What Makes A Couple Stay Together?

Authors: Suyeon Betty Hwang and Mariano Aloiso

Abstract: This study analyzes anonymous data from the How Couples Meet and Stay Together (HCMST) survey conducted in 2017, encompassing 3,510 participants and subsequent follow-up in 2020. Focused on identifying factors influencing relationship longevity, the dataset underwent rigorous cleaning, reducing 725 variables to 56 and consolidating to 2,107 observations. Employing Lasso regression, random forest, and Principal Component Analysis (PCA), the study revealed compelling insights: couples who met online were more prone to breakups, while those who initially met at school were more likely to stay together. Key determinants impacting relationship longevity included subjects' education years, attendance at religious services, and partner age and household income, with some variables showing positive correlations while others exhibited negative associations with relationship duration.

Introduction

Having solid and happy relationships is often considered key to a good life, but the specifics of what makes them successful can be tricky to pin down. Conventional studies have frequently leaned on psychological and qualitative measures, yet quantitative aspects often remain understudied. By diving into anonymous survey data, we seek to figure out what specific qualities in couples keep their relationships going strong. This report aims to delve into the factors that contribute to relationship longevity and success by utilizing objective data analysis.
Our research addresses the following questions:

1. Does the way couples meet have any impact on whether the relationship will last?
2. What makes a couple stay together?

We believe that the application of statistical methods to survey data will provide a scientific approach toward explaining dynamics that contribute to enduring and fulfilling relationships.

Overview of the study “How Couples Meet and Stay Together”

The How Couples Meet and Stay Together (HCMST) survey conducted in 2017 involved 3,510 participants, and followed up with subsets of these participants in 2020 (2,107) and 2022 (1,722). The subsequent 2020 and 2022 surveys continued the traditional HCMST relationship questions while incorporating new queries related to the impact of the COVID-19 pandemic. This research initiative was led by Michael J. Rosenfeld, Reuben J. Thomas, and Sonia Hausen and was funded by the United Parcel Service Endowment at Stanford University and the US National Science Foundation.

HCMST 2017 asked a comprehensive set of questions to participants currently in relationships (N=2862) and those who previously had partners (N=541), as denoted by the "w1_partnership_status" variable. Conducted by the online survey company Ipsos, the surveys aimed to be nationally representative by recruiting participants through the Ipsos KnowledgePanel, utilizing Address Based Sampling. The KnowledgePanel is a nationwide panel of subjects that agree to participate in surveys by Ipsos. Even subjects without home internet access were provided with the means to participate. Certain text responses and geographic specifics below regional levels were kept confidential in the data made available to the public.

Data Cleaning and Feature Engineering

The analysis was focused mainly on the data from the first survey in 2017. The data cleaning process began by eliminating the background variables from the IPSOS knowledge panel, as these had already been encoded in other features in the first survey, and also the results from the 2022 wave. Only the relationship status and relationship quality (w2_section and W2_rel_qual_reduced) were kept from the second wave, and any observations that were discontinued in the 2020 survey from 2017 were removed.

Even after the previous process, reducing the number of features was challenging. We employed correlation analysis to identify feature pairs with a perfect correlation, an indicator that they encoded the same data. This step highlighted the following variables containing the same data:

- Education: keep “w1_ppeducat” and remove “w1_ppeduc”
- Household income: keep “w1_ppincimp_cat” and remove “w2_log_real_inc” and “w1_ppincimp”
- Partner’s age: keep “w1_ppagecat” and remove “w1_ppagect4” and “w1_ppage”
- Partnership status: keep “w1_partnership_status” and remove “w1_section”, “w1_partnership_status_cohab”
- and “xpartner_type_cohab”
- Relationship quality: keep “w1_q34_reduced” and remove “w1_q34”
- Race: keep w1_ppRace_* and remove w1_interracial_5cat
- Sexual identity: keep “w1_identity_all” (more information) and remove “w1_identity_all_modified”
- State of residence: keep “w1_PPREG4” and remove “w1_ppreg9”

