Machine Learning project
March 2020

Intro

Most Americans are well aware that half of marriages end in divorce. The good news is that the divorce rate in the US is decreasing. However, so is the marriage rate. According to the Center for Disease Control, the 2018 marriage rate was 6.9 per 1,000 total population while the divorce rate was 2.9 per 1,000 population. Both rates were decreased by a tenth of a point from the previous year and the year before. It seems fewer couples are willing to gamble the risk these days and opt out of marriage. With the high risk of divorce, we found the subject of divorce to be a compelling machine learning exercise. Can divorce be predicted based on specific attributes?

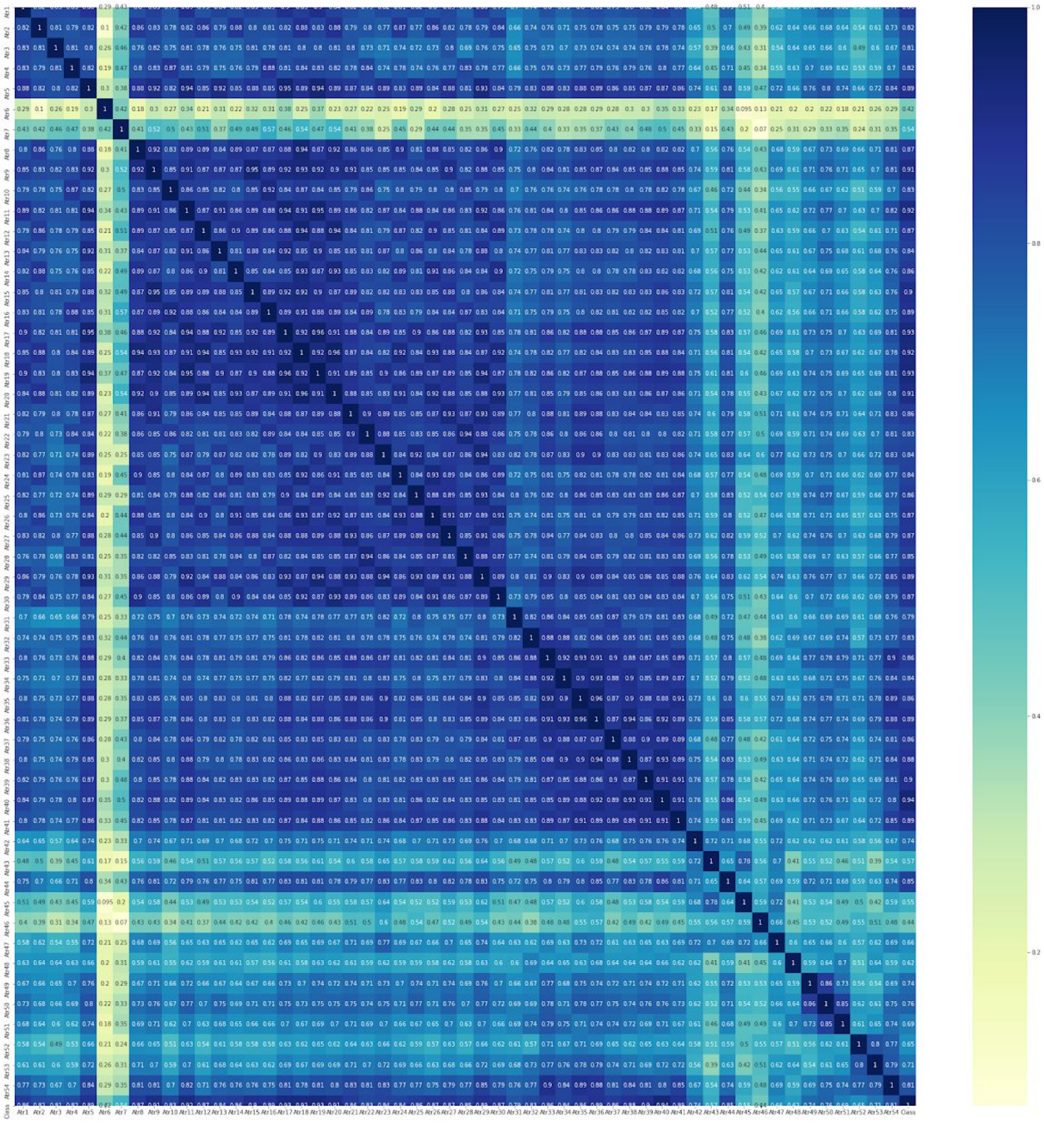
When looking for datasets around divorce we came across a study done in Turkey, where divorce rates have been increasing and are among the highest of the 33 member countries of the Organization for Economic Cooperation and Development (Yontem et al). The study used aspects of the Gottman couples therapy model, a model that derives the cause of divorce using empirical research (for more information on the Gottman Method click here). The study developed a Divorce Predictor Scale (DPS) based on the Gottman model. The DPS was composed of 54 questions exploring the interactions and feelings between couples. The study surveyed both married and divorced couples who answered the 54 questions based on a scale of 0-4, with 0 being strongly-disagree and 4 being strongly-agree. The Yontem study tested the accuracy of three machine learning methods -- Artificial Neural Networks (ANN), Radial-Basis Function Kernel (RFB) and Random Forest -- before and after feature selection. The ANN method combined with correlation based feature selection was identified to be the most accurate model to fit the data.

