–

Case studies block 2 - version 2020/2021

Riccardo Pinosio, Rob Loke, Frederik Situmeang

1 Case studies: general outline

The learning goals for the courses in the second block are:

- learn to obtain, preprocess and store data, as a necessary step to carry out analyses and building statistical and machine learning models
- learn to build statistical models to deliver business insights based on the pre-processed data

The case studies that we will cover in the four courses of the block is based on BrainBay/Funda housing data for the Netherlands in 2018 and 2019. Each course in the block will address different aspects of an end-to-end data analysis project using these case studies. In particular, in the first course (database management for business) the students will focus to learn how to process big data, both relational and non-relational. The students will start with locally stored data and will grow their skills to manage database stored in the cloud using spark. In the second course (online data mining) the students will focus on the funda case study to learn how to retrieve data from online sources. In the third course (programming machine learning for business) the students will carry out an analysis to answer the business question that the case study is centered upon. Indeed, the goal of the case studies is for the students to apply the knowledge they acquired in lectures and through studying to achieve a realistic business goal; hence, each case is based on a general story specifying the business question to be answered with data.

Students are required to refine this story into tasks that need to be completed to answer the original business question (scrum refinement and planning). These refined stories will then be taken up throughout the courses of the second block.

Here is the business question for the case study:

- Which features of a house influence its sale price and it's time to sale?
  - Original business question:

    As an online marketeer at a startup, I would like to know how the description of a house for sale on funda and the house's features influence metrics that the real estate agent cares about, like its asking price or its time-to-sale. I would also like to know whether there are differences among different geographical areas (gemeenten and provincies) in this respect. I am interested in this because we are thinking of launching a product that allows estate agents to get advice on how to position the advertisement for their objects across multiple platforms (e.g. funda.nl, facebook, ...).

  - Business stakeholder: Frederik Situmeang. Brainbay is a new NVM company that not only manages the housing data that powers funda.nl, but also applies BI and data science to provide insights on the housing market to real estate agents.

2 Case study: courses breakdown

The main stakeholder for the case study, alongside BrainBay, is Frederik Situmeang, who will give a presentation that outlines the peculiarities and challenges of the case on the first Monday (Monday Nov 2, 2020). The following is a breakdown of the students' tasks for the case study part for each of the four courses in the block:

2.1 Database management

As discussed in class, for this course we have a case study: the funda case study, which you will work on as a group. In the paragraphs below you will find updated information about these case studies.

The goal of this case study is to deliver the following:

- A working database instance that can be run on your machine and that contains (1) the funda data, and (2) the additional CBS data as tables.
  - The CBS data consists in additional information at the level of neighbourhoods that can be found here: https://opendata.cbs.

nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=84583NED&_theme=229

  - download the csv version of the data and make sure it contains at least (1) the gemeente information (2) the geslacht, leftijdsgroepen, and bevolkingsdichtheid information and (3) the gemiddeld inkomen per inwoner. You can add additional dimensions that you wish to explore to the csv if you like.

- A working python web application that does the following:
  - It connects to the running instance of the database;
  - It calculates aggregated tables using SQL and writes these tables to the database instance
- The goal of the python application is to aggregate the house price data every day into different tables that can be used as data sources for a dashboard (e.g. using tableau)
- The minimal information to be displayed on the dashboard is the following:
  - Average asking price per month for each of the gemeenten and municipalities in the Netherlands
  - Average asking price per bevolkingsdichtheid group or category (you might have to discretize this variable) for each gemeente in the Netherlands
  - Average asking price per gemeente, where the gemeenten are ordered according to the average income per inhabitant (from highest income to lowest income)
  - For every gemeente in the Netherlands and every month in 2018-2019: the percentage increase or decrease in the average house price in that gemeente compared to the previous month
  - For every gemeente in the Netherlands and every month in 2018-2019: the absolute difference between the median house price for that month in that gemeente and the median house price for the next month in that gemeente
  - The average house price in 2018-2019 according to leftijdgroep (in the whole of the Netherlands)

- The above information will have to be retrieved using SQL queries and written to appropriate database tables. If you have more time, you can investigate further aggregations of interest yourself (in addition to the above)

2.2 Online data mining

In this course we work on the funda case study, and in particular on retrieving funda data from the website via scraping. This is a common scenario where you would like to carry out a project that requires some data from the web to which you do not have direct access.

- Write a scraper for the funda website to retrieve the following information from listed houses:
  - postcode of the house
  - asking price
  - squared meters surface of living area (woonoppervlakte)
  - squared meters surface of the whole property (perceeloppervlakte)
  - year of construction
  - whether it has a garden or not
  - type of house (apartment, ...)
  - house description
  - the energielabel of the house
  - the number of rooms and number of bathrooms of the house

  A working scraper must be produced that scrapes this data from 1000 houses across all the different provinces of the Netherlands, so as to obtain a geographically representative sample for analysis. The number of houses scraped is capped at 1000 because the purpose of the course is to teach the students scraping techniques (and not to create a nuisance for Funda).

- The output of the above exercise should be (i) a working python package that implements the scraping logic, complete with unit tests, and (ii) a dump of scraped data in either json or csv format.

2.3 Programming machine learning for business

In this course we finally move to analyzing the dataset using machine learning, to approach an answer to the business case.

The goal of this module is to build models that (i) can be used to determine which variables of a house influence its asking price (whether positively or negatively) or its time to sale (again, positively or negatively) - an *inference task*, and (ii) can be used to predict the asking price and time to sale (a *prediction task*). These are the tasks that need to be carried out:

- Take the housing dataset provided by the teacher and enrich it using (i) the data mapping postcodes to gemeente and provincie, and (ii) possibly data available publicly on the website of the CBS
- fit (i) a linear regression model and (ii) a random forest model using all the features to predict (i) the asking price of a house and (ii) the time on market of the houses (for a total of four fitted models). Remember to leave part of the original data out as a test set, and calculate the predictive performance of the models using cross validation, as discussed in class.
- Answer the following questions:
  - Which model has better cross validation predictive accuracy in the two predictive tasks, the linear regression or the random forest?
  - Which model has better test predictive accuracy, the linear regression or the random forest?

    Do your models overfit or underfit the data?
  - Which model is better for the inference task, the linear regression or the random forest?
  - Are the assumptions of the linear regression satisfied?
  - How could we improve on the performance of the linear regression? And on the performance of the random forest?
- See if you can improve the predictive accuracy of your model (you can now use all the data to carry out crossvalidation). Hint: see if you can extract

more features from the text description, using e.g. simple word frequencies or topic modelling

2.4 Data Visualization for business

In this course we learn the story from the results that we had obtained from analyzing the dataset.

The goal of this module is to build a deck of slides that (i) can be used to convey the story behind the effects of various variables that the students have defined as the main subject of their investigation towards listing time and house price, and (ii) can be used to advise Funda and housing agents and brokers to improve their strategies in selling the houses faster and more profitable for their clients. These are the tasks that need to be carried out:

- Visualize the results of your analysis with based on the predictive accuracy. Identify the top performing model and provide a discussion why the model is the optimal model.
- Identify and visualize the effects of the determinants that play significant role in these models in reducing selling time and/or increasing selling price.
- Conclude your research and specify action points for Funda and the housing agents.

3 Data ownership

In the course of their work on the case studies students will produce artifacts in the form of source code. The students will retain complete ownership of the code they produce. Any code provided by the instructors will also remain in ownership of the instructors and is provided only for the purpose of learning; students do not thereby derive the right of using the code for other purposes without express permission from the instructor. Similarly, the original source data remains the property of the original parties.