We kept only one of each pair. Correlation analysis also revealed that there is a repeated variable for race, encoded as w1_pprace* and w1_ppRace_*. The features with an underscore before the number were self-reported by the participant, while the features without an underscore came from IPSOS KnowledgePanel. We removed the IPSOS KnowledgePanel features since it has a higher cardinality (15 classes vs 6). We performed correlation analysis a second time, but this time we look at high correlation values instead of perfectly correlated features. We decided to remove an additional 4 variables.

Another method applied for reducing dimensionality was looking at the top features by the number of unique values. This helped us identify features that contained specific dates such as day, month, and years that were not relevant to the analysis. Some of these variables were:

- “W1_year_fraction_met”
- “W1_met_month”
- “W1_met_after_2009”
- “W1_relate_duration_in2017_years”
- “W1_year_met_from_1990”
- “W1_relationship_start_calmonth”

For missing data, we automatically dropped features that had more than 90% of missing values, and we manually reviewed features that had more than 10% of missing values. There was a small set of few subjects that refused to answer certain questions (never more than 17 per column). We dropped features that were not meaningful, such as if the subject had internet access, replaced NaNs in binary variables with a 0 and used median value in other features.

At the end of the data cleaning process, the number of variables was reduced to 56 from 725 variables, with a total of 2107 observations.

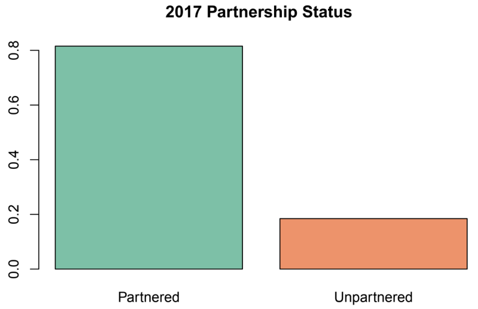
In the preliminary stages of data cleaning, we discovered that question 24 (how the couple met) was a free text field and had over 40 possible answers. We separated the responses for this question and analyzed them independently in the “How do couples meet?” section of this report. With this dataset, we extracted the subset of participants who were in relationship as of 2017 (1716 observations) and furthermore we obtained a subset of the partnered participants who remained with their partner as of 2020 (1488 observations).

The possible answers were manually grouped into 6 categories:

- Friends: subject’s friend, partner’s friend, subject’s neighbor, partner’s neighbor, subject’s significant other, etc.
- Family: subject’s family, partner’s family
- Online: social media, dating websites, internet games, phone apps, chat services, etc.
- Social events: church, voluntary organization, blind date, bar/restaurant, vacation, etc.
- Work: subject’s coworker, partner’s coworker, business trip, etc.
- School: high school, college.

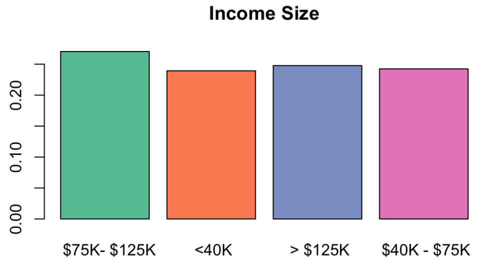
Exploratory Data Analysis

To answer our first question – the most prominent features of successful relationships– we decided to take a subset of participants who had partners as of 2017 and continued to stay with the same partner as of 2020. Furthermore, we took the respondent’s partnership status, which was originally made of four different categories–married, not married but with a partner, no partner but had a partner, and never had a partner– into a simple binary response: 1 as “partnered” and 0 as “not partnered”. 80% of the 2020 respondents were partnered and the rest of the analysis was done on these 80% of the 2020 dataset: a subset of people that are in a relationship as of 2017.

