Identification of Promising Couples using Machine Learning

Yao Liu
liuyaoly@umich.edu

Chris Labiak
clabiak

Matt Kliemannn
mtkliema

Kaustubh Srivastava
kaustubh

Yao Xiao
xyaoinum

Abstract

Divorces, or more generally, failed relationships, are tremendous social issues that not only affect the two people involved, but also friends and family. In this project, we try to address this social issue by finding out what features are indicative of an unpromising relationship and use these features to predict the outcome of a romantic relationship over a 4 year time frame. The existing social science research is mostly based on simple statistical models. Our project aims to address this issue by applying machine learning algorithms. Specifically, using a variety of feature selection and classification methods, we have been able to achieve predictions with accuracy in the low eighty percent range and discovered several new important factors that affect a relationships outcome, such as meeting on the internet having a negative impact.

1 Introduction

1.1 Motivation

This project aims to predict whether or not a couple (either married or in a romantic relationship) will break in the next four years. We also aim to identify the features that are most indicative of a breakup.

The motivation of pursuing this project is to tackle the widespread social issue of failed relationships. According to current social studies, divorce rates have doubled over the past two decades among persons over age 35. In addition, younger couples have become increasingly more critical about choosing their significant other, resulting in a low marriage rate among younger people which can have significant socioeconomic consequences [Kennedy and Ruggles, 2014]. As a result, the identification of important factors predicting a long lasting relationship is an important task.

By solving this problem, we can get great insight into what the most significant features are that make a couple stay together for a long period of time. These features could then be used by individuals as guidelines when looking for a partner to increase the chances of being successful couples. As a result, this study may help alleviate increasing divorce rate issues and make it easier for people to find their most compatible partner.

1.2 Challenges

The dataset is from the survey How Couples Meet and Stay Together [Rosenfeld et al., 2011]. This collection consists of the initial survey given in 2009 and three follow up surveys in 2010, 2011, and 2013 (referred to as wave 2, 3, 4), respectively.

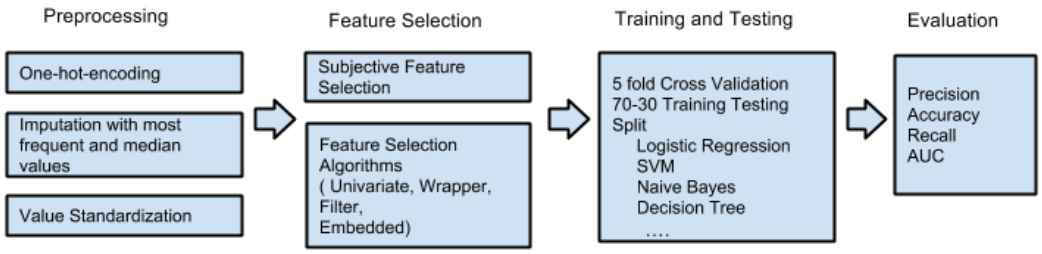
The initial survey is terminated early if the respondent is not in an active relationship - therefore our baseline is respondents who were in a relationship during the initial survey. The follow up surveys were also no longer conducted if a respondent broke up or got divorced in the prior survey.

We classify a respondent's relationship outcome as positive if they indicated they divorced or ended the relationship during one of the follow up surveys. If they responded that they were still in the same relationship during the 4th wave, we consider the respondent relationship to be a negative example.

We have 1874 examples, 450 of which are positive examples (failed relationships). This limited size of data set makes it hard to train predictive models. In addition, among the dataset, the ratio between positive and negative examples is imbalanced (1:4 ratio). A limited number of failed relationship examples makes it hard to discover important factors that caused break-up in the relationship.

Furthermore, to investigate the social issues on same-sex couples, the survey recruits more same-sex couples intentionally. However, the number of same-sex couples samples is not enough for us to train models to learn the different behaviors in relationships for two types of couples [Weisshaar, 2014].

2 Method



2.1 Assumptions

We do not consider the length of time that couples been together prior to the the start of the study aside from as a feature. Our model considers the 4 year outlook at any point of time in the relationship. We do not currently consider the weighting information provided by the survey firm to account for oversampling of parts of the population in the survey. For example, approximately 15% of respondents were same-sex couples while this is closer to 2% in the US general population. We do not consider the variables that may evolve after wave 1 such as income and education levels because followup surveys are not conducted once a couple has broken up. We assume the respondents answers to be unbiased and reflect the true nature/ reality of their relationship (brought about by culture difference) [Thomas, 2011]

2.2 Feature Preprocessing

The dataset requires some significant preprocessing in order to train easier and more understandable models. Besides pre-existing binary data, are two major types of data included in this dataset which special preprocessing: continuous and categorical.

