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 Retention project proposal

Overview

This project aims to enhance employee retention at Salifort Motors by leveraging data analytics to identify key factors influencing turnover. The analysis will involve exploring a dataset with variables such as job satisfaction, performance scores, project involvement, and other relevant factors, leading to the development of a predictive model.

Milestones	Tasks	PACE stages
Project Initiation	**Define project scope and objectives**	**Plan**
	Identify stakeholders and their expectations	**Plan**
Data Exploration	**Familiarize with the HR dataset**	**Plan**
	Explore variables and initial observations	**Analyze**
Analysis & Modelisation	**Analyze key factors influencing turnover**	**Analyze**
Development	**Select appropriate statistical and ML models**	**Analyze, Construct**
	Build predictive model for employee turnover	**Construct**

Evaluation	Assess model performance and accuracy	Analyze
	Fine-tune model based on evaluation results	Construct
Recommendations	Derive data-driven suggestions for HR	Execute
Implementation	Support HR in implementing recommendations	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?

  The primary audience includes the HR department at Salifort Motors and the senior leadership team. Additionally, other relevant stakeholders within the organization may be interested in the findings.
- 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 main objective is to identify factors influencing employee turnover and provide data-driven suggestions to enhance retention. The impact is expected to result in improved employee satisfaction, reduced turnover, and cost savings for the company.
- What questions need to be asked or answered?
  - *What are the key factors contributing to employee turnover?*
  - *Can we predict which employees are likely to leave the company?*
  - *What actionable recommendations can be provided to improve retention?*
- What resources are required to complete this project?
  - *The HR dataset from Salifort Motors.*
  - *Data analysis tools such as Jupyter Notebook.*
  - *Statistical and machine learning libraries for model development.*
  - *Communication channels for stakeholder interaction.*
- What are the deliverables that will need to be created over the course of this project?
  - *Project proposal document.*
  - *Exploratory Data Analysis (EDA) report.*
  - *Trained predictive model.*
  - *Recommendations for improving employee retention.*

Get Started with Python

- How can you best prepare to understand and organize the provided information?
  Careful review of the dataset documentation, understanding the business context, and breaking down the problem into manageable components.
- What follow-along and self-review codebooks will help you perform this work?
  Codebooks related to data exploration, preprocessing, model development, and evaluation will be essential for a systematic approach.
- What are a couple additional activities a resourceful learner would perform before starting to code?

  Conducting background research on employee turnover trends in the industry and exploring best practices for retention strategies.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
  Relevant variables include satisfaction level, last evaluation score, number of projects, average monthly hours, time spent with the company, work accidents, promotions, department, and salary.
- What units are your variables in?
  Variables like satisfaction level and last evaluation score are on a scale from 0 to 1, while others have different units (e.g., hours, years).
- 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?
  Presumptions may include expecting a negative correlation between satisfaction level and turnover, and a positive correlation between the number of projects and turnover.
- Is there any missing or incomplete data?
  This needs to be confirmed during the EDA stage.
- Are all pieces of this dataset in the same format?
  Confirming consistency in data formats is crucial during the EDA.
- Which EDA practices will be required to begin this project?

  Descriptive statistics, data distribution analysis, correlation analysis, and data visualization to gain insights into the dataset.

The Power of Statistics

- What is the main purpose of this project?
  The main purpose is to identify factors influencing employee turnover and provide actionable recommendations for improving retention.
- What is your research question for this project?
  "What factors contribute to employee turnover at Salifort Motors, and how can the company use this information to improve employee retention?"
- What is the importance of random sampling? In this case, what is an example of sampling bias that might occur if you didn't use random sampling?

  Random sampling ensures a representative subset of data is used for analysis. Without it, sampling bias might occur if, for example, only data from a specific department or level of employment is considered, leading to skewed results.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
  The HR department and senior leadership team at Salifort Motors.
- What are you trying to solve or accomplish?
  Identify factors influencing employee turnover and provide data-driven suggestions to enhance retention.
- What are your initial observations when you explore the data?
  Variables like satisfaction level, last evaluation score, and number of projects seem particularly relevant for predicting turnover.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
  The dataset and my jupyter notebook.
- Do you have any ethical considerations at this stage?