In this project, we decided to do a similar analysis using different machine learning methods. In addition, our analysis included an expanded data set, augmented with responses from close family and friends. The survey instrument was created in Google Forms and can be found here.
Our analysis was conducted using the following technologies: python, pandas: matplotlib, sklearn, seaborn, and numpy.

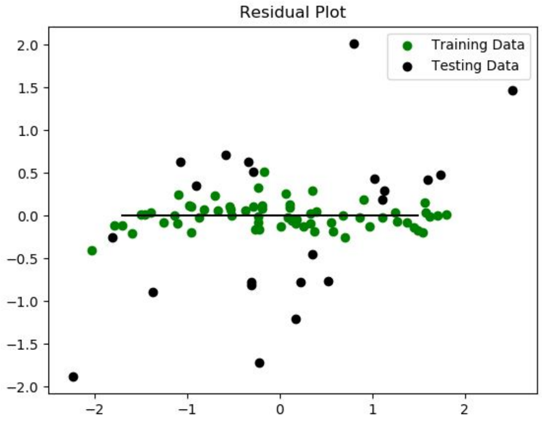
Analyses

Linear Regression

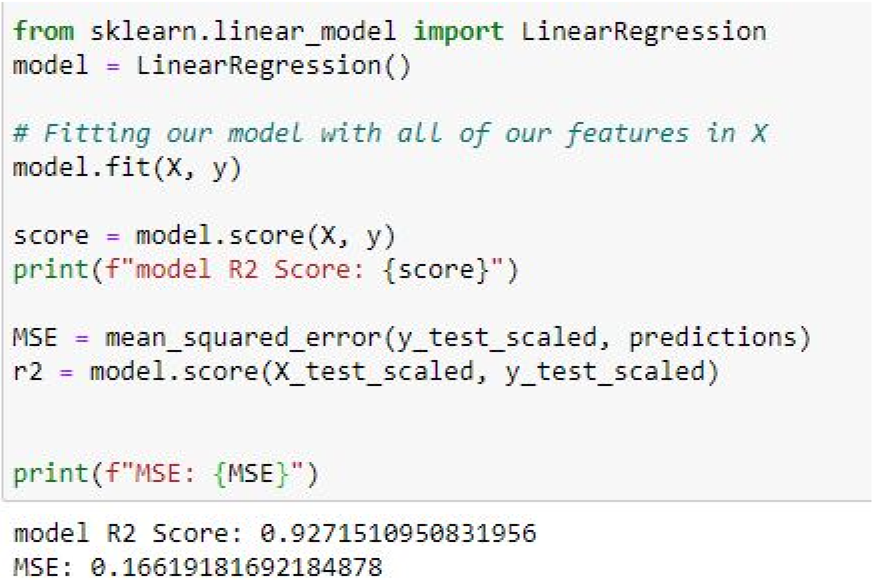
For the original Turkish dataset, we chose to do Lasso and Ridge linear regressions to reduce model complexity. Ridge puts constraints on the coefficients which in effect, shrinks the coefficient. Lasso differs from Ridge in that it helps reduce over-fitting by taking into account magnitudes. Both regression methods find the first point where the elliptical hits the region of constraints. For more information on these analyses click here.
When splitting our data into test and training sets, we got an R2 score of 92% indicating that our data is predictable and linear regression is a good fit. Both the Lasso and Ridge analyses yielded R2 scores above 80% and low mean squared error rates below 0.17 indicating our model explains a good amount of the variability around our means. These statistics suggest that we can produce predictions that are reasonably precise. Our lower MSE indicates we are close to finding the line of best fit for our model. It is no surprise that our R2 value decreased once running the analysis on our dataset as opposed to the training set. In addition to linear regression, we also used a heatmap to get an overall look into the distribution of each attribute in our dataset. We used seborn to create the heatmap looking at all 54 survey questions and average values.



