Divorce Prediction: Man vs. Machine

Case Study Narrative Part 1

Objective:

If you have read the book Blink by Malcolm Gladwell, then you might be familiar with Dr. John Gottman who is renowned in the field of psychology for being able to predict, with a high percentage of accuracy, couples that will divorce. By simply meeting with a couple and observing their interaction Dr. Gottman can identify negative communication patterns that reveal feelings of criticism, contempt, defensiveness, and stonewalling. So, is there a way that we can use data science to identify who will divorce with more accuracy than Dr. John Gottman? Dr. Gottman boasts an accuracy rate of over 94%. The goal of my case study is to learn if machine learning can be used to predict who will divorce, with higher accuracy than the human expert.

Dataset:

The dataset being used was retrieved from UCI Machine Learning Repository.

With Dr. Gottman's research in mind a group of other researchers created a dataset using a list of 54 statements that were meant to measure the respondent's feelings toward their spouse or ex-spouse and the relationship. The statements were evaluated during an interview of 170 individual respondents containing both married and divorced participants and measured on a Likert style scale. 0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always. One thing to keep in mind is that the researchers and respondents were from Turkey, so there are some places where the translation does not make grammatical sense in English, but you will be able to understand the gist of the statements.

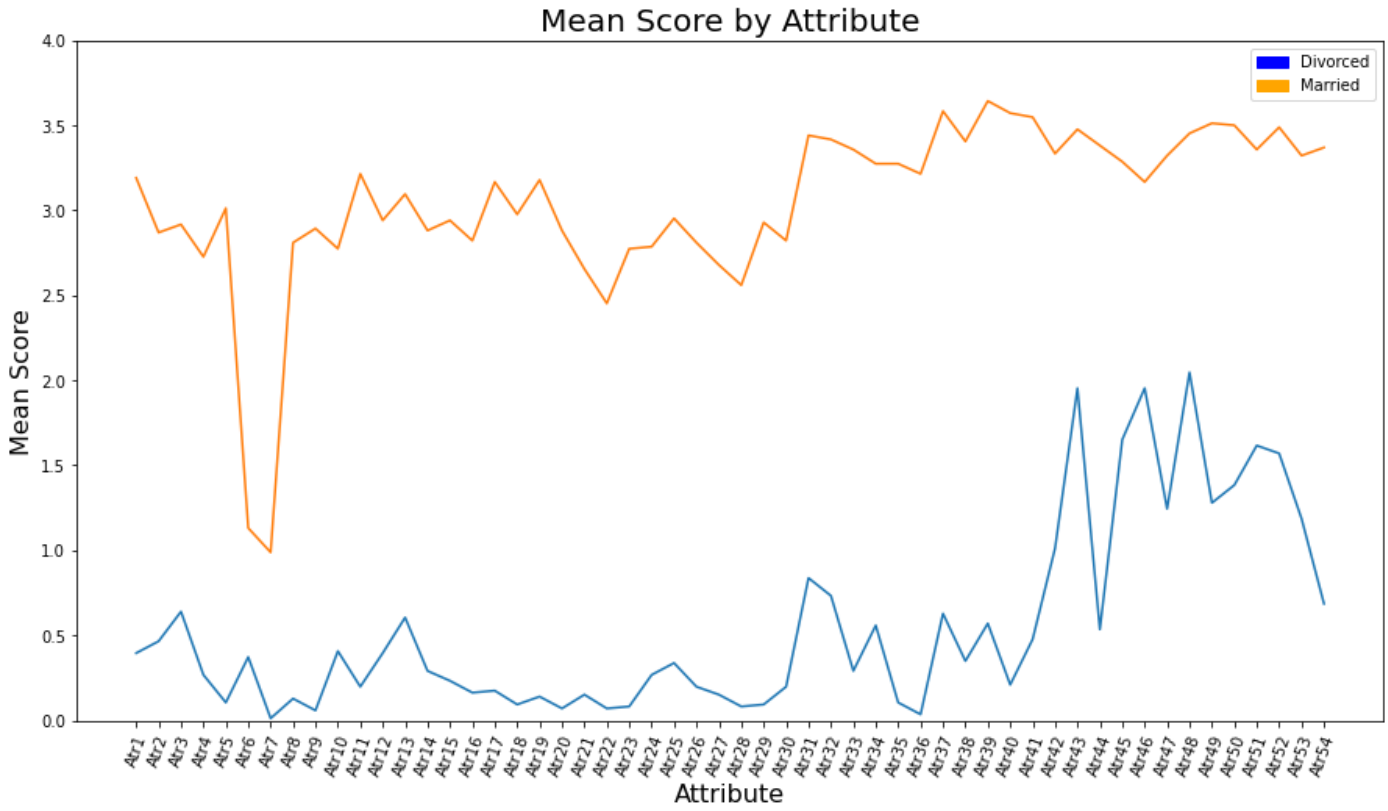
Attributes:

1. If one of us apologizes when our discussion deteriorates, the discussion ends.
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 spouse from the beginning and correct it.
4. When I discuss with my spouse, to contact him will eventually work.
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. Most of our goals are common to my spouse.
11. I think that one day in the future, when I look back, I see that my spouse and I have been in harmony with each other.
12. My spouse and I have similar values in terms of personal freedom.
13. My spouse and I have similar sense of entertainment.
14. Most of our goals for people (children, friends, etc.) are the same.
15. Our dreams with my spouse are similar and harmonious.
16. We're compatible with my spouse about what love should be.

17. We share the same views about being happy in our life with my spouse
18. My spouse and I have similar ideas about how marriage should be
19. My spouse and I have similar ideas about how roles should be in marriage
20. My spouse and I have similar values in trust.
21. I know exactly what my wife likes.
22. I know how my spouse wants to be taken care of when she/he sick.
23. I know my spouse's favorite food.
24. I can tell you what kind of stress my spouse is facing in her/his life.
25. I have knowledge of my spouse's inner world.
26. I know my spouse's basic anxieties.
27. I know what my spouse's current sources of stress are.
28. I know my spouse's hopes and wishes.
29. I know my spouse very well.
30. I know my spouse's friends and their social relationships.
31. I feel aggressive when I argue with my spouse.
32. When discussing with my spouse, I usually use expressions such as ‘you always’ or ‘you never’ .
33. I can use negative statements about my spouse's personality during our discussions.
34. I can use offensive expressions during our discussions.
35. I can insult my spouse during our discussions.
36. I can be humiliating when we discussions.
37. My discussion with my spouse is not calm.
38. I hate my spouse's way of open a subject.
39. Our discussions often occur suddenly.
40. We're just starting a discussion before I know what's going on.
41. When I talk to my spouse about something, my calm suddenly breaks.
42. When I argue with my spouse, ı only go out and I don't say a word.
43. I mostly stay silent to calm the environment a little bit.
44. Sometimes I think it's good for me to leave home for a while.
45. I'd rather stay silent than discuss with my spouse.
46. Even if I'm right in the discussion, I stay silent to hurt my spouse.
47. When I discuss with my spouse, I stay silent because I am afraid of not being able to control my anger.
48. I feel right in our discussions.
49. I have nothing to do with what I've been accused of.
50. I'm not actually the one who's guilty about what I'm accused of.
51. I'm not the one who's wrong about problems at home.
52. I wouldn't hesitate to tell my spouse about her/his inadequacy.
53. When I discuss, I remind my spouse of her/his inadequacy.
54. I'm not afraid to tell my spouse about her/his incompetence.

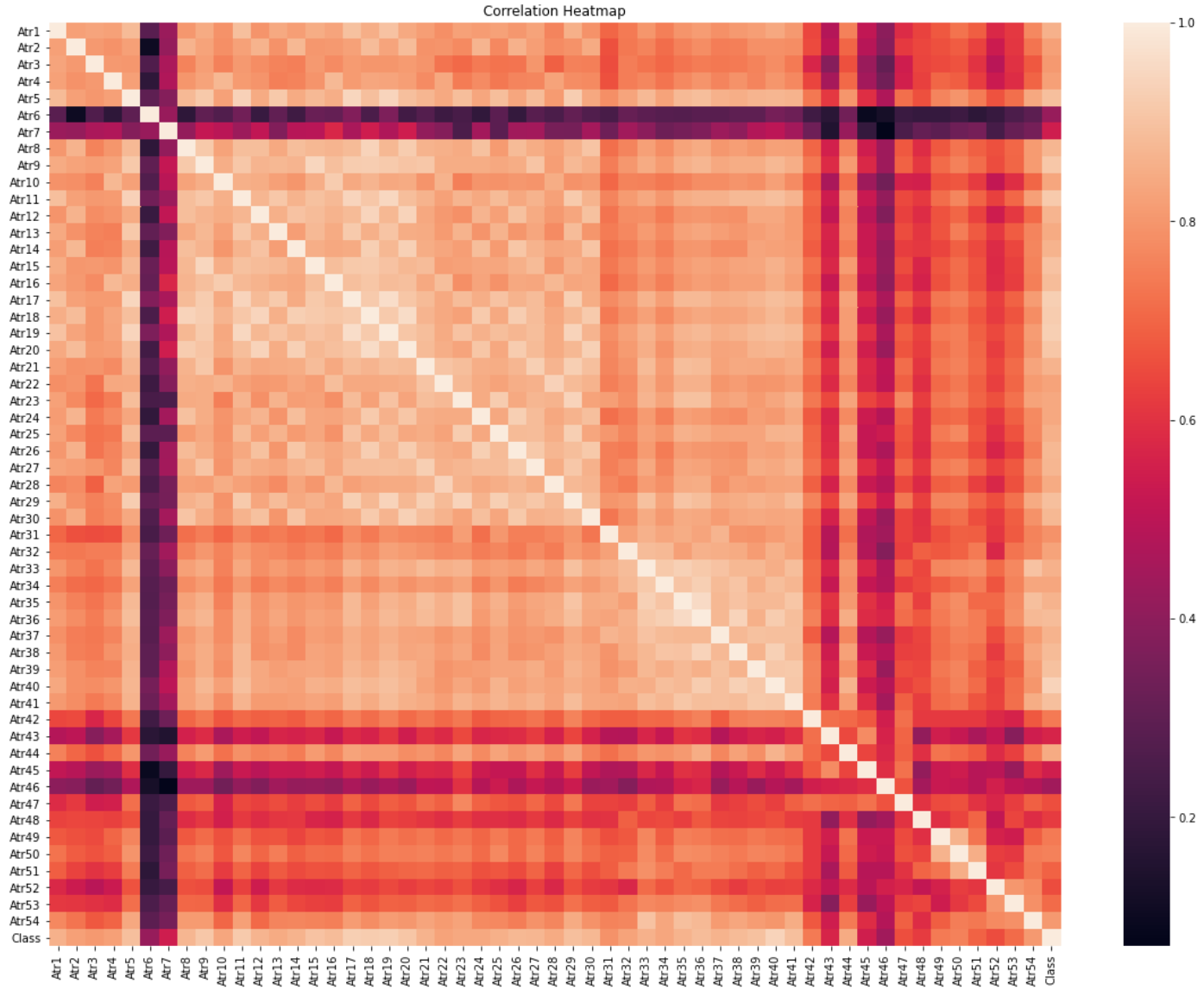
EDA Graphical Analysis:

