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

Predictive Turnover Model Project Proposal

Overview

The objective of this project is to develop a machine learning model that will predict whether an employee will leave the company, and discover the reasons behind employees' departures.

Milestones	Tasks	PACE stages
1	Outline project workflow, collect relevant data, and identify any software/hardware needs	Plan
2	Perform EDA on the data - clean, convert, and format the data in order for the data to be ready for modeling	Analyze
3	Finalize which modeling strategy will be used and build the machine learning models, then test the models to ensure they are accurate	Construct
4	Format results in an executive summary to be shared with stakeholders. Create visualizations that will clearly show results. Listen to feedback and incorporate it into the analysis.	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?
    - *Salifort's leadership 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?
    - *Trying to build a predictive model for employee turnover. I anticipate the impact of this work to be significant as it can help leaders know how to better retain employees. This will save the company time and money.*
- What questions need to be asked or answered?
    - *They're all in this doc :)*
- What resources are required to complete this project?
    - *Salifort experience and operational data about employees who have left the organization vs those who have not. Also will need to use Python and Tableau to conduct the analysis. Will be referring back to past projects in this program to ensure analysis is conducted correctly.*
- What are the deliverables that will need to be created over the course of this project?
    - *Project proposal, a clean and formatted dataset, a machine learning model (or regression analysis), visualizations to support the analysis, and an executive summary.*

Get Started with Python

- How can you best prepare to understand and organize the provided information?
    - *Conduct descriptive analysis on the dataset to understand the mean, spread, and quality of the data.*
- What follow-along and self-review codebooks will help you perform this work?
    - *The Google Analytics course has provided codebooks about arrays, dictionaries, functions, and python operators that will help in this analysis.*
- What are a couple additional activities a resourceful learner would perform before starting to code?

- *Ensuring they understand the business case by communicating with stakeholders.*

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
  - *There are many so I won't list them here, but the most important variable would be the outcome variable of this analysis which is whether or not an employee has left the company.*
- What units are your variables in?
  - *Integers or strings*
- 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?
  - *So far, it looks like we have a mix of numerical and operational data. I imagine the variable "left" will be our dependent variable. I also presume that satisfaction_level, time_spend_company, salary, and promotion_last_5years will be the most predictive variables.*
- Is there any missing or incomplete data?
  - *There is no missing data, but there are duplicates.*
- Are all pieces of this dataset in the same format?
  - *No, most are integers but 2 variables are strings and 2 are floats*
- Which EDA practices will be required to begin this project?
  - *Reformatting the categorical variables into numerical, standardizing column name format, dropping duplicates, checking and dealing with outliers*

The Power of Statistics

- What is the main purpose of this project?
  - *The objective of this project is to develop a machine learning model that will predict whether an employee will leave the company, and discover the reasons behind employees' departures.*
- What is your research question for this project?
  - *What experiential and operational data best predicts whether an employee will leave 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 important because it lessens the chance of bias and can ensure that the dataset is representative of the population. Selection bias is an example of what could occur if the data was sampled from only one department.*

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
  - *Salifort Motors senior leadership team*
- What are you trying to solve or accomplish?
  - *Lessen the time and money it costs to hire new employees by predicting who is likely to quit and the reasons why.*
- What are your initial observations when you explore the data?
  - *So far, it looks like we have a mix of numerical and operational data. I imagine the variable "left" will be our dependent variable.*
- What resources do you find yourself using as you complete this stage? (Make sure to include the links.)
  - *Will need to use Python and Tableau to conduct the analysis. Will be referring back to past projects in this program to ensure analysis is conducted correctly.*
- Do you have any ethical considerations in this stage?
  - *The only ethical consideration I have is ensuring that this model is not used to wrongly terminate employees. Employees marked as likely to quit should be given proper attention, but not singled out in a negative way.*

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
  - *The most effective model for predicting employees likely to quit, as well as finding the main reasons people leave.*
- What resources do you find yourself using as you complete this stage?
  - *Will need to use Python and Tableau to conduct the analysis. Will be referring back to past projects in this program to ensure analysis is conducted correctly.*
- Is my data reliable?

