Predicting subjective life satisfaction from indicators representing 7 key areas of life

Introduction:

This research is inspired from the book, "Oola, find balance in an unbalanced world" by Dave Braun and Troy Amdahl. The authors have recommended balancing and growing the 7 key areas of life in order to live one's dream life. These are– fitness, finance, family, field, faith, friends and fun.

To study this, an extensive social survey data was needed that could represent these key areas and was publically available for use. Hence European social survey data was used which provided a compilation of several modules of social surveys in Europe. The features from different survey modules were shortlisted and mapped to the 7 key areas mentioned in the book. Refer **Appendix** for features used.

Purpose:

According to a social study, nearly 40% of Europeans suffer each year from a mental disorder such as depression, anxiety, insomnia or dementia. It is a known fact that chronic stress and mental disorders manifest into physical illness over time. Apart from the huge economic costs, there are social consequences as sufferers become too unwell to work, personal relationships break down and the negative effects pass on to future generations.

The purpose of this study is to identify some of the risk factors associated with unhappiness and factors that could help enhance the quality of life of individuals.

Data:

Source: The European Social Survey (ESS) is an academically driven multi-country survey aimed at -1) monitoring changing public attitudes and values and to investigate how they interact with Europe's changing institutions, 2) improving methods of survey measurement and 3) developing a series of social indicators, including attitudinal indicators.

The current survey used in this study (ESS7-2014, ed2), is the seventh round and covers 22 countries across Europe. The survey involves random probability sampling, a minimum target response rate of 70% and hour- long face to face interviews with participants on different modules: 1) Media and social trust, 2) Politics, 3) Subjective wellbeing, social exclusion, religion, national and ethnic identity, 4) Immigration, 5) Health and Inequality, 6) Gender, Year of birth and Household grid, 7) Socio demographics, 8) Human values. Most responses are captured on Likert scale.

Pre-processing: The dataset has over 40,000 rows representing individuals aged 15 and over who are residents within private households in participating countries. ~55 features were shortlisted out of ~400 available for the purpose of this analysis. Then following preprocessing was done:

1) Missing data (10K rows) was removed.
2) The response variable "stflife" which indicates the degree of one's satisfaction with their life on a scale of 0 (Extremely dissatisfied) to 10 (extremely satisfied) was divided into 2 categories: Dissatisfied (0 to 5) and Satisfied (6 to 10) so that classification method can be applied.
3) Some features such as "tvtot" were condensed from numeric into meaningful categories like less than 2 hours, 3+hours, etc.

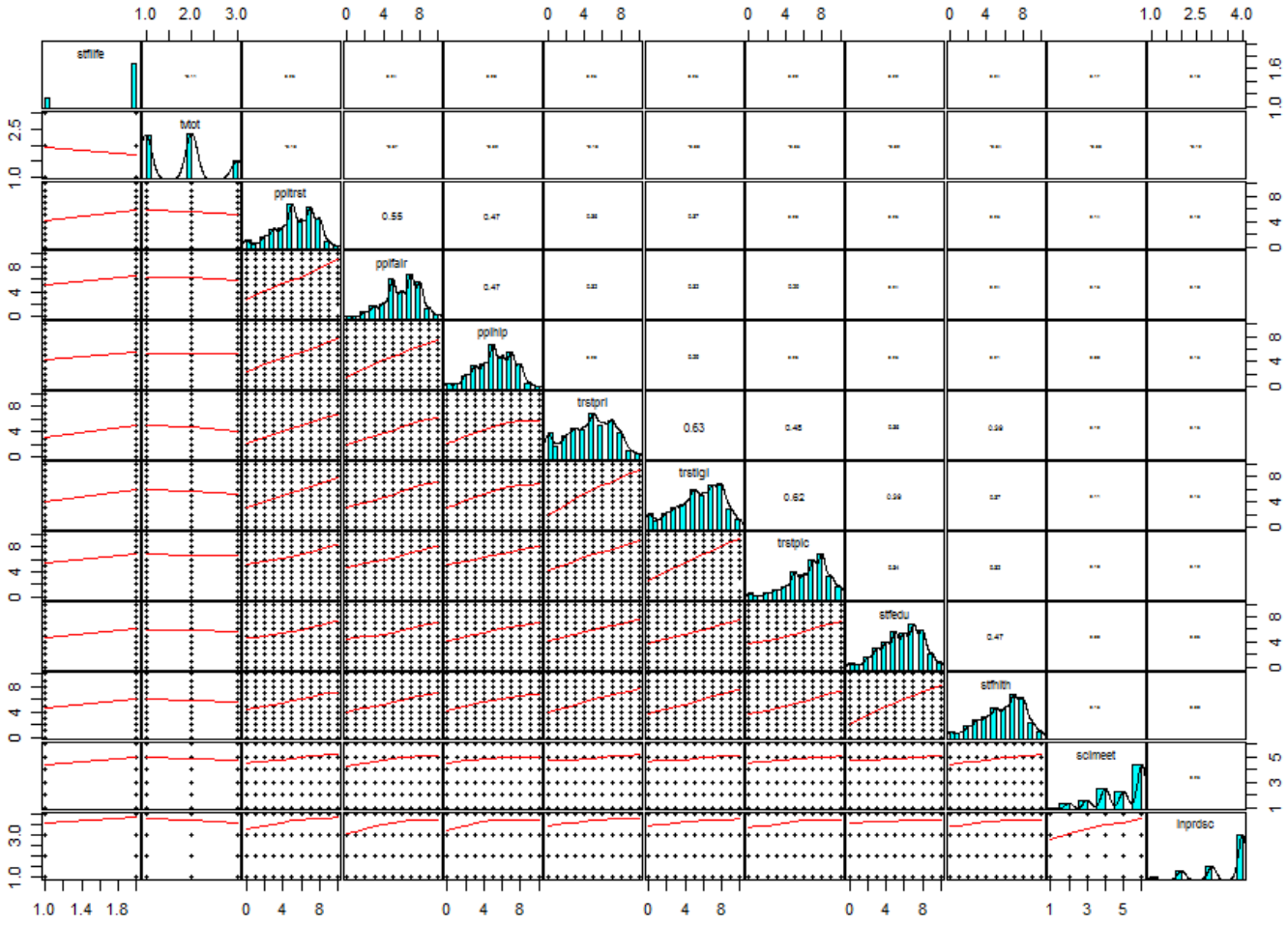
4) Features representing presence or not of different health conditions like diabetes, heart issues, etc. were combined into 1 flag feature called “Chronic” with 1 meaning presence of one or more chronic health issues and 0 meaning absence thereof.
5) The data was divided into training and test sets in the ratio of 3:1, stratified by geography so that the population with similar characteristics can be grouped together.

