What is data science

Portfolio assignment 101

Avans Hogeschool

March 9th, 2022

Author:

Frank Gabršek – 2171626

Lecturer:

Erco Argante

Table of Contents

[Data science 3](#_Toc97728751)

[The data science lifecycle 3](#_Toc97728752)

[Business understanding 3](#_Toc97728753)

[Data acquisition & understanding 3](#_Toc97728754)

[Modelling 3](#_Toc97728755)

[Deployment 4](#_Toc97728756)

[Data science and BI 4](#_Toc97728757)

[Data science and AI 4](#_Toc97728758)

# Data science

Data science is an all-encompassing term for a multidisciplinary field that seeks to provide meaningful insights from sets of structured and unstructured data.

Data science is an intersection of data within a domain with the mathematics used to derive information and patterns from it and the programming skills used to find and visualise it. These 3 areas can be found within the data science lifecycle and are applied in the portfolio assignments

# The data science lifecycle

The data science lifecycle is a set of phases that describe how one can find insights in a dataset and the relations between these phases. The lifecycle consists of Business understanding, Data acquisition & understanding, Modelling and Deployment

## Business understanding

In order to provide any meaningful comments on a dataset, a scientist needs to understand the data they’re working with. The first step in data science is therefore to learn about and understand the dataset and the encompassing domain.

In the portfolio this was performed in *assignment 4*. Alongside the choice of the dataset, research was done into the set and its values within the domain. This included among other things, researching the meaning of the values and acronyms found in the satcat dataset and researching how values such as ‘rank’ or ‘metascore’ were calculated within the imdb dataset.

## Data acquisition & understanding

With sufficient domain knowledge one can start to understand a dataset. This is done by doing various univariate and bivariate analyses on the dataset to identify distributions, correlations, and anomalies within the data. On their own, these analyses only provide a limited insight into the dataset and the superset it was sampled from, however, a valuable descriptive analysis can still be issued.

For the portfolio, univariate analyses were conducted in *assignments 6*, and in *assignment 7* a confidence interval for Revenue and Ratings were calculated to provide an insight on the superset of imdb movies. In *assignment 10* a correlation between Rating and Runtime was identified and in *assignment 12* it was found that there is an anomaly in the movies’ ranks by year. Understanding the dataset continued in *assignment 14* with a categorical analysis of genres by year, but no correlation was found.

## Modelling

In order to create predictive models, machine learning techniques can be employed within data science. Although various methods with a wide range of complexity exist, all are rooted in a similar principle. One or more feature attributes from the dataset are chosen to build a model, which is then created or ‘trained’ using part of or all the dataset as a training set. This results in a model that can predict targets or cluster datapoints within a certain degree of accuracy.

In assignments 16 and 18, predictive models were created to predict Year and Rating based on features in the dataset. Rank in assignment 18 could predicted within a deviation of 10%. Although the decision tree in assignment 16 could not accurately predict a year, it still proved useful in demonstrating the limits of this particular algorithm. In assignment 19 a clustering model was created to attempt to distinguish penguins based on a number of features.

## Deployment

The final stage for any data science is the deployment phase, wherein projects are finalised and used in a production environment. Additionally, this includes monitoring the results of the project and evaluating the models for accuracy and performance.

Although deployment isn’t incredibly applicable to the portfolio, a few aspects of this stage were employed. The decision and regression trees in assignments 16 and 18 were evaluated on their accuracy and RSME respectively. This was done in order to determine whether these models were in any way usable for predictive analyses.

# Data science and BI

At a superficial level, Business Intelligence (BI) and Data science seem similar in their methods and purpose. However, as one gathers experience within both these analytical fields, differences will become clear.

The difference between data science and BI is a difference in the type of analysis they provide. Bi seeks to perform a descriptive and diagnostic analysis. In simple terms: ‘what happened’ and ‘why did it happen’. An overlap with data science can be seen here, as methods to analyse data in BI can be the same ones we see in portfolio assignments 6 through 14. By seeking out correlations and anomalies, BI provides an insight into the past. However, the conclusions that can be derived from this past are left to the human element. This is where BI and data science diverge. While descriptive analysis is an important element, data science goes further than this by providing a predictive, and at an advanced level, a prescriptive analysis. Using various mathematical algorithms as seen in assignments 15 through 19, data science attempts to extrapolate what could happen and what could make it happen from a dataset.

# Data science and AI

Artificial intelligence, often abbreviated as AI, is intrinsically connected to Data science. This has everything to do with the intermediary field Machine Learning. In assignments 15 through 18 we’ve seen that using data science, we can teach a computer to predict an output when given a certain input. This becomes AI when we program the computer to not just predict an outcome, but to act on it. This creates a loop where a computer continuously gathers input on its situation, predicts outcomes for all possible actions and then acts to reach the most desirable outcome. This requires training the AI using vast amounts of data, which connects the field of AI to data science.