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:

- Understand your data in the problem context
- Consider how your data will best address the business need
- Contextualize and understand the data and the problem
- Perform EDA (understand the variables and analyze relationships between them)
- Create visualizations
- Determine which models are most appropriate
- Construct the model
- Confirm model assumptions
- Evaluate model results to determine how well your model fits the data
- Interpret model performance and results
- Share actionable steps with stakeholders

Project proposal

Salifort Motors project proposal

Overview

Salifort Motors aims to leverage employee data to identify the key factors driving turnover and better understand the reasons behind employee departures.

Milestones	Tasks	PACE stages
1	Understand the business scenario and define the problem	Plan
2	Data exploration and data cleaning	Plan, Analyze
3	Determine which models are most appropriate	Analyze,Construct
4	Construct the model	Construct
5	Confirm model assumptions	Analyze, Construct
6	Evaluate model results	Analyze
7	Interpret results and share actionable steps with stakeholders	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of Data Science

- Who is your audience for this project?
- 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?
- What questions need to be asked or answered?
- What resources are required to complete this project?
- What are the deliverables that will need to be created over the course of this project?

Answers:

- **Audience**: The audience includes Salifort Motors' leadership team and Human Resources department.
- **Objective**: To analyze employee turnover trends and factors, create predictive models, and provide actionable recommendations to reduce turnover rates, saving costs and improving corporate culture.
- **Questions**: What factors contribute most to turnover? What interventions would improve retention? How accurate are the predictions across different models?
- **Resources**: Python (Pandas, NumPy, Scikit-learn), Jupyter Notebook, Tableau, HR_capstone_dataset.csv, Google and IBM certifications knowledge, ethical data practices.
- **Deliverables**: Jupyter Notebooks for EDA and modeling, Tableau visualizations, executive summary.

Get Started with Python

- How can you best prepare to understand and organize the provided information?
- What follow-along and self-review codebooks will help you perform this work?
- What are a couple additional activities a resourceful learner would perform before starting to code?

Answers:

- **Preparation:** Organize the dataset by checking data types, completeness, and standardization. Review code samples and past projects for feature engineering practices.
- **Codebooks:** Resources like Python documentation, Scikit-learn guides, and Tableau tutorials will ensure a structured and efficient workflow.

- **Additional Activities:** Study similar predictive modeling case studies, read up on turnover management strategies, and ensure familiarity with machine learning concepts.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
- What units are your variables in?
- 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?
- Is there any missing or incomplete data?
- Are all pieces of this dataset in the same format?
- Which EDA practices will be required to begin this project?

Answers:

- **Relevant Variables**: Satisfaction level, number of projects, monthly hours, time spent at the company, promotions, department, salary.
- **Units**: Satisfaction levels and evaluations are measured on a scale of 0-1, hours and tenure in numerical values, and salary/department as categorical variables.
- **Presumptions**: Employees with low satisfaction and high workload are likely to leave; tenure and salary could also play key roles.
- **Data Assessment**: Check for missing values, inconsistent formats, and outliers.
- **EDA Practices**: Visualizing distributions, correlations, trends, and interactions between variables.

The Power of Statistics

- What is the main purpose of this project?
- What is your research question for this project?
- 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?

Answers:

- **Purpose**: Predict employee turnover and offer solutions to improve retention and satisfaction.
- **Research Question**: What factors most significantly predict turnover?

- **Random Sampling Importance**: Ensures unbiased data representation. Without it, sampling bias may favor certain departments or tenure groups, skewing insights.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
- What are you trying to solve or accomplish?
- What are your initial observations when you explore the data?
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
- Do you have any ethical considerations at this stage?

Answers:

The primary stakeholders are:

- **Salifort Motors' Leadership Team**: Interested in actionable insights and strategic recommendations to reduce employee turnover.
- **Human Resources Department**: Focused on understanding factors driving turnover and implementing retention strategies based on the findings.
- **Employees**: Indirect stakeholders, as the outcomes of the analysis will impact policies, work culture, and job satisfaction.

What Are You Trying to Solve or Accomplish?

The goal is to:

1. Identify key factors that influence employee turnover at Salifort Motors.
2. Develop a regression model to predict the likelihood of an employee leaving the company.
3. Use insights from the analysis to recommend actionable strategies that enhance retention, reduce costs, and foster a positive corporate culture.
4. This work aims to create a data-driven approach to turnover management, aligning business objectives with employee needs.

Initial Observations When Exploring the Data

Upon initial exploration of the dataset, the following observations emerged:

- **Satisfaction Levels**: A significant range is visible, with potential clusters at high and low satisfaction levels.
- **Workload**: Employees with a higher number of projects and monthly hours seem to correlate with higher turnover rates.
- **Salary**: Employees in the "low" salary tier may be disproportionately represented among those who have left.
- **Tenure**: Longer tenures appear to show trends of dissatisfaction, requiring further investigation.
- **Promotion and Evaluation**: Lack of promotions and high performance reviews for employees working extensive hours suggest potential cultural or workload issues.