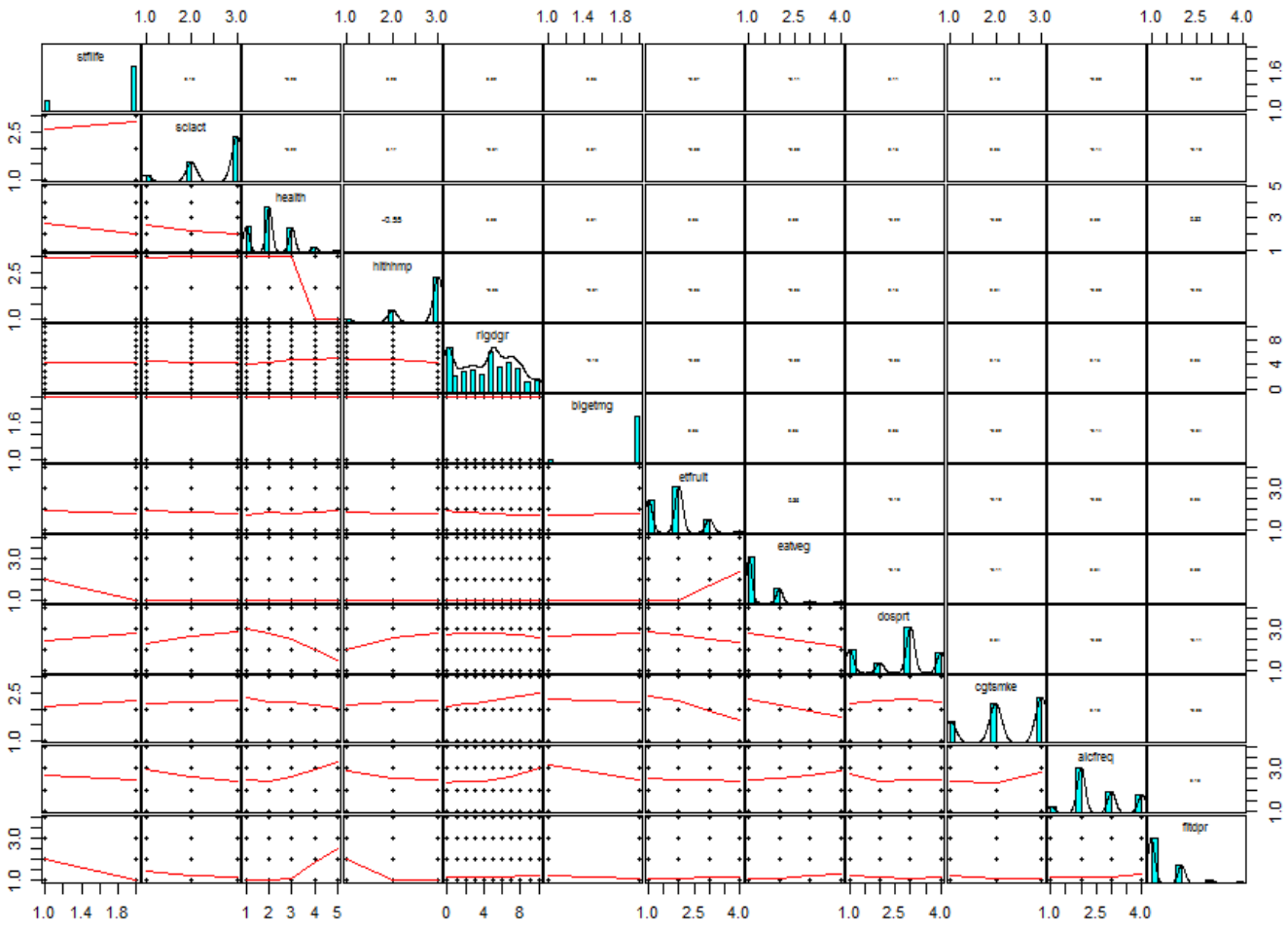
Data exploration:

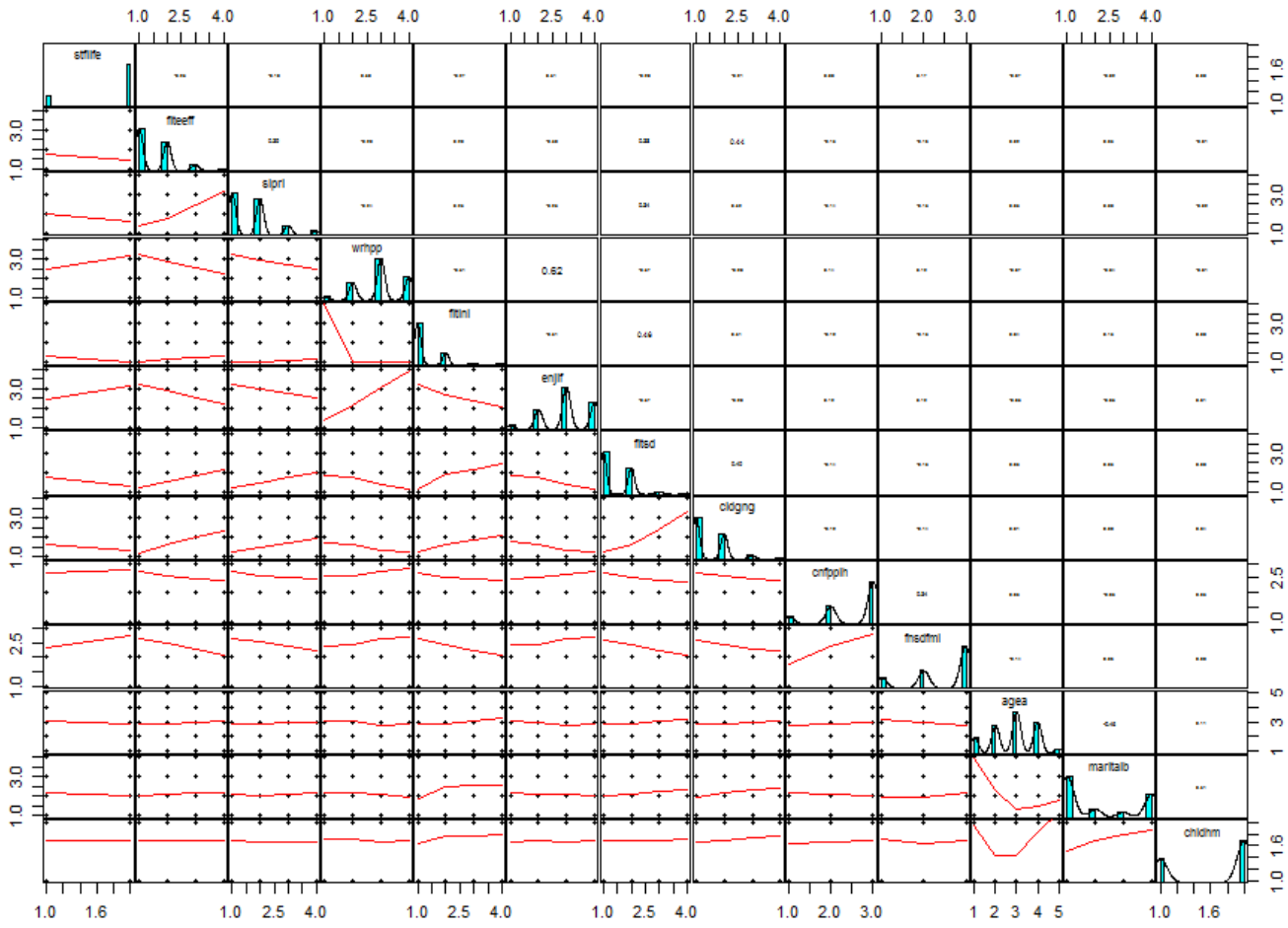
Refer **Appendix: 1** for feature explanations

