

Elements of AI: Capstone Project  
OZNAL February – June 2025

The capstone project serves as a pivotal assignment, bringing together the knowledge and skills you've developed throughout your studies. It involves an in-depth exploration of data to identify a real-world challenge and prompts you to ask: How can I design a machine learning system that effectively resolves this issue — for myself or for a company dealing with it? Embracing this type of challenge is crucial if you aim to distinguish yourself from the competition.

This project also acts as a benchmark for your ability to meet the selected learning outcomes set by the AHEP (Accreditation of Higher Education Programmes) framework[[1]](#footnote-1), established by the Institution of Engineering and Technology.

1. Project Outline

Pick a meaningful problem that you can tackle using the machine learning techniques covered in class and publicly available datasets. Decide whether to go with a classification or regression problem, choose the right method, and assess how well it performs. Then, dive into some technical literature to find a method that could improve the one you started with. Apply the new method to the same data and compare the results.

Back up your choices with a requirement analysis, exploratory data analysis (EDA), and anything else that supports your reasoning. Finally, present your project to the examiners, showing off your ability to explain and justify each step of model selection and data processing. Wrap it all up with a clear summary that outlines rationale for solving the problem your way.

2. Selecting Data

The first step for your project is to dig into public repositories and find some cool data to work with. Check out platforms like GitHub, Kaggle, or public APIs — like data.bratislava.sk — where you’ll find loads of datasets covering all sorts of topics. Pair up with someone you’d like to work with and share the workload to get the job done smoothly. Once you’ve picked a dataset, make sure to add it to the spreadsheet:

<https://docs.google.com/spreadsheets/d/1Xc5hy6af_4l_VlqnQIP3tPcAkr65kVEx/edit?usp=sharing&ouid=116584584157405496240&rtpof=true&sd=true>.

Each dataset can only be used by two teams, so double-check what others have already snagged before locking in your choice.

3. Hypothesis-Driven Project

After gathering your data, the next move is to craft a solid hypothesis aligned with your questions. For instance, you might explore whether social media activity impacts stock prices. Or you might be wondering whether football player’s skill are features good enough to classify players into offensive, defensive and midfielders. Dive into the data, analyse basic stats, and check if its features, format, quality, and quantity are suitable for testing your hypothesis using machine learning. If things don’t line up, take the time to get new data or preprocess the data to meet the requirements of your chosen method. Keep experimental reproducibility in mind and document every step along the way. Make sure to explain the reasoning behind your decisions. For example,

* if **data transformation** is necessary, clearly explain why this step is important and how it contributes to your analysis.
* **handle any missing values** in your dataset by using appropriate methods to fill them, ensuring the data remains consistent and usable. However, consider whether filling the gaps is essential — could using only complete cases provide enough data for applying machine learning? Support your decision with a power analysis to validate your approach.
* when required by the model prerequisites, think about **removing outliers**. Be cautious when removing outliers from already scaled data and critically evaluate whether this step genuinely improves the model's performance.
* if **dimension reduction** is carried out prior classification or regression, provide information why this step is necessary and what are you trying to achieve.

4. Requirements Analysis

Analyse the feature statistics to determine which regression or classification techniques align best with your dataset. Support your selection with insights derived from your exploratory data analysis (EDA). If multiple methods appear equally suitable, justify your final choice by outlining your decision-making process and highlighting the advantages of the selected approach. Incorporate a table in your report to summarize all the requirements of the chosen model and demonstrate how your feature data fulfils these criteria.

5. Data Spending

Choose an appropriate cross-validation method to ensure the reliability of your machine learning model while avoiding overtraining. Allocate your data based on the dataset's complexity and its relevance to the objectives of your analysis. Justify your choice by explaining why this cross-validation method is most suitable for a given scenario.

6. Model Fitting and Performance Evaluation

Based on your chosen method, perform function fitting to your data and assess model performance. Explain why you have selected function fitting or feature partition approach. Carefully document each step of the process, emphasizing the formulation of the model and avoiding overcomplication — such as refraining from using a polynomial fit for linearly dependent data. Consider why using a single comprehensive model might be more effective than combining multiple simple ones. Evaluate which approach to model error estimation best fits your scenario. Use model object to perform such analysis.

For classification models that estimate probabilities (logistic regression, linear/quadratic discriminant analysis), conduct a ROC analysis to identify the optimal cut-off point and translate this probability into a specific feature value used for classification. When employing a support vector machine, determine the appropriate margin rather than applying arbitrary values. Additionally, clarify whether false negatives and false positives carry equal importance in your case, and address similar considerations for other specific aspects of your chosen method. Summarize your results using confusion matrices and provide a detailed interpretation of the outcomes. Finally, compare your model's accuracy to the "No Information Rate" to evaluate its effectiveness.

7. Researching Technical Literature

Analyse the advantages and limitations of your selected method, using your data to exemplify each point. Investigate online resources to identify an alternative model — whether for classification or regression —that effectively addresses the identified weaknesses.