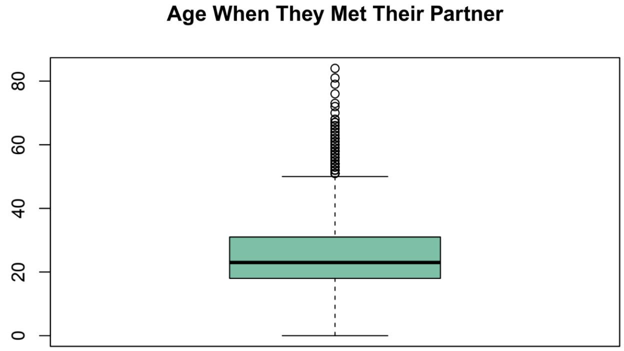


Due to the nature of most of the predictor variables being categorical data, we did barplots for most of the predictor variables. Through exploration, there were a few interesting observations. First, most respondents who are partnered in 2017 have a bachelor’s degree or higher and most unpartnered people only had up to high school degree.

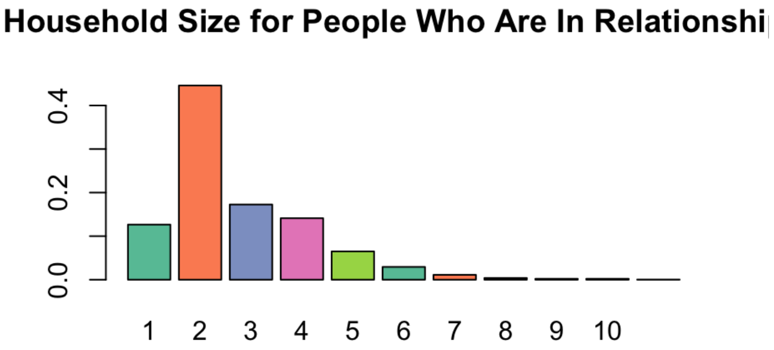
Income size for partnered people was evenly distributed among four categories whereas more than 50% of the unpartnered people had income less than $40,000.



Additionally, half of the respondents have met their now-partners in their twenties.

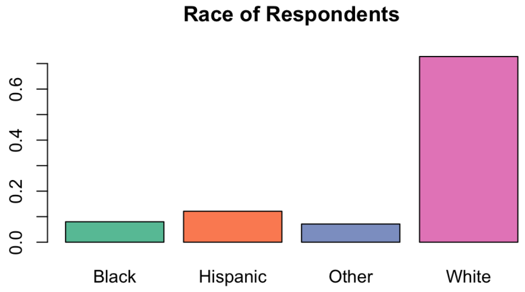


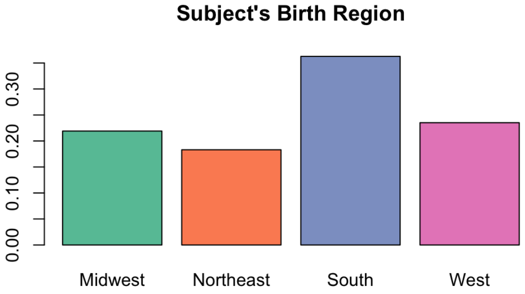
The household size for partnered people was skewed to the right where most people live with their partner or have a child or two.



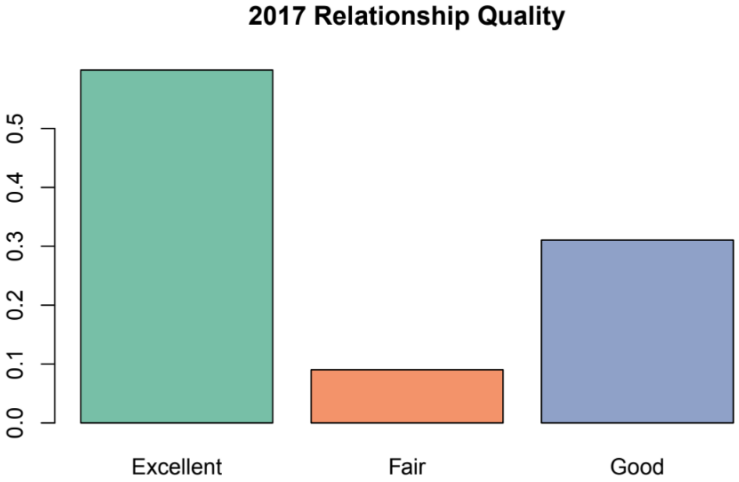
Next, for partnered participants, their partners were predominantly full-time employed or retired.

