Final Exam Project

Applied Machine Learning and Data Engineering in Business Context (LA) CDSCV1008U

Samay Mir {smi.digi@cbs.dk}
Raghava Mukkamala {rrm.digi@cbs.dk}

November 10, 2023

Instructions

1. Please note that you have to upload your solutions before the deadline to the digital exam http://exam.cbs.dk/}.
2. This final exam description is purposefully underspecified to represent the real-world scenario. Therefore, make suitable assumptions about the hypothetical case study and specify your assumptions clearly in your solution.
3. Please use PowerPoint slides for answering this assignment as the main submission. In case the exam portal does not allow you to upload a PowerPoint presentation, then convert it to PDF and upload the PDF. Also, if you want to attach the original PowerPoint (+ jupyter notebook if any), then zip it and attach it as an additional file or so.
4. {**Optional**} If you would like, write a report explaining your assumptions/findings/business insights/functionality, etc., to understand your assignment and data analysis in a better manner and upload it along with your slides.

Use case

Company XYZ has a global presence in 30 countries with a revenue of 25 Billion USD and a total of 40.000 employees. Company XYZ is a leading manufacturing company, and for the last 4 years, they have had a two-digits growth.

Company XYZ is interested in predicting sales, i.e., child car seat sales in 400 locations based on the number of predictors, numeric and categorical.

Currently, the dataset is residing in on-prem data sources SAP and Oracle. Company XYZ wants a scalable ETL, MLOps cloud solution where the data can be migrated to Azure Cloud using its appropriate services such that the solution can be made available for all their departments.

The business justification is that by improving the sales prediction and applying scalable Azure services, they can significantly improve the sales, hence improving the bottom line with 5. Moreover, the dataset and description of the problem are reused from [1, 2, 3], so if you want more information, you can refer to these links.

Assignment

Even though the main goal of this assignment is to train machine learning for sales prediction, but the company requires a well-structured presentation of your proposal where it is expected that you will propose:

1. Your understanding

2. Your Approach
3. Your Findings
4. Your Architecture for ETL and MlOPS
5. Output of your machine learning models
6. Tangible recommendation

Deliverable

You are expected to submit your solution in a PowerPoint (or PDF version), not exceeding 15 slides, optionally accompanying a report specifying your assumptions, business insights, etc. There is no restriction on how many slides you can attach as an appendix, and it is also expected that you also attach your R/Jupyter Notebook or any software of your choice. Please note that you have to upload your solutions to the digital exam.

Info

Again, please note that the assignment is underspecified purposefully to represent the real-world scenario. Therefore, please make suitable assumptions for all the things that are not specified here.

References

[1] F. Gomes, "Carseats eda and basic decision tree." `https://www.kaggle.com/code/gomes555/carseats-eda-and-basic-decision-tree`.

[2] T. H. Cho, "Ensemble modeling with carseats data." `https://www.kaggle.com/code/teresahykim/ensemble-modeling-with-carseats-data/input`.

[3] P. Dasilva, "Carseat sales multiple linear regression." `https://www.kaggle.com/code/poonamdasilva/carseat-sales-multiple-linear-regression/notebook`.