The heatmap showed significant attributes of the dataset, specifically some of the more negative survey questions. The heatmap was a great way of looking at the dataset as a whole. It shows how much each feature affects our predictions, with higher numbers (darker colors) having the most effect.

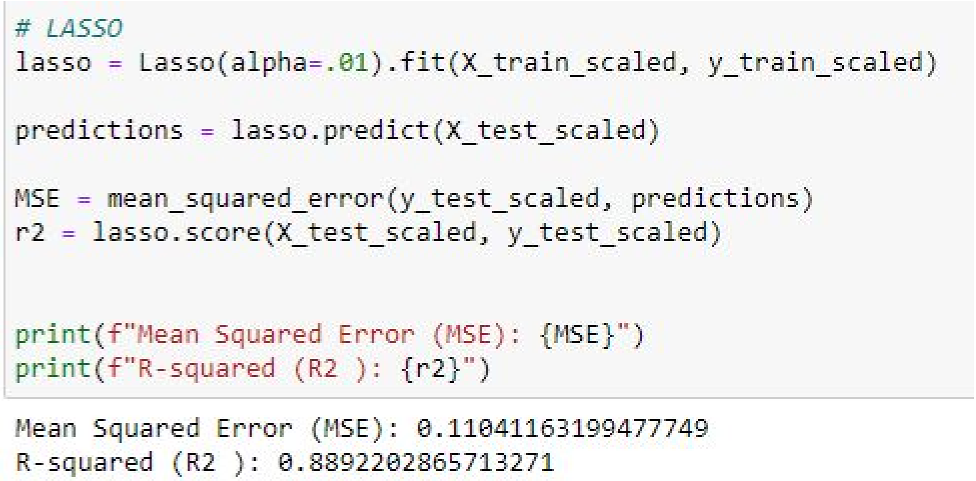


Residual plot of test and training data sets looked promising



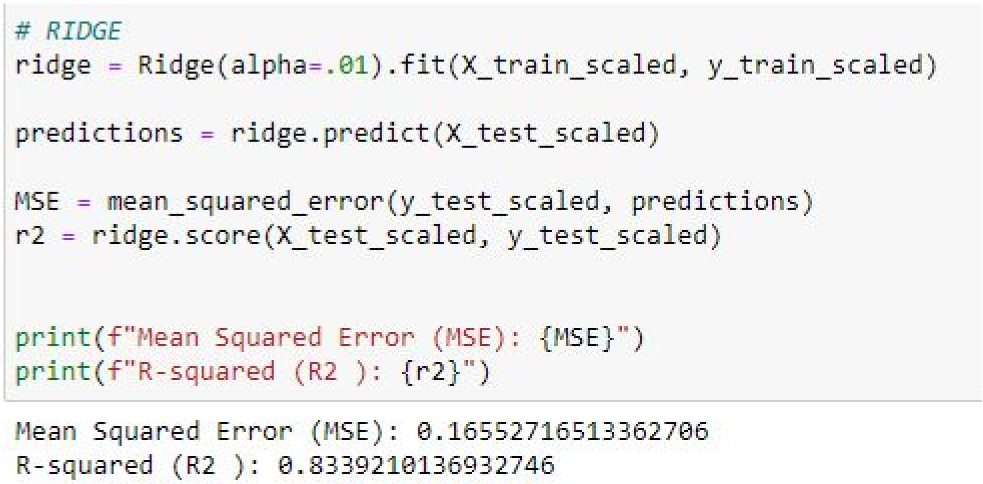
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()

# Fitting our model with all of our features in X
model.fit(X, y)

score = model.score(X, y)
print(f"model R2 Score: {score}")

MSE = mean_squared_error(y_test_scaled, predictions)
r2 = model.score(X_test_scaled, y_test_scaled)


print(f"MSE: {MSE}")
```

```
model R2 Score: 0.9271510950831956
MSE: 0.16619181692184878
```



```
# LASSO
lasso = Lasso(alpha=.01).fit(X_train_scaled, y_train_scaled)

predictions = lasso.predict(X_test_scaled)

