Project 2: Milestone 3 – White Paper

DSC680

Rohan Valder

**Topic**

**Divorce Prediction –**Analysis about negative communication pattern.

**Business Problem**

If you have read the book by Malcolm Gladwell, then you will be familiar with Dr.John Gottman, who is renowned in the field of psychology and predicts with accuracy the couple who would divorce. By meeting the couple, Dr.Gottman can understand the negative communication and the criticism and defensiveness. So is there a way through data science we can identify who will divorce with accuracy better than Dr.John Gottman? Dr.Gottman boasts an accuracy rate of 94%. The goal of my case study is if machine learning can predict divorce with higher accuracy than a human expert.

**Background**

Divorce is so common these days, especially after the pandemic when people worked from home and the couple got to stay the longest together. The divorce rates have gone high. It is interesting to analyze the cause and if possible can be prevented.

**Data Explanation**

The dataset being used is retrieved by the UCI machine learning repository.

<https://archive.ics.uci.edu/ml/machine-learning-databases/00497/>

With Dr.Gottman in mind, researchers responded to 54 statements to find out their views on the spouse, ex-spouse, and the relationship. The results were responses of 170 individuals, which consisted of married and divorced participants.

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.

**Methods**

Use sklearn’s train\_test\_split to divide the dataset into training and validation sets. Since the dataset is small, a random forest classification model will be used. After training the model, the resulting set will be evaluated against the confusion matrix and classification report. Run the same data against the logistic regression model to see if the performance is better.

**Analysis**

The first visualization I plot to do is the mean scores for all attributes by class.

Chart, line chart

Description automatically generated

Figure 1: Mean score for all attributes by class

We can see a trend that the married person scored in the higher range except for statement 6 and 7. Upon reviewing, those two statements seem to be negative whereas most of the other statements are positive. But it is important to know 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 divorced persons.

The next visualization is heatmap correlation.

A picture containing text

Description automatically generated

Figure 2: Heatmap correlation

As you can see, due to the sheer number of attributes, the heatmap correlation is not valuable. When looking at the correlation matrix, some of the features such as attribute 9 and attribute 15 have a score of around 0.95. This will be handy for feature selection, but for exploratory data analysis I was looking for features correlated with class.

That’s why next, I choose a visualization corelated with class.

Graphical user interface

Description automatically generated with low confidence

Figure 3: Correlation with class

This visualization gives me a place to start. There are several variables that score pretty high, but for the next visualization I will select top five features.Chart, bar chart

Description automatically generated

Figure 4: Top 5 features

At least with the top 5 features that are correlated with class, there is a clear difference and hardly any overlap with the married and divorced person based on the answers for the statements. This may be why they were identified as correlated.

This information is a great place to start, with some points to ponder. Feature selection would be the best to identify the people who would be divorced.

**Conclusion**

98% prediction is the best I would achieve. The one mistake could be considering the married individual as divorced. There are many reasons this mistake would have happened. Perhaps the individual scored negative, regardless of the feelings, the person was still married. This would not be an odd situation, as the person would be unhappy but still married for religious beliefs, for the kids, or the cultural reasons. The participants of the dataset were from Turkey and 96 out of the 170 respondents had an arranged marriage. Achieving 98% is amazing, but trying to get 100% is naïve. Accuracy could be better by having more observation on religion, children, arranged marriage, etc

**Assumptions**

As it is related to the human emotion and action, assume the human would not change drastically on the responses to the statements. Also, 170 respondents have answered all 54 questions.

**Limitations**

Dealing with some other prediction, like a failure of a machine part you would get better accuracy, but with human feeling and human action, it is difficult.

**Challenges/Issues**

With the researchers and respondents being Turkish, the grammatical translation in English does not make sense. Sometimes when you translate the sentence word by word, it would have a different meaning or is sometimes meaningless. But you will be able to understand the gist of the statement.

**Future Uses/Additional Applications**

More statements can be included.

**Recommendations**

In case all statements are not answered, the person should be disqualified.

**Implementation Plan**

I used the sklearn’s train\_test\_split to divide the data in training and validation set. My dataset is quite small, having around 170 observations. The split was conducted as below:

Text

Description automatically generated

My prediction was one of classification, predicting to check if a person would remain married or divorced and the dataset was small, hence went with random forest classification model. After training the model, the resultant model was evaluated based on the confusion matrix and classification report. I then chose to run the data against logistic regression to see if the performance is better. To my surprise the results were identical. In attempt of making a better model, I used SearchGridCV for the range of hyperparameters in random forest classifier to have a model better. The results are as follows:

Chart

Description automatically generated Figure 5: Random Forest classifier

These results are great. The model identified all the divorced individuals accurately except for two married individuals claiming 96% prediction accuracy which is better than Dr.Gottman’s 94% accuracy.

**Ethical Assessment**

Since the response is from the current spouse and ex-spouse, the researchers should keep the data confidential about the relationship.

**References**

Nurdialit, Dwi Gustin (2020, September). Netflix Movies and TV Shows — Exploratory Data Analysis (EDA) and Visualization Using Python. - <https://medium.com/analytics-vidhya/netflix-movies-and-tvshows-exploratory-data-analysis-eda-and-visualization-using-python-80753fcfcf7>

Pradhan, Karan (2021, July). Visualizing Netflix Data Using Python! - <https://www.analyticsvidhya.com/blog/2021/07/visualizing-netflix-data-using-python/>

Ramos, Leodanis Pozo (2022, Feb). Python's zipfile: Manipulate Your ZIP Files Efficiently - <https://realpython.com/python-zipfile/>

Ebrahim, Mokhtar (2022, Feb). Python time module (Simple Examples) - <https://likegeeks.com/python-time-module/>

**10 Questions**

1. What was the reason for selecting this topic?

- Divorce is a familiar topic, especially after the pandemic when the couple got to stay indoors for almost two years. I was interested to find out how the divorce predictions are possible.

1. Any reason to select the 54 statements?

– The researchers came up with these questions, which would give a person's feelings about their spouse or ex-spouse. It also explains their relationship.

1. What was the reason for the selected data set?

- During an interview, 170 respondents were evaluated, which included both married and divorced individuals having a scale of 0=Never, 1=Seldom, 2= Averagely, 3=Frequently and 4=Always

1. Did you have to make any adjustments to the dataset for the insights?

- No adjustments were required to the dataset. For plotting purposes, a transpose was done on the data frame.

1. What are you going to do with the insights you found?

Going forward, I’ll keep note of the statements spoken. I may be able to make predictions in my mind with my close ones.

1. Are there any areas of the project you would like to improve?

Yes, probably religion and type of marriage such as arranged or love marriage could be considered.

1. Who is going to benefit from this project?
2. How do we evaluate if the project has succeeded?
3. Why did you not check if it was an arranged marriage or a love marriage?
4. Did you have any of respondents fill the data partially?

**Appendix**

Figure 1. Mean score for all attributes by class.

Figure 2. Heatmap correlation.

Figure 3. Correlation with class.

Figure 4. Top 5 features.

Figure 5. Random forest classifier.