The first visualization I choose to do was to plot the mean scores for all attributes by Class.



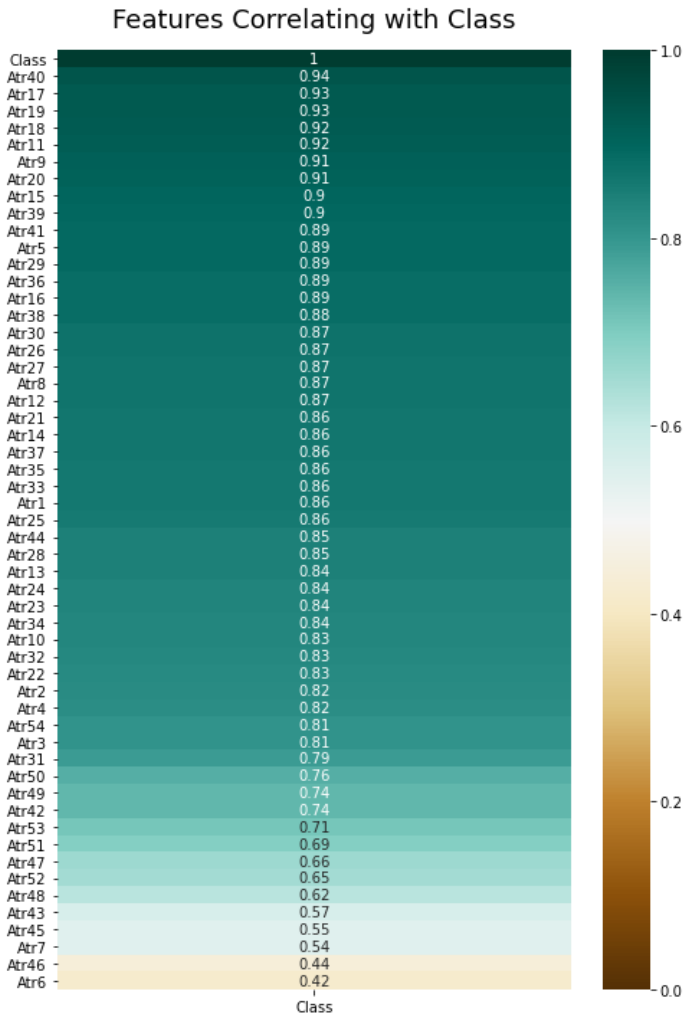
We can see that the general trend is that married persons scored statements in the higher ranges, with the exception of statements 6 and 7. Upon reviewing those statements, they do appear as negative statements, were most of the rest of the statement are more positive. But it's important to note that all statements are not created equal. 0 is not always equal to bad and 4 is not always equal to good. We also see that statements 43, 46, and 48 scored the highest among the divorced persons.

The next visualization I choose was a correlation heatmap.