MSE = mean_squared_error(y_test_scaled, predictions)
r2 = lasso.score(X_test_scaled, y_test_scaled)


print(f"Mean Squared Error (MSE): {MSE}")
print(f"R-squared (R2 ): {r2}")
```

```
Mean Squared Error (MSE): 0.11041163199477749
R-squared (R2 ): 0.8892202865713271
```



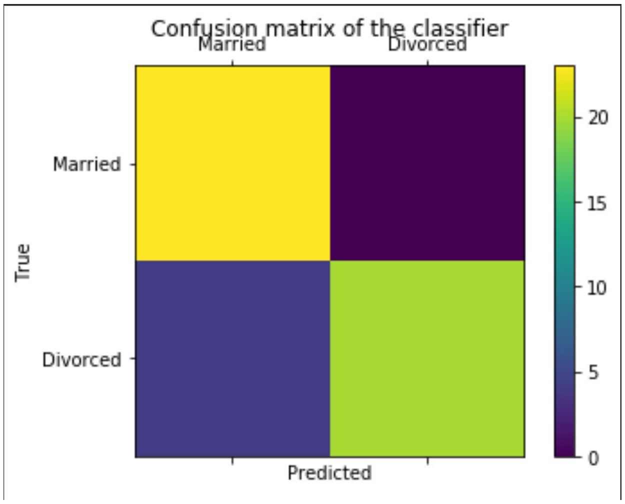
```
# RIDGE
ridge = Ridge(alpha=.01).fit(X_train_scaled, y_train_scaled)

predictions = ridge.predict(X_test_scaled)

MSE = mean_squared_error(y_test_scaled, predictions)
r2 = ridge.score(X_test_scaled, y_test_scaled)


print(f"Mean Squared Error (MSE): {MSE}")
print(f"R-squared (R2 ): {r2}")
```

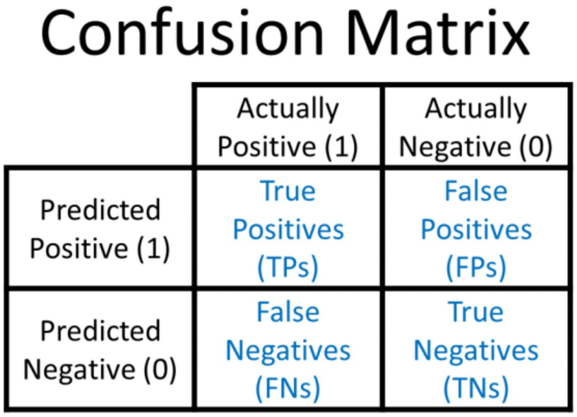
```
Mean Squared Error (MSE): 0.16552716513362706
R-squared (R2 ): 0.8339210136932746
```

Classification

For this next model, a Random Forest was utilized to determine feature importance. This type provides a strong modeling technique that is better than using a single decision tree. A con to a decision tree is overfitting which allows the model to learn from the “noise” in the training data which could ultimately have a negative impact on the model’s performance. Since the Random Forest model aggregates a multitude of decision trees, it limits overfitting as well as error due to bias which yielded more useful results. This analysis identified the feature importance (i.e., which of the 54 questions) carried the most weight in predicting outcomes (married or divorced). The six questions below were identified by the Random Forest model as top features in priority order. Our analysis identified a different set of priority features when compared to the top features identified in the Yontem study. A confusion matrix analysis was conducted to visualize the performance of the Random Forest model and was determined to have an 89% average f1-score from our train test split data.

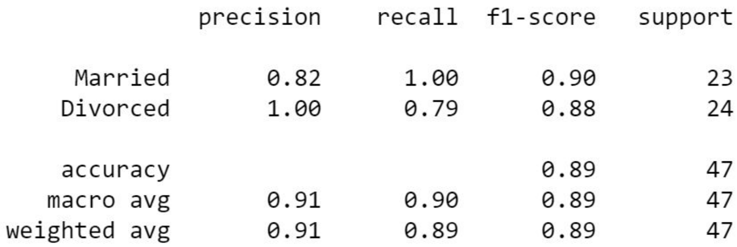
Feature	**Feature Importance Score**
Atr40. We just start to talk before I know what's going on (sudden argument)	0.1300475021600784
Atr38. I hate my spouse's way of starting a discussion	0.0828653199353468
Atr36. I sometimes humiliate my spouse during arguments	0.07302511016590253
Atr41. When I argue with my spouse about something, I get angry quickly	0.07168136837213293
Atr37. My arguments with my spouse are not calm.	0.07168136837213293
Atr17. We share the common views about being happy in our life	0.043870608469662634



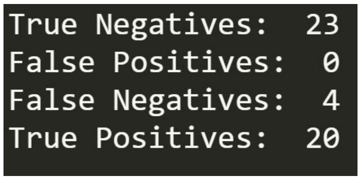


Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)



