Table of Contents

Final Assignment: Data Science for Business Project – DATDRD05-T03

Objective:

The goal of this assignment is for you to apply the CRISP-DM methodology and the concepts learned from the first period of the minor to solve a business problem using data science techniques. By completing this assignment, you will demonstrate your ability to structure a data science project, conduct analyses, and derive actionable insights from data.

The techniques used in this course are the techniques that you have learned during the course. You may use other techniques, though. However, this must be supported by proper reasoning as to why you have chosen techniques that are not part of this course. For example, you may have studied and trained yourself to use these technique. Since it is a group project, the choice must be supported by all the group members.

The lecturer will decide the group size.

Assignment Description:

You are tasked with conducting an end-to-end data science project for a hypothetical or real-world business problem of your choice. The project should follow the CRISP-DM process and include a well-defined business objective, data understanding, data preparation, modelling, evaluation, and deployment strategy.

You may use publicly available datasets or a dataset you have access to (e.g., from Kaggle, UCI Machine Learning Repository, or other sources).

Some topics you may consider:

- Data Science for Good Societies
- Future of Business with AI
- AI and The Future of Work
- AI's Influence on Global Power Structures
- Predicting customer churn for a telecom company.
- Recommending products for an e-commerce platform.
- Classifying fraudulent transactions in a banking dataset.
- Segmenting customers for a retail company to personalize marketing campaigns.
- Forecasting sales based on historical sales data.

Requirements:

The final submission should include both a written report and a Jupyter Notebook (this facilitates transferability). The report must describe each phase of the CRISP-DM process in detail, while the Jupyter Notebook should provide the actual code and any necessary visualizations. In the Jupyter file, make use of the markdowns to explain the choices made, expectations, struggles, etc. The python packages used and reasons behind that must also be stated in the markdowns.

Assignment Instructions:

You must follow the CRISP-DM framework to structure your project:

1. Business Understanding

Define the business problem you are trying to solve.

Questions to answer:

What is the objective of the analysis?

Why is this problem important from a business perspective?

What are the potential benefits of solving this problem?

What will success look like for the business?

2. Data Understanding

Investigate and describe the dataset(s) you are using.

Questions to answer:

What data is available to you, and where does it come from?

What are the key variables that might be useful in solving the business problem?

Are there any potential biases or issues with the data?

Deliverables: Summary statistics, visualizations, initial data exploration, identification of data quality issues (e.g., missing data, duplicates).

3. Data Preparation

Prepare the data for modelling by cleaning, transforming, and feature engineering.

Tasks to include:

Handling missing values, outliers, or incorrect data points.

Feature selection or engineering.

Data transformations (scaling, encoding, etc.).

Splitting the dataset into training and test sets.

4. Modelling

Choose and apply appropriate machine learning models for your business problem. You should compare different models and evaluate their performance.

Questions to answer:

What models did you choose to try and why (classification, regression, clustering, etc.)?

How do these models relate to the business objective?

How are you tuning the models (e.g., hyperparameter tuning)?

Deliverables: Training results, performance metrics (e.g., accuracy, precision, recall, RMSE, etc.).

5. Evaluation

Evaluate your model in terms of both its performance and its ability to meet the business objectives.

Questions to answer:

How well did your model perform on the test data?

How does this performance translate into business value?

Are there any potential risks, weaknesses, or limitations of the model?

6. Deployment Strategy

Describe a plan for deploying the model into a business environment.

Questions to answer:

How will the model be used in practice?

How would the business make decisions based on the model's predictions?

What are the maintenance and update strategies for the model?

Deliverables:

Written Report (approx. 2000-3000 words)

Structured according to the CRISP-DM phases.

Must include your rationale for the choices made, results, interpretations, and recommendations for the business.

Clear explanation of your findings and how they impact the business.

Jupyter Notebook.

A video summarises your work and includes clear rationale as regards each decision and/choice made. The video is max 7 minutes.

Clean and well-documented code for data exploration, preparation, modelling, and evaluation.

Visualizations and outputs to support your findings.

HANDING IN:

Al work must be handed in in HANDIN. You may submit your report, Jupyter notebook, and link to the video.

The video must be handed in as follows:

As for the Video's for HANDIN, first go to https://videotoetsing.han.nl/. log in and upload the video. You will receive a URL from the site. Copy and paste this URL unto your HANDIN page. By doing so, we will have the archive for the programme.