Utrecht University

School of Economics

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, 1082477, c.li9@students.uu.nl

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

Word count: 2605

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: Grammarly AI ChatGPT Copilot

We have used these tools for: Helping modify error 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 & Chu Li
2 Data	Casper Collet
3 Descriptive Analysis	Chu Li
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 & Chu Li

1 Introduction

Divorce remains a social challenge. The high rate of divorce worldwide (OCEO,2022) indicates the need to predict factors of marriage instability. The research question of this study 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. 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.

While prior research has explored sociological demographic predictors of divorce, few studies have used machine learning to analyze attitudinal factors such as marital interactions(Hilpert et al., 2023). To achieve this goal, we applied four separate 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, and the prediction accuracy of every model is high. 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.

2 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).

In the data processing stage, we confirmed no missing values through data cleaning. To enhance code clarity and avoid problems caused by special characters, we renamed “Divorce” to “Divorce_Y_N” and renamed variables containing special characters to more efficient identifiers such as “I_Am_Right”. Next, We split the data into 80% training set “seendata” and 20% “Unseendata” for evaluation and robustness testing. And finally, for the Lasso and Ridge models, the predictor variables use glmnet to standardize the features to ensure comparable scales.

3 Descriptive Analyses

We analyzed the dataset as a whole for outcome distributions, overall predictor characteristics, predictor correlations, and visualizations to highlight features relevant to our machine learning model. The outcome variable Divorce_Y_N (0 = not divorced, 1 = divorced) was close to balanced with 86 undivorced couples (50.6%) and 84 divorced couples (49.4%). This balance is essential for unbiased model training, as both classes will not be dominant, thereby ensuring that our four models are able to effectively differentiate between divorce outcomes.

The 54 questions were rated on a 0-4 scale, with the mean range of responses across all couples ranging from 0.494 to 2.741, and the standard deviation ranging from 0.899 to 1.842. The large range of standard deviations suggests significant variability. For example, responses like “idk_what_is_going_on” (mean = 1.871, standard deviation = 1.796), “happy” (mean = 1.653, standard deviation = 1.615), “roles” (mean = 1.641, standard deviation = 1.641), such that specific questions exhibit high variability, suggesting that there are significant differences in how different couples answer these questions, which benefits Random Forests and Bagging in utilizing diverse features to construct decision boundaries, as well as follow-up scaling of Lasso and Ridge.

We analyzed the correlations between the variables and “Divorce_Y_N”, and the results showed that most of the variables had a correlation of nearly 1 with divorce, which is a good insight into the highly separable nature of the dataset. We extracted the five variables with the highest correlations, plotted the boxplot, and attempted to analyze the predictive power. The results showed that divorced couples scored near the top of the scale (2-4), indicating strong identification with negative traits (e.g., “idk_what_is_going_on”), while non-divorced couples scored close to 0. (Figure 1) In contrast, weaker predictors showed lower predictive value. (Figure 2)

The relationships between the predictors provide further insight. For the top five questions (idk_what_is_going_on , happy, roles, marriage, harmony), the two-by-two correlation coefficients ranged from 0.876 to 0.964 (e.g., happy and roles, r = 0.964), suggesting significant multicollinearity. This suggests that there is overlap in the measures of these problems, which is well handled by our model. lasso and Ridge ease multicollinearity by regularizing and reducing redundant features, while Random Forest and Bagging are robust to correlated predictors. This model informs our methodology, ensuring that the model takes into account interrelated marital interactions when identifying key factors.

