Divorce Prediction

Shea Yee, Khoo

15 June 2024

Agenda

- Introduction
- Objective
- About Dataset
- Insights Gathered (EDA)
- ML Models
- Model Performances with Lesser Features
- Summary & Suggestion

Introduction

- In today's society, marital stability is a cornerstone of family well-being, yet many marriages face challenges that can lead to divorce.
- **Understanding the factors that contribute to divorce** can help in developing interventions to promote healthier relationships and reduce the incidence of divorce.
- This project leverages data science techniques to **predict the likelihood of divorce** based on survey responses, providing valuable insights into the predictors of marital dissolution.

Objective

Problem Statement

- Divorce is a complex issue influenced by many factors. The challenge is **to predict divorce accurately** and **to identify the most significant predictors.**

Solution

- Predict divorce using machine learning models based on survey responses.

Stakeholders

- **Couples and Individuals**: Gain insights to strengthen their relationships.
- **Marriage Counselors and Therapists**: Tailor interventions based on predictions.
- **Researchers and Academics**: Enhance understanding of marital stability.
- **Policy Makers**: Design programs to reduce divorce rates.

Value Proposition

- This project offers numerous benefits, including **early identification of at-risk couples** to facilitate timely help, providing data-driven insights for counselors, contributing valuable data to marital studies, and aiming to reduce divorce rates along with the associated emotional and financial costs.

About Dataset

- The dataset comprises responses to survey questions collected from 170 participants, consisting of both divorced and married couples (86 & 84 participants respectively).
- Dataset is retrieved from https://www.kaggle.com/datasets/rabieelkharoua/split-or-stay-divorce-predictor-dataset/data
- **Features**: 54 columns (labeled Atr1 to Atr54) corresponding to Question 1 to Question 54 (sample questions on next slide)
- **Target variable**: The last column "Class" is the status column. 'Married' is represented as '1' and 'Divorced' as '0'."

Features

Target variable

	Atr1	Atr2	Atr3	Atr4	Atr5	Atr6	Atr7	Atr8	Atr9	Atr10	...	Atr46	Atr47	Atr48	Atr49	Atr50	Atr51	Atr52	Atr53	Atr54	Class
0	2	4	4	1	0	0	0	0	0	2	...	2	1	3	3	3	2	3	2	1	1
1	4	4	4	3	0	0	0	4	4	4	...	3	0	3	4	4	4	4	2	4	1
2	3	2	0	2	1	2	2	1	4	0	...	3	2	3	1	4	2	2	0	2	1
3	3	2	3	2	3	0	3	3	1	2	...	2	2	4	4	4	1	2	3	1	1
4	4	3	0	1	1	1	0	2	0	0	...	2	1	2	3	2	2	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165	0	0	0	0	0	0	0	0	0	0	...	1	0	2	1	1	4	2	2	2	0
166	3	0	3	3	0	0	0	0	0	0	...	4	1	0	2	4	2	3	3	2	0
167	1	1	4	4	0	0	0	0	0	1	...	1	0	2	0	1	1	0	0	0	0
168	0	2	0	0	0	0	0	3	0	0	...	3	3	2	4	3	2	4	3	0	0
169	0	0	0	0	0	0	0	0	0	0	...	3	4	4	4	1	2	4	4	0	0

170 rows × 55 columns

About Dataset (Sample of survey questions)