Resources Used During This Stage

To complete this stage effectively, the following resources are being utilized:

- **Python Libraries**:
  - Pandas and NumPy for data exploration and cleaning.
  - Seaborn and Matplotlib for visualizations.
  - Scikit-learn for regression model building and evaluation.
- **Documentation and Guides**:
  - Scikit-learn User Guide
  - Pandas Documentation
  - Seaborn Documentation
- **Best Practices**: Google Advanced Data Analytics coursework, IBM AI Developer certification materials.
- **Additional Tools**: Tableau for advanced data visualization and presentation.

Ethical Considerations at This Stage

Key ethical considerations include:

- **Data Privacy**: Ensuring employee data is anonymized to protect individual identities.
- **Bias Avoidance**: Carefully examining the dataset for potential biases that could lead to unfair conclusions or recommendations, such as overrepresentation of certain departments or roles.

- **Transparency**: Communicating limitations of the analysis and ensuring stakeholders understand the assumptions and constraints of the regression model.
- **Fair Policy Recommendations**: Ensuring the insights derived from the model do not inadvertently reinforce inequities, such as unfair evaluation systems or discriminatory practices.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
- What resources do you find yourself using as you complete this stage?
- Is my data reliable?
- Do you have any additional ethical considerations in this stage?
- What data do I need/would I like to see in a perfect world to answer this question?
- What data do I have/can I get?
- What metric should I use to evaluate success of my business objective? Why?

Answers:

What Am I Trying to Solve?

The primary objective is to develop predictive machine learning models to identify employees likely to leave the company. This involves understanding the key factors contributing to turnover and using these insights to guide data-driven decisions aimed at improving retention, reducing costs, and enhancing employee satisfaction.

What Resources Do You Find Yourself Using as You Complete This Stage?

To complete this stage, the following resources are being utilized:

- **Python Libraries**:
  - Scikit-learn: For building and evaluating machine learning models.
  - Pandas and NumPy: For data cleaning, transformation, and feature engineering.
  - Matplotlib and Seaborn: For visualizing results and exploring feature importance.
- **Documentation and Tutorials**:
  - Scikit-learn Documentation
  - Pandas Documentation
  - Seaborn Documentation
  - Google Advanced Data Analytics course material for model evaluation strategies.
- **Visualization Tools**: Tableau for presenting key findings to stakeholders.
- **Guided Frameworks**: The provided Python notebook with pre-structured coding templates.

Is My Data Reliable?

The dataset is generally reliable, with self-reported data covering satisfaction, work hours, tenure, and other attributes. However, there are considerations:

- **Potential Bias**: Self-reported data might contain inherent biases, such as over- or underestimation of satisfaction levels.
- **Completeness**: Missing values or inconsistencies need to be identified and handled appropriately.
- **Recency**: Ensure the dataset reflects current trends and patterns relevant to turnover today.

Do You Have Any Additional Ethical Considerations in This Stage?

Key ethical considerations at this stage include:

- **Fairness in Model Outcomes: Ensuring the model does not disproportionately affect specific departments, salary tiers, or other employee groups.**
- **Data Privacy: Protecting employee anonymity by not exposing sensitive or identifiable information in the analysis.**
- **Transparency: Clearly communicating the model's limitations and potential biases to stakeholders to avoid misinterpretation or misuse of results.**

What Data Do I Need/Would I Like to See in a Perfect World to Answer This Question?

In an ideal scenario, additional data that would strengthen the analysis includes:

- **Employee feedback or sentiment analysis from surveys.**
- **Historical turnover trends by department, team, or region.**
- **External benchmarks or industry data on turnover.**
- **Detailed information on overtime policies and workload distribution.**

What Data Do I Have/Can I Get?

The available data includes:

- **Satisfaction levels, performance evaluations, tenure, average work hours, number of projects, salary, and promotion history.**
- **Categorical data on departments and salary tiers.**
  While this dataset provides a solid foundation, supplemental data collection (e.g., feedback surveys) could further enhance the analysis.

What Metric Should I Use to Evaluate Success of My Business Objective? Why?

The following metrics will evaluate the success of the predictive models:

- **AUC (Area Under the ROC Curve): To measure the model's ability to distinguish between employees likely to leave and those who stay. A high AUC indicates good performance.**
- **Precision: To focus on how many of the predicted positive cases (employees likely to leave) are actually correct. This is vital for targeted interventions.**

- **Recall: To ensure the model identifies as many employees likely to leave as possible, avoiding missed cases.**
- **F1-Score: To balance precision and recall, especially when turnover cases might be imbalanced in the dataset.**
- **Accuracy: Useful as a general measure, but secondary to precision and recall in this context due to class imbalance concerns.**

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?