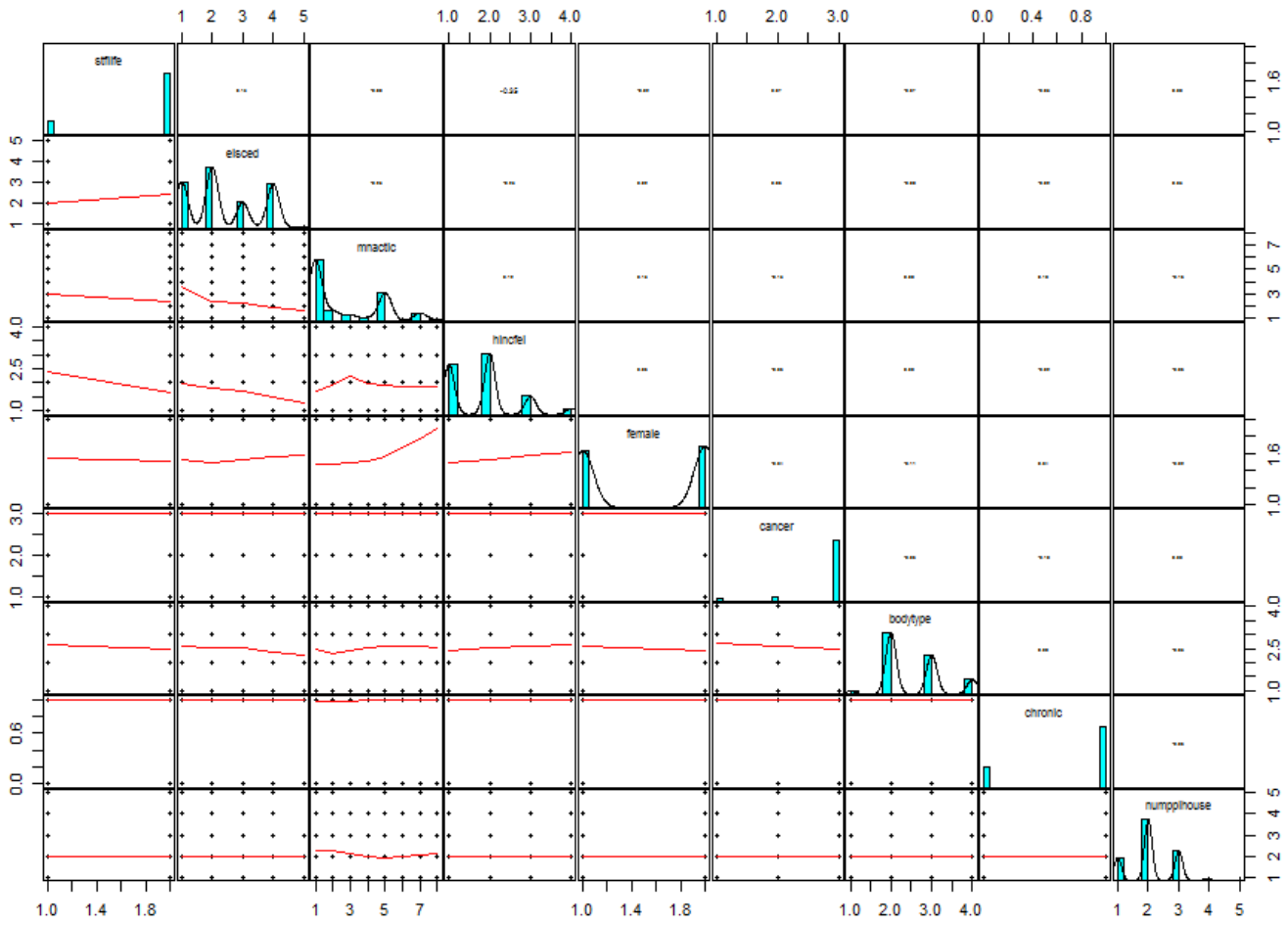
```
      tvtot          ppltrst         pplfair          pplhlp         trstprl
 0-1.5hrs:11958   Min.   : 0.00   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
 1.5-3hrs:12027   1st Qu.: 4.00   1st Qu.: 5.000   1st Qu.: 4.00   1st Qu.: 3.000
 3+hrs   : 4951   Median : 5.00   Median : 6.000   Median : 5.00   Median : 5.000
                  Mean   : 5.36   Mean   : 5.956   Mean   : 5.25   Mean   : 4.667
                  3rd Qu.: 7.00   3rd Qu.: 8.000   3rd Qu.: 7.00   3rd Qu.: 7.000
                  Max.   :10.00   Max.   :10.000   Max.   :10.00   Max.   :10.000

    trstlgl         trstplc          stfedu          stfhlth            sclmeet
 Min.   : 0.00   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Nevr      :  364
 1st Qu.: 4.00   1st Qu.: 5.000   1st Qu.: 4.000   1st Qu.: 4.000   <1mth     : 1923
 Median : 6.00   Median : 7.000   Median : 6.000   Median : 6.000   1_mth     : 2722
 Mean   : 5.52   Mean   : 6.407   Mean   : 5.731   Mean   : 5.782   many_mth  : 6011
 3rd Qu.: 8.00   3rd Qu.: 8.000   3rd Qu.: 7.000   3rd Qu.: 8.000   1_wk      : 5324
 Max.   :10.00   Max.   :10.000   Max.   :10.000   Max.   :10.000   >1wk-daily:12592

 inprdsc           sclact          health            hlthhmp          rlgdgr
 0 : 1035   very<most : 2711   Min.   :1.000   Yesalot    : 1673   Min.   : 0.000
 1 : 3861   <most     : 7840   1st Qu.:1.000   Yes-somewht: 5647   1st Qu.: 2.000
 2 : 5587   same->most:18385   Median :2.000   no         :21616   Median : 5.000
 3+:18453                      Mean   :2.126                       Mean   : 4.424
                               3rd Qu.:3.000                       3rd Qu.: 7.000
                               Max.   :5.000                       Max.   :10.000

    blgetmg         etfruit           eatveg       dosprt               cgtsmke
 Min.   :1.000   2+day :10019   1mny_day :21336   0  : 6699   1Daily        : 5851
 1st Qu.:2.000   4-7wk :14295   2.1-3X_wk: 6877   1  : 2766   2LessOftn_past:10749
 Median :2.000   <1-3wk: 4268   3.<1_wk  :  606   3-6:13332   3Never        :12336
 Mean   :1.949   Nvr   :  354   0        :  117   7  : 6139
 3rd Qu.:2.000
 Max.   :2.000

            alcfreq          fltdpr          flteeff          slprl            wrhpp
 1daily         : 2075   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 2few.mth-manywk:14315   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
 3.<1_mth       : 6712   Median :1.000   Median :1.000   Median :2.000   Median :3.000
 4Never         : 5834   Mean   :1.389   Mean   :1.599   Mean   :1.754   Mean   :2.994
                         3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:4.000
                         Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000

     fltlnl           enjlf           fltsd           cldgng           cnfpplh
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   1Alwys_ofn: 3183
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   2Smtims   : 7537
 Median :1.000   Median :3.000   Median :1.000   Median :1.000   3No-Rarely:18216
 Mean   :1.342   Mean   :2.994   Mean   :1.488   Mean   :1.507
 3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:2.000
```

```
Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :4.000

      fnsdfml         agea               maritalb     chldhm            eisced
1Alwys_ofn: 4314   14-25: 3693   1Married_cvlu:15064   1:10557   <=lowerSec  :7659
2Smtims   : 7395   25-40: 6559   2sep_divor   : 3060   0:18379   2upperSec   :9933
3No-Rarely:17227   40-60:10152   3widowd      : 2096             3advVoc     :4189
                   60-80: 7472   4Notmarried  : 8716             4BAandhigher:7066
                   80+  : 1060                                   5other      :  89


                mnactic             hincfel        stflife       female
1Paidwork           :14988   1comfortable:10607   Min.   : 0.000   0:14002
5retired            : 6902   2coping     :13148   1st Qu.: 6.000   1:14934
2education          : 2432   3difficult  : 3985   Median : 8.000
7housework_childcare: 1964   4v.difficult: 1196   Mean   : 7.263
3umemployed         : 1571                        3rd Qu.: 9.000
4Sick_disabled      :  739                        Max.   :10.000
(Other)             :  340
   weights           cancer           bodytype        chronic       numpplhouse
Min.   : 0.008538   Yes : 1190   1Underwt :  726   Min.   :0.0000   1    : 5754
1st Qu.: 0.283685   Past: 1881   2NormalWt:15009   1st Qu.:0.0000   3    :15273
Median : 0.452177   Nvr :25865   3Overwt  : 9605   Median :1.0000   4-6  : 7574
Mean   : 0.979088                4Obese   : 3596   Mean   :0.7486   7-9  :  305
3rd Qu.: 1.232637                                  3rd Qu.:1.0000   10-13:   30
Max.   :11.186967                                  Max.   :1.0000
```