2.2.1 Categorical Features

Many of the features are in categorical form. For example, the feature indicating race holds values: 1 for Caucasian, 2 for Black, 3 for Hispanic, 4 for Asian, and 5 for other races. This sort of representation lacks an ordering to the categories and would distort interclass distances. First, the missing data is imputed using the most frequent value of existing data for that feature. Most frequent value is used since other basic approaches like median or mean arent built for categorical data. Next, the categorical features are transformed into binary values through the one-hot encoding (one-of-K encoding) method provide by Scikit Learn. This method transforms a categorical feature with [M] possible values into [M] binary features where only one is true.

2.2.2 Continuous Features

For continuous features, we rectify missing values using median value imputation based on the median of values present for that feature. We also do feature standardization on continuous features by scaling them with zero mean and unit variance. This enables us to use different models such as SVM with gaussian kernel. After this process, we have 138 features to utilize. The entire feature set will serve as the baseline.

2.3 Feature Selection

One of the major challenges in our project is that we have a large number of features compared to our small sample size. To address this issue, we propose two feature selection methods. The first method is to select features based on subjective evaluation from relevant social science academic research. The other method is to apply and evaluate various feature selection algorithms.[Iavindrasana et al., 2009].

2.3.1 Subjective Feature Selection

There have been several social studies working on the same data set [Weisshaar, 2014][Rosenfeld and Thomas, 2012] [Thomas, 2011]. We incorporate features from conclusions and statistical analysis process of these studies as subjective features.

The features we selected are household income [Weisshaar, 2014], same sex couples [Weisshaar, 2014], existing divorce [Kennedy and Ruggles, 2014] , cohabitating couple [Thomas, 2011], years of education [Thomas, 2011], met through friends [Thomas, 2011], met at school [Thomas, 2011], met at church [Thomas, 2011] and met through internet [Rosenfeld and Thomas, 2012].

2.3.2 Algorithmic Feature Selection

We do univariate feature selection using chi-squared score. The top 20 features based on score are selected. [Pedregosa et al., 2011] We do multivariate feature selection using a variety of algorithms. We use filter methods such as fisher score based feature selection, correlation-based feature selection. fast correlation-based filter feature selection. [Zhao et al., 2010] We also use embedded methods such as L1-SVM with linear kernel feature selection and L1-logistic regression feature selection. L1 regularization enforces sparsity which causes only a few features to be selected. [Pedregosa et al., 2011] [Iavindrasana et al., 2009] Another embedded method used is recursive feature elimination (evaluated using linear SVM) [Pedregosa et al., 2011]

2.4 Model Training

After feature selection, the dataset was classified by the following supervised learning classifiers: Gaussian naive Bayes, Support Vector Machine with linear kernel, k-Nearest Neighbors, Decision Tree, Ridge Regression, L1 Regularized Logistic Regression, and Adaboost using decision tree. Each classifier model was trained using 5-fold cross validation and is evaluated using a 70-30 training-testing split to avoid overfitting issues.

2.5 Model Evaluation

1. Overall Accuracy
2. Precision and Recall
3. Area Under ROC Curve (AUC)

3 Related Works

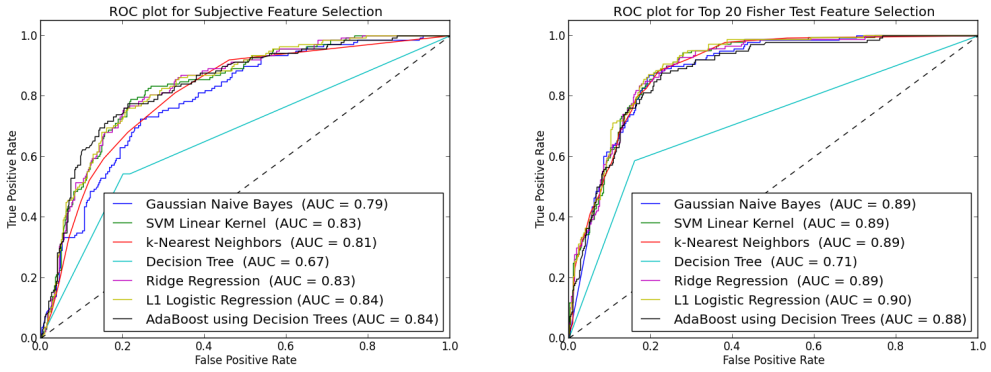
One category of related work is social science work. We refer to the statistical prediction process and conclusion of these works as our first subjective method of features selection [Weisshaar, 2014][Rosenfeld and Thomas, 2012] [Thomas, 2011] . These works are based on social

science, so they only do simple regression and test the statistical significance, which may induce overfitting problems. In our case, we combine knowledge from these research work and train the models through machine learning algorithms.

