

Team 32

Casper Collet and Chu Li

**Abstract:**

You probably know someone who has been tackled by divorce, and know most of the time it brings a lot of adverse effects with it. Issues around rights over children, money, and mental health are some of the popular ones. What if divorce can be predicted and evitable if dealt with the correct issues? In this paper, we will try to 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 Abstract/Introduction | Casper Collet |
| 2 Data | Casper Collet |
| 3 Descriptive Analysis | Casper Collet |
| 4 Empirical approach | Chu Li |
| 5 Results | Casper Collet: Random Forest, bagging  Chu li: Lasso, Ridge |
| 6 Conclusions | Casper Collet: Random Forest, bagging  Chu li: Lasso, Ridge |
| References | Casper Collet |

# Introduction

The research question for our paper is: How strongly do different answers to questions about values in marriage forecast divorces? This question was a result of two questions we would like to have answered. These are: 1. What is the predictive power of machine learning models when predicting divorce? And 2. What are the most important factors that drive divorce? These are the questions that we are going to find out, and the research question is allowing us to test and try to find the answers. 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 employed four different machine learning models to find the variables and their importance: Random Forest, Bagging, Lasso, and Ridge regression. They all showed different rankings of importance for the variables, but we focus on the ones that came up in all four models. In this way, we can state that these variables with their associated questions are most likely of great importance for couples to discuss when entering marriage.

# 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 valid for the couple and 4 indicates that it is very valid. In addition to the 54 questions, there is a 55th question that indicates whether the couple is divorced (1) or not (0). The questions were answered by 170 different couples, of which 84 were divorced (1) and 86 were not divorced (0).

# Descriptive Analyses

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. Some variables had backticks in their names, which we had to remove due to problems with the coding. For instance, the variable “I’m\_right” became “I\_am\_right”. This solved the coding errors that occurred when working with these variables. At last, 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. It removed a lot of the individual predictive power of the variables since they were put into categories, and this was not what we were looking for. As a result, we decided to abandon this grouping method and proceed with alternative machine learning techniques. These models were the Random forest, Bagging, Lasso, and Ridge models.

To start testing these techniques, we decided to remove 20% of the 170 rows of data. The 20% consisting of 34 rows became the dataset “unseen\_data”, and the remaining 80% consisting of 136 rows became the dataset “seen\_data”. The seen\_data became our testing data for the 4 named models. The internet and different AI tools suggested that we split the 80% again in 80/20% to test for Random Forest and Bagging. We deliberately choose not to do this since our dataset is fairly small, and this might change some of the predictive power of our models.

# 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** | 100% 95% CI : (0.8972, 1) | 100% 95% CI : (0.8972, 1) |  |  |
| **(Important variables ranked on highest mean decrease accuracy) 1** | Humiliate | Anxieties |  |  |
| **2** | Anxieties | Humiliate |  |  |
| **3** | Marriage | Marriage |  |  |
| **4** | Sudden\_discussion | Harmony |  |  |
| **5** | Trust | Idk\_what\_is\_going\_on |  |  |
| **6** | Idk\_what\_is\_going\_ on | Hopes\_wishes |  |  |
| **7** | Harmony | Sudden\_discussion |  |  |
| **8** | Roles | Roles |  |  |
| **9** | Enjoy\_travel | Happy |  |  |
| **10** | Happy | Enjoy\_travel |  |  |

The results of the random forest model had a predictive power of 100%. This is great, but it might have some underlying reasons. The dataset is relatively small, so the predictive power might not be the same on a bigger dataset. It is possibly a result of overfitting, which is why we added Lasso to our model choices. The random forest model's confusion matrix resulted in 21 true negatives and 13 true positives from the unseen dataset. It also showed a Kappa score of 1, which means perfect agreement between prediction and reality.  
The results of the bagging model also had a predictive power of 100%, possibly due to the same reasons as the random forest model. The bagging model’s confusion matrix resulted in 21 true negatives and 13 true positives from the unseen dataset. Same as the random forest model, it showed us a Kappa score of exactly 1.

Comparing the random forest and bagging models, they are identical to each other. This is not that surprising since they almost work the same way. Although this is the case, the importance plots showed different results, as shown in the table above. The overlapping variables from these two models in importance order are: “Humiliate”, “Anxieties”, “Marriage”, “Harmony”, “Sudden\_discussion”, “Idk\_what\_is\_going\_on”, “Roles”, “Enjoy\_travel” and “Happy”. The random forest model also showed “Trust” and the bagging model “Hopes\_wishes”.

The questions that come with these overlapping variables from the random forest and bagging model are (in the same order as given above):

* I can be humiliating when we discussions.
* I know my spouse's basic anxieties.
* My spouse and I have similar ideas about how marriage should be.
* I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
* Our discussions often occur suddenly.
* We're just starting a discussion before I know what's going on.
* My spouse and I have similar ideas about how roles should be in marriage.
* I enjoy traveling with my wife.
* We share the same views about being happy in our life with my spouse.

# Conclusion

Although we have to keep in mind that our dataset is small and the models might not work as well on bigger datasets, we can conclude that our machine learning models are giving a great prediction on divorce. Random Forest and bagging scored a 100% accuracy on the unseen data, which indicates great predictive power. The overlapping variables with the associated questions are an indicator of what should be discussed before marriage to reduce the chances of getting divorced.

To come back to our research question (How strongly do different answers to questions about values in marriage forecast divorces?), we can see that those answers to the questions have a huge impact on forecasting divorce since the accuracy of the models is extremely high. Our research shows that chances of divorce can be lowered if the right questions are asked. This can contribute to a lower divorce rate around the world if the research is applied to a larger scaled data set. Our suggestion focuses mainly on new couples to think about these questions before getting into marriage and the already married couples to rethink their choices.

The main question we have left after the research we conducted is what would have happened to the predictive power if we had a dataset that was so much larger than the 170 rows we have right now. If further projects allow us to dive deeper into this issue, it would be an extremely interesting topic.

# References

Dataset:

Author username: Csafrit, Last updated: 3 years ago,

<https://www.kaggle.com/datasets/csafrit2/predicting-divorce>

Use of AI:

Grammarly AI:

The use of Grammarly was conducted by Casper Collet. It is a built-in AI tool for your computer that helps you with the spelling of words and sentence building.

ChatGPT:

The use of ChatGPT was conducted by both team members. It is a tool that helps you with different things. Mainly, it helped us write and solve code when we did not know what to do. We made sure we understood what the AI tool did and wrote the right inputs to get the right code that we desired. It also gave us some ideas on what to do for the project, but these were terrible (like the classification idea of the questions in three categories), so that wasted a lot of our time.

Copilot AI:

The use of Copilot AI was conducted by Chu Li. It is a built-in tool for Rstudio to help write the code.

# 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.]