stflife
tMot
pplfrst
0.55
pplfair
0.47
pplhlp
trstprl
0.63
trstlgl
0.62
trstplc
stfedu
0.47
stfhlth
sclmeet
inprdsc


etlife
sclact
health
-0.55
hlthhmp
rlgdgr
bigetmg
etfruit
eatveg
dosprt
ogtsmke
alcfreq
fltdpr


stflife
fifteeff
slpri
wrhpp
fitinl
enjlf
fitsd
cldgng
cntppln
fmothml
agea
maritalb
chldhm



Data Analysis and Interpretation: Two classifiers namely K nearest neighbors and Random Forests are used. Classification method was chosen over Regression because of the high imbalance between observations over different predictor levels.

1) K-Nearest Neighbors:

KKNN package was used to implement K-Nearest neighbor classifier with inverse prevalence penalty due to the presence of rare class (i.e. Dissatisfied). A 2D tuning with k, kernel (optimal/triangular) and distance=1, and inverse prevalence penalty resulted in the following "winning" setting:

```
  k  kernel   inverr   flaterr
  9 optimal    0.355     0.166
```

Refer **Appendix: 2** for KNN tuning error rate plot

Confusion matrix for training cross-validation using winning setting:

```
             Dissatisfied Satisfied ClEr
Dissatisfied         1243      2394 0.66
Satisfied             811     14842 0.05
```

```
Sensitivity Specificity
       0.95        0.34
PPV = 0.86
```

Confusion matrix for test data predictions using winning setting:

```
             Dissatisfied Satisfied ClEr
Dissatisfied          630      1230 0.66
Satisfied             424      7362 0.05

Sensitivity Specificity
       0.95        0.34
PPV = 0.86
```

