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 Employee Retention Study

Overview

Analyzing the data to come up with ideas for how to increase employee retention. The proposed idea is to design a model that predicts whether an employee will leave the company based on their department, number of projects, average monthly hours, and any other data points you deem helpful.

I will deploy various models to analyze a dataset and generate business insights for Salifort stakeholders. In particular, I will build and evaluate a logistic regression model or the following machine learning models: decision tree, random forest and XGBoost. I will also update my stakeholders through an executive summary, demonstrating my ability to organize and communicate key information.

Milestones	Tasks	PACE stages
Formulate the goal and estimate work	Formulate the goal of the project Perform initial overview of the data source Select the appropriate EDA and modeling methods and tools	Plan
Provide the data quality overview	Ask and answer questions about the source data and expectations	Plan
Summary with the initial data insights	Perform EDA of the dataset Perform feature engineering tasks (as necessary) Evaluate features of the dataset	Analyze, Construct

Summary of the statistical insights	Perform statistical analysis	Analyze
Model results	Select models and fine tune hyperparameters Create, train and test the model	Construct
Executive Summary	Create model performance overview, data interpretation and visualizations	Execute
Executive Summary	Prepare executive summary with the modeling results and business insights	Execute

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- The primary audience of the project is the Salifort company leadership that are concerned by the high level of their employees turnover.
- The idea is to develop an accurate enough machine learning model based on the data available that is capable of predicting employee retention. As a result, the developed prediction app should be utilized by Salifort HR to improve the satisfaction level in the company for the current and future employees.
- What questions need to be asked or answered?
  - Why has the leadership contacted me?
  - What does your stakeholder want from this interaction?
  - What's important to them, their team, or their organization?
- What resources are required to complete this project?
  - Spreadsheets
  - Python programming language with the data processing, analysis, modeling and visualization libraries.
- What are the deliverables that will need to be created over the course of this project?
  - Global-level project document
  - Data files ready for EDA
  - EDA report
  - Analysis of testing results
  - Determine success of the evaluated models
  - Select final model
  - Report to all Stakeholders

Get Started with Python

- How can you best prepare to understand and organize the provided information?
  - Import necessary libraries

  - Perform high-level data analysis with a summary
  - Identify dataset feature types, ranges
  - Check for empty values, duplicates and any anomalies in the data
- What follow-along and self-review codebooks will help you perform this work?
  - Previously completed course assignments and code in Jupyter notebooks
  - Python help articles and function cheat-sheets
  - Code pieces from the previous data science projects.
- What are a couple additional activities a resourceful learner would perform before starting to code?
  - Note the data source location path
  - List out the libraries to be used

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
  - tenure,
- What units are your variables in?
  - Most of the features, both categorical and numerical, are represented by the integer numbers. The only two categorical string features are ‘department’ and ‘salary’.
- 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 presumption is that the ‘high’ salary and ‘promotion_in_5yrs’ should lead to ‘left’ = 0 and the higher ‘satisfaction_level’
- Is there any missing or incomplete data?
  - There is some NaN and duplicated data, and it would be good to have more detailed information like start and end dates to perform feature engineering and track possible seasonality.
- Are all pieces of this dataset in the same format?
  - Most of the data are in float or int format, and there are two features (‘salary’ and ‘department’) that are in str format.
- Which EDA practices will be required to begin this project?
  - Feature engineering (data collection and transformation)
  - High-level statistical analysis

    - Feature correlation and variance evaluation
    - Residuals normality evaluation

The Power of Statistics

- What is the main purpose of this project?
    - To develop a model that is capable of predicting the employees retention status
- What is your research question for this project?
    - What models and features are useful for accurate prediction?
- 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?
    - The random sampling provides potentially unbiased and balanced training and test samples. Since the original dataset is imbalanced, the sequential data selection may result in a trivial subset represented by a single value of the target variable, which makes no sense for modeling.

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
    - HR of the Salifort
    - Salifort Management
    - Development Team
- What are you trying to solve or accomplish?
    - Determine what model to be used for machine learning model development
- 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 in this stage?

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
    - The problem of accurate prediction of the target variable with the selected model.
- What resources do you find yourself using as you complete this stage?

    - Machine learning model Python libraries
    - Set of hyperparameters
- Is my data reliable?
    - From the data provided, it is not evident that the data is absolutely reliable. First of all, the data is imbalanced. Also, the may not represent the entire population of the company's employees worldwide because we do not know from what region the have collected the data and whether departments provided represent the entire org structure.
- Do you have any additional ethical considerations in this stage?
    - Since no pii or other sensitive information is provided in the dataset, there are no ethical considerations like GDPR, HIPAA or other regulations compliance, except possible bias.
- What data do I need/would I like to see in a perfect world to answer this question?
    - We most likely need
- What data do I have/can I get?
    - The satisfaction level, salary range and the employment duration are available and were used for modeling.
- What metric should I use to evaluate success of my business objective? Why?
    - Since my dataset is imbalanced, the f1 score combining precision and recall will be used to evaluate the model performance.

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?

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?

The Power of Statistics

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

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 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?
- 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?