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 seeking a method to use employee data to gauge what makes them leave the company.

Milestones	Tasks	PACE stages
1	Understand the business scenario and define the problem	Plan
2	Data exploration and data cleaning	Plan, Analyze
3	Determine which models are most appropriate	Analyze,Construct
4	Construct the model	Construct
5	Confirm model assumptions	Analyze, Construct
6	Evaluate model results	Analyze
7	Interpret results and share actionable steps with stakeholders	Execute

Data Project Questions & Considerations

PACE: Plan Stage

- What are you trying to solve or accomplish?

Analyze the data collected by the HR department and to build a model that predicts whether or not an employee will leave the company.

- Who are your stakeholders for this project?

Salifort Motors.

- What resources do you find yourself using as you complete this stage?

Package for data manipulation such pandas and numpy.
Package data visualization such matplotlib and seaborn.

- Do you have any ethical considerations at this stage?

Salifort Motors want to discover the reasons behind employees that leaving the company and will cost more about the high turnover rate.
- The worst case for predict employees as turnover but they are not turnover is causing work management not effective also make more dissatisfaction.
- The worst case for predict employees as not turnover but they are turnover is causing make misadjust work and cost more about the high turnover rate.

- What are your initial observations when you explore the data?

They all are numeric columns except department and salary column that not.

PACE: Analyze Stage

- What did you observe about the relationships between variables?

number of projects, monthly hours, and evaluation scores all have some positive correlation with each other.

- What do you observe about the distributions in the data?

The mean and median satisfaction scores of employees who left are lower than those of employees who stayed.

- What transformations did you make with your data? Why did you chose to make those decisions?

Transform tenure for compairs between stay/left and short/long.

- What are some purposes of EDA before constructing a predictive model?

For prepare data that can cause impact model.

- What resources do you find yourself using as you complete this stage?

Package for data manipulation such pandas and numpy.
Package data visualization such matplotlib and seaborn.

- Do you have any ethical considerations in this stage?

It's importance to find correlation between value because it can help idenify what value that can lead to the reasons behind employees that leaving the company such as Overwork? Ungratifying to work? Burn out?, or getting fired.

PACE: Construct Stage

- Do you notice anything odd?

No.

- Which independent variables did you choose for the model, and why?

Left for predict employees who left / stay.

- Are each of the assumptions met?

This is classification task logistic regression assumptions met.

- How well does your model fit the data?

All models performs very well in round 1 and round 2.

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

All models currently performs very well, there is no need to do more for now.

- What resources do you find yourself using as you complete this stage?

Packages for build model and confusion matrix such sklearn Packages for saving models such pickle

- Do you have any ethical considerations in this stage?

The worst case for predict employees as turnover but they are not turnover is causing work management not effective also make more dissatisfaction. The worst case for predict employees as not turnover but they are turnover is causing make misadjust work and cost more about the high turnover rate.

PACE: Execute Stage

- What key insights emerged from your model(s)?

All models performed very well. - The logistic regression model achieved precision of 80%, recall of 83%, f1-score of 80% (all weighted averages), and accuracy of 83%, on the test set. - For tree-based machine learning after conducting feature engineering, the decision tree model achieved AUC of 93.8%, precision of 87.0%, recall of 90.4%, f1-score of 88.7%, and accuracy of 96.2%, on the test set. The random forest modestly outperformed the decision tree model.

- What business recommendations do you propose based on the models built?

Recommendation how to retain employees to stakeholders.

- What potential recommendations would you make to your manager/company?

- Cap the number of projects that employees can work on. - Consider promoting employees who have been with the company for atleast four years, or conduct further investigation about why four-year tenured employees are so dissatisfied. - Either reward employees for working longer hours, or don't require them to do so. - If employees aren't familiar with the company's overtime pay policies, inform them about this. If the expectations around workload and time off aren't explicit, make them clear. - Hold company-wide and within-team discussions to understand and address the company work culture, across the board and in specific contexts. - High evaluation scores should not be reserved for employees who work 200+ hours per month. - Consider a proportionate scale for rewarding employees who contribute more/put in more effort.

- Do you think your model could be improved? Why or why not? How?

All models currently performs very well, there is no need to do more for now.

- Given what you know about the data and the models you were using, what other questions could you address for the team?

This model helps predict whether an employee leave and identify which factors are most influential. These insights can help HR make decisions to imporve employee retention.

- What resources do you find yourself using as you complete this stage?

Packages for build model and confusion matrix such sklearn

- Do you have any ethical considerations in this stage?

The worst case for predict employees as turnover but they are not turnover is causing work management not effective also make more dissatisfaction. The worst case for predict employees as not turnover but they are turnover is causing make misadjust work and cost more about the high turnover rate.