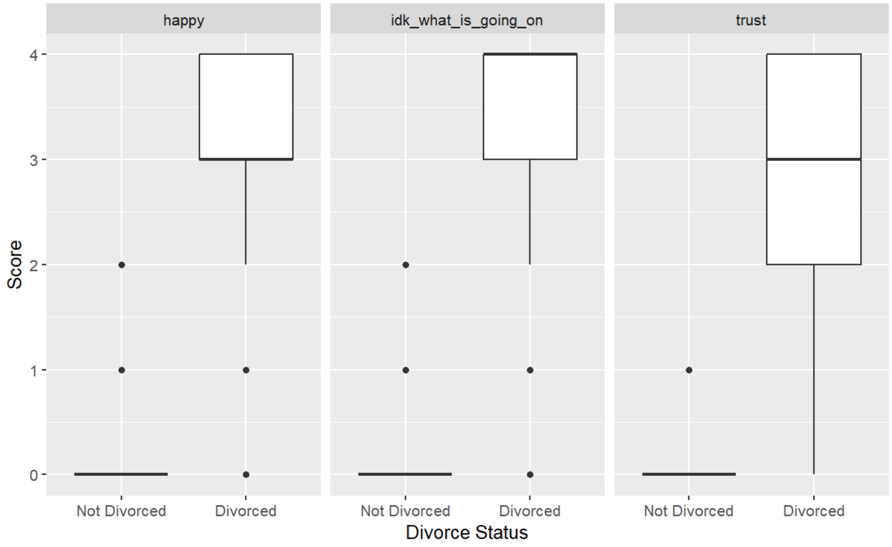

Figure 1: Boxplots of idk_what_is_going_on, happy, and roles by divorce status.

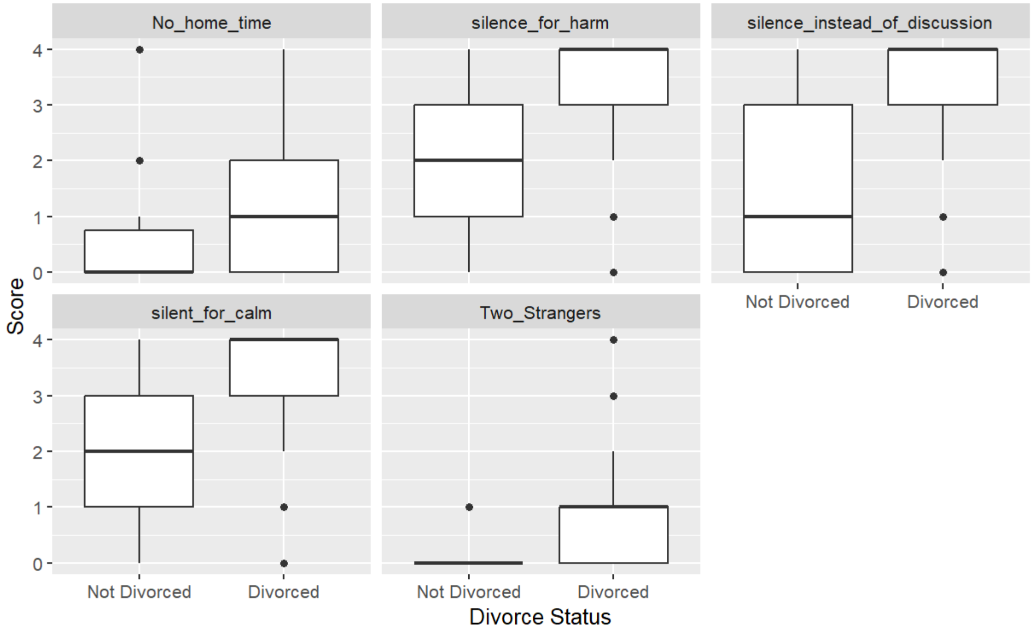

Figure 2: Boxplots of No_home_time, silence_for_harm, Two_Strangers, silent_for_calm, and silence_instead_of_discussion by divorce status.

4 Empirical Approach

In order to achieve predictive divorce, as mentioned in the previous section, based on the high separability and strong correlation of the dataset, we used four machine learning methods: Random Forest, Bagging, Lasso Regression, and Ridge Regression. In the following Empirical Approach, you will see our methodology, model specification, principles of each method, and robustness considerations.

The divorce prediction dataset was split into a training set (seendata, 136 couples) and a test set (Unseendata, 34 couples) using an 80/20 ratio. We use random seeds to ensure randomness. The training set was used to fit all models and the test set was used for performance evaluation (Section 5). All models were implemented in R, using suitable packages for classification and regularization.

