Data Science 1 - Final Project

Updated 16 April 2024

Dr.-Ing. Christian Debes
M.Sc Pertami Kunz

1 Introduction

The objective of this final project is to demonstrate all your acquired knowledge in data science throughout a practical problem to be solved. This includes a proper problem formulation and motivation, determining the suitable algorithms and presenting your outcome. It is also a perfect preparation for the project seminar Data Science II.

2 Task Instructions

- Dataset selection
  Begin by selecting a dataset that captures your interest. This dataset should be relevant to a question or problem you are eager to explore. The dataset does not have to be big, however it should be rich enough to support comprehensive analysis, including hypothesis testing and predictive modeling.
- After selecting your dataset, you will need to decide on your analytical strategy. You have two distinct options to choose from:

Option A **Separate Hypothesis Testing and Predictive Modeling:**
Hypothesis Testing: Formulate a specific hypothesis related to your dataset that you wish to test. This hypothesis should be based on an observed phenomenon or a theoretical premise that can be empirically tested using statistical methods.
Predictive Modeling: Independently from your hypothesis test, identify a response variable within your dataset that you aim to predict. Utilize the remaining variables as predictors in your model. You can choose to apply either regression or classification techniques based on the nature of your response variable.

Option B **Integrated Approach - Hypothesis Testing in Predictive Modeling:**
Model Selection: Begin by defining a response variable that you intend to predict using some or all other variables in your dataset.
Hypothesis Testing on Models: Instead of directly comparing model performance through metrics like accuracy or R-squared, conduct hypothesis tests to statistically evaluate the difference in performance among multiple predictive models. This approach allows you to assess whether one model significantly outperforms others, ensuring that the observed differences in performance are not merely due to random chance.

3 Report Checklist

Your final report should include the items below (ones with ✓). They do not need to be always in this given order. Each item does not need to be in a separate section/subsection. The important point is to make your report **concise and fluent**.

- Introduction

  - ✓ (5 pts) Objective.
    Introduction to the objective of your report. Formulate it in such a way that it is understandable by a non-technical audience. Describe the business/societal or other benefits arising from it.
  - ✓ (5 pts) Dataset Introduction. Comprehensive overview of the dataset, including its origin, structure, and relevance to the objective.
  - ✓ (5 pts) Problem statement. Clear formulation of the problem the report addresses, highlighting its significance and challenges.
- ✓ (10 pts) Some preprocessing.
  Adequacy and justification for the preprocessing steps undertaken to prepare the dataset for analysis.
- Exploratory Data Analysis
  - ✓ (20 pts) Some **useful** plot(s) and/or table(s).
    Effectiveness and relevance of visualizations and tables in uncovering insights and guiding the research question or hypothesis. This includes clarity, accuracy, and the ability to communicate complex information in an accessible manner.
- ✓ (10 pts) Some feature extraction/engineering
  Creativity and technical rationale behind the feature extraction/engineering steps, including any transformations, imputations, or derivations of new features.
- Hypothesis Testing
  - ✓ (3 pts) Clear statement of the hypothesis being tested.
  - ✓ (4 pts) Explanation of the choice of test statistic and its relevance to the hypothesis.
  - ✓ (8 pts) Comprehensive detailing of the hypothesis testing steps, including the derivation of relevant values (p-value, confidence interval, etc.) and the final conclusion drawn.
- One/some regression or classification models.
  - ✓ (3 pts) Explanation of the data splitting strategy for training and testing.
  - ✓ (5 pts) Clear mathematical expression of the model(s) used and justification for their selection.
  - ✓ (4 pts) Explanation of how hyperparameters were chosen (if any).
  - ✓ (8 pts) Detailed analysis of model performance on training and test sets, including metrics used for evaluation.
- ✓ (5 pts) Conclusion
  Concise and insightful summary of the key findings, implications, and potential for future work.
- ✓ (5 pts) Proper citation
  Accuracy and completeness in citing all sources, datasets, and Python libraries referenced in the report.

4 Useful links

Here are some links to choose a dataset from (you are free to choose from other sources):

1. UCI Machine Learning Repository
2. Kaggle Datasets

Examples:

1. Amazon's Books EDA and Hypothesis Test

2. A Statistical Analysis & ML workflow of Titanic
3. More hypothesis tests examples
4. Regression examples
5. Classification examples

References for hypothesis test on machine learning algorithms:

1. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms
2. Statistical Comparisons of Classifiers over Multiple Data Sets
3. Inference for the Generalization Error