Almost 80% of the respondents were white while the actual national white proportion is about 55%. 36% of the participants were born from the southern part of America and the participants' birth regions are not evenly distributed.



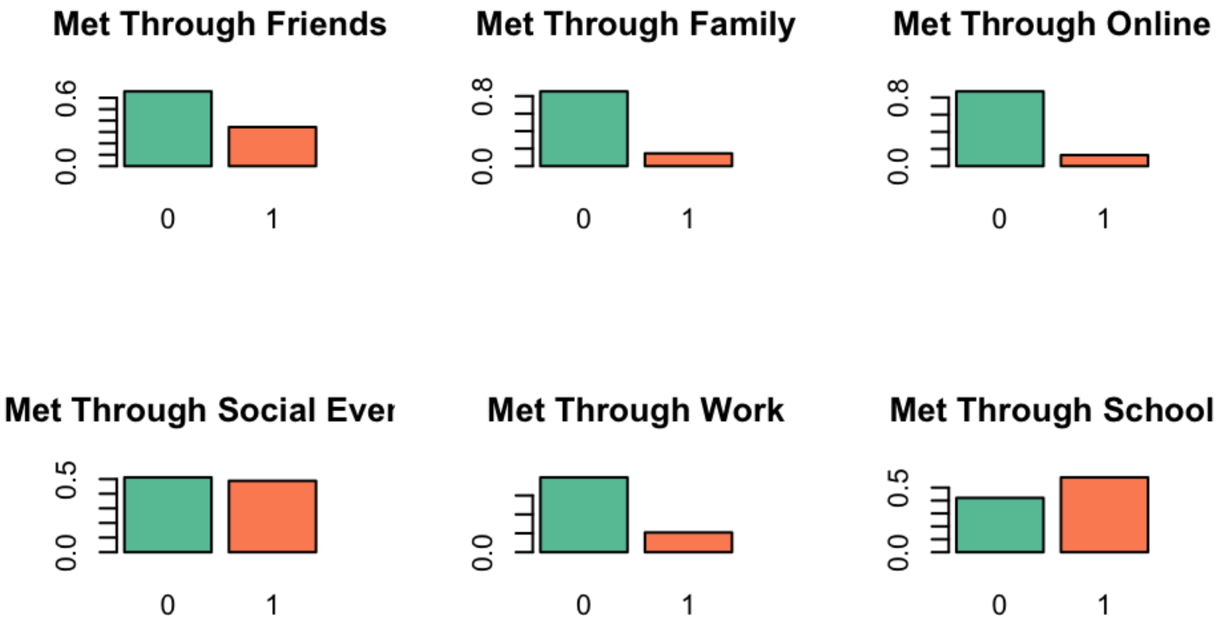


Lastly, most of the participants reported that their relationship quality was either excellent or good as of 2017. Since we are measuring whether couples will stay together and our dataset is potentially biased with people who are in a good relationship, this has to be kept in mind throughout the analysis.



How do couples meet?

Based on the How Couples Met Data, which is based on Question 24 and participants who were in relationships as of 2017, social events and school had the highest proportions and online was the platform that had the least proportions of how couples met.



