Choose Your Own Project - Divorce Predictors

Luke Straughan

22/04/2020

Introduction

This report is part of an assignment for the Capstone course in Harvardx's Data Science programme. The assignment is to choose one's own project. It was recommended that one use an existing dataset. This project used the Divorce dataset by Noah Weber on kaggle.com. The goal of this project was to create a machine learning model that will predict whether a couple would get divorced, based on the attributes available in the dataset. The objective is to have at least one model that has at least 90% accuracy. The dataset itself will be expanded upon in the Data Wrangling and Data Analysis sections below.

In the Data Wrangling section, the data is downloaded and cleaned/formatted in order for easier interpretation of the data. Before any work is done on it, the data set is separated into a set that can be used for analsysis and training and the validation set that will only be used for the final results. 90% of the data is assigned to the 'dat' set and 10% to the 'validation' set. Then, in the Data Analysis section, the data will be separated again so that training and testing can be done without involving the validation set. There is not a large amount of data in the set to begin with so, to avoid overfitting, 60% of the 'dat' data is assigned to the training set and 40% to the testing set. A prediction algorithm is then used to train for initial accuracies. The accuracies are then tested and reported to find the attributes that have the greatest impact on whether a couple will get divorced. The Modelling section will then follow. This section will continue to use the training and testing sets that were created in the Data Analysis section. An ensemble model is utilised here to combine many of the various machine learning models to improve the results. Once the testing is successful, the 'dat' set that contains 90% of the original data will be used for the final training. The Results section will then be used to test and report on this final training using the validation set. Finally, the Conclusion section will reflect on the results and suggest ways on which the project could be improved.

Data Wrangling

Naturally, the first step is to access the data. The data set is only approximately 18KB, so downloading it will be the simplest and most efficient option. The code below will download the file directly into the working directory. The first line checks to make sure that the file is in in fact in the working directory.

```
download.file("https://raw.githubusercontent.com/LDStraughan/Harvard_DS_Capstone_CYO/master/divorce.csv
file.exists("divorce.csv")
```

Then, the data must be inspected. One can see below that the raw data is difficult to interpret, however, it seems to be relatively clean already. Each column is conveniently separated by a semi-colon.

```
## [1] "Atr1;Atr2;Atr3;Atr4;Atr5;Atr6;Atr7;Atr8;Atr9;Atr10;Atr11;Atr12;Atr13;Atr14;Atr15;Atr16;Atr17;At
## [2] ""
## [3] "2;2;4;1;0;0;0;0;0;0;1;0;1;1;0;1;0;0;0;1;0;0;0;0;0;0;0;0;0;1;1;2;1;2;0;1;2;1;3;3;2;1;1;2;3;2;1;3
```