Summarize the content of a relevant online source concisely, explaining your rationale for selecting it. Provide an overview of its key findings and construct a persuasive argument demonstrating how the suggested model enhances specific attributes such as accuracy, speed, or interpretability. Remember that numerous advanced models and extensions exist beyond the scope of this machine learning primer, and this task is designed to encourage independent exploration. **The use of deep learning, including neural networks, is strictly prohibited in this project.**

8. Model Comparison

Conduct a thorough requirements analysis before implementing a second model to compare its performance against the initial model. Identify the specific area where the new model provides enhancements and justify your choice of that area over others. Clearly explain why certain aspects, such as explainability over performance, hold greater value for your case, and provide detailed documentation to support the relevance of these priorities to your analysis.

9. Summarise and Visualize

Prepare your presentation using a combination of one-pager summary in lazy language, technical project documentation in R markdown (without length restriction, just make sure that you have every key point explained) and Shiny application for your oral presentation.

* **One-Pager summary:** Provide a concise overview of your domain background, key findings, and proposed solution, limited to one A4 page. Clearly explain why the chosen models are the most suitable for testing the hypothesis and how your research contributed to identifying an improved model. Summarize your conclusions regarding model development and domain insights in accessible, non-technical language to ensure widespread understanding. Submit this as a Word document or PDF.
* **Project documentation:** Create thorough project documentation using R Markdown. Offer detailed explanations for each step of your code to ensure clarity and accessibility for a technical audience. Comments all model parameters, inputs, and outputs. Provide detailed interpretation of the results.
* For the oral presentation**,** create **a Shiny application** that showcases the data and fitted model to the client. The application should feature a user interface that enables data loading, function fitting, and dynamic model parameter adjustments. Any modifications made should automatically update the model. A working example of such an application was demonstrated in the last lecture, but you are free to explore Shiny's feature demos (<https://shiny.posit.co/r/gallery/#feature-demos>) for additional inspiration.

10. Code, ChatGPT and Proper Referencing of Scientific Literature

Although the use of LLMs like ChatGPT is not entirely restricted, you must be able to explain every piece of content they generate, particularly when it comes to R code. The project idea about which hypothesis to test on data of your interest must be entirely yours. So should every explanation provided in the context of experimental setup and execution.

Failure to provide adequate explanations will result in no or reduced points being awarded for the corresponding task. Also be sure to properly reference online sources and any scientific or technical literature used in your project.

11. Teamwork

You will collaborate in teams of two. Open and effective communication with your partner is essential when dividing the workload, assigning tasks, and ensuring responsibilities are clearly understood. Take the time to plan the project's timeline, establish checkpoints, and monitor progress to keep the work on schedule. Actively support each other throughout the process, address challenges together, and stay aligned with the shared goal of delivering the project successfully and on time.

12. Deliverables

Submit your project report in three distinct formats, demonstrating various skills acquired throughout the course. The one-page summary and R Markdown report must be uploaded to the **AIS** system before the deadline. The corresponding web link will be available once the site is accessible. The Shiny app should be prepared to run on your notebook and ready for the oral presentation.

13. Grading

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **AoL** | **Points** |
| **One-pager** | Lay summary |  | 5 |
| **Project report** | Refine the problem statement, select relevant data, and conduct exploratory data analysis (EDA). Perform requirements and research analysis, followed by the implementation of a second model and a comparison of results, culminating in a comprehensive summary. Includes 2 models and their through comparison. |  | 25 |
| **Research** | Clearly justify the chosen area for improvement and explain how this research contributes to addressing the identified shortcomings. |  | 3 |
| **Code** | Demonstrate a thorough understanding of the project's underlying code. The LLM code will undergo rigorous testing to assess comprehension and will influence other sections of the work. |  | 2 |
| **Oral presentation** | Develop and present a Shiny application that integrates the implemented models, providing compelling evidence to support their adoption. |  | 5 |

14. Important Deadlines

* Project documentation and one-pager summary: must be submitted to AIS before 10:00 AM 5th May 2025.
* Oral presentation with application in Shiny:

Selected projects: 7 May 2025, from 10:00 AM to 12:00 PM.

All other projects: 9 May 2025, from 10:00 AM to 3:00 PM.

15. Last Update

This document was last updated on Tuesday, 15 April 2025.

1. Areas of Learning (AoLs) that you will be tested in the capstone project.

   |  |  |
   | --- | --- |
   | **M1:** Comprehensive application of mathematics, statistics and engineering principles to solve complex problems.  **M2:** Formulate and analyse complex problems to reach substantial conclusions, even with incomplete information.  **M4:** Conducting research to address complex challenges. | **M5:** Select and apply appropriate tools, techniques and resources for engineering tasks.  **M7:** Communicate effectively to both technical and non-technical summary about engineering concepts used in your project. |

   [↑](#footnote-ref-1)