LAU
الجامعة اللبنانية الأميركية
Lebanese American University

Divorce Prediction Analysis Report

By Haidar Saleh

December 13, 2025

This report is submitted to Dr. Leila Issa in fulfillment of the final project for DSC 604: Statistics for Data Science.

Abstract

This study investigates the predictors of divorce using a dataset of relational, psychological, behavioral, and demographic variables from 5,000 couples. After preprocessing and one-hot encoding of categorical variables, Principal Component Analysis (PCA) was used to reduce dimensionality and identify latent patterns. Logistic regression was then applied to the PCA-transformed dataset to determine which variables most strongly influence divorce likelihood. The results revealed that financial stress, infidelity, trust, communication, and social support were the strongest predictors of divorce. The model achieved approximately 60% accuracy, with strong performance in correctly identifying non-divorced couples but limited recall for divorced couples. These findings highlight the importance of psychological and relational factors in marital stability and demonstrate the value of combining PCA with logistic modeling in social science prediction tasks.

1 Introduction

Divorce is a multifaceted social and psychological process influenced by various relational, emotional, financial, and demographic factors. Identifying the factors that most strongly predict divorce is essential for designing preventive interventions, supporting couples in therapy, and informing public policy. Thus, the research question of this study is: *Which factors, such as communication, financial stress, trust, and infidelity, most strongly predict the likelihood of divorce?*

Traditional statistical methods provide limited insight when datasets include numerous correlated variables. Machine learning techniques such as Principal Component Analysis (PCA) and logistic regression offer powerful tools for uncovering hidden structures within marital datasets and for modeling divorce outcomes with improved interpretability. This study applies these methods to evaluate which relational and demographic characteristics contribute most to the likelihood of divorce.

The primary objective is to determine the strongest predictors of divorce by analyzing PCA components and logistic regression coefficients. A secondary objective is to validate these findings through formal hypothesis testing.

2 Literature Review and Contribution to Existing Research

The existing literature on divorce often focuses on psychosocial determinants, demographic characteristics, and relational satisfaction. The cited study, "Intention to Divorce and Its Determinants in Young Newly Marrieds: A Field-Based Cross-sectional Study from Iran," examines factors associated with intention to divorce among newly married couples. Their findings highlight the role of perceived stress, marital satisfaction, and conflict in shaping early thoughts about divorce. While valuable, this prior work differs substantially from the present study in two important ways: research objective and methodology.

First, the Iranian study investigates intention to divorce, not actual divorce outcomes. Intention is influenced by attitudes and perceptions and does not always translate into behavior. In contrast, the present research examines real divorce outcomes, providing a more concrete understanding of which measurable factors actually contribute to marital dissolution. This shift from intention to outcome represents a significant advancement in practical relevance, as intention-based surveys cannot reliably predict which marriages ultimately end in divorce.

Second, the literature relies on traditional statistical techniques—primarily descriptive analysis and regression—to identify correlates of divorce intention. These approaches, while informative, do not incorporate predictive modeling or latent structure identification. The present study introduces Principal Component Analysis (PCA) to uncover underlying dimensions within relational and demographic variables and applies logistic regression to evaluate the predictive significance of these components. This integration of machine learning and inferential statistics provides a more robust analytical framework than the correlational models used in earlier work.

Overall, the present research contributes added value by:

- Moving from intention to actual outcome prediction, improving practical relevance.
- Applying PCA and logistic regression, extending beyond traditional correlational methods.

Together, these contributions position the current study as an advancement over existing

literature, offering deeper analytical insight and more actionable findings regarding the predictors of divorce.

3 Methods

3.1 Dataset and Preprocessing

The dataset consisted of 5,000 couples and included a wide range of variables capturing demographic characteristics (age at marriage, marriage duration, number of children, income, education, employment status, religion, and cultural background), psychological and relational factors (communication score, trust score, conflict frequency, conflict resolution style, mental health issues, financial stress, social support, and shared hobbies), and behavioral indicators (infidelity, domestic violence history, pre-marital cohabitation, and counseling attendance). The outcome variable was divorce status (0 = not divorced, 1 = divorced). The dataset has 2 limitations. The dataset is not collected from a specific country (or society). Also, the number of divorced couples (1991) is less than the number of undivorced couples (3009). Preprocessing steps included:

- One-hot encoding of categorical variables such as education level, marriage type, conflict resolution style, religious compatibility, and employment status.
- Standardization (scaling) of numerical variables using `StandardScaler` to ensure equal weighting before PCA.
- Splitting features and the target variable (`divorced`) for modeling.