Random Forests are robust in classification tasks and are able to handle possible nonlinear relationships in the dynamics of the marriage. Using the “randomForest” package, we trained a model containing 500 trees (ntree = 500) to ensure stability, with “mtry” being the number of variables tried per split, which we set to 3. The “importance = TRUE” parameter computes variable importance, which is extracted via “importance(rf_model)”. “MeanDecreaseAccuracy” in Figures 3 and 4 indicates how much the accuracy decreases if the variable is “substituted.MeanDecreaseGini” indicates how much the Gini impurity decreases when using the variable. We realize the variable importance assessment by evaluating these two graphs. In addition, we analyze the Random Forest OOB error to provide validation and reduce the risk of overfitting. Random forests directly address our second sub-question: what are the most important factors that drive divorce? by ranking predictors and identifying key marital factors.

Immediately following this, we employ Bagging as a complementary tree-based method to enhance robustness. Bagging constructs 500 trees (aligned with random forests) and uses all 54 predictors in each split, unlike random forests where mtry = 3. This method reduces the variance by averaging the samples, thus improving the stability of a single tree. Variable importance was computed via “importance(rf_model)”, and similarly, the “MeanDecreaseAccuracy” and the “MeanDecreaseGini” in Figure 4 can also be ranked for predicting factors that may lead to divorce. Comparing the two graphs we can easily see that questions such as “idk_what_is_going_on” and humiliate are consistently important and have high correlations. In addition, Higher values (e.g., 12 vs. 8) suggest Bagging may emphasize certain variables more due to full variable sampling. In other words, the inclusion of Bagging helps to address the potential overfitting in the small dataset (170 vs. couples) in a small dataset (170 couples). It contributes to our robustness test by validating the nonlinear pattern across methods.