Gottman is the expert in marriage outcome research. He was able to construct a short term divorce prediction model (7- years) for married couples which is 93% accurate. [Gottman and Levenson, 2000] His model primary uses features about marital interaction and emotion. He found that negative affect during marital conflict most predicted breakup. Our model differs in that the feature set we use primarily consists of demographic data. In addition, we predict the outcome of a relationship in general, not only married couples. Finally, we do extensive cross validation on our dataset, which is not typically done in research about relationship outcomes. [Heyman and Slep, 2001]

The second category of related work we consider is clinical machine learning problem because of the similar issues such as unbalanced dataset, no clear boundary, large number of features and factor analysis issues etc. We refer to this work on feature selection and evaluation methods [Iavindrasana et al., 2009].

4 Experimental Results



A machine learning feature selection approach will provide better classification power than subjectively selecting features, as is conventionally done in social science research. In addition, the features selected by machine learning algorithms have significantly increased the power to correctly classify positive examples (higher AUC). This is important to identify the most contributing factors in an unsuccessful relationship.

The most predictive feature of a relationship success turned out to be a high relationship quality (1). Another very predictive feature of relationship success include living together (2) for a long time (4), which confirms with the social science work [Thomas, 2011].

We can also confirm that the longer a couple is in a relationship, they are less likely to breakup [Thomas, 2011]. Specifically, features 6, 10, 11, 12, 15. The only downside is meeting at an older age (16), which has a negative weight. This is likely related to being divorced (9), as divorced people who are back on the market at an older age do not find success.

Other features that reflect positively on a couples outcome include positive parental approval (14), high household income (8).

Meeting on the internet turns out surprisingly to reflect negatively on relationship success (17), which is indicated as an increasing intermediary to look for partners in recent years in social science work [Rosenfeld and Thomas, 2012].

Finally, if the respondent is paying for their home in some manner (3,5), that has a positive impact on their chance for increasing relationship success, as compared to someone who lives for free.

Classifier \ Feature Selection		No Selection	Subjective	Univariate		Multivariate			
						Embedded			Wrapper
				Top 20 Chi-Square	Top 20 Fisher-Test	L1-LogReg	L1-SVM	RFE	CFS
Naive Bayes	Accuracy	0.518	0.772	0.792	0.815	0.811	0.817	0.407	0.824
	Precision	0.964	0.399	0.775	0.87	0.804	0.819	0.985	0.819
	Recall	0.333	0.55	0.554	0.583	0.584	0.592	0.291	0.604
SVM	Accuracy	0.817	0.701	0.813	0.827	0.811	0.811	0.815	0.817
	Precision	0.84	0.848	0.601	0.775	0.761	0.761	0.783	0.768
	Recall	0.589	0.443	0.624	0.618	0.59	0.59	0.593	0.599
Nearest Neighbors	Accuracy	0.802	0.792	0.819	0.82	0.811	0.806	0.824	0.833
	Precision	0.326	0.449	0.522	0.536	0.5	0.464	0.543	0.594
	Recall	0.714	0.602	0.667	0.667	0.651	0.646	0.676	0.683
Decision Tree	Accuracy	0.776	0.737	0.774	0.778	0.776	0.783	0.797	0.794
	Precision	0.609	0.542	0.558	0.587	0.449	0.493	0.666	0.63
	Recall	0.538	0.469	0.538	0.544	0.554	0.567	0.575	0.572
Ridge	Accuracy	0.842	0.776	0.822	0.826	0.833	0.83	0.831	0.829
	Precision	0.623	0.225	0.58	0.609	0.623	0.616	0.63	0.616
	Recall	0.699	0.62	0.656	0.656	0.672	0.669	0.664	0.664
Logistic Regression	Accuracy	0.827	0.792	0.822	0.829	0.833	0.831	0.835	0.831
	Precision	0.616	0.304	0.572	0.616	0.609	0.601	0.616	0.609
	Recall	0.659	0.667	0.658	0.664	0.677	0.675	0.68	0.672
Adaboost	Accuracy	0.82	0.829	0.811	0.82	0.824	0.831	0.801	0.82
	Precision	0.594	0.565	0.543	0.609	0.58	0.623	0.572	0.601
	Recall	0.646	0.684	0.636	0.641	0.661	0.667	0.598	0.643

Results over selected feature selection algorithms and different classifiers have been listed above. Most models have achieved classification accuracy in the low 80% range. SVM model with features selected by fisher scores provides the best combination of accuracy, recall, and precision results.

