

Team 32

Casper Collet and Chu Li

**Abstract:**

You probably know someone who has been tackled by divorce, and know most of the times it brings a lot of negative effects with it. Issues around rights over children, money and mental health are some the popular ones. What if divorce can be predicted and evitable if dealt with the correct issues? In this paper, we are going to try and predict if couples are divorced using different machine learning methods. We will use a dataset with different questions for couples that were answered using a scale from 0-4, 0 being not applicable and 4 being applicable. We will try categorizing the questions into groups, finding the most important variables, and using all variables on their own so we can find out what method results in the highest accuracy and predictability of divorce.

Course: Data Science Lab (ECB3DSL), Academic year 2024/2025

Team: 32

Casper Collet, 2860422, c.b.collet@students.uu.nl

Chu Li

Supervisor(s): Dr. Ulrich Zierahn-Weilage, Dr. Tina Dulam

Word count:

**Statement of Originality**

We, the aforementioned students, herewith declare to have written this document and that we are responsible for the content of it. We declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it.

Utrecht University School of Economics is responsible solely for the supervision of completion of the work, not for the content.

**Disclosure Statement**

In this project, we have made use of the following Generative AI tools:

We have used these tools for: Helping write the code, grammar checks in the paper and brainstorming.

**Division of Work**

We, the aforementioned students, herewith declare that we have divided the work on this project and this project paper as stated in the following table:

|  |  |
| --- | --- |
| Section and Content | Name and Student Number |
| 1 Introduction | All students |
| 2 Data | [Student Name & Number 1] |
| … | … |
| 5 Conclusions | All students |

Signatures

[copy scans of your signatures here, add your student name and number]

# Introduction

The research question for our paper is: How strongly do different answers to questions about values in marriage forecast divorces? We chose this question because it allows us to experiment with different machine learning models to assess their predictive power on the dataset and identify which specific answers are most significant for the likelihood of divorce. Our goal is to provide these insights to couples considering marriage. By reflecting on the questions and ideally reaching a mutual understanding before getting married, couples may reduce the risk of divorce and increase the likelihood of a happier marriage.

In our analysis, we identified the three most important variables in our dataset, which correspond to the following questions. We employed four different machine learning models to evaluate the importance of these variables, including Random Forest, Bagging, Lasso, and Ridge regression. The variables we highlighted are those that ranked the highest overall across all models in the importance plots.

# Data

The dataset used in this paper is the divorce\_data dataset from Kaggle.com. It consists of 54 questions, ranging from “If one of us apologizes when our discussion deteriorates, the discussion ends” to “I'm not afraid to tell my spouse about their incompetence.” Each question is answered with a number from 0 to 4, where 0 indicates that the statement is not true for the couple and 4 indicates that it is very true. In addition to the 54 questions, there is a 55th question that indicates whether the couple is divorced (1) or not (0).

Initially, we began by cleaning the data. This was not a significant task, as the dataset was fairly clear and contained few, if any, missing values. We ensured that the 55th question was formatted as binary to facilitate our analysis.

Next, we experimented with different methods to determine which yielded the highest predictability regarding divorce. We divided the questions into three distinct groups: Communication\_data, Value\_alignment\_data, and Knowledge\_of\_spouse\_data. Communication\_data included questions about communication skills, Value\_alignment\_data focused on whether the life values of the couples were aligned, and Knowledge\_of\_spouse\_data contained questions about the couples' understanding of each other.

However, we quickly realized that splitting the questions into three groups did not effectively address our research question or help us identify the most significant variables. As a result, we decided to abandon this grouping method and proceed with alternative machine learning techniques. These where Random Forest, Bagging, Lasso and Ridge.\

# Descriptive Analyses

[Provide those descriptive analyses, which are necessary to understand the relevant features of your data that matter for the research project]

# Empirical Approach

[Explain the empirical approach of your project – how this should look like strongly depends on your specific type of project]

# Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random Forest | Bagging | Lasso | Ridge |
| Accuracy |  |  |  |  |
| Important variable 1 |  |  |  |  |
| 2 |  |  |  |  |
| 3 |  |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| 6 |  |  |  |  |
| 7 |  |  |  |  |
| 8 |  |  |  |  |
| 9 |  |  |  |  |
| 10 |  |  |  |  |
| \ |  |  |  |  |

# Conclusion

[Conclude: What do you find? How does that contribute to the research question? How does that contribute to the relevant scientific and societal debate? What remains open?]

# References

Last Name 1, First Name 1, and Last Name 2, First Name 2 (year): “Title”, Journal Name, Volume(Issue):PageX-PageY.

# Appendix

[The appendix is optional and you do not necessarily need it. How to decide whether to put something in the appendix or main text? That is a subjective decision. Broadly speaking, the main text should be self-contained and everything that matters for the main line of arguments should be in the main text. The appendix instead is for smaller arguments and checks that are not necessary for the main story of your project.]