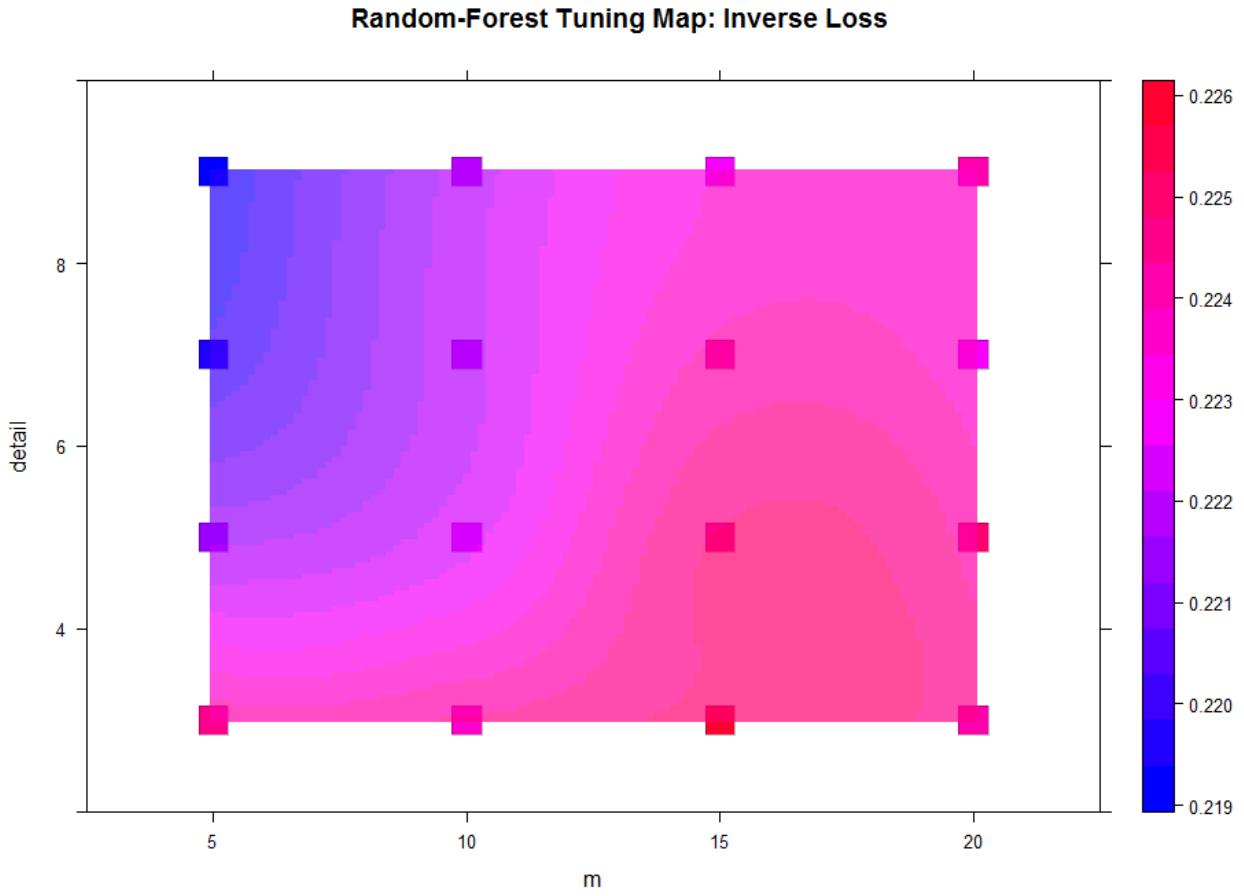
2) Random Forests:

randomForest package is used for this paper. Initial 1D tuning with ntree resulted in the following error rate:

```
ntree 100    200   300   400    500
      3.2      2   1.2   0.7      0
```

Then, a 3D tuning with settings ntree=500 (chosen from above step), penalty = inverse prevalence penalty, nodesize (detail) and number of features (mtry) resulted in the following "winning" setting:

```
    m detail  flat       inv
    5      9 0.224     0.219
```