Therefore, separating the raw data into columns by the semi-colon will fix all issues. The data is stored into the object 'divorce' to be easily accessed.

```
divorce <- read_delim("divorce.csv", ";")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
head(divorce)
```

```
## # A tibble: 6 x 55
##    Atr1  Atr2  Atr3  Atr4  Atr5  Atr6  Atr7  Atr8  Atr9 Atr10 Atr11 Atr12 Atr13
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     2     2     4     1     0     0     0     0     0     0     1     0     1
## 2     4     4     4     4     4     0     0     4     4     4     4     3     4
## 3     2     2     2     2     1     3     2     1     1     2     3     4     2
## 4     3     2     3     2     3     3     3     3     3     3     4     3     3
## 5     2     2     1     1     1     1     0     0     0     0     0     1     0
## 6     0     0     1     0     0     2     0     0     0     1     0     2     1
## # ... with 42 more variables: Atr14 <dbl>, Atr15 <dbl>, Atr16 <dbl>,
## #   Atr17 <dbl>, Atr18 <dbl>, Atr19 <dbl>, Atr20 <dbl>, Atr21 <dbl>,
## #   Atr22 <dbl>, Atr23 <dbl>, Atr24 <dbl>, Atr25 <dbl>, Atr26 <dbl>,
## #   Atr27 <dbl>, Atr28 <dbl>, Atr29 <dbl>, Atr30 <dbl>, Atr31 <dbl>,
## #   Atr32 <dbl>, Atr33 <dbl>, Atr34 <dbl>, Atr35 <dbl>, Atr36 <dbl>,
## #   Atr37 <dbl>, Atr38 <dbl>, Atr39 <dbl>, Atr40 <dbl>, Atr41 <dbl>,
## #   Atr42 <dbl>, Atr43 <dbl>, Atr44 <dbl>, Atr45 <dbl>, Atr46 <dbl>,
## #   Atr47 <dbl>, Atr48 <dbl>, Atr49 <dbl>, Atr50 <dbl>, Atr51 <dbl>,
## #   Atr52 <dbl>, Atr53 <dbl>, Atr54 <dbl>, Class <dbl>
```

```
##   Atr1 Atr2 Atr3 Atr4 Atr5 Atr6 Atr7 Atr8 Atr9 Atr10 Atr11 Atr12 Atr13 Atr14
## 1    2    2    4    1    0    0    0    0    0     0     1     0     1     1
```

```
## 2    4    4    4    4    4    0    0    4    4     4     4     3     4     0
## 3    2    2    2    2    1    3    2    1    1     2     3     4     2     3
## 4    3    2    3    2    3    3    3    3    3     3     4     3     3     4
## 5    2    2    1    1    1    1    0    0    0     0     0     1     0     1
## 6    0    0    1    0    0    2    0    0    0     1     0     2     1     0
##   Atr15 Atr16 Atr17 Atr18 Atr19 Atr20 Atr21 Atr22 Atr23 Atr24 Atr25 Atr26 Atr27
## 1     0     1     0     0     0     1     0     0     0     0     0     0     0
## 2     4     4     4     4     3     2     1     1     0     2     2     1     2
## 3     3     3     3     3     3     2     1     0     1     2     2     2     2
## 4     3     3     3     3     3     4     1     1     1     1     2     1     1
## 5     1     1     1     1     2     1     1     0     0     0     0     2     1
## 6     2     0     2     1     0     1     0     0     0     0     2     2     0
##   Atr28 Atr29 Atr30 Atr31 Atr32 Atr33 Atr34 Atr35 Atr36 Atr37 Atr38 Atr39 Atr40
## 1     0     0     1     1     2     1     2     0     1     2     1     3     3
## 2     0     1     1     0     4     2     3     0     2     3     4     2     4
## 3     2     3     2     3     3     1     1     1     1     2     1     3     3
## 4     1     1     3     2     3     2     2     1     1     3     3     4     4
## 5     2     1     1     1     1     1     1     0     0     0     0     2     1
## 6     0     0     0     4     1     1     1     1     1     1     2     0     2
##   Atr41 Atr42 Atr43 Atr44 Atr45 Atr46 Atr47 Atr48 Atr49 Atr50 Atr51 Atr52 Atr53
## 1     2     1     1     2     3     2     1     3     3     3     2     3     2
## 2     2     2     3     4     2     2     2     3     4     4     4     4     2
## 3     3     3     2     3     2     3     2     3     1     1     1     2     2
## 4     2     2     3     2     3     2     2     3     3     3     3     2     2
## 5     0     2     3     0     2     2     1     2     3     2     2     2     1
## 6     2     1     2     3     0     2     2     1     2     1     1     1     2
##   Atr54 Class
## 1     1     1
## 2     2     1
## 3     2     1
## 4     2     1
## 5     0     1
## 6     0     1
```

