Data Science in Action

Master Thesis Project

Data Science & Society

Course Number: 880502
Department of Cognitive Science & Artificial Intelligence
Tilburg University
Academic Year 2021-2022

Thesis Coördinator: Dr. Marijn van Wingerden
Email: msc-dss-thesis@tilburguniversity.edu

Program Director: Dr. Andrew Hendrickson
Email: A.Hendrickson@tilburguniversity.edu

Objectives

The purpose of the Data Science in Action projects is for the students to practice with the data science methodology, including but not necessarily limited to data wrangling, exploratory data analysis, preprocessing, data visualization, data mining and machine learning. For their projects, students use the R and/or Python programming language with accompanying libraries in an appropriate and correct manner. They are expected to approach the data scientific problems and questions pertaining to their project with curiosity, creativity, in an analytical manner, and as analytical thinkers. Students are required to translate complex and often extensive practical requirements (for instance, those of a commercial or governmental organization, or a research institution) into a work plan for developing, improving, or extending a data science solution. The proposed solution will support specific decision making and problem-solving processes and generalize to other similar contexts and new data.

Using an existing data set provided by a third party, students identify a substantive research question that can be addressed using the specific large data set. In order to formulate an appropriate research goal, students will actively develop in-depth knowledge about a specific application area that will be discussed in the theoretical background for their thesis. Students are supported by experts in the domain provided by the data set owner (external supervisor) and are advised to build on their prior expertise in a particular domain (e.g., their Bachelor studies) as much as possible.

The first stage in the project should be a well-crafted individual thesis proposal that provides the evaluating staff members with a clear view on the feasibility of the project. The thesis proposal is presented both in writing and orally during a presentation round organized by the evaluating staff members. If the thesis proposal is successful (receives a “pass”), students continue with the actual research. The end-product of the Master Thesis Data Science in Action (DSiA) project is the Master thesis. Next to that, students present the outcomes of their project to the general public during an open graduation session on a scientific poster accompanied by a short presentation.

Thesis Content Requirements

Length

The length of the manuscript should be 8,000 words ±10%, excluding references and appendices. Students are required to list the number of words on their Title Page. Theses that do not fall in this range will be automatically failed.

Elements of your thesis

The thesis consists of the following sections:

- Title page
- Preface
- Abstract
- Data Source/Code/Ethics Statement
- Introduction
- Related Work
- Methods
- Experimental Setup
- Results
- Discussion
- Conclusion
- Acknowledgements
- References
- Appendices and Supplementary Materials

Title page

Contains the title, author and other standard information. See template for details. The title summarises the substance of your thesis. Typically, it informs readers about what the research topic is and how it is being investigated; findings and other details are usually left out. Ideally, it should be less than 1215 words. Here are some rules of thumb for formulating the title:

- Be clear and avoid ambiguity

  ✗ LCA of behavioural characteristics among TB patients

  ✓ Latent Class Analysis of behavioural characteristics among tuberculosis patients
- Avoid being overly general or vague

  ✗ Why do people evade taxes?

  ✓ Personality correlates of tax evasion behaviour among Dutch
- Be succinct; the finer details should be included in the Abstract (see section below)

  ✗ Support Vector Machines outperform other classification methods in forecasting stock market movement

  ✓ Forecasting Stock Market Movement with Support Vector Machines

Abstract

The summary is a very brief but self-contained account of your thesis. It should be around 150-250 words. The following points should be addressed:

- What problem is being investigated?
- What is your research question? The research question should follow from how other researchers addressed the problem (i.e., in terms of approach, focus, etc.) in the past.
- What distinguishes your approach from theirs? What are the essential features of your method?
- What dataset are you using?
- What are the main findings?