These steps ensured that all variables were numeric and on comparable scales for PCA and logistic regression.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure of the dataset, assess data quality, and identify initial patterns relevant to divorce prediction. The dataset contained 5,000 couples and included demographic variables (age at marriage, marriage duration, number of children, income, education, employment status), relational factors (communication, trust, conflict frequency, conflict resolution style), psychological

measures (mental health issues, financial stress), and behavioral indicators (infidelity, domestic violence history, counseling attendance). The outcome variable was divorce status (0 = no, 1 = yes).

Missing Values and Data Quality

Inspection of the dataset revealed no missing values in any variable, allowing for complete-case analysis without imputation. All categorical variables were properly encoded, and numerical variables fell within expected ranges (e.g., `communication_score` and `trust_score` ranged from 1 to 10). No duplicated observations were found.

Distribution of Numerical Variables

Numerical variables such as `communication_score`, `trust_score`, `social_support`, and `financial_stress_level` exhibited approximately uniform or moderately skewed distributions. `combined_income` displayed a much larger scale compared to other features, confirming the need for normalization before PCA or logistic regression. Scaling was applied later to ensure comparability across predictors.

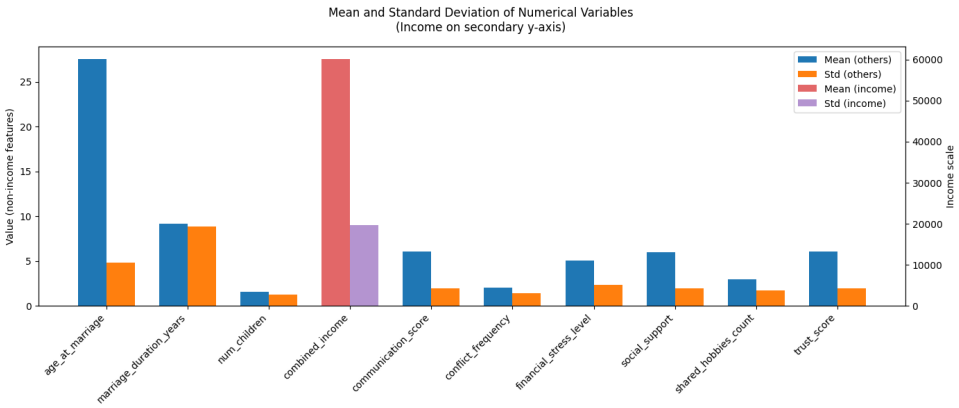

Figure 1: Mean and standard deviation of numerical variables.

Distribution of Categorical Variables

Categorical features such as `marriage_type`, `conflict_resolution_style`, and `religious_compatibil` exhibited balanced distribution across categories. However, `infidelity_occurred` and `domestic_violence_history` were relatively infrequent events, though their distributions still allowed for meaningful analysis.

One-hot encoding expanded these categorical variables into multiple binary indicators for use in PCA and logistic regression.

