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:

Introduction to Python

- Demonstrate an understanding of the form and function of Python programming.

Utilization of Python by Data Professionals

- Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions.

Data Organization and Analysis

- Demonstrate the ability to organize and analyze a dataset to uncover the underlying "story."

Exploratory Data Analysis (EDA)

- Create a Jupyter notebook dedicated to exploratory data analysis.

Data Visualization

- Create visualizations using Tableau to represent the findings clearly.

Statistical Analysis

- Use Python to compute descriptive statistics and conduct hypothesis testing.

Model Building

- Build a multiple linear regression model and perform ANOVA testing on the results.

Model Evaluation

- Evaluate the performance and accuracy of the regression model.

Machine Learning Models

- Demonstrate the ability to use a notebook environment to create a series of machine learning models on a dataset to address a specific problem.

Executive Summary

- Articulate the findings in an executive summary intended for external stakeholders.

Project proposal

Salifort Motors Employee Retention Project

Overview

The goal is to identify the causes of employee turnover and create a model to predict it.

Milestones	Tasks	Deliverables	PACE stages
1	Establish structure for project workflow (PACE)	• **Global-level project document.**	**Plan**
1a	Write a project proposal		**Plan**
2	Compile summary information about the data	• **Data files ready for EDA.**	**Analyze**
2a	Begin exploring the data		**Analyze**
3	Data exploration and cleaning	• **EDA report.**	**Plan / Analyze**
3a	Visualization building	• **Tableau dashboard/visualizations.**	**Analyze / Construct**
4	Compute descriptive statistics	• **Analysis of testing results between important variables.**	**Analyze**
4a	Conduct hypothesis testing		**Analyze / Construct**
5	Build a regression model		**Analyze / Construct**

5a	Evaluate the model	• **Determine the success of the model.**	**Execute**
6	Build a machine learning model	• **Final model.**	**Construct**
6a	Communicate final insights with stakeholders	• **Report to all stakeholders.**	**Execute**

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?
    - The Human Resources department is a key stakeholder in this project. The insights gained from the analysis will ideally enable the company to reduce employee turnover while improving retention and satisfaction.
- 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 objective is to understand the reasons behind employee turnover and ultimately create a model that can predict it.
- What questions need to be asked or answered?
    - How was the data collected?
    - How far back does the data go?
- What resources are required to complete this project?
    - The employee data.
- What are the deliverables that will need to be created over the course of this project?
    - A project proposal, including descriptive statistics, EDA reports, visualizations, model results, and an executive summary outlining accomplishments and next steps.

Get Started with Python

- How can you best prepare to understand and organize the provided information?
    - Complete some preliminary analysis of the data to ensure it looks complete. Reference the data dictionary, and ensure all required tools are available and ready to be used.
- What follow-along and self-review codebooks will help you perform this work?
    - All previous workbooks at this point are relevant, and can be referenced as guidelines for completing the various stages of the project.
- What are a couple additional activities a resourceful learner would perform before starting to code?

  - Reviewing related material to acquire an understanding of how the flow of the analysis should go. Reference a data analysis template to get an idea of the general structure of a data analysis project.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
  - At this point, it appears the most relevant variables are:
    - Dependent variable: ‘left’
    - Features: ‘satisfaction_level’, ‘number_project’, ‘average_monthly_hours’, ‘promotion_last_5years’, ‘salary’.
- What units are your variables in?
  - There is a mix of float64, int64 and object variables.
- 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?
  - Without having completed any analysis yet, my assumption at this point is that ‘satisfaction_level’ and ‘salary’ features will have predictive signal.
- Is there any missing or incomplete data?
  - There does not appear to be any missing or incomplete data.
- Are all pieces of this dataset in the same format?
  - No, there are three types of variables in this dataset, floats, ints and objects (categorical variables.) The objects have multiple classes, and the floats and ints are discrete and continuous numeric variables.
- Which EDA practices will be required to begin this project?
  - Checking the data distributions for skew, outliers, missing and incomplete data, imputation or dropping of rows with missing values. Additionally checking for duplicate data and removing those rows is key to ensure the models built aren’t skewed.

The Power of Statistics

- What is the main purpose of this project?
  - To identify the key variables associated with employee churn.
- What is your research question for this project?

