with attrition.

It can highlight potential issues (e.g., multicollinearity, missing values).

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

Python libraries (Pandas, NumPy, Scikit-learn, Seaborn, Matplotlib)

Online tutorials and documentation

- Do you have any ethical considerations in this stage?

Avoiding bias in data selection and model training.

Ensuring transparency in model decisions and fairness.

Protecting employee data privacy and confidentiality.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?

average_monthly_hours: 201.05 - This average looks unusual. If we assume a typical full-time employee works about 160-170 hours per month (based on a 40-hour work week), 201 hours suggest that employees might be working more hours than usual. This could indicate high workload or overtime.
Work_accident: 0.145 - About 14.5% of employees have experienced a work accident, which seems high but could be industry-specific.
left: 0.238 - Approximately 24% of employees have left the company, which could indicate a high turnover rate.
promotion_last_5years: 0.021 - Only 2.1% of employees have been promoted in the last 5 years. This seems quite low and might suggest limited career advancement opportunities within the company.

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

Number of unique departments is 10, number of unique salary levels is 3.

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:

Histograms: For distributions of numerical variables like satisfaction level, last evaluation, and average monthly hours.
Box Plots: To identify outliers and compare distributions across different groups (e.g., department, salary).
Scatter Plots: To explore relationships between pairs of variables, such as satisfaction level and last evaluation.
Heatmaps: For correlation analysis among numerical variables.
Bar Charts: For categorical data such as department and salary.

Machine Learning Algorithms:

Logistic Regression: To predict employee attrition (binary classification).

Decision Trees and Random Forests: For feature importance and prediction.

Clustering (e.g., K-means): To identify patterns or groupings within the data.

Support Vector Machines (SVM): For classification tasks.

Neural Networks: For complex prediction tasks, if the data volume and complexity warrant it.

Other Data Outputs:

Descriptive Statistics Summaries: Means, medians, standard deviations, etc.

Model Performance Metrics: Accuracy, precision, recall, F1 score, ROC-AUC.

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

Data Cleaning:

Handle missing values.

Convert categorical variables to numerical if necessary.

Standardize or normalize data if required.

Exploratory Data Analysis (EDA):

Summarize data using descriptive statistics.

Visualize distributions and relationships.

Identify and handle outliers.

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

satisfaction_level

last_evaluation

number_project

average_monthly_hours

time_spend_company

Work_accident

left

promotion_last_5years

Department

salary

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

missing_data = df0.isnull().sum()

print(missing_data)

I found no missing data.

The Power of Statistics

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

Null Hypothesis (H0): There is no significant difference in satisfaction levels between employees who left and those who stayed.

Alternative Hypothesis (H1): There is a significant difference in satisfaction levels between employees who left and those who stayed.

- What conclusion can be drawn from the hypothesis test?

We reject the null hypothesis.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?

Significance of Predictors (P-values):

Significant Predictors: Variables such as satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, Work_accident, promotion_last_5years, Department_RandD, Department_hr, Department_support, Department_technical, salary_low, and salary_medium have very low p-values (< 0.05), indicating they are significant predictors.

Non-significant Predictors: Department_accounting, Department_management, Department_marketing, Department_product_mng, and Department_sales have high p-values (> 0.05), suggesting they might not be significant in predicting employee attrition.

Coefficient Signs:

Negative Coefficients: Predictors such as satisfaction_level, number_project, Work_accident, promotion_last_5years, and Department_RandD have negative coefficients, indicating an inverse relationship with the likelihood of leaving the company.

Positive Coefficients: Predictors like last_evaluation, average_monthly_hours, time_spend_company, Department_hr, Department_support, Department_technical, and salary levels (low, medium) have positive coefficients, suggesting a direct relationship with the likelihood of leaving the company.

Standard Errors and Z-values:

Check if any predictor has a large standard error relative to its coefficient, which might indicate multicollinearity or instability in the model. Most predictors here seem to have reasonable standard errors.

Residual Plot:

The residual plot shows a pattern where residuals are mostly clustered along a diagonal line. This is somewhat expected in logistic regression but should ideally be randomly scattered around the horizontal axis at zero. This pattern may indicate some model misspecification or issues with how the data fits the model.

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

Remove Non-significant Predictors:

Consider removing predictors like Department_accounting, Department_management, Department_marketing, Department_product_mng, and Department_sales since they do not significantly contribute to the model. This can simplify the model and potentially improve its performance.

Check for Multicollinearity:

Calculate the Variance Inflation Factor (VIF) for each predictor to identify multicollinearity issues. High VIF values (typically > 10) indicate multicollinearity.

Transform Variables if Necessary:

For non-linear relationships, consider transforming the variables (e.g., log transformation) to better fit the model.

The Nuts and Bolts of Machine Learning

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

Problems Identified:

Non-significant Predictors:

There are predictors with high p-values indicating they may not significantly contribute to the model.

Residual Patterns:

The residual plot shows a non-random pattern, suggesting potential model misspecification.

Fixes:

Remove Non-significant Predictors:

Simplify the model by removing predictors with high p-values.

Check and Address Multicollinearity:

Use VIF to identify and address multicollinearity.