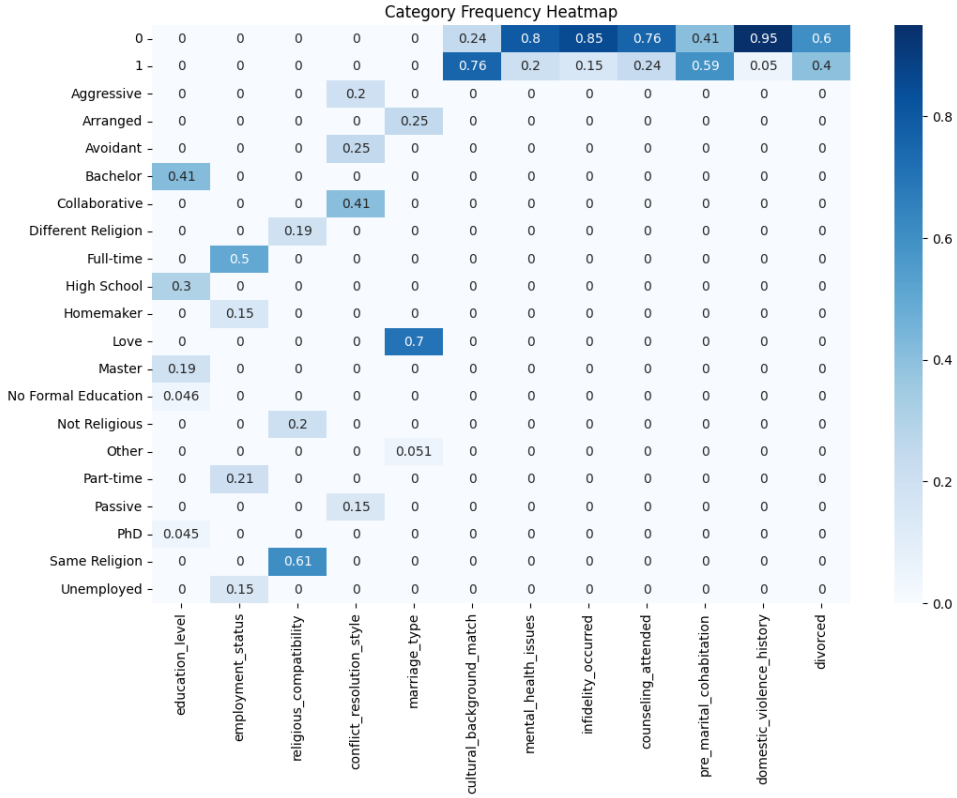

Figure 2: Frequency of categorical variables.

Outlier Detection

Outlier assessment was performed on numerical variables. Most features showed minimal extreme values, except `combined_income` and `social_support`. While these outliers were not removed, scaling mitigated their impact on PCA and prevented them from disproportionately influencing model coefficients.

Relationship Between Predictors and Divorce

Initial visualizations (boxplots, histograms, and bar charts) revealed strong separation between divorced and non-divorced couples on relational and emotional variables. Divorced couples had:

- Lower communication scores.
- Lower trust scores.

- Higher financial stress.
- Higher conflict frequency.
- Higher probability of infidelity and domestic violence.
- Lower social support.

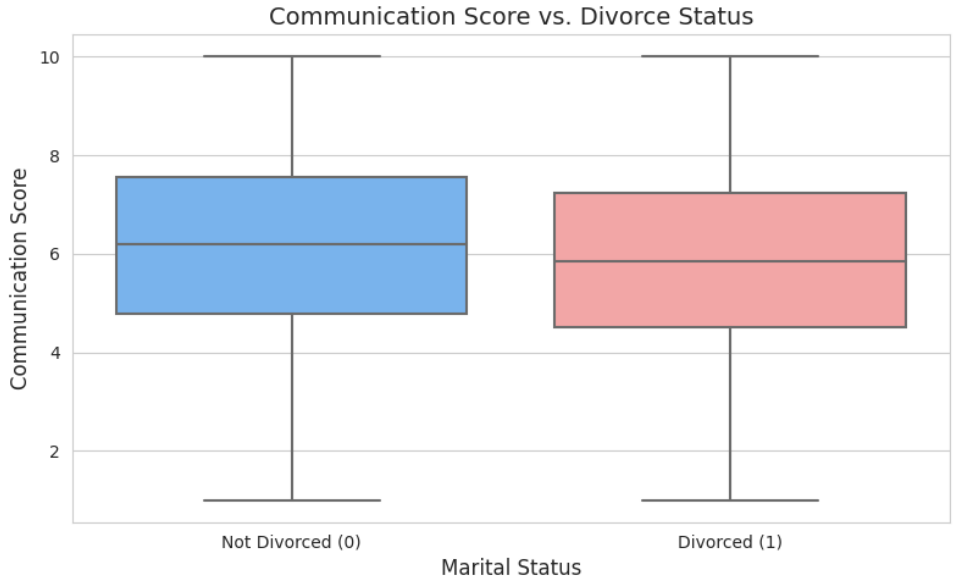

Figure 3: Communication Score Box Plot.