  Ensure the privacy and confidentiality of employee data, avoid biased practices, and communicate the purpose and implications of the analysis to stakeholders clearly.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?

  Identify factors influencing employee turnover and provide actionable recommendations for retention.
- What resources do you find yourself using as you complete this stage?

  The dataset and my jupyter notebook.
- Is my data reliable?

  This needs to be assessed during the data exploration (analyze stage).
- Do you have any additional ethical considerations in this stage?

  Maintain privacy and confidentiality of employee data and avoid biased practices.
- What data do I need/would I like to see in a perfect world to answer this question?

  Detailed data on employee experiences, reasons for leaving, and external factors influencing turnover.
- What data do I have/can I get?

  The provided dataset includes variables such as satisfaction level, last evaluation score, number of projects, average monthly hours, time spent with the company, work accidents, promotions, department, and salary.
- What metric should I use to evaluate success of my business objective? Why?

  Success can be measured by the reduction in employee turnover and an improvement in overall employee satisfaction, contributing to long-term company success.

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?

  Last evaluation, number of projects, tenure, somehow a measure of the amount of hours the employee works.

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?

  Get a grip of the data, clean it and explore it. Engineer features if necessary. Sample, visualize and prepare the data.

- 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.?

  Filtering duplicate employees and outliers. Rename columns' names to be more descriptive.

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

  Boxplots, barplots, histograms, scatterplots, correlation heatmaps, confusion matrices, Decision Tree Map.

  Barplots are mostly for executives, the rest is for the data team.

The Power of Statistics

- Why are descriptive statistics useful?

  Descriptive statistics play a crucial role in summarizing and presenting the main features of a dataset. They provide a concise and informative overview, allowing for a better understanding of the data

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

The null hypothesis is a statement of no effect, no difference, or no change in the population parameter under investigation. It represents the status quo or the default assumption. Often denoted as H0, it is what the researcher aims to test against or challenge. Whereas the alternative hypothesis is the statement that contradicts the null hypothesis. It represents what the researcher is trying to show or demonstrate through statistical analysis. It can take various forms, such as asserting a difference, an effect, or a change in the population parameter.

Regression Analysis: Simplify Complex Data Relationships

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

  EDA:

  - helps in understanding the distribution of each variable, identifying outliers, and assessing the need for data transformations to meet the assumptions of linear regression,

  - explores relationships between the independent variables and the dependent variable,

  - allows for the identification and handling of missing data, ensuring that the dataset is complete before building the regression model,

  - helps in detecting outliers that might disproportionately influence the regression coefficients,

  - assesses whether the relationships between the independent variables and the dependent variable are linear, a fundamental assumption for multiple linear regression,

  - examines the presence of multicollinearity, where independent variables are highly correlated, which can affect the stability and interpretability of regression coefficients,

  - guides decisions on variable transformations, such as log transformations, to improve linearity and meet assumptions of normality,

  - helps in deciding which variables to include in the model based on their relationships with the dependent variable and their significance,

  - informs data preprocessing steps, such as handling categorical variables, scaling, or standardizing variables, ensuring that the data is suitable for regression analysis.

- Do you have any ethical considerations in this stage?

- Ensure that sensitive information is handled with care, and privacy laws and regulations are adhered to. Avoid the use of personally identifiable information unless absolutely necessary.

- Stay mindful of potential biases in the data and algorithms. Analyze and address biases that might arise from historical data, sampling methods, or algorithmic decisions to prevent unfair outcomes.

- Maintain transparency in the analysis process. Clearly communicate the methods, assumptions, and limitations of the analysis to stakeholders and ensure that results are interpretable.

- Implement robust data security measures to protect against unauthorized access, data breaches, or misuse of information. Follow best practices for data encryption and storage.

- Respect data ownership rights and restrictions. Obtain proper authorization before accessing or using datasets, especially when dealing with proprietary or confidential information.

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
- To predict employees likely to quit
- To identify factors that contribute to their leaving

  Yes but the data needs to be tweaked slightly.
- Does the data break the assumptions of the model? Is that ok, or unacceptable?

  Homoscedasticity assumption is violated but that's not a problem if we use Random Forest or a Decision Tree.
- Why did you select the X variables you did?