Transform Variables:

Apply transformations to handle non-linear relationships.

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

Chosen Independent Variables:

satisfaction_level: Significant predictor of employee satisfaction.

last_evaluation: Reflects performance, potentially influencing attrition.

number_project: Workload indicator, impacting stress and satisfaction.

average_monthly_hours: Indicator of work-life balance, affecting satisfaction.

time_spend_company: Tenure, which can influence loyalty and attrition.

Work_accident: Reflects workplace safety, impacting satisfaction.

promotion_last_5years: Career growth opportunities, influencing satisfaction and attrition.

Department: Encoded to capture departmental differences in attrition.

salary: Encoded to account for financial satisfaction.

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

It fits the data well.

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

Feature Engineering:

Create new features based on domain knowledge.

Hyperparameter Tuning:

Use GridSearchCV to find the best parameters for the model.

Ensemble Methods:

Combine multiple models (e.g., Random Forest, Gradient Boosting) to improve performance.

- Do you have any ethical considerations in this stage?

Bias and Fairness:

Ensure the model does not unfairly discriminate based on sensitive attributes such as age, gender, or race.

Transparency:

Make model decisions interpretable to stakeholders.

Privacy:

Protect the confidentiality of employee data.

Fair Use of Data:

Ensure data is used ethically and in accordance with privacy regulations and company policies.

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?

Check for missing values and understand why they are missing.

- What data initially presents as containing anomalies?

Outliers or unusual values in satisfaction level, average monthly hours, and promotion_last_5years.

- What additional types of data could strengthen this dataset?

Employee demographics (age, gender, education level)

Job roles and levels

Reasons for leaving (if available)

Employee feedback or survey data

Go Beyond the Numbers: Translate Data into Insights

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

High Attrition Factors: Low satisfaction levels, high average monthly hours, and lack of promotion are significant factors contributing to attrition.

Departmental Differences: Some departments have higher attrition rates, indicating possible issues within those departments.

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

Improve Job Satisfaction: Implement programs to boost employee satisfaction, such as flexible working hours and recognition programs.

Monitor Work Hours: Address high average monthly hours by promoting work-life balance initiatives.

Promotion Opportunities: Increase opportunities for career advancement to retain employees.

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

How does employee satisfaction vary by department and role?

What are the predictors of high performance in employees who stay?

How do salaries impact employee satisfaction and retention?

- How might you share these visualizations with different audiences?

Executives: Summarize key findings in a concise report with visual highlights.

HR Managers: Provide detailed visualizations with actionable insights.

Employees: Share high-level insights in town hall meetings or internal newsletters.

The Power of Statistics

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

Impact of Interventions: Whether interventions such as flexible hours or recognition programs have a significant impact on employee satisfaction and retention.

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

Implement Successful Interventions: Roll out successful interventions company-wide.

Monitor and Adjust: Continuously monitor the impact of these interventions and make adjustments as needed.

Regression Analysis: Simplify Complex Data Relationships

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

Understanding Relationships: Beta coefficients help understand the relationship between predictors and the outcome variable.

Making Informed Decisions: Knowing which factors have the strongest impact on attrition helps prioritize interventions.

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

Targeted Interventions: Focus on factors with the highest impact on attrition, such as improving satisfaction and reducing work hours.

Department-Specific Strategies: Develop tailored strategies for departments with high attrition rates.

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

Feature Selection: Remove non-significant predictors to simplify the model.

Data Enrichment: Add more relevant features to improve model accuracy.

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

Focus on Key Predictors: Implement strategies to improve key predictors like job satisfaction and promotion opportunities.

Monitor High-Risk Employees: Use the model to identify and support employees at risk of leaving.

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

Predictors of Attrition: Low satisfaction and high work hours are strong predictors of attrition.

Impact of Promotions: Lack of promotions significantly increases the likelihood of leaving.

- Do you have any ethical considerations at this stage?

Bias and Fairness: Ensure the model does not discriminate based on protected attributes.

Transparency: Make the model's decisions interpretable to stakeholders.

Privacy: Protect employee data privacy.

The Nuts and Bolts of Machine Learning

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

Important Predictors: Satisfaction level, average monthly hours, and promotions are critical predictors of attrition.

- What are the criteria for model selection?

Accuracy and Performance: Choose models with high accuracy and performance metrics.

Interpretability: Prefer models that are easy to interpret for stakeholders.

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

Model Fit: Check if the model makes logical sense and aligns with business understanding.

Error Analysis: Understand the implications of model errors in the business context.

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

Unimportant Features: Remove features that do not significantly contribute to the model.

Impact on Model: Test the model performance after removing these features.

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

How do different job roles impact attrition?

What are the predictors of high performance among retained employees?

How does employee engagement correlate with satisfaction and retention?

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

Python Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Documentation and Tutorials: Official documentation and online tutorials

- Is my model ethical?

Bias Mitigation: Ensure the model does not reinforce existing biases.

Transparency: Make the model's decision process clear to stakeholders.

- When my model makes a mistake, what is happening? How does that translate to my use case?
  Error Implications: Understand what happens when the model misclassifies an employee.

  Business Translation: Translate model errors to potential business impacts and address them.