Answers:

- **Sufficiency**: The dataset is rich and diverse, providing enough information to establish predictive trends. Additional data such as employee feedback could further strengthen analysis.

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

Answers:

- **EDA Steps**: Analyze variable distributions, correlations, and trends. Filter, sort, and clean data for consistency.
- **Joining Additional Data**: Adding external employee sentiment or industry benchmarks would enhance context.
- **Visualization Assumptions**: Heatmaps for correlations, bar charts for categorical data, and scatter plots for workload impact.

The Power of Statistics

- Why are descriptive statistics useful?
- What is the difference between the null hypothesis and the alternative hypothesis?

Answers:

- **Utility of Descriptive Statistics**: Summarize key trends like satisfaction and tenure distributions to inform models and stakeholders.
- **Hypotheses**:
    - Null Hypothesis: Employee turnover is random and not affected by the analyzed variables.
    - Alternative Hypothesis: Turnover is influenced by satisfaction level, workload, tenure, and salary.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
- Do you have any ethical considerations at this stage?

Answers:

- **EDA Purpose**: Identify variable relationships, remove redundancies, and ensure assumptions for modeling.
- **Ethical Considerations**: Anonymize sensitive data and maintain fairness in interpretation.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
- Why did you select the X variables you did?
- What are some purposes of EDA before constructing a model?
- What has the EDA told you?
- What resources do you find yourself using as you complete this stage?
- Do you have any ethical considerations in this stage?

What Am I Trying to Solve? Does It Still Work? Does the Plan Need Revising?

The goal is to develop machine learning models to predict employee turnover at Salifort Motors. This involves identifying key factors driving turnover and providing actionable insights to improve employee retention.

- **Does It Still Work?** Yes, the current approach—focusing on predictive modeling using key employee data—continues to align with the objectives of the project.

- **Does the Plan Need Revising?** The plan remains effective, but adjustments may be required if certain features (e.g., missing data or unexpected correlations) impact model reliability. If the data reveals unexpected patterns, additional EDA or feature engineering may be necessary.

Does the Data Break the Assumptions of the Model? Is That Acceptable?

- **Breaking Assumptions**:
    - Logistic regression assumes linearity between predictors and the log-odds of the dependent variable. If non-linear relationships exist, they may violate these assumptions.
    - Tree-based models (e.g., Decision Tree, Random Forest, XGBoost) are more flexible and do not rely on linearity, making them more robust for this dataset.
- **Acceptability**: Violations of assumptions in linear models may necessitate feature transformations or selection of non-linear models. For example, if the data shows skewness, normalization or log transformations might be required.

Why Did You Select the X Variables You Did?

The X variables were selected based on their potential influence on employee turnover, supported by domain knowledge and EDA results. Key variables include:

- **Satisfaction Level**: A critical indicator of employee morale.
- **Number of Projects**: Helps gauge workload.
- **Average Monthly Hours**: Reflects work-life balance.
- **Time Spent at the Company**: Indicates employee tenure and potential burnout.
- **Department and Salary**: Provides insight into team dynamics and financial incentives.
  These variables were chosen for their direct relevance to the research question and their likely impact on turnover.

What Are Some Purposes of EDA Before Constructing a Model?

EDA plays a crucial role in:

1. **Understanding the Dataset**: Identifying trends, distributions, and correlations to inform model building.

2. **Detecting Issues**: Spotting outliers, missing values, or inconsistent formats that could compromise model performance.
3. **Feature Selection**: Recognizing which variables are most impactful for predicting turnover.
4. **Checking Assumptions**: Ensuring data meets the requirements for the planned models, such as linearity in logistic regression.
5. **Hypothesis Formation**: Generating ideas about relationships between variables and turnover.

What Has the EDA Told You?

EDA has revealed several insights:

- Employees with low satisfaction levels and high workloads are more likely to leave.
- Long-tenured employees without promotions show higher turnover rates, suggesting dissatisfaction.
- Departments with "low" salary employees are at higher risk of turnover.
- High monthly hours correlate with higher performance reviews but also higher turnover, indicating potential burnout.

What Resources Do You Find Yourself Using as You Complete This Stage?

The following resources are instrumental:

- **Python Libraries**: Pandas, NumPy (for data preprocessing), Scikit-learn (for model building and evaluation), Seaborn/Matplotlib (for visualization).
- **Documentation and Tutorials**:
    - Scikit-learn Documentation
    - Pandas Documentation
    - Seaborn Documentation
- **Google Capstone Materials**: Templates and coding frameworks provided in the course.
- **Visualization Tools**: Tableau for stakeholder presentations.