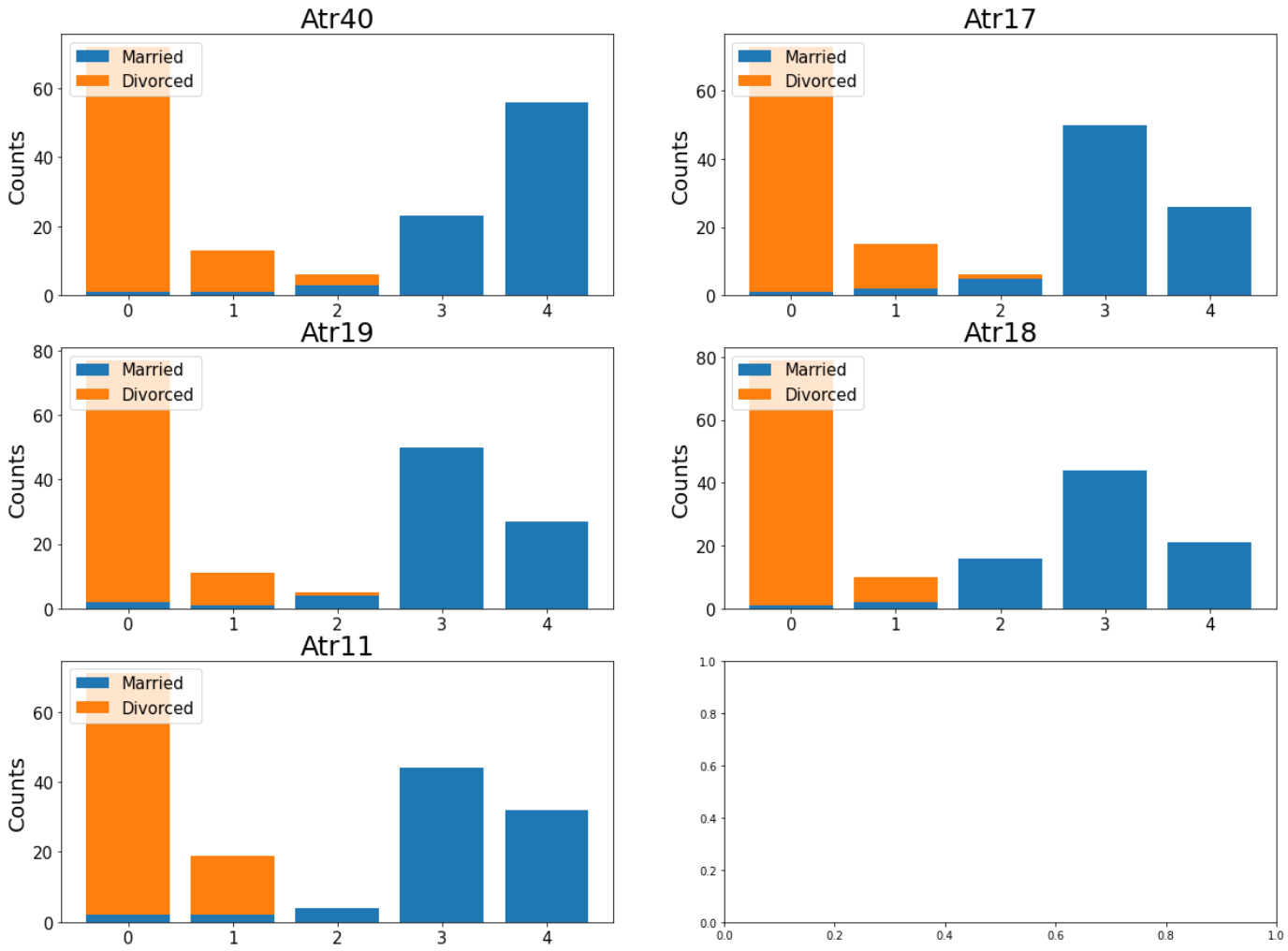


As you can see, due to the sheer number of features this correlation heatmap is not very valuable. When looking at the correlation matrix, I can see that there are some features that have higher correlation scores, like Atr 9 and Atr 15 at ~0.95. This information may come in handy when conducting feature selection, but for exploratory data analysis I was more interested if there were any features that correlated with Class.

That's why I choose a visualization to display features correlated with class next.



This visualization at least gives me a place to start. There are a number of variables that score pretty high, but for the next visualization I'm going to take a look at just the top five features.



At least with the top five features that are correlated with Class, we can see that there is a clear divide and very little overlap between divorced and married counts of answers on statements. This may be why they were identified as correlated.

This information has given me a place to start, some ideas to ponder, but it's really going to come down to feature selection to create the best model to predict who will get divorced.