To find out if there is a relationship between the way the subject met their partner and whether they are with the same partner in wave 2, logistic regression was used. Based on the training data, which was 80% of the How Couples Met Data, we weigh the dataset to address imbalance and results are shown in the below table:

Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	1.80341	0.18441	9.779	< 2e-16
friends	-0.2091	0.18832	-1.11	0.267
family	0.41869	0.26206	1.598	0.11
online	-0.93929	0.22594	-4.157	3.22e-05 **
social_events	0.2762	0.16918	1.633	0.103
work	-0.06624	0.22388	-0.296	0.767
school	0.22479	0.16413	1.37	0.171

Null deviance	504.61 on 1371 degrees of freedom
Residual deviance	490.50 on 1365 degrees of freedom
AIC	259.68
Deviance Test Pr(>Chi):	0.02843125

The logistic regression with null deviance of 504.61 and residual deviance 490.50 indicates that the model is slightly better with the predictors. Based on the logistic regression summary, social events and work are statistically significant predictors, which means couples who met at social events and work have a higher probability of staying together as a couple. School seems to have the highest p-value, which means couples who met at school are least likely to stay together in a relationship. To further diagnose the significance of the improvement in model fit, Chi-Square statistic was used if the improvement of the model with predictors is indeed statistically significant.

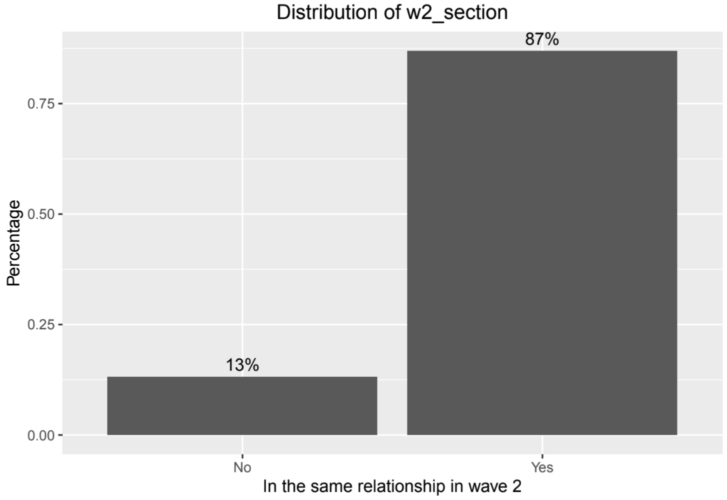
The table below measures the quality of the model. Based on the confusion matrix, the model achieves an accuracy of 59.99% with a sensitivity of 60.87%, a specificity of 59.73%, and a balanced accuracy of 60.30%. This tells us that the current model makes a slightly better prediction than random guessing. The balanced accuracy being close to the overall accuracy suggests that the model is relatively consistent in predicting both classes.

	Reference	
Prediction	No	Yes
No	28	120
Yes	18	178

Confusion matrix of How They Met model

Dimensionality Reduction

After feature engineering, the clean dataset had a total of 84 categorical and numerical features. There were also columns with missing values and high class imbalance, with 87% of the data belonging to couples that stayed together.