  The independent variables chosen for the model include features related to employee satisfaction, performance evaluation, number of projects, average monthly hours worked, time spent with the company, work accidents, promotion status, department, and salary level. These features were selected based on their potential relevance to predicting employee turnover.
- What are some purposes of EDA before constructing a model?

  - Understand the distribution and characteristics of individual variables.

  - Identify patterns, trends, and relationships between variables.

  - Detect outliers or anomalies that may impact model performance.

- Inform feature engineering and selection.

- Validate assumptions and check data quality.

- Guide the selection of appropriate modeling techniques.

- What has the EDA told you?

  Poor management seems to be the main reason why employees are quitting the company. They have to deal with long working hours, multiple projects, and low satisfaction levels. This can make them feel unappreciated and frustrated. Many employees in this company may be suffering from burnout. Moreover, employees who have stayed for more than six years are less likely to quit.

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

  - **Seaborn Documentation:** https://seaborn.pydata.org/

  - **Matplotlib Documentation:** https://matplotlib.org/

  - **Pandas Documentation:** https://pandas.pydata.org/pandas-docs/stable/

- Do you have any ethical considerations in this stage?

  - Report results accurately and responsibly. Avoid sensationalism and ensure that the interpretation of findings is unbiased and supported by the data.

  - Be open to feedback from stakeholders and the broader community. If ethical concerns are raised, be willing to revisit and revise your analysis in response to valid ethical considerations.

  - Be accountable for the implications of the analysis. Consider the potential impact of your findings on individuals, communities, or organizations and take responsibility for ethical decision-making.

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
  Without the specific data and variable details, it's challenging to pinpoint unusual averages. However, during the analysis stage, we looked at various visualizations and statistics to understand the distribution of variables and identify any anomalies.
- How many vendors, organizations or groupings are included in this total data?

  There are 10 departments in the company.

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 such as histograms, bar charts, boxplots, and correlation heatmaps. Additionally, machine learning algorithms, including logistic regression, random forest, SVM, XGBoost, and a neural network, could be implemented for predicting employee turnover.
- What processes need to be performed in order to build the necessary data visualizations?
  data preprocessing, encoding categorical variables, and scaling features, training and evaluation.
- Which variables are most applicable for the visualizations in this data project?
  Variables such as satisfaction level, last evaluation, number of projects, average monthly hours, time spent at the company, work accident status, promotion status, department, and salary were crucial for visualizations and model building.
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

  The approach to handling missing data wasn't explicitly mentioned in the provided context. However, typical strategies include imputation, removal of missing values, or using models that can handle missing data.

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
  - *Null Hypothesis (H0): The average satisfaction level of employees who left the company is the same as the average satisfaction level of employees who stayed.*
  - *Alternative Hypothesis (H1): The average satisfaction level of employees who left the company is different from the average satisfaction level of employees who stayed.*
- What conclusion can be drawn from the hypothesis test?

  The conclusion is that we reject the null hypothesis. There is enough evidence to suggest a difference in satisfaction levels between employees who left the company and those who stayed.

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?
  The residuals are not randomly scattered around the horizontal axis, which suggests that the logistic regression model may not be a good fit for the data. The predicted values are not evenly distributed along the horizontal axis, which suggests that the model may not be capturing the relationship between the variables well. Logistic Regression model's assumptions may not hold, or there may be influential points or outliers in the data that are affecting the model's performance.
- Can you improve it? Is there anything you would change about the model?
  It might be beneficial to investigate these issues further and consider alternative modeling approaches or data transformations.

The Nuts and Bolts of Machine Learning

- Is there a problem? Can it be fixed? If so, how?
  There are outliers that need to be filtered but most importantly, there is multicollinearity among independent variables.
- Which independent variables did you choose for the model, and why?
  'satisfaction_level', 'last_evaluation', 'number_project', 'average_monthly_hours', 'tenure' because they have the most impact (given the available data) on employee turnover.
- How well does your model fit the data? (What is my model's validation score?)
  The best random first model achieves a 93.84% AUC score with a recall and accuracy higher than 90% with 90.36% and 96.16% respectively while precision was slightly lower at 87%.
- Can you improve it? Is there anything you would change about the model?