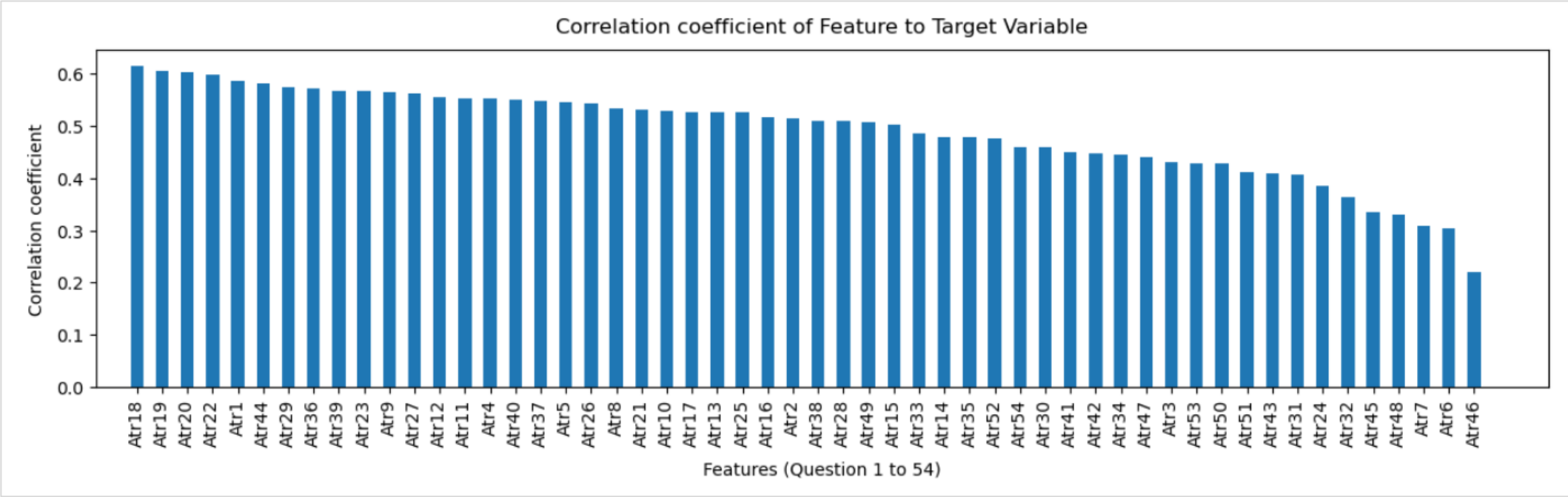
1. When one of our apologies apologizes when our discussions go in a bad direction, the issue does not extend.
2. I know we can ignore our differences, even if things get hard sometimes.
3. When we need it, we can take our discussions with my wife from the beginning and correct it.
4. When I argue with my wife, it will eventually work for me to contact her.
5. The time I spent with my wife is special for us.
6. We don't have time at home as partners.
7. We are like two strangers who share the same environment at home rather than family.
8. I enjoy our holidays with my wife.
9. I enjoy traveling with my wife.
10. My wife and most of our goals are common.
11. I think that one day in the future, when I look back, I see that my wife and I are in harmony with each other.
12. My wife and I have similar values in terms of personal freedom.
13. My husband and I have similar entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams of living with my wife are similar and harmonious.
16. We're compatible with my wife about what love should be.
17. We share the same views with my wife about being happy in your life.
18. My wife and I have similar ideas about how marriage should be.
19. My wife and I have similar ideas about how roles should be in marriage.
20. My wife and I have similar values in trust.

⋮

53. When I discuss it, I remind her of my wife's inadequate issues.
54. I'm not afraid to tell her about my wife's incompetence.

Insights Gathered (EDA)

- Performed point-biserial correlation on each of the features against target variable (binary class)



- Each feature contributes to the target variable to a certain extent, without any single feature showing notably stronger correlation compared to the others.
- We will train **supervised classification models** that captures contribution of these features.

ML Models | Model Selection

- Modeling : This is a supervised binary classification problem.
- Sixteen classification models are trained to predict Divorce/ Married based on 54 features.
- Before training each model, the dataset is divided into 3 parts (or folds) for cross-validation. Accuracy scores are then computed based on these cross-validation folds.
- Among models achieving the highest accuracies, **Logistic Regression was chosen for divorce prediction** due to its simplicity in both implementation and processing time.