Data Source/Code/Ethics Statement

It is important to acknowledge the source of your data and/or code, to indicate under which ethical permission experiments were performed and/or data were obtained and where your project code can be found. Also indicate the license under which images used in the thesis were cleared/obtained. You can use this sample text:

- "Work on this thesis (did/did not) involve collecting data from human participants or animals. [if experiments were carried out: the data collection policy and experimental design was evaluated by the TSHD Research Ethics and Data Committee to be in compliance with requirements of the GDPR EU, approval number:].
- Student research with human participants only requires evaluation by the REDC ***if*** the collected data will be used by supervisors ***or*** with a serious intend to publish the results in a journal that requires formal ethical clearance.
  - See: https://www.tilburguniversity.edu/sites/default/files/download/TSHD_Flowchart_studentresearch.pdf.
- The contrary is also the case: Student research with human participants does not have to be evaluated by the REDC ***if*** the collected data will not be used by the supervisor ***and*** will not be used for publications in journals that require ethical clearance. In that case, self-evaluation with the checklist below is sufficient.
- The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.
- [In case the data or code used in the thesis was obtained from an external source:] The author of this thesis acknowledges that they do not have any legal claim to this data or code.
- [In case the data or code used in this thesis was obtained from an external source was added to, changed, or was the basis for the acquisition of new data or code, and if the additional data was collected from human participants or animals]: The author of this thesis has evaluated his/her project according to the "Ethics checklist Student research with human participants".
- [If the data will be used for research projects for TiU-based researchers] the supervisor of this research project obtained an evaluation of compliance from the TSHD Research Ethics and Data Committee."
- The code used in this thesis is/is not publicly available [list repository]
- Images used in this thesis, when not produced by the author, were licensed under [list creative commons or other licensing mechanism / permission image owner]

Introduction

Explain briefly what your research questions are, why they are important, how you approached them, and what your findings were.

- Context
    - Start with the goal of your research and how this addresses your problem statement
    - Describe the context of your thesis in a very concise manner. Briefly explain the research domain, what the state-of-the-art is, and why the subject matter is interesting. Your readers should be left feeling that your thesis deals with an issue that is both important and interesting.
    - You do not need to go into great lengths to describe every relevant prior study yet; that should be reserved for the section on related work. For now, it is fine to state something along the lines of: *This issue has been addressed extensively (see Section 6).*
    - Devote one paragraph of the introduction explicitly to the scientific relevance of your project. Note that scientific relevance could be derived from the domain-specific research question addressed in your research, or in the proposal of a new algorithm or approach.
- Research questions
    - Once the context is established, specify you research questions. If necessary, describe very briefly how each research question will be answered.
- Findings
    - Give a brief (one paragraph) overview of your main findings.

Related work

Related work, sometimes labeled as theoretical framework or background, is a crucial element of your thesis. Explain the larger scientific context of the problem: what is the theory behind it if any, what previous research has been done related to it, and how your work builds on this related research. Below are some step-by-step guidelines for writing this section:

- Specify the area of research in which your work belongs and provide a context for the research focus. What research issue is your work focused on? Why is it an issue of importance?
- Describe relevant work conducted in the same research area (with proper references). Has this issue been addressed in the literature? By whom? What have they done and found? What are the relevant theories? Are there any contradicting findings or competing models/theories? What is the state of the art?
- Identify research gaps and/or shortcomings of existing method; define research problems:
    - What is missing from prior research? What are the limitations of existing models? Could there be alternative approaches to solving the same problems?
    - Specifically, what research problems are left unanswered? What insights or implications will tackling these research problems bring about?
- Specify the research questions and goals of current work; announce the methodology you are going to adopt. What are the research questions? What are you trying to achieve with

the current work? How are you going to fill the research gaps? What sets your work apart from prior research? How did the literature review inform your methodological choices? What dataset are you going to use?

Common problems and remedies:

✗ Failure to maintain focus on the research question, by including references to studies that are only remotely related to yours.
✓ Make sure you are not trying to impress the readers by the broad scope of your knowledge, thereby forgetting that they are interested in your current research only.

✗ Failure to support statements with adequate references.
✓ Always give credit where credit is due. If you are making a statement along the lines of: *It has been established in prior research that…*, make sure you follow the statement with references.

✗ Failure to express arguments or ideas in your own words.
✓ It is not acceptable to simply paraphrase the work of someone else by changing a few words here and there, without acknowledging the source. If you must include a direct quote, enclose it with quotation marks and specify the page number in your reference. Failure to do so is a case of plagiarism and can lead to severe consequences!

✗ Failure to include references to recent work.
✓ Whilst certain dated works remain important and are still widely cited (e.g., Gold, 1967, if the research concerns grammatical inference), try to stay on top of developments in the field and refer to the more recent literature.