- *Yes, it comes from the company's HR department*

- Do you have any additional ethical considerations in this stage?
  - *None beyond what I mentioned above.*
- What data do I need/would I like to see in a perfect world to answer this question?
  - *The data provided is pretty good, but we could use more variables such as benefits provided, the amount of individual development employees do, or satisfaction with their manager.*
- What data do I have/can I get?
  - *The data provided by Salifort Motors*
- What metric should I use to evaluate success of my business objective? Why?
  - *A mix of accuracy, precision, recall, and f1 score will indicate the success of the model. As for the metric used to evaluate the success of lessening costs, you can look at the employee turnover rate.*

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?
  - *Yes, it looks like some variables should be explanatory of whether or not an employee left.*

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?
  - *Describe and get basic info about the data, check for missing data and outliers, create visualizations to understand distributions and relationships.*
- 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.?
  - *No, this data set looks fairly complete asides from reformatting the categorical variables as numerical.*

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?
    - *Since the audience is not a data team, I would use simple charts such as a histogram to clearly display the data.*

The Power of Statistics

- Why are descriptive statistics useful?
    - *Descriptive statistics allow you to get a general understanding of the data at a quick glance.*
- What is the difference between the null hypothesis and the alternative hypothesis?
    - *The null hypothesis is always no effect/no relationship while the alternative hypothesis states that there is an effect.*

Regression Analysis: Simplify Complex Data Relationships

- What are some purposes of EDA before constructing a multiple linear regression model?
    - *EDA allows you to get a base understanding of relationships between variables. You can also check if some assumptions are met.*
- Do you have any ethical considerations in this stage?
    - *None further than the ethical considerations I stated above.*

The Nuts and Bolts of Machine Learning

- What am I trying to solve? Does it still work? Does the plan need revising?
    - *Build a model that can predict employee turnover. Yes, I think the plan still works, but we will know for certain once the model is built.*
- Does the data break the assumptions of the model? Is that ok, or unacceptable?
- Why did you select the X variables you did?
    - *Those variables seemed to have higher correlation as shown in the EDA*
- What are some purposes of EDA before constructing a model?
    - *Check assumptions for the model and have a base understanding of the relationships between variables.*
- What has the EDA told you?
    - *The assumptions are met to conduct a Logistic Regression*
- What resources do you find yourself using as you complete this stage?

    - *Again, referring back to past projects.*
- Do you have any ethical considerations in this stage?
    - *None further than the ethical considerations I stated above.*

Data Project Questions & Considerations

PACE: Construct Stage

Get Started with Python

- Do any data variables averages look unusual?
    - *No, the data looks relatively normal*
- How many vendors, organizations or groupings are included in this total data?
    - *One company employee base with 10 departments represented*

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?
    - *A heatmap to view correlations between variables, a logistic regression model will need to be fit to training data.*
- What processes need to be performed in order to build the necessary data visualizations?
    - *The processes have already been discussed.*
- Which variables are most applicable for the visualizations in this data project?
    - *Avg monthly hours, tenure, salary, evaluation score, etc.*
- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
    - *If there is not much missing data then simply remove it from the dataset.*

The Power of Statistics

- How did you formulate your null hypothesis and alternative hypothesis?
    - *Based on the instructions given by senior leadership. Created the hypotheses based on what they want explored.*
- What conclusion can be drawn from the hypothesis test?

Regression Analysis: Simplify Complex Data Relationships

- Do you notice anything odd?

- *The model is not great at predicting employees who leave, only employees who will stay.*

- Can you improve it? Is there anything you would change about the model?
  - *Yes, I would resample the data so the classes are more balanced.*

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?
  - *I kept in all of the variables provided because I thought them all potentially impactful on leaving employees.*
- How well does your model fit the data? (What is my model's validation score?)
  - *Fairly well, the model has an avg precision of .79, recall of .82, and f1 score of .8*
- Can you improve it? Is there anything you would change about the model?
  - *Yes, I would resample the data so the classes are more balanced.*
- Do you have any ethical considerations in this stage?
  - *No, none further than mentioned earlier*

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?