```
              precision    recall  f1-score   support

     Married       0.82      1.00      0.90        23
    Divorced       1.00      0.79      0.88        24

    accuracy                           0.89        47
   macro avg       0.91      0.90      0.89        47
weighted avg       0.91      0.89      0.89        47
```

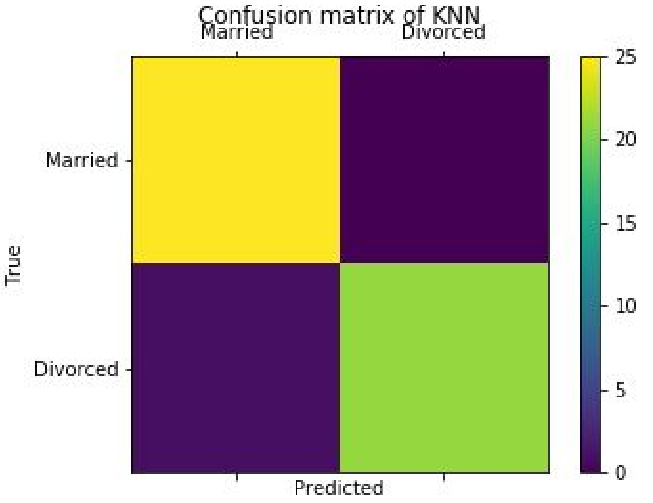


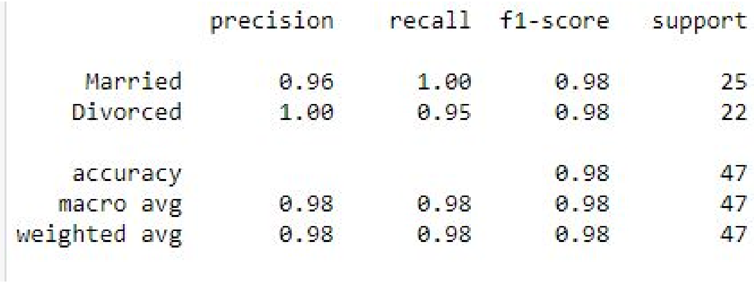
```
True Negatives:   23
False Positives:   0
False Negatives:   4
True Positives:   20
```

Similar to the Yontem study, our team analyzed the accuracy of three machine learning models on the augmented data set. The models used were K-Nearest Neighbor (KNN), LinearSVC, and SGDClassifier. Each model was run with and without feature selection (using the top six feature importance scores identified in the Random Forest model). The same analysis was run on data from the original study only.

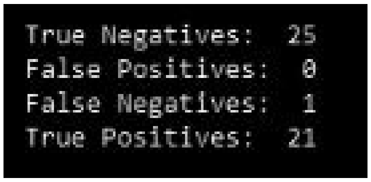
Feature Selection	**Classification**	**Number of Features**	**Accuracy Combined Data**	**Accuracy Turkey Data Only**
None	KNN	54	.979	.953
	Linear SVC		.936	1.0
	SGD Classifier		.86	.96
Random Forest	KNN	6	.936	.953
	Linear SVC		.893	.953
	SGD Classifier		.83	.76

Our analysis found the best performing model to be KNN without feature selection on the combined data set, while the Linear SVC was the top performer on the original data set (Turkey only data). Except for KNN, the models achieved better accuracy overall with the original data set. This suggests that cultural differences may be reflected in responses received from friends and family vs. responses from the original data. The worst performing model was the SGD Classifier with feature selection for both data sets. In all cases, the performance of the models without feature selection was better when compared to the same model run with feature selection. This reflects relatively low feature importance scores from the Random Forest model. Shown below are the classification report and confusion matrix for the best and worst performing models on the combined data set.

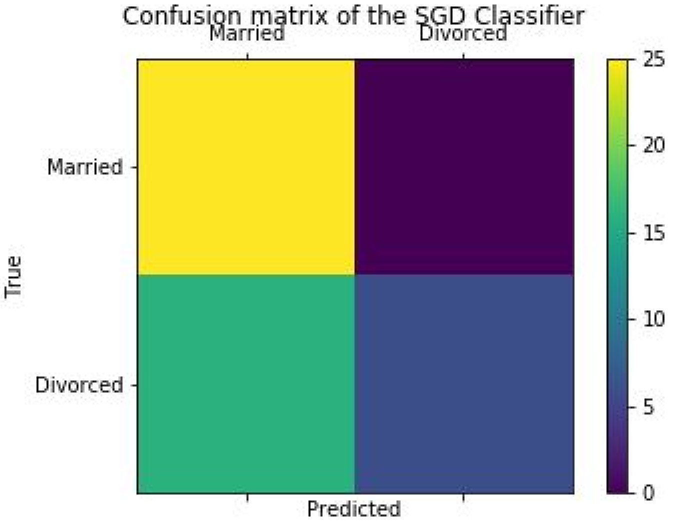


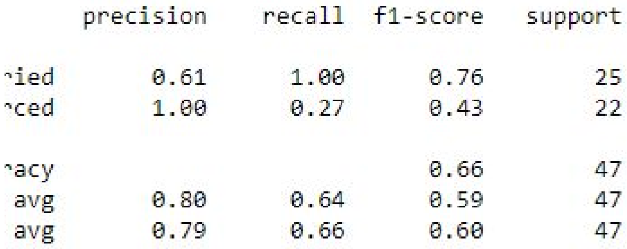