✗ Failure to critically reflect on the literature.
✓ Demonstrate awareness of relations among existing models or studies by specifying any relevant commonality, contradiction, or inconsistency among them.

✗ Failure to give a convincing rationale for conducting the current study.
✓ Explain how the current work continues and improves upon previous lines of enquiry. Be explicit about the contribution of the current work.

Methods

In this section you describe your general approach, for example which mathematical models or computational algorithms you used. This is different from your experimental setup: here you should provide the description of your modeling approach, for example describe formally a new type of model that you are proposing, or a modification of an existing model, in a general way, that is without reference to the specific way you tested it empirically.

You would usually explain the methods using a combination of mathematical formulas, diagrams, and verbal descriptions. Make sure that you provide a justification for the methods used, compared to the alternatives.

Common problems and remedies:

✗ Symbols in formulas are not defined or explained.

✓ Make sure that it's clear what the notation stands for.

Experimental Setup

Experimental setup is a section where you describe in detail your dataset and the experimental procedure. Other researchers should have sufficient information to replicate your work based on this section alone. The following information should be covered:

- Description of your raw dataset: the organization offering the dataset, sample size, how and when the data was collected, which features could be found in the data, and any other relevant information
    - Where appropriate, report (descriptive) statistics to offer a better impression of the dataset or selected results of exploratory data analysis.
    - Cleaning / preprocessing of your data; was there any oddity (e.g., error) in the dataset and what was done about it, which parts of the data were discarded and why, whether or not certain features were transformed and why, what was done about the missing values and why, and any other preprocessing done.
    - Always justify your decisions with theoretical and/or statistical arguments
- Description of the experimental procedure: what was the task, which algorithm was used, which parameters were chosen and how.
- Description of the actual implementation, i.e., programming languages and versions, packages used, proprietary applications supporting the coding, etc.
- Description of evaluation criteria: for example, which error measure was used (e.g., classification accuracy, mean squared error, f1 score).

Common problems and remedies:

✗ Failure to perform the correct task.

✓ Make sure that the tasks you choose to perform are in accordance with the type of data. For example, you cannot perform Principal Component Analysis (PCA) on categorical features.

✗ Failure to list all important details.

✓ Always write with other researchers in mind and include all relevant details. When in doubt, ask yourself: If I were to leave this piece of information out, would other researchers be able to reproduce my work?

✗ Failure to justify choices made.

✓ Always be explicit about the rationale for making certain choices; they should be made on theoretical (e.g., prior research), methodological (e.g., algorithmic bias) or empirical grounds (i.e. tuning on validation data). For example, it is better to use only one or two algorithms properly than trying out every algorithm under the sun without proper justification and in a superficial way.

Results

In this section, you report your results, often with the help of statistics, tables, and figures. Below are some guidelines:

- Present your results in a structured manner, often with the help of tables or figures.
    - In your text, do not simply restate the information listed in the tables or figures. Try to make sense of the results, highlighting important or interesting findings that you might revisit in the discussion section. The figures are not the presentation of your results but their illustration.
    - Do not leave information presented in tables or figures unexplained. You have included information there for a reason, so take the time to go through it (e.g., explain what each column is about).
    - Provide high quality clear figures with well-sized legends and informative captions.
    - Do not cluster tables and figures – there should always be some text in between.
    - Larger tables (with more than ca. 10-15 rows) should be placed in the appendix.
- Where appropriate, explore the results further by means of statistical analysis, confusion matrices, or visualizations
    - The goal is to obtain a more fine-grained understanding of your results, uncovering patterns that might not be obvious from the overall results (for example, does the overall pattern of results hold across ages and genders?

Or, in the case of a AI model, does the predictive performance of the model differ greatly between classes).

Common problems and remedies:

✗ Failure to report the baseline performance.

✓ Always report the baseline performance, as it is difficult to interpret the results without knowing the basis for comparison (e.g., previous research, chance-level performance, etc.)

✗ Failure to interpret information listed in tables and figures.

- ✓ Instead of simply restating what is listed in the tables and figures, explain the substantive meaning of your findings such that your readers know what they should focus on .

✗ Failure to use the correct type of figure.

- ✓ Consult scholarly articles and books to see which type of figure is appropriate for which visualization purpose.

✗ Failure to format numbers according to English-language conventions.