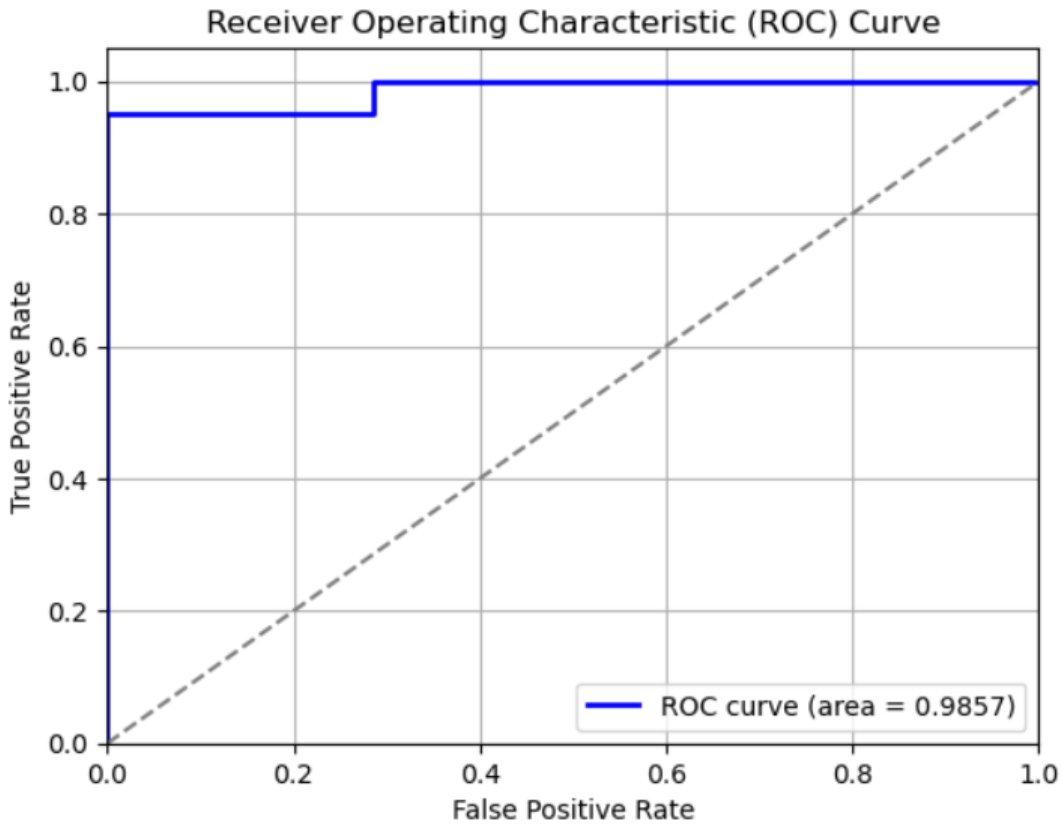
Model	Accuracy	Stdev	Processing time (s)
Logistic Regression	**0.9765**	**+/- 0.0082**	**0.021**
Bagging LR	0.9825	+/- 0.0143	0.093
AdaBoost LR	0.9765	+/- 0.0082	0.020
SVC	0.9766	+/- 0.0219	0.012
Bagging SVC	0.9764	+/- 0.0086	0.063
AdaBoost SVC	0.9708	+/- 0.0298	0.115
K-NN	0.9764	+/- 0.0167	0.015
GaussianNB	0.9707	+/- 0.0298	0.008
Bagging GNB	0.9765	+/- 0.0082	0.034
AdaBoost GNB	0.9004	+/- 0.0577	0.111
Decision Trees	0.8413	+/- 0.0133	0.007
Bagging DT	0.9001	+/- 0.0162	0.036
AdaBoost DT	0.8530	+/- 0.0077	0.010
Gradient Boosting	0.9291	+/- 0.0389	0.156
Random Forest	0.9706	+/- 0.0084	0.197
Stacking Classifier	0.9704	+/- 0.0170	0.864

ML Models | Model performances

- Metric #1 : **Accuracy score 0.9765 +/- 0.0082**

- Metric #2 : **Area under ROC curve ~0.99** (Curve is very close to top-left corner)

  Model has outstanding performance in terms of distinguishing between positive and negative classes.