Number	Weight	Feature
1	-2.279	Self Described Relationship Quality
2	-1.411	Couple is cohabiting
3	-0.923	Respondent owns or mortgages home
4	-0.8	How long ago couple began cohabiting
5	-0.792	Respondent rents home
6	-0.512	Age of partner
7	-0.48	Respondent is married
8	-0.454	Household Income
9	0.413	Respondent is divorced
10	-0.345	Age of respondent
11	-0.277	How long ago couple were first romantic
12	-0.248	How long ago couple met
13	0.232	Respondent has never been married before
14	-0.213	Level of parental approval
15	-0.204	How long ago relationship began
16	0.158	Age of respondent when couple met
17	0.13	Couple met on the internet

Using features obtained from the Fisher test and trained on a linear SVM classifier, we obtained the weights each feature is allocated in the model. Some features with very low weights have not been shown.

5 Conclusion

From the experiment results above, the introduction of machine learning algorithms to this social science field has increased the predictive power of models. Moreover, it helps discover new important factors such as negative impact of meeting couple through internet and how a divorced person has a reduced chance of having a new successful relationship.

We are able to calculate a four year relationship success rate with approximately 82% accuracy.

Even though the data set itself only has few subjective and emotional evaluation of the relationship from the survey respondent, the self described relationship quality has been listed as most important factor, than other demographic statistics, by the best performance training model (SVM with top 20 features selected based on fisher scores). This also confirms the important role

of emotional factors that play in a romantic relationship, which is revealed by Gottmans work [Gottman and Levenson, 2000].

Based on the comparison between social science work and our approach, various changes can be done in the future to help better understand this social problem. The first aspect of future work may focus on decreasing the bias of our study by referring to social science methodology, such as using multiple imputation through SVD to fill in missing values [Howell, 2007], taking into account of weighting information [Winship and Radbill, 1994] and considering evolving values during the period with event history models [Weisshaar, 2014]. Another aspect of future work can focus on constructing new features based on social science domain knowledge on this problem, such as difference in race. Lastly, our work doesnt solely consider married couples or dating couples as is conventionally done by social science researchers. We considered both and our models tended to select marriage status as a useful feature and weight it highly. This informs us we should consider both instances separately - evaluating success prior to marriage and during marriage individually.

[Kennedy and Ruggles, 2014] [Weisshaar, 2014] [Rosenfeld and Thomas, 2012] [Thomas, 2011] [Heyman and Slep, 2001] [Rosenfeld et al., 2011] [Gottman and Levenson, 2000] [Iavindrasana et al., 2009] [Guyon and Elisseeff, 2003] [Pedregosa et al., 2011] [Zhao et al., 2010] [Metz, 1978] [Hanley and McNeil, 1982]

References

[Gottman and Levenson, 2000] Gottman, J. M. and Levenson, R. W. (2000). The timing of divorce: Predicting when a couple will divorce over a 14-year period. *Journal of Marriage and Family*, 62(3):737–745.

[Guyon and Elisseeff, 2003] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

[Hanley and McNeil, 1982] Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

[Heyman and Slep, 2001] Heyman, R. E. and Slep, A. M. S. (2001). The hazards of predicting divorce without crossvalidation. *Journal of Marriage and Family*, 63(2):473–479.

[Howell, 2007] Howell, D. C. (2007). The treatment of missing data. *The Sage handbook of social science methodology*, pages 208–224.

[Iavindrasana et al., 2009] Iavindrasana, J., Cohen, G., Depeursinge, A., Müller, H., Meyer, R., and Geissbuhler, A. (2009). Clinical data mining: a review. *Yearb Med Inform*, 2009:121–133.

[Kennedy and Ruggles, 2014] Kennedy, S. and Ruggles, S. (2014). Breaking up is hard to count: The rise of divorce in the united states, 1980–2010. *Demography*, 51(2):587–598.

[Metz, 1978] Metz, C. E. (1978). Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Rosenfeld and Thomas, 2012] Rosenfeld, M. J. and Thomas, R. J. (2012). Searching for a mate the rise of the internet as a social intermediary. *American Sociological Review*, 77(4):523–547.

[Rosenfeld et al., 2011] Rosenfeld, M. J., Thomas, R. J., and Falcon, M. (2011). and 2014.how couples meet and stay together. waves 1, 2, and 3: Public version 3.04, plus wave 4 supplement version 1.02 [computer files].

[Thomas, 2011] Thomas, R. J. (2011). How americans (mostly don't) find an interracial partner: Race and ethnic differences in the use of social foci and networks for couple formation.

[Weisshaar, 2014] Weisshaar, K. (2014). Earnings equality and relationship stability for same-sex and heterosexual couples. *Social Forces*, 93(1):93–123.

[Winship and Radbill, 1994] Winship, C. and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257.

[Zhao et al., 2010] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., and Liu, H. (2010). Advancing feature selection research–asu feature selection repository. *School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe*.