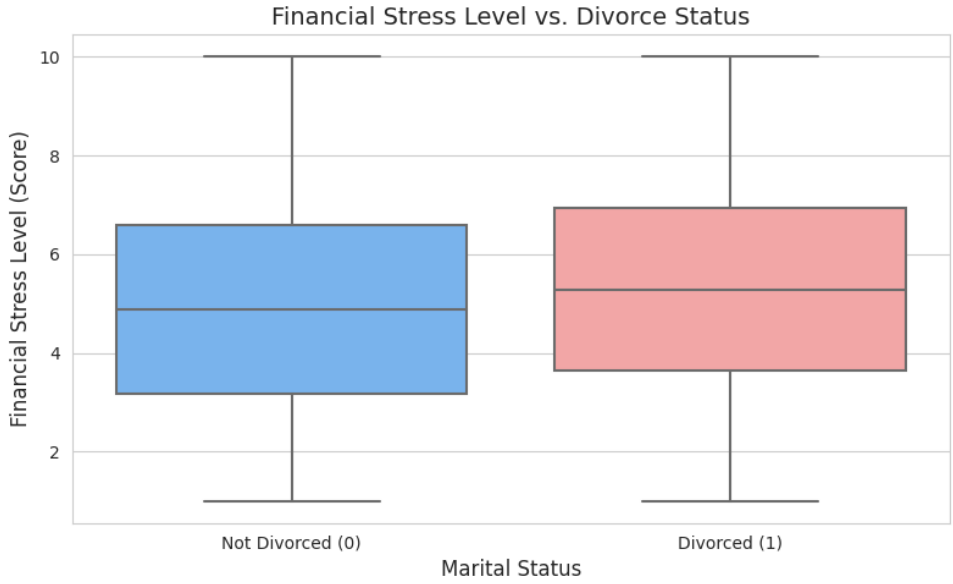

Figure 4: Financial Stress Box Plot.

These early patterns aligned with psychological theories of marital breakdown and provided preliminary evidence supporting later hypothesis testing.

Justification for Further Modeling

Overall, the exploratory analysis provided a clear initial understanding of the dataset, confirmed the need for preprocessing steps, and highlighted relational and behavioral variables as likely key predictors of divorce. These insights guided the modeling and hypothesis testing stages that followed.

3.3 Principal Component Analysis (PCA)

Because many of the predictors in the dataset were correlated—for example, communication, trust, conflict frequency, and financial stress—Principal Component Analysis (PCA) was applied to reduce dimensionality and uncover latent structures in the data. PCA transforms the original variables into a new set of uncorrelated components (PCs) that capture the maximum possible variance in decreasing order. Before applying PCA, all numerical features and encoded categorical variables were scaled to ensure that variables measured on different scales contributed equally to the analysis.

A total of 23 principal components were retained, collectively explaining approximately 86% of the total variance in the dataset.

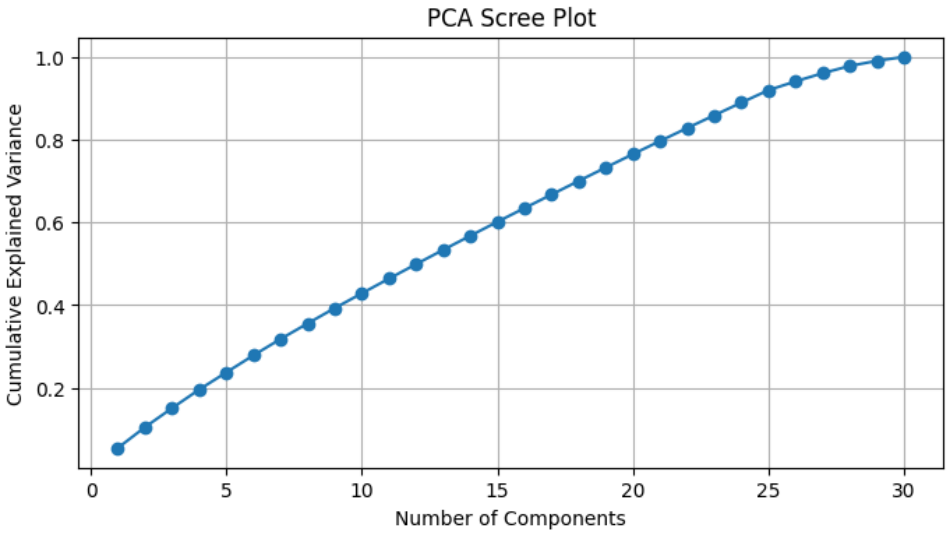

Figure 5: PCA scree plot.

Key PCA findings included:

- The first 23 principal components captured approximately 86% of total variance.
- PC1 represented religious compatibility.

- PC2 captured conflict resolution style (collaborative vs. avoidant tendencies).
- PC3 represented marriage type and education.
- The most predictive components, however, were PC21, PC9, and PC10, which were driven by variables such as financial stress, infidelity, trust, communication, and social support.

These PCA components were then used as inputs to logistic regression.