The attributes depicted by the columns are as follows:

```
## [[1]]
##  [1] "1. If one of us apologizes when our discussion deteriorates, the discussion ends."
##  [2] "2. I know we can ignore our differences, even if things get hard sometimes."
##  [3] "3. When we need it, we can take our discussions with my spouse from the beginning and correct i
##  [4] "4. When I discuss with my spouse, to contact him will eventually work."
##  [5] "5. The time I spent with my wife is special for us."
##  [6] "6. We don't have time at home as partners."
##  [7] "7. We are like two strangers who share the same environment at home rather than family."
##  [8] "8. I enjoy our holidays with my wife."
##  [9] "9. I enjoy traveling with my wife."
## [10] "10. Most of our goals are common to my spouse."
## [11] "11. I think that one day in the future, when I look back, I see that my spouse and I have been
## [12] "12. My spouse and I have similar values in terms of personal freedom."
## [13] "13. My spouse and I have similar sense of entertainment."
## [14] "14. Most of our goals for people (children, friends, etc.) are the same."
## [15] "15. Our dreams with my spouse are similar and harmonious."
## [16] "16. We're compatible with my spouse about what love should be."
## [17] "17. We share the same views about being happy in our life with my spouse"
```

```
## [18] "18. My spouse and I have similar ideas about how marriage should be"
## [19] "19. My spouse and I have similar ideas about how roles should be in marriage"
## [20] "20. My spouse and I have similar values in trust."
## [21] "21. I know exactly what my wife likes."
## [22] "22. I know how my spouse wants to be taken care of when she/he sick."
## [23] "23. I know my spouse's favorite food."
## [24] "24. I can tell you what kind of stress my spouse is facing in her/his life."
## [25] "25. I have knowledge of my spouse's inner world."
## [26] "26. I know my spouse's basic anxieties."
## [27] "27. I know what my spouse's current sources of stress are."
## [28] "28. I know my spouse's hopes and wishes."
## [29] "29. I know my spouse very well."
## [30] "30. I know my spouse's friends and their social relationships."
## [31] "31. I feel aggressive when I argue with my spouse."
## [32] "32. When discussing with my spouse, I usually use expressions such as ‘you always’ or ‘you neve
## [33] "33. I can use negative statements about my spouse's personality during our discussions."
## [34] "34. I can use offensive expressions during our discussions."
## [35] "35. I can insult my spouse during our discussions."
## [36] "36. I can be humiliating when we discussions."
## [37] "37. My discussion with my spouse is not calm."
## [38] "38. I hate my spouse's way of open a subject."
## [39] "39. Our discussions often occur suddenly."
## [40] "40. We're just starting a discussion before I know what's going on."
## [41] "41. When I talk to my spouse about something, my calm suddenly breaks."
## [42] "42. When I argue with my spouse, i only go out and I don't say a word."
## [43] "43. I mostly stay silent to calm the environment a little bit."
## [44] "44. Sometimes I think it's good for me to leave home for a while."
## [45] "45. I'd rather stay silent than discuss with my spouse."
## [46] "46. Even if I'm right in the discussion, I stay silent to hurt my spouse."
## [47] "47. When I discuss with my spouse, I stay silent because I am afraid of not being able to cont
## [48] "48. I feel right in our discussions."
## [49] "49. I have nothing to do with what I've been accused of."
## [50] "50. I'm not actually the one who's guilty about what I'm accused of."
## [51] "51. I'm not the one who's wrong about problems at home."
## [52] "52. I wouldn't hesitate to tell my spouse about her/his inadequacy."
## [53] "53. When I discuss, I remind my spouse of her/his inadequacy."
## [54] "54. I'm not afraid to tell my spouse about her/his incompetence."
```

The ‘divorce’ data set is then, as aformentioned, separated into the ‘dat’ and ‘validation’ sets. The validation set should not be used for anything except the final reporting of the results. Note that this will only be succesful if the data is a data frame.

```
y <- divorce$Class
set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y, times=1, p=0.9, list=FALSE)
dat <- divorce[test_index,]
validation <- divorce[-test_index,]
```

Data Analysis

Now, one is ready to analyse the data. As can be seen below, the original data set contains 170 rows and 55 columns. The rows exclude the headings which contain the attribute numbers. Therefore, the data was collected from 170 couples. The 55 columns contain the 54 attributes while the last column, column 55, is the "Class" column which depicts whether the couple is divorced or not. A "1" depicts a divorce while a "0" depicts that the couple remains together.

```
dim(divorce)
```

```
## [1] 170  55
```

The 'dat' set that will be used for the primary training and testing in the project contains data from 153 couples while the 55 columns remain intact. This is of absolute importance as the predictors (attributes) and results (class) cannot change - only the amount of data for predictors & results.

```
dim(dat)
```

```
## [1] 153  55
```

As aformentioned, one still has to train and test the data. Therefore, the 'dat' set is separated further into the 'train' and 'test' sets that will be used for training and testing, respectively. 60% of the 'dat' data is assigned to the training set and 40% to the testing set.

```
y <- dat$Class
set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y, times=1, p=0.6, list=FALSE)
train <- dat[test_index,]
test <- dat[-test_index,]
```

The training set now contains data from 92 couples. This is 60% of 90% of the original data, meaning it is approximately 54% of the original data.

```
dim(train)
```

```
## [1] 92 55
```