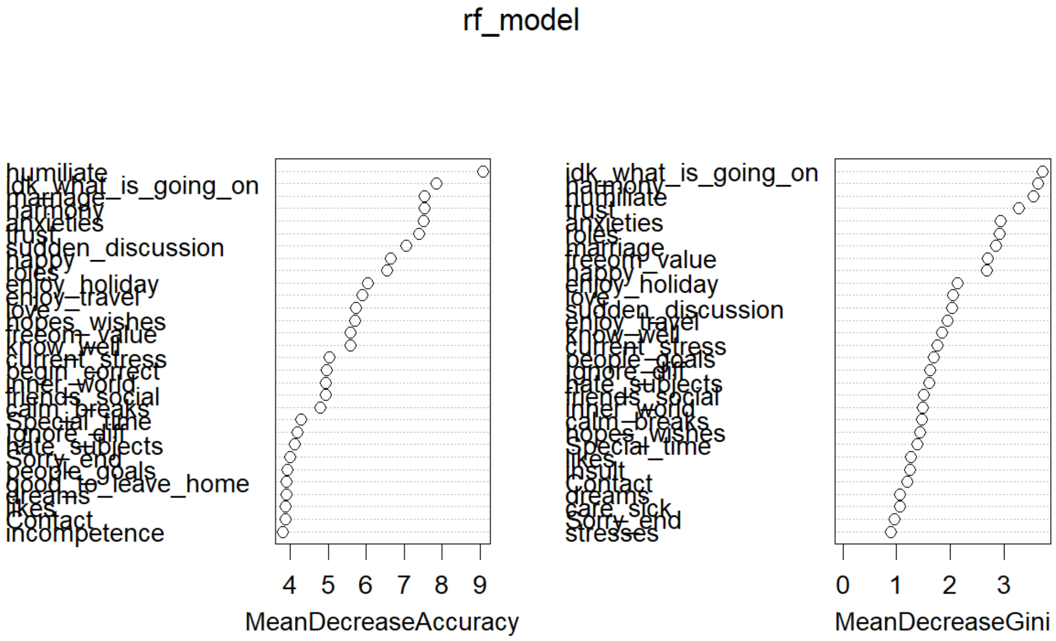

Figure 3: Variable Importance-RF Model

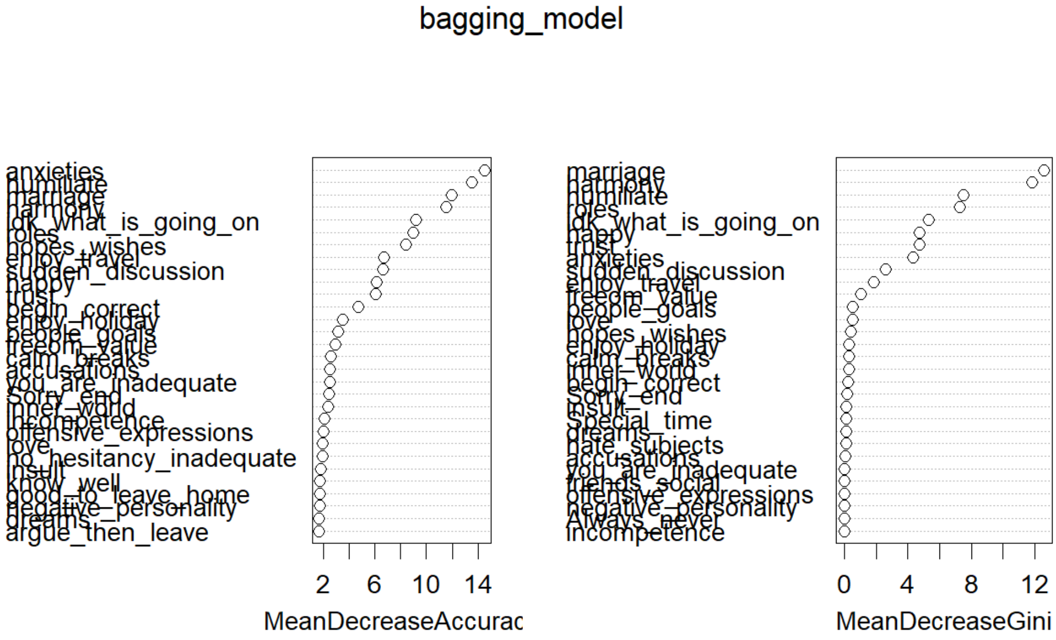

Figure 4: Variable Importance-Bagging Model

Since our dataset has multicollinearity, we also chose to use Lasso Regression. We first create the predictor and outcome matrices. “alpha = 1” specifies the use of lasso, and feature selection is performed by applying L1 regularization via “cv.glmnet”. Coefficients were extracted at the optimal lambda (lasso_model$lambda.min), as a way to keep the important divorce correlation coefficients and reduce the less important ones to zero. coefficients down to zero, thus reducing redundancy. Overall, Lasso models the odds of divorce as a linear function of 54 predictors, with the regularization parameter (lambda) selected by cross-validation to highlight influential predictors. Lasso's unique methodology helps us to identify interactions with marital values that drive divorce, such as “No_home_time”, and exclude relevant predictors such as “harmony” and “know_well”.

Ridge Regression is also implemented by “glmnet”, which applies L2 regularization, shrinking the coefficients but not setting them to zero. This method appropriately preserves all predictors while reducing their variance, thus ensuring stable estimates. Similar to Lasso, Ridge also linearly models the log probability of divorce, with lambda values chosen by “ridge_model$lambda.min”. Ridge includes all predictors, and the smallest coefficient, “I_Am_Right= -0.012”, is retained. Lasso's sparsity contrasts with the sparsity of Lasso, improving the comprehensiveness of the predictions. Together, Lasso and Ridge provide a linear perspective that complements the previous two nonlinear methods, ensuring that our robustness tests capture both linear and nonlinear effects.

Robustness Considerations : These four methods constitute the robustness test and their complementary strengths ensure the reliability of the results. Random Forest and Bagging are able to capture nonlinear patterns, while Lasso solves the multicollinearity problem by sparsity and Ridge by shrinkage. The rankings of “idk_what_is_going_on” and “anxieties” yielded high results for all four models, proving the robustness of the models and ensuring a comprehensive machine learning and analysis in line with our research goals. machine learning and analysis.