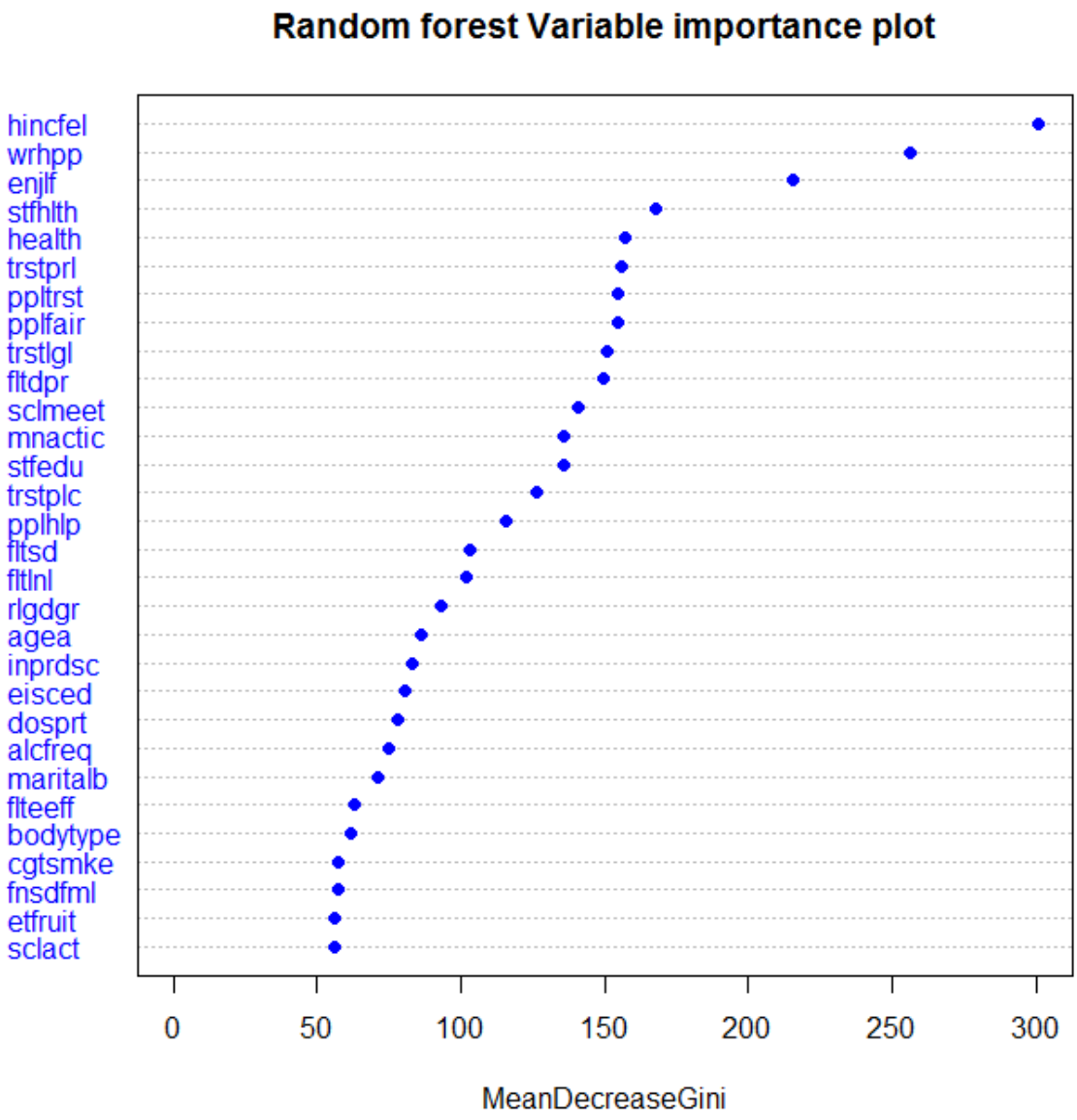
Confusion matrix for winning setting on training data:

	Dissatisfied	Satisfied	ClassErr
Dissatisfied	2864	773	0.212
Satisfied	3541	12112	0.226

Refer **Appendix: 3** for Random Forest winning setting confidence interval vs missed observations plot

Confusion matrix for winning setting on test data predictions:

	Dissatisfied	Satisfied	ClassErr
Dissatisfied	1476	384	0.260
Satisfied	1779	6007	0.296



Refer **Appendix: 1** for variable explanations

Limitations: A major limitation of this study is personal bias of the respondents when using Likert scales. More than 80% of respondents have reported higher satisfaction with life which is contradictory to an earlier study, reporting nearly 40% of Europeans suffering from mental disorder such as depression, anxiety, insomnia or dementia.

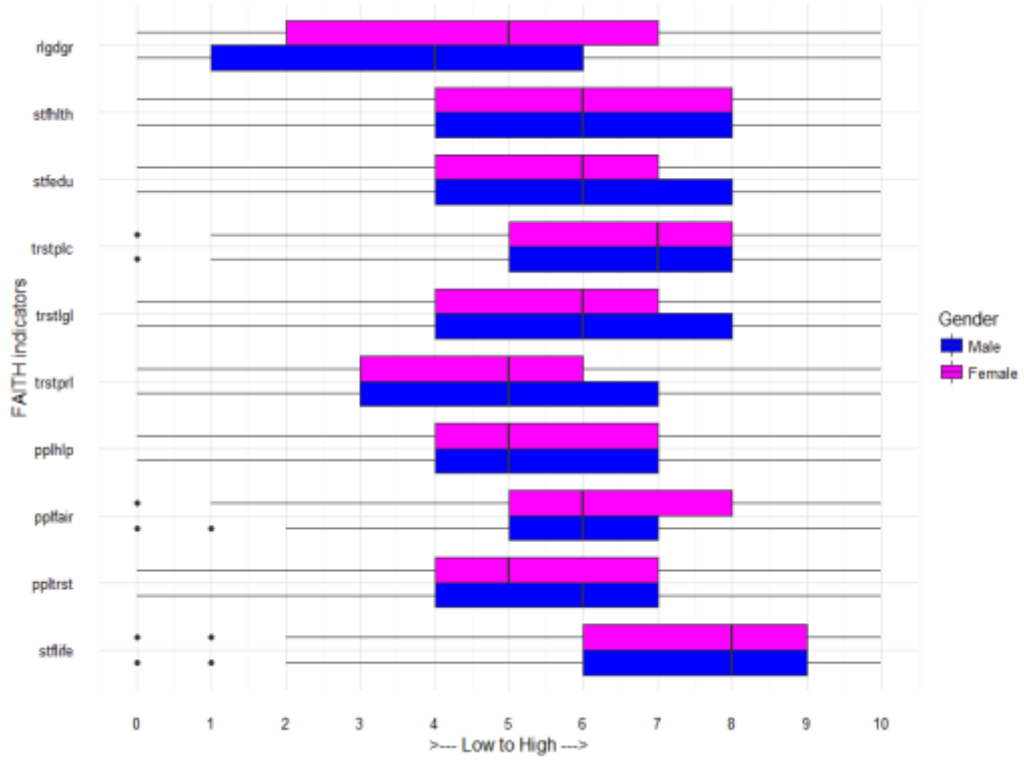
Conclusion: Looking at the variable importance plot, it can be concluded that the contentment in life is related to financial security, finding joy in daily activities, taking care of health, having faith in humanity and nurturing good relationships. In short, it seems the book rightly advises people to balance the 7 F's of life.