Accuracies

One can finally begin to train and test the data without worry of corrupting the results with the validation set. The code below creates a function with which will calculate the impact of the attributes (x) on whether Class is a 1 or a 0 (y). Predictions are made, which are then used to find the attributes with the maximum accuracy. In the function, Class was not eliminated. So, naturally, Class will be most accurate. Therefore, the next most accurate is used to find the optimum cutoff. This cutoff is the number reported by the attribute that best dictates whether the Class is a 1 or a 0. All of this allows one to find how accurate these attributes with said cutoff are in predicting divorce.

```
func <- function(x){
  rangedValues <- seq(range(x)[1],range(x)[2],by=1)
  sapply(rangedValues,function(i){
    y_hat <- ifelse(x>i,'1','0')
    mean(y_hat==train$Class)
  })
}
predictions <- apply(train,2,func)
acc <- sapply(predictions,max)
max_acc <- order(acc, decreasing = TRUE)[2]
predictions <- func(train[,max_acc])
rangedValues <- seq(range(train[,max_acc])[1],range(train[,max_acc])[2],by=1)
cutoffs <-rangedValues[which(predictions==max(predictions))]
```

When tested, the average accuracy so far is:

```
## [1] 1
```

This is incredibly positive. At the very least, this shows that there is an absolute correlation between the answers given in the study and whether a couple is divorced, At most, it shows this test alone can be used to predict a couple's divorce.

A similar function as the one created above is used on the test set to produce the following accuracies:

```
##      Atr1      Atr2      Atr3      Atr4      Atr5      Atr6      Atr7      Atr8
## 0.9016393 0.9508197 0.9016393 0.9344262 0.9508197 0.6721311 0.8196721 0.9344262
##      Atr9     Atr10     Atr11     Atr12     Atr13     Atr14     Atr15     Atr16
## 0.9508197 0.9016393 0.9508197 0.9344262 0.9016393 0.9344262 0.9508197 0.9508197
##     Atr17     Atr18     Atr19     Atr20     Atr21     Atr22     Atr23     Atr24
## 0.9672131 0.9508197 0.9508197 0.9508197 0.9180328 0.9016393 0.9016393 0.9016393
##     Atr25     Atr26     Atr27     Atr28     Atr29     Atr30     Atr31     Atr32
## 0.9344262 0.9672131 0.9180328 0.9344262 0.9180328 0.9180328 0.9344262 0.9016393
##     Atr33     Atr34     Atr35     Atr36     Atr37     Atr38     Atr39     Atr40
## 0.9016393 0.9180328 0.9180328 0.9508197 0.9344262 0.9180328 0.9672131 0.9672131
##     Atr41     Atr42     Atr43     Atr44     Atr45     Atr46     Atr47     Atr48
## 0.9180328 0.8524590 0.8032787 0.9508197 0.7213115 0.7213115 0.7868852 0.8524590
##     Atr49     Atr50     Atr51     Atr52     Atr53     Atr54
## 0.8360656 0.9016393 0.8524590 0.7868852 0.8688525 0.8852459
```

The average accuracy is:

```
## [1] 0.900425
```

This is incredibly positive as it already achieves the objective for this assignment - however, just barely. The ensemble model should find at least one model that has a higher accuracy.

Top 10 Accuracies

```
##     Atr16     Atr18     Atr19     Atr20     Atr36     Atr44     Atr17     Atr26
## 0.9508197 0.9508197 0.9508197 0.9508197 0.9508197 0.9508197 0.9672131 0.9672131
##     Atr39     Atr40
## 0.9672131 0.9672131
```

From this we can tell that the attributes that contribute most to a couple's divorce are:

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 my spouse's basic anxieties.
22. I can be humiliating when we discussions.
23. Our discussions often occur suddenly.
24. We're just starting a discussion before I know what's going on.
25. Sometimes I think it's good for me to leave home for a while.

Highest Accuracy

```
##     Atr17     Atr26     Atr39     Atr40
## 0.9672131 0.9672131 0.9672131 0.9672131
```

The accuracies of attributes 17, 26, 39 and 40 are all 0.9672131.

17. We share the same views about being happy in our life with my spouse.
18. I know my spouse's basic anxieties.
19. Our discussions often occur suddenly.
20. We're just starting a discussion before I know what's going on.

It is important to note, however, that not all of the couples are being used here, as well as the fact that different seeds might yield slightly different results.