5 Results

After training on “seendata”, we evaluated model performance on a test set (Unseendata, 20%, 34 couples). Performance metrics (accuracy, area under the AUC-ROC curve, MSE) reflect predictive power, while top-ranked predictors highlight key marital values. We will discuss these findings, their implications, and potential limitations, and analyze them in the context of our research objectives.

	Random Forest	**Bagging**	**Lasso**	**Ridge**

Accuracy	100%
	95% CI : (0.8972, 1)
	AUC-ROC=1
	MSE=0.00465
		100%
		95% CI : (0.8972, 1)
		AUC-ROC=1
		MSE=0.00788
			100% accuracy
			AUC-ROC=1
			MSE=0.00097
				100% accuracy
				AUC-ROC=1
				MSE=0.00137

(Important variables ranked on highest mean decrease accuracy) 1	Humiliate	Anxieties	Idk_what_is_going_on	No_home_time
2	Anxieties	Humiliate	Anxieties	Begin_correct
3	Marriage	Marriage	No_home_time	Idk_what_is_going_on
4	Sudden_discussion	Harmony	Accusations	Anxieties
5	Trust	Idk_what_is_going_on	Begin_correct	Accusations
6	Idk_what_is_going_ on	Hopes_wishes	Marriage	Good_to_leave_home
7	Harmony	Sudden_discussion	You_are_inadequate	Sudden_discussion
8	Roles	Roles	Sudden_discussion	No_hesitancy_inadequate
9	Enjoy_travel	Happy	No_hesitancy_inadequate	Aggro_argue
10	Happy	Enjoy_travel	Hopes_wishes	Ignore_diff

The results of the random forest model had a predictive power of 100%. The AUC-ROC curves (Figures 5-6) confirm perfect predictions, with all models having an area of 1 and low MSE values (Lasso：0.00097，Ridge：0.00137，Random Forest：0.00465，Bagging：0.00788). 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. The OOB error rate is 2.94%, which is a very small. It also showed a Kappa score of 1, which means perfect agreement between prediction and reality.