<u>Case Study Narrative Part 2:</u>

Feature Selection:

My dataset consisted of 55 categorical features. 54 of those features had values that were categorical, but already in ordinal numeric form since they were values given to statements on a five-point scale (0=Never, 1=Seldom, 2=Averagely, 3=Frequently, 4=Always). The target feature 'Class' is a binary categorical feature which is 0 for divorced and 1 for married. Due to the nature of this data, feature extraction is not possible and feature selection becomes very important for reducing the dimensionality.

Since both features and target are categorical, I decided to use the chi-square(X^2) statistic to test the independence of the features from the target variable. Independent variables have little effect on the target variable and can be removed.

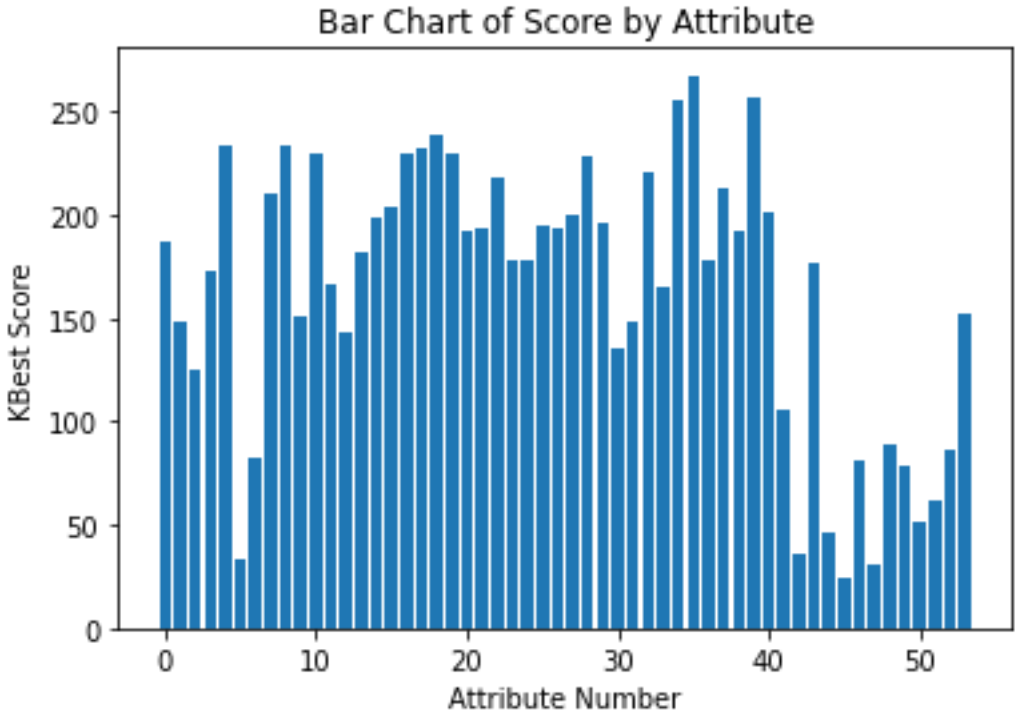
With the chi-square test, you must figure out what k value to designate. Since the features were numerical, I decided to use sklearn's f_classif to calculate the ANOVA F-value statistic for features and target and use that in combination with sklearn's SelectPercentile() to identify a percentile of best features. Designating a percentile of 75% reduced the number of features from 54 to 40.

KBest scores by feature (higher = better):

```
Feature 1: 186.8848063482285
Feature 2: 148.5646677937914
Feature 3: 124.86272609819119
Feature 4: 173.30274965283266
Feature 5: 233.04633621600595
Feature 6: 32.7564330621463
Feature 7: 82.000006591784
Feature 8: 210.3169282777606
Feature 9: 234.06994829064408
Feature 10: 151.00039875373955
Feature 11: 229.05224165673073
Feature 12: 166.52922248101427
Feature 13: 143.62384643779984
Feature 14: 181.5312586841199
Feature 15: 198.39810079593195
Feature 16: 203.4334621646305
Feature 17: 230.17948081326384
Feature 18: 232.75429017830314
Feature 19: 239.13623048619291
Feature 20: 230.19977315757507
Feature 21: 191.86480329222738
Feature 22: 193.4412440711256
Feature 23: 218.19773209671462
Feature 24: 178.25841229623688
Feature 25: 178.36037116550926
Feature 26: 194.78408817337026
Feature 27: 193.88834836260082
Feature 28: 199.8353860505023
Feature 29: 228.67468019985873
Feature 30: 195.78990634891568
Feature 31: 135.61506121484865
Feature 32: 148.6996124031008
Feature 33: 221.26349735409656
Feature 34: 164.94121932738824
Feature 35: 255.47433632804575
Feature 36: 267.489029243182
Feature 37: 177.74183862867125
Feature 38: 213.48922018027105
Feature 39: 192.1763136961303
Feature 40: 256.79032156961074
Feature 41: 200.9566015608411
```

```
Feature 42: 106.1025706017785
Feature 43: 36.41244161972168
Feature 44: 177.31870196986478
Feature 45: 46.174124529081666
Feature 46: 24.498374593648393
Feature 47: 80.75414852047575
Feature 48: 30.63950399003795
Feature 49: 88.92747768070765
Feature 50: 78.33819471817102
Feature 51: 52.002619928819996
Feature 52: 62.11258551453615
Feature 53: 86.45775237397646
Feature 54: 152.052419485406
```