Do You Have Any Ethical Considerations in This Stage?

Ethical considerations include:

- **Bias in Data**: Ensuring the model does not unintentionally favor or disadvantage specific departments, salary tiers, or tenures.
- **Data Privacy**: Anonymizing employee data to protect identities.
- **Transparency**: Clearly communicating model limitations and potential errors to avoid misuse of predictions.
- **Fair Outcomes**: Ensuring insights and recommendations are equitable and do not reinforce systemic biases.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
- How many vendors, organizations or groupings are included in this total data?

Answers:

- **Variable Averages**: Spot anomalies such as unusually high monthly hours or evaluations.
- **Groupings**: Department categories, salary tiers, tenure ranges.

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?
- What processes need to be performed in order to build the necessary data visualizations?
- Which variables are most applicable for the visualizations in this data project?
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

Answers:

- **Visualizations Needed**: Tableau dashboards for turnover trends and heatmaps for correlation analysis.
- **Processes**: Filter out irrelevant features, aggregate data for grouping, and normalize numerical variables.
- **Missing Data**: Use imputation or flag missing values for further analysis.

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
- What conclusion can be drawn from the hypothesis test?

Answers:

- **Hypotheses Formulation**: Grounded in initial data review; test variables against turnover likelihood.
- **Conclusions**: Statistical significance for satisfaction and workload factors influencing turnover.

Regression Analysis: Simplify Complex Data Relationships

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

Answers:

- **Odd Observations**: Disproportionate turnover in certain departments; explore root causes.
- **Improvements**: Optimize regression parameters or add relevant features.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?
- Which independent variables did you choose for the model, and why?
- How well does your model fit the data? (What is my model's validation score?)
- Can you improve it? Is there anything you would change about the model?
- Do you have any ethical considerations at this stage?

Answers:

- **Problems and Fixes**: Address overfitting or underfitting through hyperparameter tuning.
- **Independent Variables**: Selected based on correlation and domain relevance (e.g., satisfaction level, workload).
- **Validation Scores**: Achieve high accuracy and precision while balancing recall.

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?
- What data initially presents as containing anomalies?
- What additional types of data could strengthen this dataset?

Answers:

- **Initial Recommendations**: Investigate workload distribution and satisfaction ratings before modeling.
- **Anomalies**: Flag extreme monthly hours or unusually low satisfaction levels.
- **Additional Data**: Surveys on employee motivation or feedback on work culture.

Go Beyond the Numbers: Translate Data into Insights

- What key insights emerged from your EDA and visualizations(s)?
- What business recommendations do you propose based on the visualization(s) built?
- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
- How might you share these visualizations with different audiences?

Answers:

- **Key Insights**: Overwork and lack of promotions significantly contribute to turnover.
- **Business Recommendations**: Adjust workload, reward tenure with promotions, and improve clarity on expectations.
- **Further Questions**: Explore retention strategies for low-performing departments.
- **Visualization Sharing**: Tailor Tableau dashboards for leadership presentations or HR workshops.

The Power of Statistics

- What key business insight(s) emerged from your A/B test?
- What business recommendations do you propose based on your results?

Answers:

- **A/B Test Insights**: Low workload groups may have higher satisfaction levels, influencing turnover rates positively.
- **Recommendations**: Adopt proportional workload policies.

Regression Analysis: Simplify Complex Data Relationships

- To interpret model results, why is it important to interpret the beta coefficients?
- What potential recommendations would you make to your manager/company?
- Do you think your model could be improved? Why or why not? How?
- What business recommendations do you propose based on the models built?
- What key insights emerged from your model(s)?
- Do you have any ethical considerations at this stage?

Answers:

- **Beta Coefficients**: Crucial for understanding variable impact (e.g., how satisfaction influences turnover likelihood).
- **Model Improvements**: Remove potential data leakage features (e.g., last evaluation).

The Nuts and Bolts of Machine Learning

- What key insights emerged from your model(s)?
- What are the criteria for model selection?
- Does my model make sense? Are my final results acceptable?
- Were there any features that were not important at all? What if you take them out?
- Given what you know about the data and the models you were using, what other questions could you address for the team?
- What resources do you find yourself using as you complete this stage?
- Is my model ethical?
- When my model makes a mistake, what is happening? How does that translate to my use case?

Answers:

- **Insights**: XGBoost performs optimally, highlighting employee workload and satisfaction as primary turnover drivers.
- **Criteria for Selection**: AUC and recall metrics indicate robustness in prediction accuracy.
- **Ethical Considerations**: Ensure fairness in recommendations across departments and salary tiers.
- **Mistakes in Model**: Explore impacts of errors on prediction to fine-tune real-world use cases.