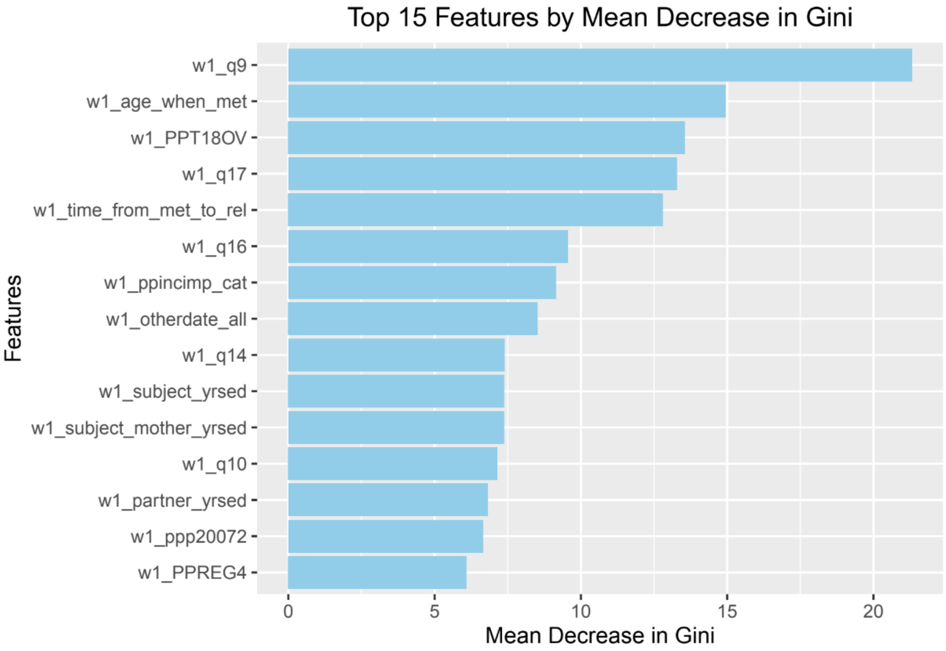


Random forest was used to find the features that contained the most variance before further analysis, as this model can handle highly correlated variables, missing values and skewness in the distribution of the classes.

Weights were created to address the disparity in class distributions, by ensuring more equitable representation of minority classes (relationship ended) during training. The weights were calculated by dividing the total number of samples by the product of the number of unique classes and the respective class counts. The resulting weights assign greater importance to underrepresented classes, which allows the random forest model to mitigate the effects of class imbalance and enhance its predictive power across all classes.

The Random forest model had a sensitivity of 0.07, specificity of 0.99, Cohen's kappa of 0.2024 and balanced accuracy of 0.532. These metrics are significantly better than the performance of the majority class model (a baseline model that always predicts the majority class).

Mean decrease accuracy is a metric that represents how much the accuracy of the model decreases when a particular feature is not available. Higher values indicate more important features. Based on the plot below, the top 15 important features are w1_q9, (partner's age), w1_age_when_met (age when they met), w1_PPT18OV (number of household members), w1_q17 (times the participants married), w1_time_from_met_to_rel (years it took them to start dating from the time they met), w1_q16 (how many relatives they see each month), w1_ppincimp (income category), w1_otherdate_all (whether they met someone for dating in the past year), w1_q14 (subject's mother's educational attainment), w1_subject_yrsed (highest educational degree received), w1_subject_mother_yrsed (subject's mother's years of education), w1_q10 (partner's educational attainment), w1_partner_yrsed (partner's educational attainment), w1_ppp20072 (how often subject attends religious services), w1_ppreg4 (subject's state of residence). These were the features that had the highest impact on couples staying together.



Mean Decrease in Gini: Gini importance measures the total decrease in node impurity that a feature causes. A higher Gini importance suggests a more influential feature in terms of making splits and decisions within the trees.

Why do couples stay together?

Even after reducing dimensionality with random forest, there is a high risk of dealing with multicollinearity in the data. To answer the question of what makes couples stay together, we selected two models that are robust to multicollinearity: Lasso regression and principal component regression. New train and test sets were resampled for these models.

Baseline Model

A baseline model gives us a context to interpret the results and capabilities of more complex models. In our case, we use as a baseline the model that always predicts the majority class.

	Reference	
Prediction	No	Yes
No	0	0
Yes	41	276

Confusion matrix of baseline model

The baseline model achieves an accuracy of 87.07%. However, it has a sensitivity of 0, a balanced accuracy of 50% and a Cohen's Kappa of 0. The goal of the Lasso model is to improve these metrics.

Lasso Regression

As Lasso may also be sensitive to class imbalance, weights were reapplied to this model. The following is the performance of the model.