I wanted to visually inspect the KBest scores for all features, so I plotted them on a bar chart for visual evaluation and printed out the array of scores in descending order.



```
array([267.48902924, 256.79032157, 255.47433633, 239.13623049,
       234.06994829, 233.04633622, 232.75429018, 230.19977316,
       230.17948081, 229.05224166, 228.6746802 , 221.26349735,
       218.1977321 , 213.48922018, 210.31692828, 203.43346216,
       200.95660156, 199.83538605, 198.3981008 , 195.78990635,
       194.78408817, 193.88834836, 193.44124407, 192.1763137 ,
       191.86480329, 186.88480635, 181.53125868, 178.36037117,
       178.2584123 , 177.74183863, 177.31870197, 173.30274965,
       166.52922248, 164.94121933, 152.05241949, 151.00039875,
       148.6996124 , 148.56466779, 143.62384644, 135.61506121,
       124.8627261 , 106.1025706 ,  88.92747768,  86.45775237,
        82.00000659,  80.75414852,  78.33819472,  62.11258551,
        52.00261993,  46.17412453,  36.41244162,  32.75643306,
        30.63950399,  24.49837459])
```

Using these two visual methods, I determined that a score of 100 is where a division can be made. Features scoring <100 would be removed due to a lack of significance. In total, there were 12 features (Atr6, Atr7, Atr43, Atr45, Atr46, Atr47, Atr48, Atr49, Atr50, Atr51, Atr52, Atr53) scored below 100 in the chi-square test. These 12 were also 12 of the 13 lowest scoring features in the correlation with target matrix and removing them leaves k=42, which nearly aligned with sklearn's SelectPercentile() at 75% number of 40.

On the positive side. Of the top five highest scoring features on the chi-square test (Atr36, Atr40, Atr35, Atr19, and Atr9), two of them were on the top five highest scoring features in the correlation with target matrix (Atr40 & Atr19).

Case Study Narrative Part 3:

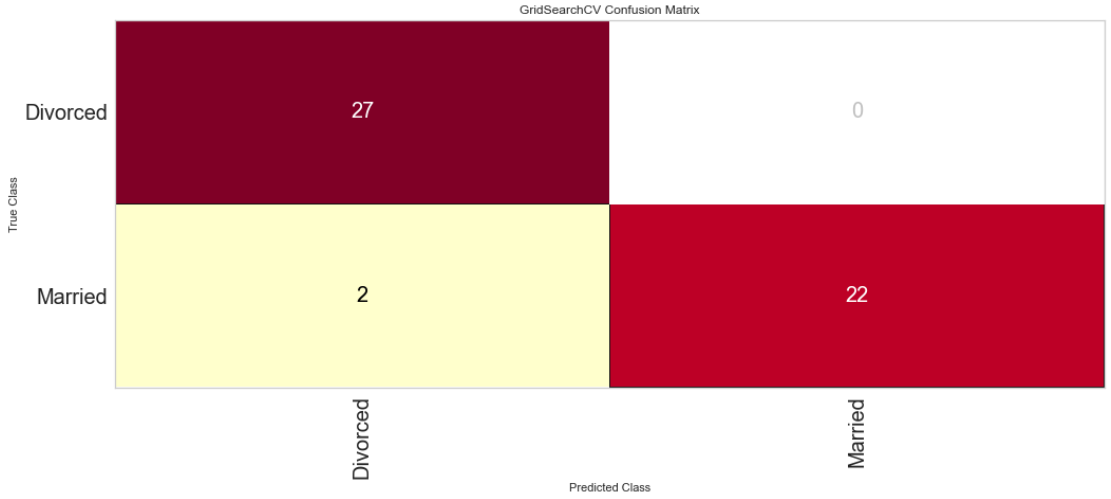
I used sklearn's train_test_split to divide the dataset into training and validation sets. My dataset is quite small, containing only 170 observations total. The split was conducted as below:

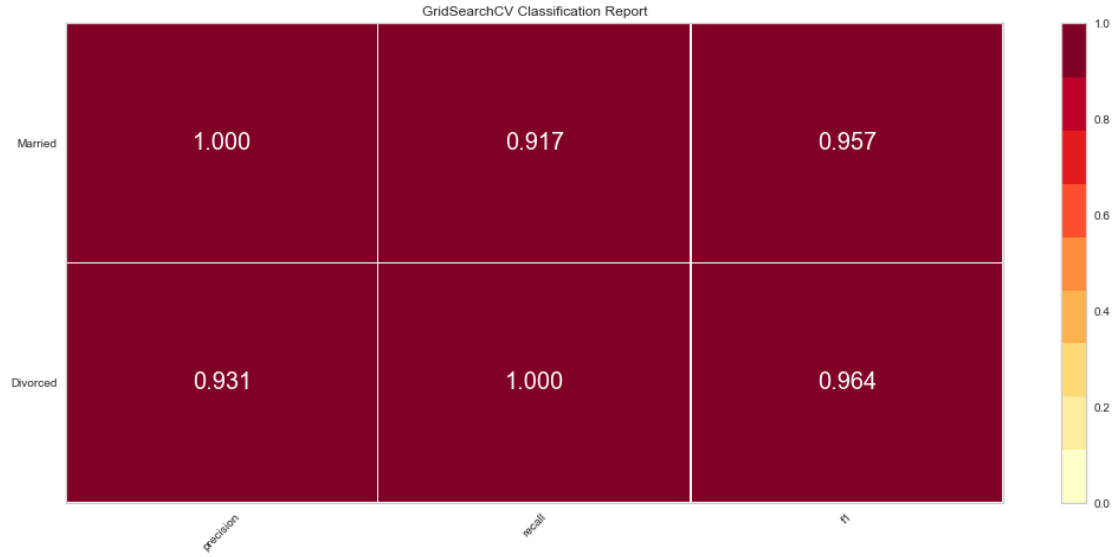
```
No. of samples in training set:  119
No. of samples in validation set: 51


No. of married and divorced in the training set:
1    60
0    59
Name: Class, dtype: int64


No. of married and divorced in the validation set:
0    27
1    24
Name: Class, dtype: int64
```

My prediction problem was one of classification, predicting whether a person would stay married or be divorced and my dataset was small, which lead me to choose a random forest classification model. After training the model, the resulting model was evaluated using results of a confusion matrix and classification report. I then chose to run the same data through a logistic regression model to see if that performed any better. To my surprise the results were identical. In attempt to further better my model, I used GridSearchCV to search over a range of hyperparameters for the random forest classifier to identify the best model. The results were as follows:

```
Best n_estimators: 10
Best max_features: 0.05
```

Using the random forest classification model with the fine-tuned hyperparameters provided the following results:

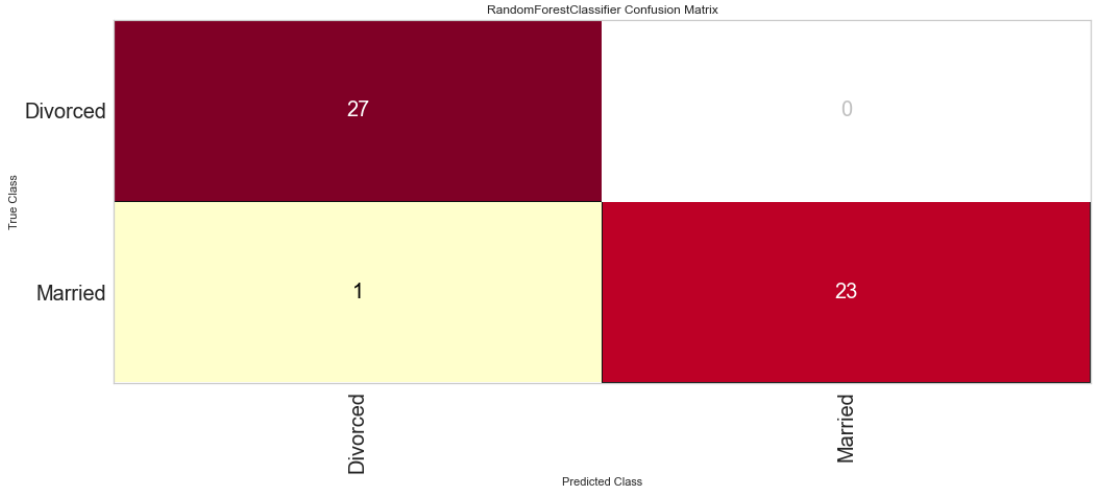


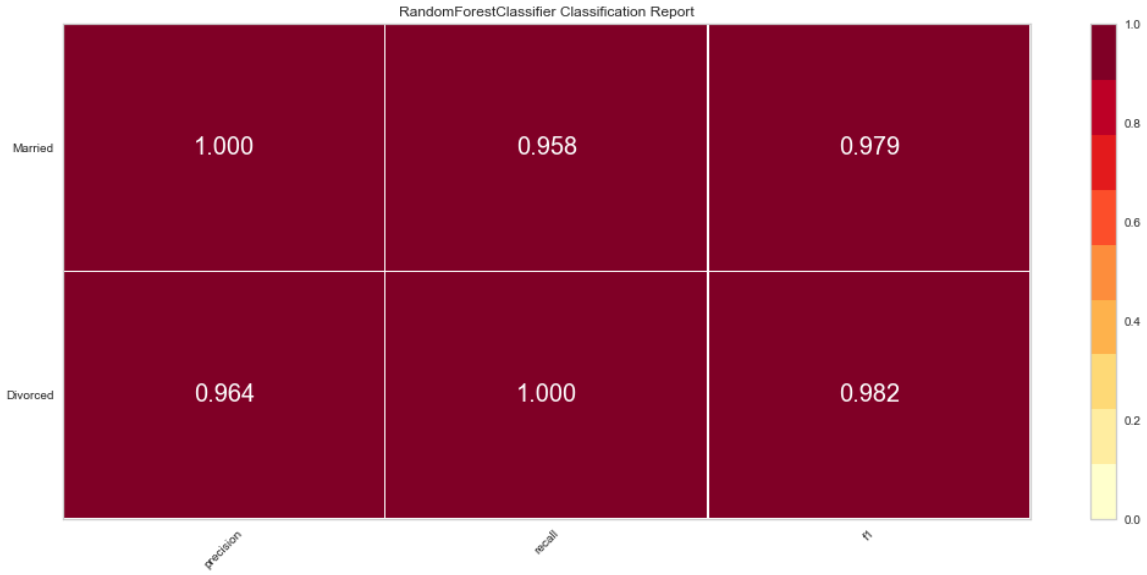


These results are great. The model identified all divorced individuals correctly and only misidentified two married individuals as divorced and can confidently claim 96% prediction accuracy, which does beat Dr. Gottman's 94% accuracy rate. One thing interesting to note is that when using different values for the random_state parameter, this was the worst score achieved. Scores ranged from 96-100% accuracy, with most being 100% correct.

Case Study Final:

After the results from part 3, I decided to go back and experiment with feature selection a bit more. I reduced the number of features used for modeling using the SelectKBest percentile of 25% which provided 14 features (Atr5, Atr9, Atr11, Atr17, Atr18, Atr19, Atr20, Atr23, Atr29, Atr33, Atr35, Atr36, Atr38, Atr40) instead of the 40 features used previously. This change improved the model by two full points from 96% prediction accuracy to 98%.





Seeing that feature selection might be the key to prediction accuracy improvement, I chose to use recursive feature elimination using cross validation (RFECV). RFECV results returned that 5 features (Atr11, Atr17, Atr18, Atr20, Atr40) would provide the best model. I reduced the dataset down to those features and ran the model again. Prediction accuracy remained the same. I once again worked with RandomSearchCV and GridSearchCV in attempt to tune hyperparameters and find better results, but concluded that the default settings for the RandomForestClassifier() worked the best.

Conclusion:

98% prediction accuracy was the best I could achieve. The model made only one mistake on the test data which is a fantastic result considering this is a case study based on social science. With a different situation like, expected rate of failure of a machine part, you might strive for a better accuracy, but measuring human feelings and human actions is not an exact science. The one mistake made was predicting a married individual as divorced. There are many reasons this mistake could have occurred. Perhaps the individual scored negatively on the assessment, but regardless of feelings, was still married. This would not be an odd situation, a person may be unhappy but remain married for religious beliefs, for the kids, or cultural reasons. In fact, the participants of this dataset were from Turkey and 96 of the 170 respondents had an arranged marriage. So, 98% is amazing and to think we could achieve 100% would be naïve, but I do believe that the results of this case study could possibly be improved upon by adding more observations and additional variables like religion, education level, kids or not, arranged marriage or not, etc.

References:

Lisitsa, E. (2013, February 18). The Research: Predicting Divorce from an Oral History Interview. The Gottman Institute. Retrieved from https://www.gottman.com/blog/the-research-predicting-divorce-from-an-oral-history-interview/.

UCI Machine Learning Repository. (2019, July 24). Divorce Predictors data set Data Set. Center for Machine Learning and Intelligent Systems. Retrieved from https://archive.ics.uci.edu/ml/datasets/Divorce+Predictors+data+set.

Yöntem, M., Adem, K., İlhan, T., Kılıçarslan, S. (2019). DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS. Nevşehir Hacı Bektaş Veli University SBE Dergisi, 9 (1), 259-273. Retrieved from https://dergipark.org.tr/tr/download/article-file/748448.