  - What is causing employee churn at Salifort Motors? And, armed with this knowledge, can a predictive model be built to mitigate it?
- 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?
  - If random sampling is not utilized, you may inadvertently collect data from one group over another, which would skew the results of the project.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
  - Salifort Motors HR department, who will utilize this information to implement measures to reduce employee churn.
- What are you trying to solve or accomplish?
  - Identify the cause of employee churn at Salifort Motors so HR can implement measures to improve retention.
- What are your initial observations when you explore the data?
  - The work culture at Salifort Motors appears to be quite intense, with a large portion of employees consistently working overtime and managing more than 3-4 projects at a time.
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
  - Previous project notebooks.
- Do you have any ethical considerations in this stage?
  - Through initial analysis, it became clear many employees are overworked at the company. Although the the point of the analysis is not to have an opinion on work culture, it does appear it's an issue that needs to be addressed.

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
  - Build a machine learning model that can effectively predict employee churn.
- What resources do you find yourself using as you complete this stage?
  - Referencing previous notebooks.

- Is my data reliable?
  - It appears so. Throughout EDA, any outstanding issues were addressed, such as missing values, outliers and so forth.
- Do you have any additional ethical considerations in this stage?
  - Only the ethical considerations I mentioned prior, regarding the work culture of the company itself.
- What data do I need/would I like to see in a perfect world to answer this question?
  - More granular data pertaining to work habits, daily work routines, etc.
- What data do I have/can I get?
  - I already have access to higher-level data, such as average monthly hours, assigned projects, satisfaction levels, etc. More granular data would certainly be ideal to make analysis potentially even more thorough and to develop even more accurate models.
- What metric should I use to evaluate success of my business objective? Why?
  -

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?
  - It appears at this stage, the available data will be sufficient to at the very least gain basic insight into employee retention.

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?
  - Iteratively completing the steps necessary to handle outliers, missing data, normalizing features and so forth.
- 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.?
  - So far, additional data has not been necessary to add. But, that may change after the first iteration of the model is complete. In terms of structuring, outliers were removed from the 'tenure' variable, but the underlying structure of the data has not been changed.
- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
  - Boxplots will be useful for identifying outliers and histograms will be necessary for determining distributions.

The Power of Statistics

- Why are descriptive statistics useful?
  - They allow you to acquire a granular look at proportions, such as comparing groups to identify trends or patterns. They allow you to check for medians, means, standard deviations and min/max values.
- What is the difference between the null hypothesis and the alternative hypothesis?
  - The null hypothesis is the default state of no significance. The alternative hypothesis is the alternate state in which there is significance, and if so, you can reject the null hypothesis in favour of the alternative.

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
  - EDA allows you to ensure issues are managed prior to building a regression model. Many models require certain assumptions to be met prior to model development, otherwise their outcomes will potentially be incorrect.
- Do you have any ethical considerations in this stage?
  -

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?

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
    - Considering most variables are either normally or approximately normally distributed, none of the averages looked unusual. The variable tenure contained some outliers, which skewed the average a little.
- How many vendors, organizations or groupings are included in this total data?
    - There are 10 different departments represented in the data and 3 salary groups.

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?
    - Boxplots and histograms to indicate outliers and distribution. Statistical tests to check for significance of features, among a Logistic Regression and Decision Tree/Random Forest models aimed at predicting employee churn.
- What processes need to be performed in order to build the necessary data visualizations?
    - Ideally, handle any outliers or missing data prior to building visualizations.
- Which variables are most applicable for the visualizations in this data project?
    - All variables are applicable.
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
    - Depending on the severity of the missing data, the rows will either be dropped, or the missing values will be imputed if the missing values cannot be rectified through acquiring it from the source.

The Power of Statistics

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

  - Independent T-Tests were completed between the numeric variables and the outcome variable ‘left’ to determine statistical significance. Chi-square tests were conducted for the binary features. The null hypothesis was there was no statistical significance between the feature and the outcome variable and the alternative hypothesis is that there is a statistical significance between the feature and outcome variable.
- What conclusion can be drawn from the hypothesis test?
  - Many of the discrete numeric variables were statistically significant whereas many of the boolean variables were not.

Regression Analysis: Simplify Complex Data Relationships

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

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 in this stage?

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?

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?

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?

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?

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?