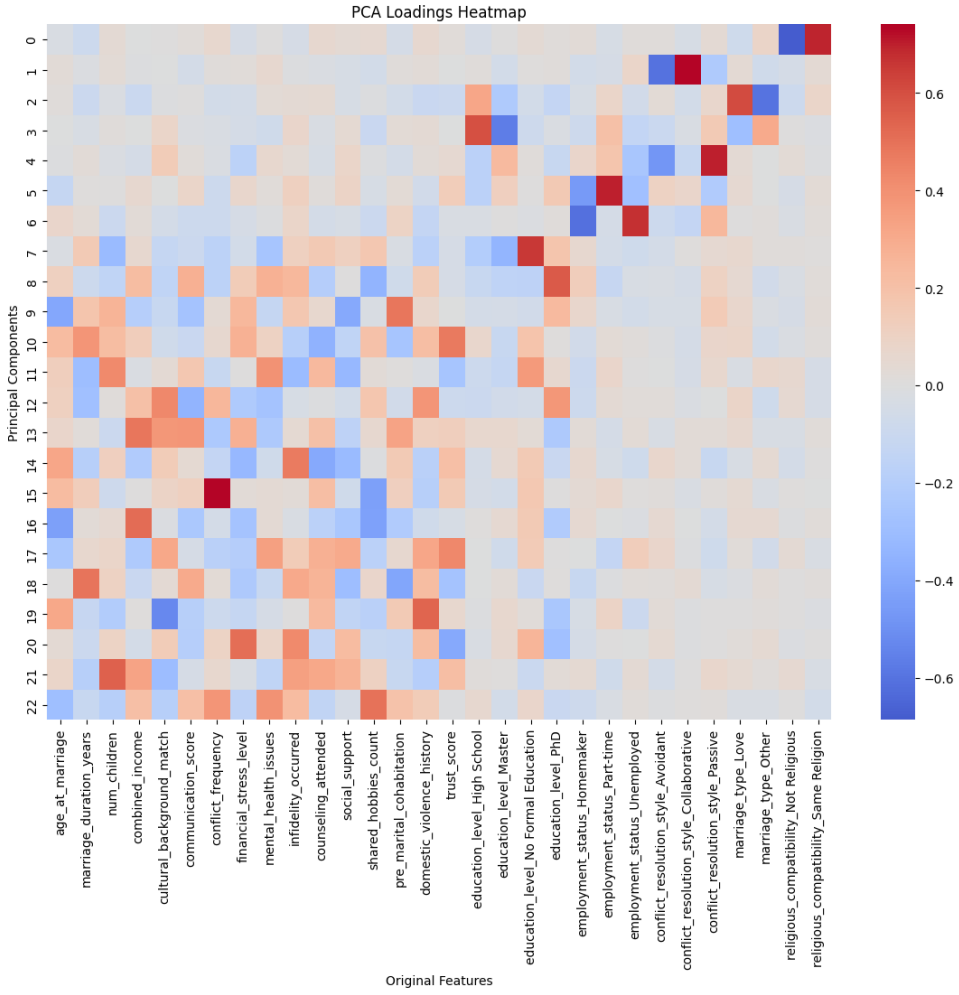

Figure 6: PCA loadings heatmap visualizing the contribution of each original feature to each principal component.

3.4 Logistic Regression Modeling

After dimensionality reduction, logistic regression was used as the primary predictive model for divorce; it was trained using the 23 components. Logistic regression estimates the probability of divorce by modeling the relationship between the transformed PCA components and the binary outcome. The model provides coefficients and odds ratios that indicate the direction and strength of the relationship between each component and the likelihood of divorce.

The model achieved an accuracy of approximately 60%, performing well at identifying non-divorced couples but showing weak recall for predicting divorced cases due to class imbalance (593 non-divorced vs. 407 divorced). Despite this limitation, the model coefficients revealed important insights. PC21 had the highest absolute coefficient, confirming that financial stress and violations of trust (e.g., infidelity) significantly elevate divorce risk.

The regression results, therefore, complement the PCA findings, demonstrating that relational quality factors—especially trust and financial strain—carry the greatest predictive weight in determining divorce likelihood. The model's performance metrics were:

- Accuracy: 0.60
- Precision (Non-Divorced): 0.61
- Recall (Non-Divorced): 0.93
- Precision (Divorced): 0.56
- Recall (Divorced): 0.14

The model performed well in identifying stable couples but had difficulty classifying couples likely to divorce. This outcome suggests either class imbalance or subtle divorce-related signals that require further modeling techniques.