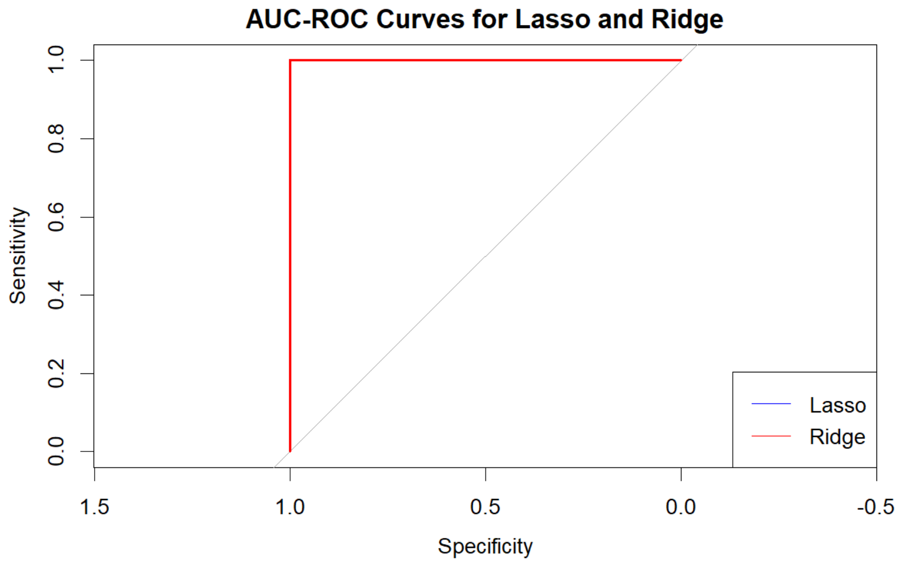

Figure 5: AUC-ROC Curves for Lasso and Ridge

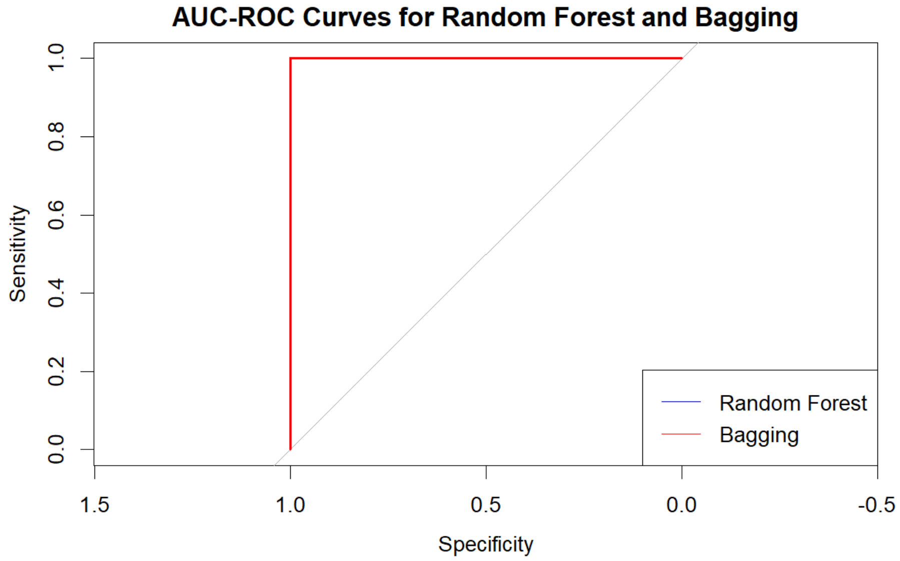

Figure 6: AUC-ROC Curves for Random Forest and Bagging

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.

Lasso and Ridge's coefficients confirm these findings. lasso selects idk_what_is_going_on (1.286), “anxieties” (0.986), and “No_home_time” (0.603) as the main predictors, while several of Ridge's largest coefficients include “No_home_time” (0.320), “idk_what_is_going_on” (0.320), and “idk_what_is_going_on” (0.320). time” (0.603) as the main predictors, while several of Ridge's largest coefficients include ‘No_home_time’ (0.320), ”idk_what_is_going_on “ (0.183), and “anxieties” (0.174). The consistency of these predictors across models highlights their role in driving divorce, reflecting the reality of communication barriers and emotional disconnect in divorced couples.

6 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. Our research can inform premarital counseling and education programs by encouraging couples to discuss key topics such as communication (idk_what_is_going_on) and emotional needs before marriage. For married couples, these insights could guide marriage counseling. If extended to larger datasets, this approach may influence public policy (Fareed et al., 2022), such as integrating predictive tools into family support systems, thereby contributing to global efforts to strengthen marriages.

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, for example, testing these models on larger, more diverse datasets that incorporate demographic variables. If these directions are pursued, the complex interactions between marital interactions and the risk of divorce could be further elucidated based on the present study. This would make for a very interesting topic.

References

Dataset:

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

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

OECD. (2022). *OECD Family database*. https://www.oecd.org/els/family/SF_3_1_Marriage_and_divorce_rates.pdf

Hilpert, P., Vowels, M. R., Mestdagh, M., & Sels, L. (2023). Emotion dynamic patterns between intimate relationship partners predict their separation two years later: A machine learning approach. *PLoS ONE*, *18*(7), e0288048. https://doi.org/10.1371/journal.pone.0288048

Fareed, M. M. S., Raza, A., Zhao, N., Tariq, A., Younas, F., Ahmed, G., Ullah, S., Jillani, S. F., Abbas, I., & Aslam, M. (2022). Predicting divorce prospect using ensemble learning: support vector machine, linear model, and neural network. *Computational Intelligence and Neuroscience*, *2022*, 1–15. https://doi.org/10.1155/2022/3687598

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 when we had errors in the code. 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:

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

Appendix

Here is the link of our GitHub Desktop, showing weekly progress and division of labor between two-person teams: https://github.com/cbcollet/DSL.git