	Reference	
Prediction	No	Yes
No	13	10
Yes	28	266

Confusion matrix of Lasso regression

The Lasso regression model manages to improve all the metrics of the baseline model. It achieves an accuracy of 87%, but most importantly, it yields a balanced accuracy of 64.04%, along with a Cohen's Kappa of 0.3454.

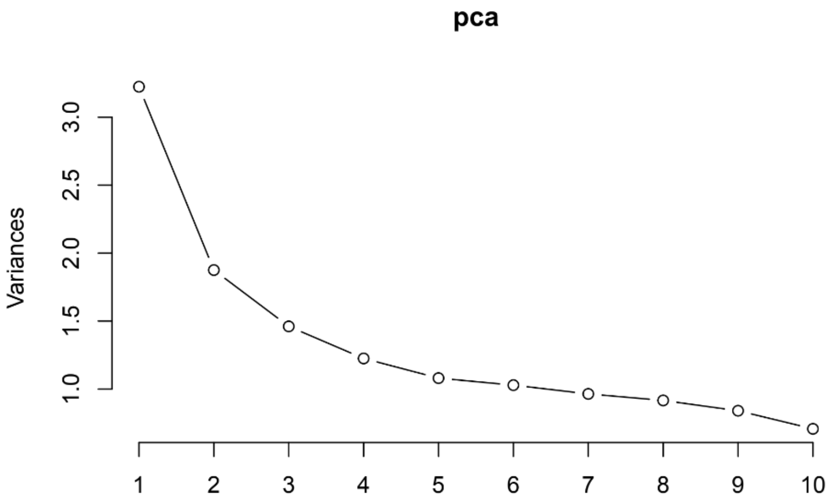
This model shrank the coefficients for w1_ppp20072, w1_PPT18OV, w1_q10, w1_subject_mother_yrsed, and w1_partner_yrsed. The remaining coefficients are as follows:

Variable	Coefficient
w1_q17	0.312
w1_ppincimp_cat	0.170
w1_subject_mother_yrsed	0.041
w1_PPREG4	0.018
w1_q9	0.011
w1_q14	-0.005
w1_q16	-0.007
w1_age_when_met	-0.023
w1_time_from_met_to_rel	-0.026
w1_otherdate_all	-0.676

The variables that are positively correlated to couples staying together are Q17 (times the participants have been married) and the income category (higher income is correlated to a higher likelihood of couples staying together). On the opposite side, subjects that have dated different people in the year prior to the first survey, and those with a longer timespan between meeting and a relationship, are more likely to have broken up.

Principal Component Regression

Principal component regression is sensitive to data in different scales. It is essential to scale the data to ensure that all variables contribute equally to the analysis and have a similar impact on the resulting principal components. In this process, the HCMST dataset was scaled by standardizing the variables (except for the outcome variable) by centering them around their means and scaling them to unit variance.



The variance plot shows that the first two principal components explain almost half of the variance in the outcome variable.

w1_ppp20072	w1_ppincimp_cat	w1_PPREG4
-0.0234755	-0.2881937	0.00251327
w1_PPT18OV	w1_q9	w1_q10
0.02812115	0.12645104	-0.4492715
w1_q14	w1_q16	w1_q17
-0.404907	0.06559015	0.08830098
w1_otherdate_all	w1_age_when_met	w1_time_from_met_to_rel
-0.0268477	0.00429276	0.01108032
w1_subject_mother_yrsed	w1_partner_yrsed	w1_subject_yrsed
-0.4012982	-0.4469879	-0.4000152

First principal component

Based on the first principal component analysis, partner's, subject's mother's, subjects' years of education, and partner's mother's years of education (all negative loadings) had the highest influences on the component. Based on the second principal component analysis, partner's age as of 2017 and when the subject (both positive loadings) has married had the highest influence on the component.