Feature importance was derived by combining PCA loadings with logistic regression coefficients. This allowed mapping PCA components back to the original psychological and relational variables.

4 Results

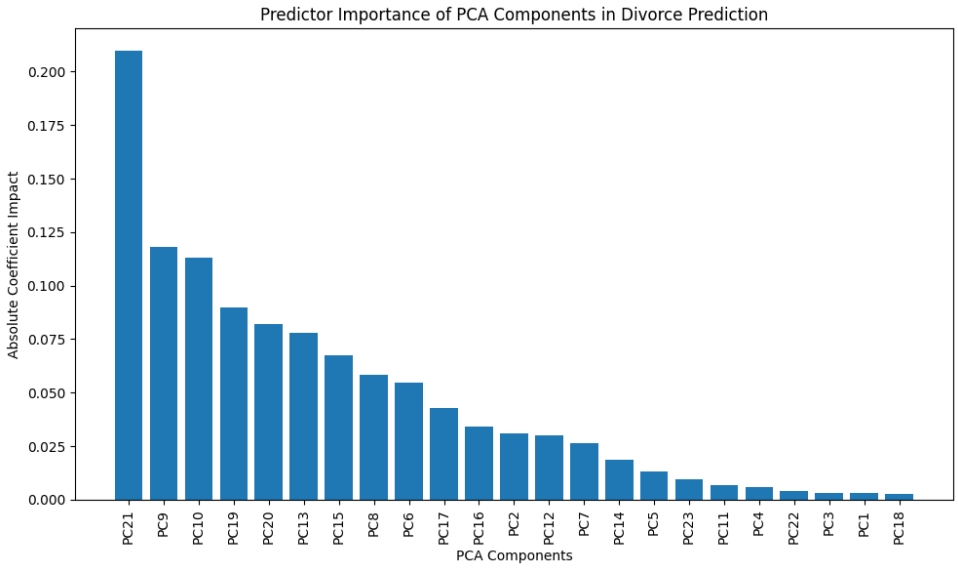

Figure 7: Most influential principal components driving divorce prediction.

The bar graph illustrates the relative contribution of each principal component (PC) to the logistic regression model predicting divorce. The height of each bar represents the absolute value of the logistic regression coefficient associated with that component, meaning components with larger absolute coefficients exert a stronger influence on the model's predictions.

The results show that PC21 is by far the most influential component, with an impact of approximately 0.21, more than double that of any other component. This indicates that the variables heavily loading onto PC21—primarily financial stress, infidelity, and trust—form the most dominant latent dimension affecting divorce likelihood. Following PC21, PC9 and PC10 also show notable influence, with impacts around 0.12 and 0.11, respectively. These components capture additional patterns related to education level, shared hobbies, communication quality, mental health, and pre-marital or demographic factors such as cohabitation history and age at marriage.

The remaining components gradually decrease in contribution, with many having minimal impact on the model (near zero). This pattern indicates that only a subset of components captures the majority of predictive signal, while the rest represent noise or low-relevance variability in the data. Overall, this distribution validates the usefulness of PCA: it compresses the predictive information into a few meaningful components while reducing redundancy among the original variables.

PC21 Top Loadings

Variable	Loading
financial_stress_level	0.510991
infidelity_occurred	0.425825
trust_score	0.393879
education_level_PhD	0.289452
education_level_No Formal Education	0.256588

PC9 Top Loadings

Variable	Loading
education_level_PhD	0.569001
shared_hobbies_count	0.347705
communication_score	0.281541
mental_health_issues	0.272854
infidelity_occurred	0.24924

PC10 Top Loadings

Variable	Loading
pre_marital_cohabitation	0.486167
age_at_marriage	0.402984
social_support	0.395877
communication_score	0.265194
num_children	0.263575

Figure 8: Dominant variables in the top three predictive principal components (PC21, PC9, and PC10).

The tables summarizing the top loadings for PC21, PC9, and PC10 highlight the variables that contribute most strongly to the principal components most associated with divorce prediction. PC21, which has the highest influence in the logistic regression model, is dominated by financial stress, infidelity, and trust, indicating that this component captures a central dimension of relational instability. These variables load heavily and in the expected direction, reinforcing the theoretical understanding that financial strain and breaches of trust significantly erode marital stability. PC9, on the other hand, is driven by education level, shared hobbies, communication, and mental health, suggesting a component representing personal compatibility and psychosocial functioning within the marriage. Lastly, PC10 is shaped primarily by pre-marital cohabitation, age at marriage,

social support, and number of children, reflecting a demographic–background dimension related to life history and social environment. Together, these components demonstrate how PCA uncovers latent structures within the dataset, grouping related predictors into meaningful dimensions that are later used by logistic regression to enhance interpretability and predictive performance.