- ✓ Make sure you use decimal points, and commas as thousand separators (i.e. 1.2 and 10,000)

Example structure for reporting results:

- First, describe in one or two sentence(s) what results will be reported in this section. For example: In this section, classification performance for the feature types described in section 3 on X dataset will be presented.
- Describe the classification performance, which should be listed in a table too. Let your readers know which table you are referring to. Do not simply repeat the information found in the table; tell your readers: Which approach yields the best performance? Which the worst? Is there any other noteworthy result? Are your results similar to those reported in prior research? For example: The results of the classification tasks on the dataset are shown in Table 1. Both approach A and approach B yield the best classification performance. Approach C performs considerably worse, a result that contradicts prior research [reference(s)] and our expectations.
- Explore your results further. For example: Does the classification performance differ across classes? What contributes to poor classification? It might be worthwhile to conduct additional classification tasks, for example on a subset of the dataset. Present the additional results in a table too and try to make sense of these additional findings. For example: “We analyzed the classifications of each Y [= what is being classified] individually to gain a better understanding of why approach C did not perform as expected. It appears that the poor classification is due to the following reasons: [list the plausible reasons here]”

Discussion

In this section, you should evaluate your results regarding the research questions listed in the introduction. The following are some recommended elements:

- Remind you readers what the goal of your study was.
- Discuss the findings, preferably in the same order in which they were presented in the results section.
- If a finding was surprising, you should offer reasonable speculations as to why this particular result was observed.
- It could be the case that your results only partially answered your research questions due to limitations of the model or the data. Acknowledge these limitations, offer possible solutions, and defend the validity of your results.
- Put your results in perspective by making links to the literature.
- Make very clear what the contribution of your study is within the existing framework.

Common problems and remedies:

✗ Failure to provide context for the results.

- ✓ The discussion section should be understandable when standing alone. It is important that you spell out your research questions and goals again, so that researchers who only read this section can still have a good idea of what your findings are.

Conclusion

Conclusion is a short section where you restate the research questions and provide the answers to them by combining the results you obtained with a very brief summary of how they can be placed in the context of existing research. Here you also explain the implications of your work for the field, and identify which directions future research could take based on the contribution of your study.

Acknowledgements

In this brief section you can acknowledge sources of funding, data, or anyone who helped you with your research.

References

It is recommended to use an author-year citation style such as APA.

Appendices and Supplementary Materials

Appendices are appropriate for extra, non-essential visualizations, examples and analyses.

It is strongly encouraged, whenever possible, to store data and source code in an online repository and refer to it in the manuscript. Github.com is a common choice.

Copyright and collaboration

Do not reuse any figures or other images without a proper license or permission of the author; if you have the permission, the author still needs to be credited. See also the Data/Code/Ethics statement section

If you collaborated with someone else on some part of the project, indicate clearly which part of the work you are building upon was done by something else.

Plagiarism/Overlap

Make sure your thesis does not contain unacknowledged quotes or paraphrases as these could be detected as cases of plagiarism. Please consult the plagiarism FAQ in the Canvas course and the policy regarding overlap at the School level:

https://www.tilburguniversity.edu/students/studying/regulations/eer/humanities

Textual/conceptual overlap indicative of plagiarism is grounds for failing a thesis submission.

Checklist

Make sure to check the following items before you send your manuscript for review to your supervisor and/or second reader.

- Is the length within the specified limits?
- Did you run a spell checker?
- Did you proofread for grammar and clarity?
- Did you use English-language conventions for number formatting (decimal point, comma thousand separator)?
- Did you round excessively precise numbers?
- Are all quotes and paraphrases from other texts properly referenced?
- Are all figures either your own or used by permission?
- Are the symbols used in formulas defined or otherwise explained?

Useful sources

Below are some useful resources for writing your thesis.

- APA style
  - APA publication manual (6th ed.). Available from the library. ISBN: 9781433805622
  - Purdue online writing lab: https://owl.english.purdue.edu/owl/section/2/10/
- General tips on writing and mechanics of style (e.g., punctuation, grammar, spelling, sentence structure, etc.)
  - *How to write a paper* (7th edition), by Ashby, M. F. (2011). Available from: http://www.grantadesign.com/download/pdf/How_to_write_a_paper_6th_editi on_2005.pdf