Artificial Intelligence
Datathon Assignment

Introduction

You are a member of the **Data Science & AI Team of Zalando**,
a leading European eCommerce Company (https://www.zalando.de/).

Zalando is a Berlin-based European online fashion and lifestyle platform founded in 2008. It sells clothing, shoes, accessories and beauty products to millions of customers across Europe, combining e-commerce, logistics and tech to personalize shopping. In 2024, Zalando generated about €10.6 billion in revenue, and its trailing 12-month revenue (TTM) is around €11.6 billion, reflecting recent growth.

Currently, you are working on the further development of your marketing and sales strategy. Recently, you received a data file from the German Zalando Online Shop with around 5.000 data points from online transactions. The data file is containing data from the web shop-systems like basket size, user satisfaction, purchased product category, etc.

The consistency of a data point depends on the quality of data gathering from the shop, CRM systems, and other factors. Thus, you have to check the consistency and completeness of the data.

The Head of eCommerce is demanding a report about the general effectiveness of the online shop, relevant insights for the improvement of user satisfaction, sales conversions, product and segmentation strategies, and other relevant factors.

Furthermore, the Head of eCommerce is interested in AI and keen to know how to use data for additional digital services or recommendations for the sales department.

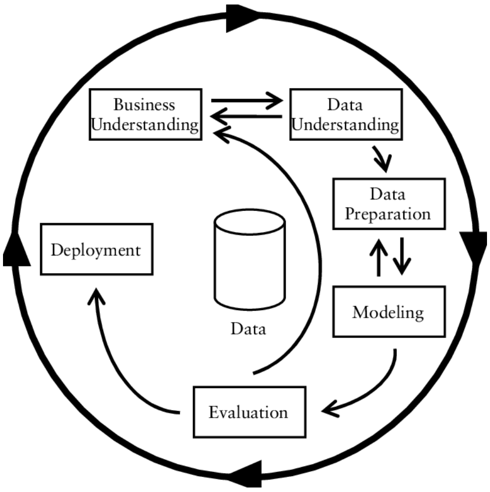
Your board presentations are starting at 15:00.
Prepare your files* for the presentation.

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 multiple variables.

Generally, you are interested in machine learning algorithms based on your data that can predict relevant target metrics for sales based on typical objectives in this functional area (e.g. satisfaction, conversion, loyalty, product-, marketing-, segmentation strategies, etc.).

*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 the following variables. Most variables are self-explaining.

- CustID
  = Outputs the ID of the customer from the CRM system.
  Numeric
- Order
  = Indicates whether the customer has purchased in this transaction or not.
  Binary
- Visits
  = Indicates the number of visits to the online shop before the current visit.
  Numeric
- Device
  = Indicates the customer's preferred device for logging into the store.
  Categorial, qualitative
- Region
  = Indicates the region from which the customer comes.
  Categorial, quantitative
- Distance
  = Indicates the distance between the delivery center (logistics) and the customer's home.
  Numeric
- Payment
  = Indicates the customer's preferred payment method.
  Categorial, qualitative
- Gender
  Binary
- FaceTime
  = A grouped variable indicating how long the customer has been on the website so far.
  Categorial, quantitative, ordinal
- NoDevices
  = Indicates the number of devices the customer uses and has registered in the store.
  Numeric
- Category
  = Indicates the preferred product category that the user has observed in the store.
  Categorial, qualitative
- CSat
  = Indicates the customer satisfaction level of the customer.
  Numeric, Likert
- MaritalState
  Categorial, qualitative
- NumAdress
  = Indicates the number of addresses that the customer has registered in the store.
  Numeric
- Complaints
  = Indicates whether the customer has complained in the last 30 days.
  Binary
- Coupons
  = Indicates how often the customer has used coupons in the past.
  Numeric

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

- NumOrders
  = Indicates the number of orders placed by this customer in the last 30 days.
  Numeric
- LastOrder
  = Indicates the days since the customer's last order.
  Numeric
- Returns
  = Indicates the costs (EUR) of returns of this customer in the last 30 days.
  Numeric, currency
- BasSize
  = Indicates the size of the shopping cart (EUR) of the current transaction.
  Numeric, currency

Stage 3: Data Preparation

The consistency of a data point depends on the quality of data gathering from the shop, CRM systems, 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

- ○ Evaluate the quality of your models based on relevant analysis and metrics.
- ○ If you run classification models report accuracy, precision, and recall and find arguments for the appropriateness of such metrics.
- ○ Report the fit of your models on the given data.
  Provide evidence that you prevent overfitting on the data.
- ○ 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.
- ○ 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 CIO 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 board 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