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

Finding the best ML model for predicting an employee ‘leaving’ the company

Classifiers used for model building - Logistic regression, naive-bayes, Decision tree, Random forest and XGBClassifier

Milestones	Tasks	PACE stages
	Load the dataset, look for outliers,correct the mistakes in the names of variables,clean the dataset by removing the duplicates	**plan**
	Using histograms, pairplot, scatterplot and boxplot find the necessary correlation between all independent variables and the target variable ‘left’	**analyze**
	Construct the various analyses and models and check for the best model for employee retention	**construct**
	Execute the best model with the help of metric scores.	**execute**

Data Project Questions & Considerations

PACE: Plan Stage

Foundations of data science

- Who is your audience for this project?
- Salifort Motors management is the stakeholder in this project.
- 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 company wants to predict whether an employee leaves or not. I'm trying to find out the best model suited for this project. The impact of this work would be an increase in profit for the company as money spent on training the employee is lost when he chooses to leave.
- What questions need to be asked or answered?
- The relationship between independent variables and the dependent variable needs to be found.
-
- What resources are required to complete this project?
- Python, jupyter notebook
- What are the deliverables that will need to be created over the course of this project?
- Prediction of an employee 'leaving' the company.

Get Started with Python

- How can you best prepare to understand and organize the provided information?
- What follow-along and self-review codebooks will help you perform this work?
- What are a couple additional activities a resourceful learner would perform before starting to code?

Go Beyond the Numbers: Translate Data into Insights

- What are the data columns and variables and which ones are most relevant to your deliverable?
- The independant variables selected are as follows
- sat_level
- last_eval
- no_project

- avg_monthly_hrs
- time_spend_company
- work_accident
- left
- promotion_last_5years
- department
- salary
-
-
- What units are your variables in?
- Sat_level - numeric value between 0 and 1. 1 being highly satisfied and 0 being highly dissatisfied.
- Last_eval - numeric value between 0 and 1.
- No_project - numerical value of set (1,2,3,4,5,6)
- Avg_monthly_hrs - number between 100 and 350
- Time_spend_company - number in set (1,2,3,4,5,6)
- Work_accident - binary categorical variable
- Left - binary categorical variable
- Promotion_last_5years - binary categorical variable.
- Department - nominal categorical variable
- Salary - ordinal categorical variable
-
- 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?
- Is there any missing or incomplete data?
- There isn't any missing data but there are 3008 rows of duplicates that needs to be removed.
- Are all pieces of this dataset in the same format?
- No.
- Which EDA practices will be required to begin this project?
- We need to visualize the relationship between different variables using countplot, boxplot and histogram.

The Power of Statistics

- What is the main purpose of this project?
- What is your research question for this project?
- 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?

Regression Analysis: Simplify Complex Data Relationships

- Who are your stakeholders for this project?
- What are you trying to solve or accomplish?
- 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.)
- Links to the resources :
  https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.htm
  https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html
  https://xgboost.readthedocs.io/en/latest/python/python_api.html
  https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
- Do you have any ethical considerations in this stage?

The Nuts and Bolts of Machine Learning

- What am I trying to solve?
- What resources do you find yourself using as you complete this stage?
- Is my data reliable?
- Do you have any additional ethical considerations in this stage?
- What data do I need/would I like to see in a perfect world to answer this question?
- What data do I have/can I get?
- What metric should I use to evaluate success of my business objective? Why?

- The main metrics required here are precision_score and recall_score since it correctly gives us the idea of our model which predicts an employee leaving the company.

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?
- Recommendations :
- 1. Since it has been found that employees working for (3,4) projects have better 'sat_level' and better 'last_eval' than employees working for 2 projects, the manager of that group (with 2 projects) should introspect and look at the managerial stability achieved by other manager (3,4) and look for the missing spark in his team. If the employees working for (3,4) projects are happier, and score better last_eval scores than those working for 2 projects then there is a serious issue that needs to be solved within the group. Also it has been found that employees working for 2 projects have higher 'avg_monthly_hrs' than those working for (3,4,5) projects. Hence this 'avg_monthly_hrs' needs to be reduced for employees working 2 projects.
- 2. Employees seem to be having a decrease in 'sat_level' as the 'time_spend_company' increases indicating the lack of enthusiasm among senior employees. This could be directly related to another variable 'promotion_last_5years'. Employees do feel that they need to be promoted.
- 3. Employees working overtime are not appreciated enough as a large section of employees working for 215 to 325 'avg_monthly_hrs' have left.
- 4. Also it has been found that a large section of employees working less 'avg_monthly_hrs' have left and therefore those employees working less need to be inspired to work hard for rewards and could ensure their stay.
- What data initially presents as containing anomalies?
- Variable ‘time_spend_company’ has outliers. Also salary and department variables could be one-hot-encoded to give better results.
- What business recommendations do you propose based on your results?
- Recommendations :

- 1. Since it has been found that employees working for (3,4) projects have better 'sat_level' and better 'last_eval' than employees working for 2 projects, the manager of that group (with 2 projects) should introspect and look at the managerial stability achieved by other manager (3,4) and look for the missing spark in his team. If the employees working for (3,4) projects are happier, and score better last_eval scores than those working for 2 projects then there is a serious issue that needs to be solved within the group. Also it has been found that employees working for 2 projects have higher 'avg_monthly_hrs' than those working for (3,4,5) projects. Hence this 'avg_monthly_hrs' needs to be reduced for employees working 2 projects.
- 2. Employees seem to be having a decrease in 'sat_level' as the 'time_spend_company' increases indicating the lack of enthusiasm among senior employees. This could be directly related to another variable 'promotion_last_5years'. Employees do feel that they need to be promoted.
- 3. Employees working overtime are not appreciated enough as a large section of employees working for 215 to 325 'avg_monthly_hrs' have left.
- 4. Also it has been found that a large section of employees working less 'avg_monthly_hrs' have left and therefore those employees working less need to be inspired to work hard for rewards and could ensure their stay.
-

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?
- Since our aim is prediction of an employee leaving, the best models suited for the purpose are LogisticRegression(), DecisionTreeClassifier(), GaussianNB(), RandomForestClassifier() and XGBClassifier().
- Does my model make sense? Are my final results acceptable?
- My model makes perfect sense though it could be improved with feature engineering.
- Were there any features that were not important at all? What if you take them out?

- Since my model based on XGBCLassifier has produced a very high score on all metrics, it would be futile to check for variables that are less important as leaving a variable out will affect at least one of the metrics. But we could find the variables of utmost importance using their Gini Index or Information Gain(IG) and then modify our model based on our findings on the 'best split'.