ZIEKENHUIS LEEGDUUR VOORSPELLEN

TUSSENTIJDSE OPDRACHT 1 – DATA ANALYSE

Deze tussentijdse opdracht 1 is het eerste deel van de grote integratieopdracht en heeft als focus:

1. **Data Analyse**: Analyseren, beschrijven, visualiseren, feature engineering en transformeren van de data om dit bruikbaar te maken voor Machine Learning.

We noemen dit ook wel EDA. Voer onderzoek uit naar wat EDA betekend en uit welke componenten dit bestaat. Bepaal aan de hand daarvan wat logisch om te analyseren en hoe. Er zijn verschillende tools hiervoor en als analist mag jij je eigen gereedschap kiezen. MAAR let op het is niet toegestaan om een handmatige analyse te doen. Zowel niet visueel als met heel veel plak en knipwerk van data. De CDO wil zien hoe dit met tools geregeld wordt (geautomatiseerd) door de dat ergens in te lezen waarna je natuurlijk wel tool opties of functies mag aanklikken.

Hierbij nog wat tips, tricks en andere eisen:

- Een EDA analyse doe je op de totale data. Dus voeg de twee bestanden (train en test) samen tot 1 bestand met de naam **Zorg-LOS.csv**
  Dit bestand moet je ook opleveren met de rest van de bewijzen.
- EDA is groot en veel en je moet zelf bepalen wat je nodig hebt om een goede analyse te kunnen maken om een goed Machine Learning model te kunnen maken (doe onderzoek en kijken bij extra toelichting wat er allemaal binnen zou kunnen vallen)

Wat de CDO minimaal wil zien is:

- Een tabel met alle records, het type data veld en of het getransformeerd moet worden om bruikbaar te zijn voor een model generatie
- Is de data goed normaal verdeeld en wat zijn de ander statistische waarden
- Wat valt er te vertellen over de datapunt (ontbrekende, afwijkingen etc.) en wat ga je daar mee doen
- Wat verteld de heat map voor story
- Kijk op basis van de zaken bij “extra toelichting” wat er nog meer logisch is om te bekijken en te beschrijven

Je mag het rapport gelijk in de gewenste einddocumentstijl opleveren en stap voor stap uitbreiden (levend document) of als losse delen opleveren. Maar het definitieve integratie document moet uit 1 geheel bestaan dus alle resultaten van de losse tussentijdse opdrachten moet hierin zijn opgenomen.

1. DATA ANALYSE

Je hebt van de organisatie al de dataset gehad die bestaat uit 2 delen, een training en een test set (train.csv en test.csv). Als analist wordt er van jouw verwacht dat je weet wat dit betekend in het kader van het selecteren en maken van machine learning modellen. Natuurlijk verwacht de CDO dat je nog wel kritisch kijk naar deze data en hier EDA op toepast naast het beschrijven van de dataset. De uitkomsten hiervan moet gedocumenteerd worden in het op te leveren rapport voor de CDO en je moet aangeven in het adviesrapport of er aan de data en de data kwaliteit in het kader van Data Governance nog iets gedaan moet worden.

VOORWAARDEN

Technisch

1. Geef aan hoe je de EDA hebt uitgevoerd en met welke tools.
2. Geef aan hoe je structureel transformatie en EDA kan uitvoeren (dus niet eenmalige handmatige handelingen).

Informatie behoefte business

1. Wat zijn de verschillende EDA uitkomsten op gebruikelijke categorieën (bv. lege velden etc.)? Doe hiervoor onderzoek want er zijn er een aantal.
2. Beschrijf het data cleaning en preparation proces indien van toepassing.
3. Beschrijf en visualiseer de conclusie en aanbeveling op business, technisch en data niveau.

Hoe in te leveren

De adviespresentatie over data en het gedetailleerde rapport en andere gebruikte bestanden toevoegen aan het zip bestand dat je dient in te leveren op 14 januari 2024 op Brightspace.

EXTRA TOELICHTING

Bronnen aanwezig in Brigtspace die wij al voor jullie hebben geselecteerd

- TW2-Book - Principles_of_Data_Mining - Max Bramer
- TW2 - EDA with Python.pdf

Overige:

- Kijk of de data normaal verdeeld is en zo niet beslis of je dit moet corrigeren via een transformatie (hiervoor bestaan technieken en tools)
- Meer info over EDA
  Data Analysis with a Single Line of Code Using Advanced Python Libraries: Automate Your EDA and Understand the Data Quickly | by Thomas Kidu | Oct, 2023 | Medium
  en
  5 Python Packages for Effortless EDA by Cornellius Yudha Wijaya
  en
  10 Automated EDA Tools That Will Save You Hours Of Work by Ritesh Gupta
- Transform skewed data
  How to use sklearn to transform a skewed label in a dataset | by Tracyrenee | MLearning.ai | Medium

Wat is EDA:

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process. When performing EDA on items, you focus on understanding the characteristics and patterns associated with specific items in your dataset. Here are some techniques and considerations for EDA related to items:

1. **Summary Statistics:** Calculate basic statistics such as mean, median, minimum, maximum, and standard deviation for item-specific variables. This helps in understanding the central tendency and variability of the data associated with each item.
2. **Data Visualization:** Use various types of charts and graphs like histograms, box plots, and bar charts to visualize item-related data. Visualization can reveal distribution patterns, outliers, and trends associated with items.
3. **Categorical Variables:** If your items have categorical attributes (e.g., item categories, brands), create frequency tables or bar charts to show the distribution of items across different categories.
4. **Correlation Analysis:** If your dataset contains multiple variables related to items, perform correlation analysis to identify relationships between them. For example, you might explore correlations between item price and customer ratings.
5. **Missing Values:** Check for missing values in item-related variables. Understanding the extent of missing data is crucial for making decisions about imputation or data handling strategies.
6. **Outlier Detection:** Identify outliers within item-specific variables. Outliers can significantly impact your analysis and might need special attention or treatment.
7. **Time Series Analysis:** If your dataset includes a time component (e.g., sales data over time), perform time series analysis to identify trends, seasonality, and patterns associated with specific items.
8. **Comparative Analysis:** Compare different items based on various attributes. Visualization techniques like side-by-side box plots or bar charts can be useful for this purpose.

Remember, the goal of EDA is to uncover insights and patterns in your data, which can guide further analysis and decision-making processes. Each dataset and analysis scenario may require different EDA techniques, so it's important to adapt your approach based on the specific characteristics of the items and the data you are working with.

Bron: ChatGPT – prompt: Data analysis eda items