Table

The attributes and their accuracy values can be seen in the table below. The table is arranged from most to least accurate.

```
## # A tibble: 54 x 2
##    Attributes                                                      value
##    <chr>                                                           <dbl>
##  1 17. We share the same views about being happy in our life with my spou~ 0.967
##  2 26. I know my spouse's basic anxieties.                         0.967
##  3 39. Our discussions often occur suddenly.                       0.967
##  4 40. We're just starting a discussion before I know what's going on.     0.967
##  5 2. I know we can ignore our differences, even if things get hard somet~ 0.951
##  6 5. The time I spent with my wife is special for us.             0.951
##  7 9. I enjoy traveling with my wife.                              0.951
##  8 11. I think that one day in the future, when I look back, I see that m~ 0.951
```

```
##  9 15. Our dreams with my spouse are similar and harmonious.               0.951
## 10 16. We're compatible with my spouse about what love should be.          0.951
## 11 18. My spouse and I have similar ideas about how marriage should be     0.951
## 12 19. My spouse and I have similar ideas about how roles should be in ma~ 0.951
## 13 20. My spouse and I have similar values in trust.                       0.951
## 14 36. I can be humiliating when we discussions.                           0.951
## 15 44. Sometimes I think it's good for me to leave home for a while.       0.951
## 16 4. When I discuss with my spouse, to contact him will eventually work.  0.934
## 17 8. I enjoy our holidays with my wife.                                   0.934
## 18 12. My spouse and I have similar values in terms of personal freedom.   0.934
## 19 14. Most of our goals for people (children, friends, etc.) are the sam~ 0.934
## 20 25. I have knowledge of my spouse's inner world.                        0.934
## 21 28. I know my spouse's hopes and wishes.                                0.934
## 22 31. I feel aggressive when I argue with my spouse.                      0.934
## 23 37. My discussion with my spouse is not calm.                           0.934
## 24 21. I know exactly what my wife likes.                                  0.918
## 25 27. I know what my spouse's current sources of stress are.              0.918
## 26 29. I know my spouse very well.                                         0.918
## 27 30. I know my spouse's friends and their social relationships.          0.918
## 28 34. I can use offensive expressions during our discussions.             0.918
## 29 35. I can insult my spouse during our discussions.                      0.918
## 30 38. I hate my spouse's way of open a subject.                           0.918
## 31 41. When I talk to my spouse about something, my calm suddenly breaks.  0.918
## 32 1. If one of us apologizes when our discussion deteriorates, the discu~ 0.902
## 33 3. When we need it, we can take our discussions with my spouse from th~ 0.902
## 34 10. Most of our goals are common to my spouse.                          0.902
## 35 13. My spouse and I have similar sense of entertainment.                0.902
## 36 22. I know how my spouse wants to be taken care of when she/he sick.    0.902
## 37 23. I know my spouse's favorite food.                                   0.902
## 38 24. I can tell you what kind of stress my spouse is facing in her/his ~ 0.902
## 39 32. When discussing with my spouse, I usually use expressions such as ~ 0.902
## 40 33. I can use negative statements about my spouse's personality during~ 0.902
## 41 50. I'm not actually the one who's guilty about what I'm accused of.    0.902
## 42 54. I'm not afraid to tell my spouse about her/his incompetence.        0.885
## 43 53. When I discuss, I remind my spouse of her/his inadequacy.           0.869
## 44 42. When I argue with my spouse, i only go out and I don't say a word.  0.852
## 45 48. I feel right in our discussions.                                    0.852
## 46 51. I'm not the one who's wrong about problems at home.                 0.852
## 47 49. I have nothing to do with what I've been accused of.                0.836
## 48 7. We are like two strangers who share the same environment at home ra~ 0.820
## 49 43. I mostly stay silent to calm the environment a little bit.          0.803
## 50 47. When I discuss with my spouse, I stay silent because I am afraid o~ 0.787
## 51 52. I wouldn't hesitate to tell my spouse about her/his inadequacy.     0.787
## 52 45. I'd rather stay silent than discuss with my spouse.                 0.721
## 53 46. Even if I'm right in the discussion, I stay silent to hurt my spou~ 0.721
## 54 6. We don't have time at home as partners.                              0.672
```

Modelling

Finally, the modelling phase can begin. It is important to note that in order for this ensemble model to succeed, the Class column must be a factor. As aforementioned, the model being used is an ensemble model that incorporates a number of other models. These models (that can be seen below) include: Generalised Linear Models (glm); Linear Discriminant Analysis (lds); Naïve Bayes classification (naive_bayes); Support-Vector Machines (svmLinear); K-Nearest Neighbour (knn); Locally Estimated Scatterplot Smoothing (gamLoess); Multinomial Logistic Regression (multinom); Random Forest (rf); and Adaptive Boosting (adaboost) to improve the performance.

```
models <- c("glm", "lda", "naive_bayes", "svmLinear", "knn", "gamLoess", "multinom", "rf", "adaboost")
```