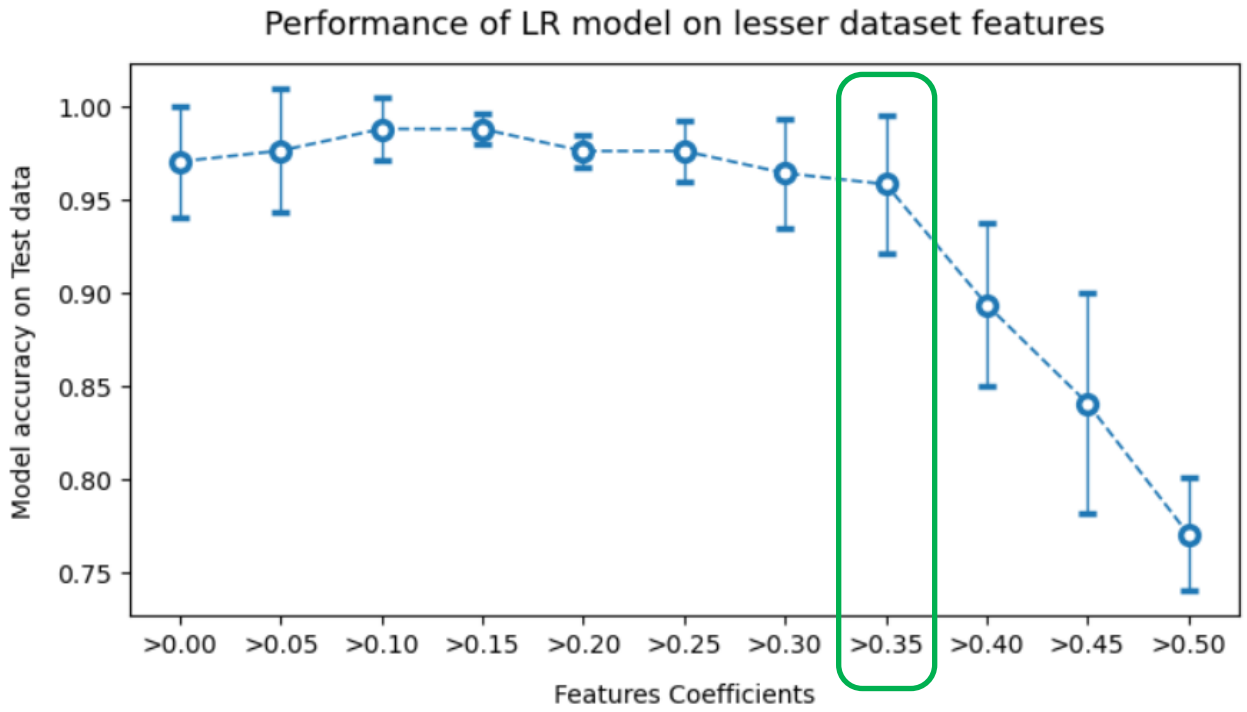
ML Models | Examine Features Importance

- Assess the importance of features (survey questions) in the trained model by examining their coefficients, as displayed in the table to the right.
- **"Are ALL survey questions truly necessary for accurately predicting Divorce/ Married?"**

	Feature	Coefficient	Coefficient(abs)
0	Atr30	0.506821	0.506821
1	Atr44	0.468043	0.468043
2	Atr40	0.457532	0.457532
3	Atr42	0.434503	0.434503
4	Atr1	0.430442	0.430442
5	Atr11	0.397022	0.397022
6	Atr18	0.392593	0.392593
7	Atr25	0.376216	0.376216
8	Atr52	0.362117	0.362117
9	Atr36	0.348962	0.348962
10	Atr6	0.317963	0.317963
11	Atr14	0.310469	0.310469
12	Atr19	0.292337	0.292337
13	Atr35	0.278165	0.278165
14	Atr34	0.271580	0.271580
15	Atr53	0.261419	0.261419
⋮	⋮	⋮	⋮
50	Atr41	-0.026727	0.026727
51	Atr32	0.018168	0.018168
52	Atr33	0.003632	0.003632
53	Atr21	0.001850	0.001850

