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

Salifort Motors project proposal

Overview

Salifort motors is looking to use employee data to a method to quantify factors that make an employee leave the company

---

Milestones	Tasks	PACE stages
1	Understand the business objectives and questions need answering	Plan
2	EDA, data cleaning, graphing	Plan, Analysis
3	Determine which models are appropriate	Analysis,Construct
4	Construct the models	Construct
5	Confirm model assumptions	Analyze, Construct
6	Evaluate model results	Analyze
7	Interpret results & share actional steps to stakeholders	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project? Management team
- 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? To discover the factors that lead to employees to leave the company. This will help the company retain its current workforce.
- What questions need to be asked or answered? Are employees stratified working at the company, what is the average work hours of an employee
- What resources are required to complete this project? Python, analysis libraries, data visualization apps
- What are the deliverables that will need to be created over the course of this project? Data visualization and written analysis

Get Started with Python

- How can you best prepare to understand and organize the provided information? Clean the data
- What follow-along and self-review codebooks will help you perform this work? Looking at notes and previous labs for the step by step process
- What are a couple additional activities a resourceful learner would perform before starting to code? Use google to help with code syntax, other websites if issues arise.

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable? Avg work hours, tenur, left/stay, departments,
- What units are your variables in? Left is binary, dummy variable categories, everything else numeric
- 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? The avg worker works over 40 hours a week, stratification levels are low for employees who leave.
- Is there any missing or incomplete data? no, some duplicate data

- Are all pieces of this dataset in the same format? There are numeric and categorical
- Which EDA practices will be required to begin this project? Clean, check for outliers,

The Power of Statistics

- What is the main purpose of this project? Discover reasoning of employee churn rate
- What is your research question for this project? What are the factors that contribute to an employee 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? To minimize bias and to make the data more generalized to make sure the data is more representative to the population of the data. I used 25% of the data for random sampling

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project? Senior leadership
- What are you trying to solve or accomplish? People leaving the company
- What are your initial observations when you explore the data? Employees work over 40 hours a week
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.) python, notes for step by step process.
- Do you have any ethical considerations in this stage? Harm of employee positions from the data(firing), transparency clear step by step of how data was analyzed. Accountability of the results of the findings

The Nuts and Bolts of Machine Learning

- What am I trying to solve? What factors lead to company Churn
- What resources do you find yourself using as you complete this stage? Personal notes
- Is my data reliable? Data was cleaned, how the data is collected can be a concern because it was done by the first party.
- Do you have any additional ethical considerations in this stage? Same as above
- What data do I need/would I like to see in a perfect world to answer this question? A high degree of accuracy for the prediction of employee staying & leaving.
- What data do I have/can I get? We have the company reports on current and past employees

- What metric should I use to evaluate success of my business objective? Why? I will use a classification report that includes precision, recall, accuracy and f1 scores. Also a confusion matrix to check the true & false predictions of positive and negative.

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? We seem to have enough variables to make a good prediction

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? Dummy variables for categories
- 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.? we will remove outliers for the logistic regression
- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience? Bar graphs, keep things simple for senior management their job its to make decisions not analysis of the data.

The Power of Statistics

- Why are descriptive statistics useful? We can find insights without making assumptions on the population
- What is the difference between the null hypothesis and the alternative hypothesis? Null is that employees leave at a normal distribution. alternative hypothesis states that there is a factor disrupting a normal distribution

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model? Outliers, dummy variables
- Do you have any ethical considerations in this stage? Make sure the data is clean and representative

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising? The logistic regression is good at predicting if an employee will stay however less so when predicting if an employee will leave. Testing with a Decision tree method
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
- Why did you select the X variables you did? X is everything contributing to left/stay
- 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? Personal notes
- Do you have any ethical considerations in this stage? Same as above

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?

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)?

The Power of Statistics

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

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?