Artificial Intelligence
Datathon Assignment

Introduction

You are a member of the **Data Science & AI Team of BSH**.

BSH Hausgeräte GmbH, a subsidiary of the Bosch Group, is a leading global manufacturer of home appliances headquartered in Munich, Germany. Established in 1967 as a joint venture between Bosch and Siemens, BSH became wholly owned by Bosch in 2015. The company offers a wide range of innovative and energy-efficient appliances under brands such as Bosch, Siemens, Gaggenau, and Neff. Operating in over 50 countries with numerous production sites and service centers, BSH emphasizes sustainability, digitalization, and user-centric design. Through continuous innovation and a strong global presence, BSH remains committed to improving everyday life through advanced home appliance solutions.

Currently you are working on the development of smart home services. Therefore, the company has introduced a large array of connected products, e.g. ovens, fridges, coffee machines. The connected products create a per day data frame on device usage, consumer preferences, app usage, and other relevant metrics.

Recently, you received a data file with around 5´400 data points from different households. Each data point is collecting the aggregated data for one full day.

The consistency of a data point depends on connectivity, device usage, and other factors. Thus, you have to check the consistency and completeness of the data.

The CEO of BSH is demanding a report on the performance of the connected products series, customer segmentation strategies, digital service ideas, decisions that can be done based on the data, and recommendations on new smart home services deployed on the BSH Home Connect App.

Your board presentations are starting at 15:00.
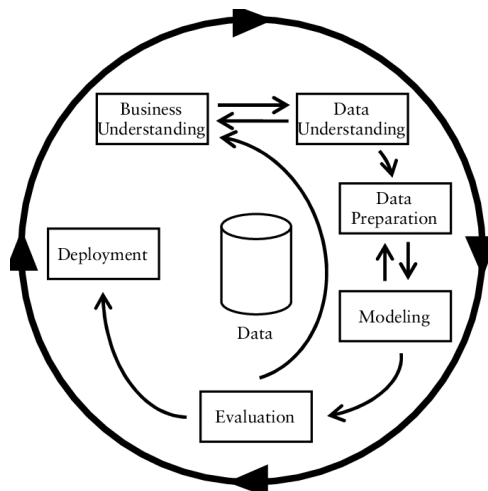Prepare your files* for the presentation.

You have a max of 10 minutes time for presentation.
Additional 10 minutes are reserved for questions.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Procedure

Your team is working around the CRISP-DM cycle.

Stage 1: Business Understanding

The data set is containing data points for different households regarding:

- ID of each data point
- Number of persons per household
- Number of adults per household
- Region of the household
- Usage of a smart fridge, smart oven, and smart coffee machine
- External data available on weather
- Vital data: Average steps per day, average hours sleep per night
- Consumption preferences: Coffee cups, coffee strength
- Three different product categories stored in the fridge (barbecue, beverages, vegetables)
- In-App Purchases (binary)
- In-App Purchases (value)
- App Usage Intensity (app face time)
- User satisfaction
- Net promoter score (NPS)

Generally, you are interested in factors driving sales and new AI based service ideas, as BSH follows a device-as-a-platform business model. Therefore, it is relevant which factors drive purchases in general and based on the sales value. You also want to integrate some consumption or device usage patterns in such models. Your CEO wants to know which recommendations for marketing and sales strategy can be derived from the data and how can such recommendations be implemented in AI based services on the BSH App. It is also relevant how such factors impact on consumer satisfaction and loyalty.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Stage 2: Data Understanding

The data set is containing data points for different households regarding:

- Number of persons per household
  Scale: 1-6

- Number of adults, children
  Scale: 1-2

- Region of the household
  Scale: Germany, UK, Netherlands

- Usage of a smart fridge
  Scale: 1-n, refers to precise frequency of device usage
  (Fridge open and fridge close are counted as 1 interaction)

- Usage of a smart oven
  Scale: 1-n, refers to precise frequency of device usage
  (Oven on and off are counted as 1 interaction)

- Usage of a smart coffee machine
  Scale: 1-n, refers to precise number of cups per day

- Coffee Strength
  Scale: 1-n, refers to average coffee strength
  (1=weak, 2=average, 3=strong)

- Steps per day
  Scale: Number of average steps per adult per day

- Hours sleep per night
  Scale: Number of average hours the adults sleep per night
  (counted beginning with evening sleep till next morning)

- External data available on weather
  Scale: Rainy, Average, Sunny

- Three different product categories stored in the fridge
  (Product 1: Barbecue, Product 2: Beverages, Product 3: Vegetables)
  Scale: Number of replacements („take in" and „take out"
  are counted seperately) in the fridge per day

- In-App Purchases (binary)
  Scale: Yes/No

- In-App Purchases (value)
  Scale: 0-n EUR (measured in Cent)

- App Usage Intensity (app face time)
  Scale: Average seconds adults using the app per day

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

- User satisfaction with devices and smart services
  Scale: Results from a customer satisfaction survey (1-10)

- Net promoter score (NPS), a customer loyalty metric
  Scale: Results from a customer satisfaction survey (0-10)
  https://de.wikipedia.org/wiki/Net_Promoter_Score

Stage 3: Data Preparation

The consistency of a data point depends on connectivity, device usage, and other factors. Thus, you have to check the consistency and completeness of the data.

In addition, synthetic data may have to be generated for missing values or due to the balance of the data set.

You should use a data preparation approach powered by Python and Jupyter Notebooks. Report your findings regarding data preparation and data quality. Present relevant descriptive statistics on relevant variables.

Stage 4: Modeling

Modeling depends on the goals of your analysis. You are requested to use state-of-the-art statistical methods and present your findings based on different types of data visualizations and different types of data analysis.

Make full use of the data and develop different models for different business problems, i.e. different target variables.

Overall, you should develop at least two different models with cluster, regression or classification analysis, e.g. binary classification, multi-class classification or regression models.

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation

Stage 5: Evaluation

o Evaluate the quality of your models based on relevant analysis and metrics.

o If you run classification models report accuracy, precision, and recall
and find arguments for the appropriateness of such metrics.

o Report the fit of your models on the given data.
Provide evidence that you prevent overfitting on the data.

o Also, report the different features and weights included in your model.
Give an interpretation on the overall quality of the model and the business
interpretation of different features and weights.

o If your model performance is poor, reflect on possible reasons
and mention areas for improvement.

Stage 6: Deployment

You do not have to create model deployments but the CEO wants to get answers
for the questions below.

- How do you plan to eventually deploy the model?
- How can the deployed model be integrated into existing IT infrastructures
  and digital services?
- What are potential limitations and assumptions?

Technical Support

Each team will receive technical support regarding comprehension questions.

Results

Prepare a compelling presentation for the CEO about your working cycle, findings,
and recommendations, and cover the following aspects:

- Specific problems and questions you want to solve.
- Data exploration and preparation.
- Suitable machine learning approaches.
- Modelling and evaluation.
- Business value of machine learning and limitations.
- Possible integrations with existing and new digital services.

The presentation needs to include information about your teamwork
in the format of a statement of work per team member (last slide).

Good luck!

*Mandatory submissions:
1) Presentation file as PDF,
2) All data files and notebooks related to the presentation