The following code trains the ensemble model using the training data. It may take some time to run.

```
fits <- lapply(models, function(model){
  print(model)
  train(Class ~ ., method = model, data = train)
})
```

Test

To test the ensemble model, one creates a prediction function that applies it to the testing data.

```
pred <- sapply(fits, function(object)
  predict(object, newdata = test))
```

The average accuracy is:

```
##         glm         lda naive_bayes   svmLinear         knn    gamLoess
##   0.9508197   0.9508197   0.9508197   0.9508197   0.9508197   0.9508197
##    multinom          rf    adaboost
##   0.9508197   0.9508197   0.9508197
```

```
## [1] 0.9508197
```

The results are then checked in the following code.

```
votes <- rowMeans(pred == "1")
y_hat <- ifelse(votes > 0.5, "1", "0")
mean(y_hat == test$Class)
```

```
## [1] 0.9508197
```

Both pieces of code record an average accuracy of 0.9508197. This is absolutely positive as it reports that the ensemble model would be approximately 95% accurate. It is important, however, to remember that this is only on 54% of the data available. More accurate results may be attained when using 90% of the data, instead.

Final Modelling

A final ensemble model will now be created on the much larger 'dat' data set that was created in the Data Wrangling section.

```
fits <- lapply(models, function(model){
  print(model)
  train(Class ~ ., method = model, data = dat)
})
```

Results

Final Test (on validation set)

Similar to the testing prior, the final test will be done by creating a prediction function that applies the ensemble model above to the validation set.

```
final_pred <- sapply(fits, function(object)
  predict(object, newdata = validation))
```

The accuracy of each model used within the ensemble model can be seen below. Each model was 100% effective. Naturally, then, theaverage accuracy would also be 100%.

```
##         glm         lda naive_bayes   svmLinear         knn    gamLoess
##           1           1           1           1           1           1
##    multinom          rf    adaboost
##           1           1           1
```

This result is checked on the following code.

```
final_votes <- rowMeans(final_pred == "1")
final_y_hat <- ifelse(final_votes > 0.5, "1", "0")
mean(final_y_hat == validation$Class)
```

```
## [1] 1
```

Another ensurance is made that the predicted results and the values in the Class column are correct/identical:

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
result <- as.numeric(apply(final_pred, 1, getmode))
identical(result, validation$Class)
```

```
## [1] TRUE
```

Therefore, it is clear that this ensemble model has been 100% effective.

Conclusion

The prediction algorithm used for the Data Analysis section was adequately successful while the ensemble model used for the actual modelling was even more so. Every model used for the final testing was 100% accurate. Although the overall ensemble model was decidedly effective and the objective of this assignment was achieved, there are certainly ways in which it can be improved upon.

According to the original study through which this data was collected, the couples were asked to fill out a form which included "...questions on gender, marital status, age, monthly income, family structure, type of marriage, happiness in marriage and divorce thought." (Yontem et al 2019: 263) Should this information have been included in the data set it would have helped to provide greater insights into divorce predictors. Should this information have been available, extra columns could have been added to each row. For example, a column could have been made for monthly income or age and optimum lambdas or cutoffs could have been found to find correlations between said variables and divorce. Additionally, family structure and type of marriage could be used as categorical data that may influence divorce. Additionally, the modelling would have been more effective with more data - meaning more couples. The final training set only had 153 couples' data to work with. Although the accuracy was 100%, this can be easier to achieve with smaller data sets as there is less variance and smaller chance of outliers. More significant accuracy could have been attained with more data.

Nontheless, the objective of this project was to create an ensemble model that would correctly predict whether a couple would get divorced based on this data set. This objective was achieved with significant success.

Bibliography

Yöntem, M.K., Adem, K., İlhan, T. and Kılıçarslan, S., 2019. Divorce prediction using correlation based feature selection and artificial neural networks. *NevşehirHacıBektaşVeliÜniversitesiSBEDergisi*, 9(1), pp.259-273. Available from: https://dergipark.org.tr/tr/download/article-file/748448 [Cited 10 April 2020]