We could try XGBoost and optimize its parameter too, create new feature as its seem to help and also maybe add more granular data about user satisfaction or rather dissatisfaction. Which elements makes employee turnover increase after 4 years in the company?

- Do you have any ethical considerations in this stage?

  The model should not be used for anything than its intended purposes.

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?

  conduct further investigation about why four-year tenured employees are so dissatisfied.

- What data initially presents as containing anomalies?

  The amount of years presented some outliers but it was in fact a class really well defined.

- What additional types of data could strengthen this dataset?

  Granular data about employee dissatisfaction points. If they have been made aware of important policies or not, etc.

Go Beyond the Numbers: Translate Data into Insights

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

  Poor management seems to be the main reason why employees are quitting the company. They have to deal with long working hours, multiple projects, and low satisfaction levels. This can make them feel unappreciated and frustrated. Many employees in this company may be suffering from burnout. Moreover, employees who have stayed for more than six years are less likely to quit.

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

□ Cap the number of projects that employees can work on.
□ Consider promoting employees who have been with the company for at least four years, or conduct
□ further investigation about why four-year tenured employees are so dissatisfied.
□ Either reward employees for working longer hours, or don't require them to do so.
□ If employees aren't familiar with the company's overtime pay policies, inform them about this. If the
□ expectations around workload and time off aren't explicit, make them clear.
□ Hold company-wide and within-team discussions to understand and address the company work
□ culture, across the board and in specific contexts.
□ High evaluation scores should not be reserved for employees who work 200+ hours per month.
□ Consider a proportionate scale for rewarding employees who contribute more/put in more effort.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
  - *why four-year tenured employees are so dissatisfied?*
  - *Are employees familiar with the company's overtime pay policies?*

- How might you share these visualizations with different audiences?

By tailoring the visualizations to the target audience based on their areas of expertise and stake in the project.

The Power of Statistics

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

  The random forest model slightly outperforms the decision tree model.

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

  Use the model to identify which factors are most influential. These insights can help HR make decisions to improve employee retention.

Regression Analysis: Simplify Complex Data Relationships

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

  Interpreting beta coefficients in a regression model is crucial as they provide insights into the relationships between the independent variables and the dependent variable. Each beta coefficient represents the change in the mean of the dependent variable for a one-unit change in the corresponding independent variable, while holding other variables constant.

- What potential recommendations would you make to your manager/company?
  - ☐ Cap the number of projects that employees can work on.
  - ☐ Consider promoting employees who have been with the company for at least four years, or conduct
  - ☐ further investigation about why four-year tenured employees are so dissatisfied.
  - ☐ Either reward employees for working longer hours, or don't require them to do so.
  - ☐ If employees aren't familiar with the company's overtime pay policies, inform them about this. If the
  - ☐ expectations around workload and time off aren't explicit, make them clear.
  - ☐ Hold company-wide and within-team discussions to understand and address the company work
  - ☐ culture, across the board and in specific contexts.
  - ☐ High evaluation scores should not be reserved for employees who work 200+ hours per month.
  - ☐ Consider a proportionate scale for rewarding employees who contribute more/put in more effort.

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

  Yes by adding more relevant data and features, maybe try new architectures to model turnover and maybe add a wider hyperoptimization parameter range.

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

Use the model to identify which factors are most influential. These insights can help HR make decisions to improve employee retention.

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

  The most relevant variables for the decision tree were: 'last_evaluation', 'number_project', 'tenure' and 'overworked'.

- Do you have any ethical considerations at this stage?

  This model should only be used to predict whether an employee will leave the company and identify which factors are most influential. These insights should help HR make decisions to improve employee retention.

The Nuts and Bolts of Machine Learning

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

  In the random forest model, `last_evaluation`, `tenure`, `number_project`, `overworked`, `salary_low`, and `work_accident` have the highest importance. These variables are most helpful in predicting turnover.

- 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?

  Only if used to predict whether an employee will leave the company and identify which factors are most influential and that these insights are used by HR to make decisions to improve employee retention.

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

  Either an employee about to leave is not detected before is departure or a loyal employee is identified as leaving. In either case it is not really worst than not having the model but its always safe for HR to verify the model's insights on the ground.