4.1 Most Influential Predictors

The following variables were determined to have the strongest predictive impact on divorce:

- Financial stress level – the strongest predictor of divorce.
- Infidelity – strongly associated with marital breakdown.
- Trust score – lower trust significantly increases divorce risk.
- Communication score – poor communication contributes to marital instability.
- Mental health issues – increase the likelihood of divorce.
- Social support – low external support is linked to higher divorce rates.
- Shared hobbies – couples who share more activities are less likely to divorce.

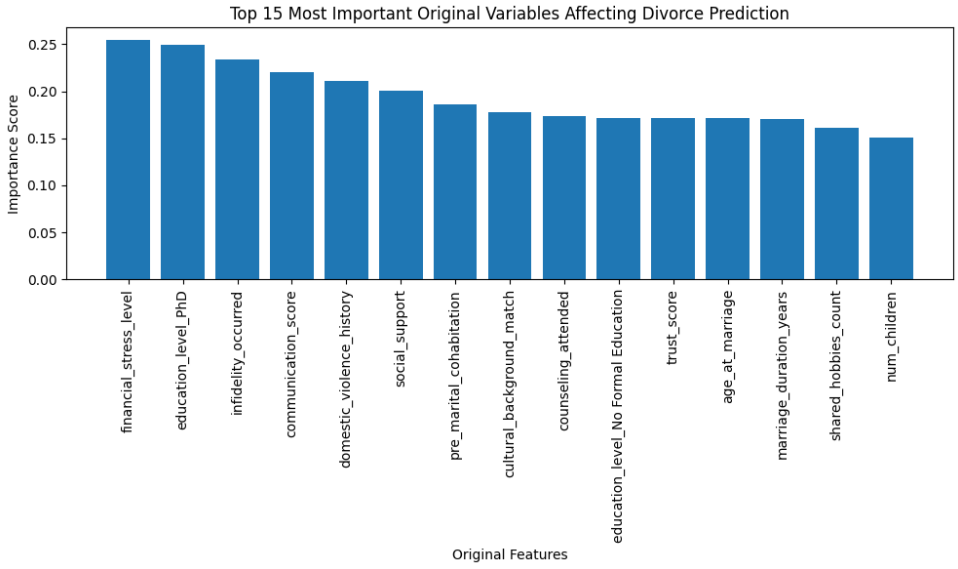

Figure 9: Most important original predictors.

These results align with extensive psychological and sociological research on marital stability.

4.2 Least Influential Predictors

Variables that showed minimal predictive impact included:

- Education level.
- Marriage type (love, arranged, other).
- Cultural background match.

Although these variables explained variance in PCA, they did not meaningfully influence divorce prediction in logistic regression.

4.3 Hypothesis Testing

The hypothesis examined in this study was:

$$H_0 : \text{Financial stress does not affect the likelihood of divorce.}$$
$$H_1 : \text{Financial stress increases the likelihood of divorce.}$$

To evaluate this hypothesis, a logistic regression model was fitted using financial stress level as the predictor and divorce status as the binary outcome. Logistic regression is an appropriate statistical test because the dependent variable is dichotomous, and the coefficient and associated p-value allow for formal hypothesis evaluation. This test follows the hypothesis testing framework outlined in Unit 3 of the course, where statistical significance is determined by comparing the p-value to a significance level of $\alpha = 0.05$.

The logistic regression results showed the following:

- Coefficient for financial stress: 0.0620
- p-value: 6.58×10^{-7}
- Odds ratio: 1.064

The coefficient is positive, indicating that higher financial stress is associated with an increased likelihood of divorce. More importantly, the p-value is extremely small and far below the significance threshold $\alpha = 0.05$. Therefore, we reject the null hypothesis (H_0) and conclude that financial stress significantly increases the likelihood of divorce.

The odds ratio of 1.064 means that:

> For every one-unit increase in financial stress, the odds of divorce increase by approximately 6.4%.

These results provide strong statistical evidence supporting the alternative hypothesis (H_1) and confirm that financial stress is an important predictor of marital dissolution.

5 Discussion

The findings reveal that relational and psychological variables are far more important predictors of divorce than demographic factors. Financial stress, infidelity, trust, communication, and mental health issues emerged as critical drivers of marital breakdown. PCA components tied to these variables also had the strongest predictive influence in the logistic model.