ML Models | Examine Features Importance

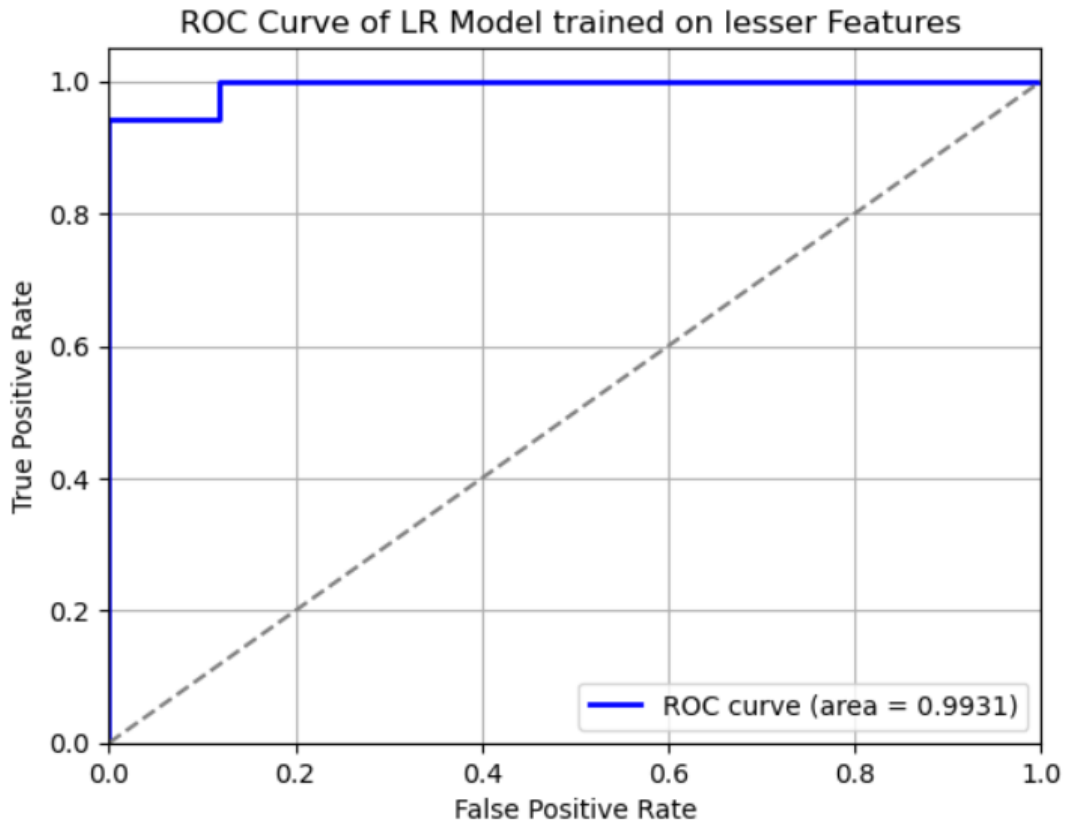
- Features Coefficients = [>0, >0.05, >0.10, >0.15, >0.20, >0.25, >0.30, >0.35 , >0.40, >0.45, >0.50]
- Train LR model with lesser features based on coefficient range defined above. Accuracy scores as follow:



- **Reducing the number of features to those with coefficients >0.35 maintains a significantly high accuracy score for the model.**

ML Models | Model performances

- Performance metrics of **new LR model** trained based on lesser dataset features.
- Metric #1 : **Accuracy score 0.9585 +/- 0.0367**
- Metric #2 : **Area under ROC curve ~0.99** (Curve is very close to top-left corner)
- Reducing the number of features to those with coefficients >0.35 maintains a significantly high accuracy score & ROC AUC.
- It is recommended to include only these questions to streamline the survey process.



The 9 Questions that matter more (FYI)

1. I know my wife's friends and their social relationships.
2. Sometimes I think t's good for me to leave home for a while.
3. We're just starting a fight before I know what's going on.
4. When I argue with my wife, it only snaps in and I don't say a word.
5. When one of our apologies apologizes when our discussions go in a bad direction, the issue does not extend.
6. I think that one day in the future, when I look back, I see that my wife and I are in harmony with each other.
7. My wife and I have similar ideas about how marriage should be.
8. I have knowledge of my wife's inner world.
9. I wouldn't hesitate to tell her about my wife's inadequacy.

Summary & Suggestion

Successful Model Creation

- We have successfully developed a robust **predictive model that accurately identifies at-risk couples early in their relationship**. By analyzing comprehensive data, our model provides valuable insights into the factors influencing marital stability and divorce risk, contributing significantly to marital studies and intervention strategies.

Suggestions

- Integrate the predictive model into counseling practices to offer data-driven insights for tailored interventions.
- Engage with community organizations and policymakers to advocate for early intervention and supportive policies.

Model Limitation & Suggestion

The divorce prediction model achieves a **high accuracy score despite not accounting for crucial aspects like family commitment** (such as having children, mortgage loans, and other financial obligations).

This could be due to the **homogeneity of the survey participants**, as all respondents are from Turkey, **leading to biased data collection.**

To improve the robustness and generalizability of our model, **suggest conducting the survey among a more diverse group of people**. Including participants from different backgrounds and countries will provide a more comprehensive dataset, leading to more reliable and applicable predictions.

Thank you