```
              precision    recall  f1-score   support

     Married       0.96      1.00      0.98        25
    Divorced       1.00      0.95      0.98        22

    accuracy                           0.98        47
   macro avg       0.98      0.98      0.98        47
weighted avg       0.98      0.98      0.98        47
```

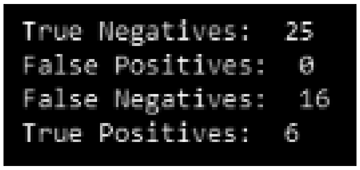


```
True Negatives:  25
False Positives:  0
False Negatives:  1
True Positives:  21
```





```
      precision    recall  f1-score   support

 ied       0.61      1.00      0.76        25
 ced       1.00      0.27      0.43        22

 acy                           0.66        47
 avg       0.80      0.64      0.59        47
 avg       0.79      0.66      0.60        47
```



```
True Negatives:  25
False Positives:  0
False Negatives:  16
True Positives:  6
```

Issues

The first major issue we came across was the translation of survey questions. The original survey was translated into English, and some of the translations didn't make a lot of sense. As a team, we evaluated the results and determined the best English translation before distributing the survey. We created a savvy survey form to share with family and friends via email. The form was easily exported to csv for quick addition to our data. Another issue with the Turkish dataset is the granularity. While we know the data is composed of 84 males and 86 females, we do not know which answers belong to which gender group. Age groups were also not included in the dataset, although the paper mentions them. The paper also mentions data on whether the couples had children and whether their marriage was for love or arranged. We feel our analysis would have been much more extensive and meaningful if we had access to these additional data points.

At first we tried to analyze the Turkish data by breaking it up into 4 categories (just the married data, just the divorced, just the positive survey questions, and just the negative survey questions). However, when doing the linear regressions we found some dataset were resulting in negative R2 values. By isolating certain datasets we were increasing any inherent biases.

Conclusion

Divorce can be predicted fairly accurately based on attributes such as criticism, contempt, stonewalling and defensiveness. These attributes combined with failure of repair attempts are the highest predictors of divorce as noted by Gottman and demonstrated in this analysis.

Resources

1. DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS (Yontem et al) https://dergipark.org.tr/tr/download/article-file/748448
2. The Gottman Method https://www.gottman.com/about/the-gottman-method/
3. Ridge and Lasso Regression https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b
4. Divorce Predictors data set Data Set http://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set#
5. Marriage Success Predictor Adapted Survey (Tammy's form) https://docs.google.com/forms/d/e/1FAIpQLSfnvxPtGGtEI_Nb1UZRecZ6lBX7Z3pQL5zoWu8fyVhLIgug2A/viewform
6. Center For Disease Control National Marriage and Divorce Rates https://www.cdc.gov/nchs/data/dvs/national-marriage-divorce-rates-00-18.pdf