Interestingly, variables such as marriage type and education level were not strong predictors, even though they contributed to general variance in the data. This highlights the distinction between variables that create structural differences in data and those that truly influence marital outcomes.

The logistic regression model demonstrated strong performance for non-divorced couples but had limited ability to identify divorcing couples. This suggests:

- Class imbalance.
- Nonlinear relationships not captured by logistic regression.
- Possible need for resampling strategies (e.g., SMOTE) or tree-based models.

6 Limitations

This study faces several important limitations that should be acknowledged when interpreting the results. First, the dataset is imbalanced, with substantially more non-divorced

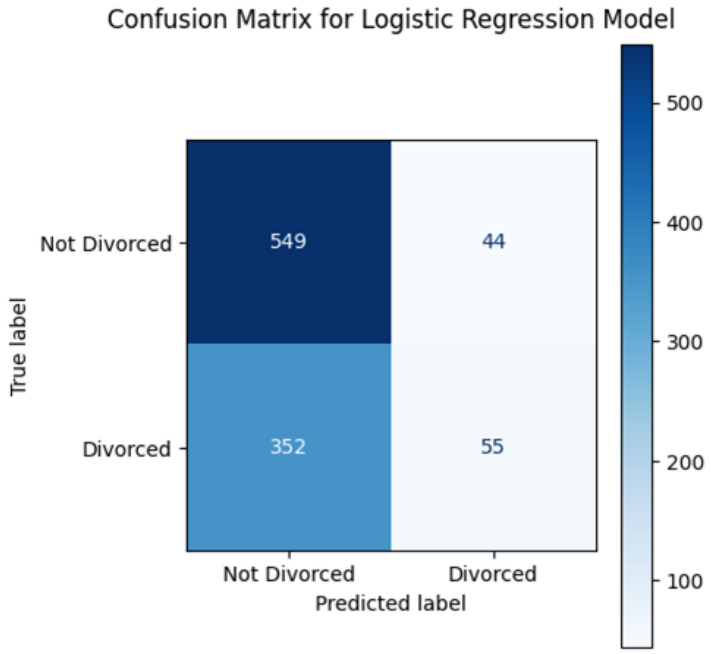

Figure 10: Confusion matrix of the logistic regression model.

couples (3,009) than divorced ones (1,991). This imbalance reduces the model's ability to accurately identify couples at risk of divorce, contributing to the high number of false negatives in the confusion matrix. Second, several key psychological constructions, such as communication quality, trust, and emotional stability, are represented using single-item numerical measures. These simplified indicators lack the depth and multidimensional structure typically required to capture the complexity of relationship dynamics, potentially weakening the explanatory power of the model. Third, the analysis is based on cross-sectional data that provide a snapshot of couples at one point in time. Because divorce is a gradual process influenced by evolving relational, financial, and emotional conditions, the absence of longitudinal observations limits the model's capacity to capture temporal changes that may precede marital dissolution. Finally, the dataset likely reflects a specific cultural and societal context, meaning that relationship norms, divorce rates, and family structures may differ substantially across regions or communities. As a result, the generalizability of the findings to broader or culturally distinct populations may be limited. Together, these limitations highlight the need for more comprehensive, balanced, and context-aware data to improve the predictive accuracy and applicability of divorce prediction models.

7 Conclusion

This study investigated the factors that most strongly predict the likelihood of divorce by applying exploratory data analysis, principal component analysis, and logistic regression to a dataset of 5,000 couples. The results showed that relational and emotional factors, particularly financial stress, infidelity, trust levels, communication patterns, and social support, play a substantial role in distinguishing couples who remain married from those who separate. PCA revealed underlying latent dimensions, such as relationship instability and marital quality, which helped address multicollinearity among predictors and improved model interpretability. Logistic regression further demonstrated that financial stress and infidelity were among the strongest contributors to divorce likelihood.

However, despite identifying meaningful patterns, the model achieved moderate predictive accuracy and struggled to correctly identify divorce cases, largely due to dataset imbalance and the simplified nature of several psychological variables. Additional limitations, including cultural specificity and the use of cross-sectional rather than longitudinal data, further constrain the generalizability of the findings. Overall, the study highlights both the potential and challenges of using statistical models to understand marital outcomes, suggesting that richer, more balanced, and temporally detailed data would allow for more accurate and actionable divorce prediction frameworks.