Questions Asked During Presentation

1. Why did you choose those models?

The choice of models in this analysis was done based on the characteristics of the dataset and the research objectives. Initially, the dataset contained 84 categorical and numerical features post feature engineering, alongside columns with missing values and significant class imbalance, with approximately 87% of the data representing couples that remained together. Random forest was used first to reduce the dimensionality of the data down to 15 variables as it effectively handles variables with high variance, correlated attributes, missing values, and skewed class distributions.

Despite using random forest to reduce dimensionality, there was still a risk of high multicollinearity in the dataset. To address this, we used Lasso regression and principal component regression, two models robust to multicollinearity, to investigate the factors influencing relationship longevity. A baseline model was created to provide a better context for interpreting the outcomes and capabilities of more complex models. This model utilized the majority class for prediction.

2. Can you talk about the accuracy of your models?

We analyzed different metrics to evaluate the performance of our models. The baseline model set a benchmark with an accuracy of 87.07%, but with several limitations. Notably a sensitivity of 0, a balanced accuracy of 50%, and a Cohen's Kappa of 0, indicating inadequate predictive ability and a lack of agreement beyond chance. The random forest model is able to increase specificity to 99%, but it suffers from a low sensitivity of 7%, leading to imbalanced predictions. The overall performance remains modest with a Cohen's Kappa of 0.2024 and a balanced accuracy of 53.2%.

After feature selection using random forest, both Lasso regression and Principal Component Regression show notable improvements. Lasso regression notably boosts the model performance, achieving an accuracy of 87%, a balanced accuracy of 64.04%, and a Cohen's Kappa of 0.3454. Regarding Principal Component Regression, the variance decreases drastically with the first 2 components (reaching 1.9), and it slowly decreases with the following components. This suggests that these two principal components captured the majority of the dataset's variance.

Conclusion

With the How Couples Met and Stay Together dataset, we aimed to investigate how couples met and what contributes to couples staying together. Data cleaning was the most challenging part of the process due to the large number of features and multicollinearity between variables. Once we reduced the dimensionality of the data by removing redundant features and selecting relevant variables to answer the question, we used a Lasso Regression and the Principal Component Analysis: to what extent does how couples meet influence their longevity of relationships? With Lasso Regression, the subject’s years of education, attending religious services, whether the partner is in marriage had an impact on the longevity of relationships. Through the Principal Component Analysis, we found out that the partner's age and household income were positively correlated with the target variable. Partner’s dating history, attending religious services, and the subject's mother’s educational level were negatively correlated with the target variable. Using the How Couples Met data, which was an extraction based on one question from the original dataset, we observed that couples that met online are more likely to break up while couples that met at school are more likely to stay together.

Shortcomings and Future Work

One notable limitation in this report lies in the nature of the dataset used. The dataset's anonymized nature might restrict the depth of the analysis. The absence of contextual details might limit the understanding of factors affecting relationship longevity. Future research could benefit from supplementary qualitative data that allows for a more nuanced exploration of relationship outcomes.

There also exist potential areas for improving the robustness of the models. The original dataset is biased towards married couples as 56% of the participants are married and most there waa high class imbalance of the observations where 87% of the participants stayed partnered. More experimentation can be conducted on different techniques for dealing with high class imbalance within the dataset, where a substantial majority of couples remained together. Employing advanced techniques to handle imbalanced datasets might refine the accuracy of predictions.

While this study revealed correlations between certain variables and relationship outcomes, work can be done on creating models that attempt to predict whether a couple will remain together over the next *n* years, or predicting relationship quality and compatibility based on the attributes of the subject and potential partner.

Sources

Rosenfeld, Michael J., Reuben J. Thomas, and Sonia Hausen. 2023. How Couples Meet and Stay Together 2017-2020-2022 combined dataset. [Computer files]. Stanford, CA: Stanford University Libraries.

Faraway, Julian J. Extending the Linear Model with R. Second Edition. Chapman & Hall/CRC, 2016.