Mandatory Assignment #01

Applied Machine Learning and Data Engineering in Business Context KAN-CDSCV1008U

Samay Mir and Raghava Mukkamala
smi.digi@cbs.dk, rrm.digi@cbs.dk

September 14, 2023

Instructions

1. Please note that you have to upload your solutions in the **Canvas** itself only. You DON'T need to interact with the digital exam in any way for the mandatory assignments.
2. Please use Python 3 or R language for answering the following questions whenever it is necessary.
3. Write a report explaining your assumptions/findings/business insights/functionality etc. to understand your assignment and data analysis in a better manner. It is also good practice to use comments extensively in your code so that it will be easy for other people to understand them.

Use case

In order to answer the assignment properly, let us imagine a hypothetical scenario that a company has hired you, as a Data Scientist in their newly formed Data Analytics division. The company is a very traditional company that does its business based on its gut feeling and implicit domain-specific tacit knowledge rather than using data-driven methods. However, the company has just started its journey in the new digitalization landscape and is trying to figure out how it could use data-driven methods to modernize its operations. As the first step in their journey, the company has hired you as a data scientist, thinking that you could help them with data crunching and educate them on what kinds of benefits the data-driven methods can bring to the table.

So, as a data scientist, your first job is to analyze the dataset to provide some good business insights and, more importantly, to build a story out of the dataset to showcase the usefulness of data analysis and the importance of your job. Therefore, as mentioned previously, focus on building a good story and business case out of the dataset rather than focusing on performing a robust analysis. It does not mean that you should not do robust analysis; you are free to do so as long as you integrate that into your story on the dataset. You are free to make any business assumptions you deem fit for your analysis. Assumptions such as that data reside in silos, accessing the data is cumbersome, you have obtained business insights from the company's experts etc, are all very valid.

Choice of Dataset

Please feel free to use any dataset of your choice and, perform the analysis, build a good story around the data as specified above. There are plenty of datasets available online (e.g., `https://www.kaggle.com/`, `https://archive.ics.uci.edu/ml/index.php` and many other sources) and therefore feel free to use any dataset you like.

Analysis

The main purpose of the following specification is to provide some guidelines for the analysis of the dataset so as to build a good story around the data. Feel free to use whatever steps/methods of your choice to build a good and coherent story around the dataset.

- **Story board**
  1. State and demonstrate clearly the task you are solving
  2. Describe challenges or complications, if any, you might have to face or encounter
  3. Show your high-level approach
  4. What business value are you expecting to achieve?
- **Approach in-depth**
  Your approach should at least include the following
  1. **Data pre-processing**
     (a) Conduct descriptive analysis of the given dataset.
     (b) Using different plots and visuvalizations to analyze the data further, to use in the report and build a good story around the data.
     (c) Also investigate missing values, outliers, or any other abnormalities that you can find while working on the given dataset. Describe your findings in the report clearly.
  2. **Pattern Recognition**
     (a) For this sub-task, first perform correlation analysis and see whether any systematic correlations exist in the dataset or not. Explain your findings in the report.
     (b) Furthermore, you can also test-principal component analysis or any other suitable method and use a biplot to see if any systematic patterns exist or not.
  3. **Prediction**
     (a) Perform a few predictions on the dataset using a suitable algorithm. Here you may choose an algorithm(s) [1] of your choice that will fit the dataset.
     (b) Explain your rationale for why you have chosen the selected algorithms.

Describe your findings/results in a PowerPoint format with adequate explanations. Again, focus more on building a good business story around the data using various plots/visualizations/insights. Explain what business value you are creating for your client, i.e. are you increasing the client's revenue, optimizing processes or reducing their cost

Your PowerPoint should be limited to 20 slides. All other materials should be placed in the Appendix. Please also upload your code to Canvas along with your submission.

Hints: The following guidelines might help you to perform your data analysis.

- Figure out whether your dataset suffers from high dimensionality, is noisy or has too many outliers. Find the primary principal components that cover most of the dataset.
- Figure out most of the important variable/feature(s) of the dataset for prediction.
- Explain your tables well enough so that everyone will understand. Please don't assume that your audience has a technical background.
- Make proper recommendations and explain how you are optimizing your client's business.

References

[1] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.