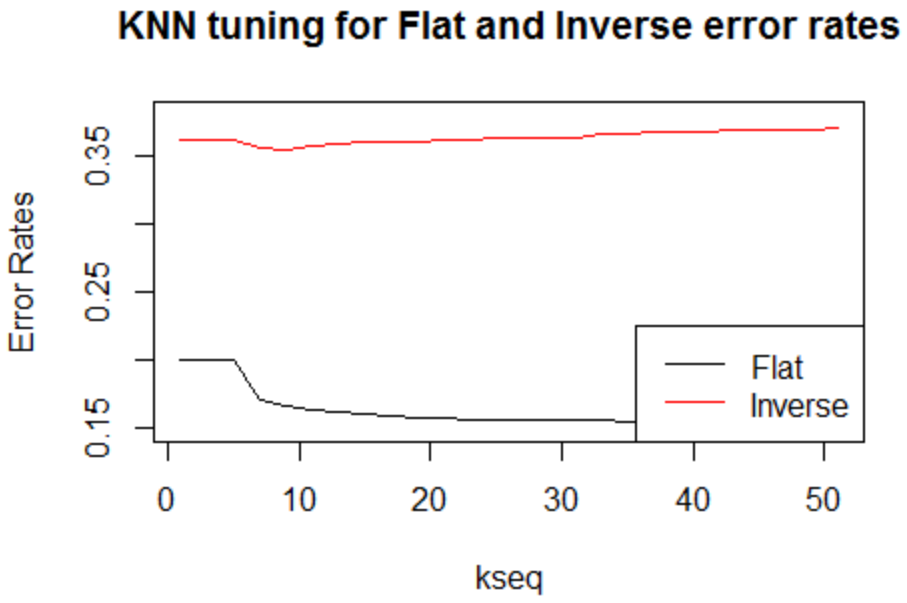
Appendix: 1

Variable	Meaning	Levels
tvtot	TV watching, total time on average weekday	"0-1.5hrs","1.5-3hrs","3+hrs"
ppltrst	Most people can be trusted, or you can't be too careful	0 (you can't be too careful) through 10 (Most people can be trusted)
pplfair	Most people try to take advantage of you, or try to be fair	0 (Most people try to take advantage of me) through 10 (Most people try to be fair)
pplhlp	Most of the time people helpful or mostly looking out for themselves	0 (People mostly look out for themselves) thru 10 (People mostly try to be helpful)
trstprl	Trust in country's parliament	0 (No trust at all) through 10 (Complete trust)
trstlgl	Trust in the legal system	0 (No trust at all) through 10 (Complete trust)
trstplc	Trust in the police	0 (No trust at all) through 10 (Complete trust)
stflife*	How satisfied with life as a whole-response	*Response variable- "Dissatisfied" and "Satisfied"
stfedu	State of education in country nowadays	0 (Extremely bad) through 10 (Extremely good)
stfhlth	State of health services in country nowadays	0 (Extremely bad) through 10 (Extremely good)
sclmeet	How often meet with friends, relatives or colleagues	"Nevr","<1mth","1_mth","many_mth","1_wk",">1wk-daily"
inprdsc	How many people with whom you can discuss personal matters	"0","1","2","3+"
sclact	Take part in social activities compared to others of same age	1 (Much less than most) through 10 (Much more than most)
health	Subjective general health	0 (Very good) through 5 (Very bad)
hlthhmp	Hampered in daily activities by illness/disability/infirmity/mental problem	"Yesalot","Yes-somewht","no"
rlgdgr	How religious are you	0 (Not at all religious) through 10 (Very religious)
blgetmg	Belong to minority ethnic group in country	1 (Yes), 2 (No)
etfruit	How often eat fruit	"2+day","4-7wk","<1-3wk","Nvr"
eatveg	How often eat vegetables	"1mny_day","2.1-3X_wk","3.<1_wk","0"
dosprt	Do sports or other physical activity, how many of last 7 days	1-7 (Number of days)
cgtsmke	Cigarettes smoking behaviour	"1Daily", "2LessOftn_past", "3Never"
alcfreq	In the last 12 months, how often have you had a drink containing alcohol?	"1daily", "2few.mth-manywk","3.<1_mth","4Never"
height*	Height of respondent (cm)	*Hidden - used to calculate "bodytype"
weight*	Weight of respondent (kg)	*Hidden - used to calculate "bodytype"
bodytype*		*calculated - "1Underwt", "2NormalWt","3Overwt","4Obese"

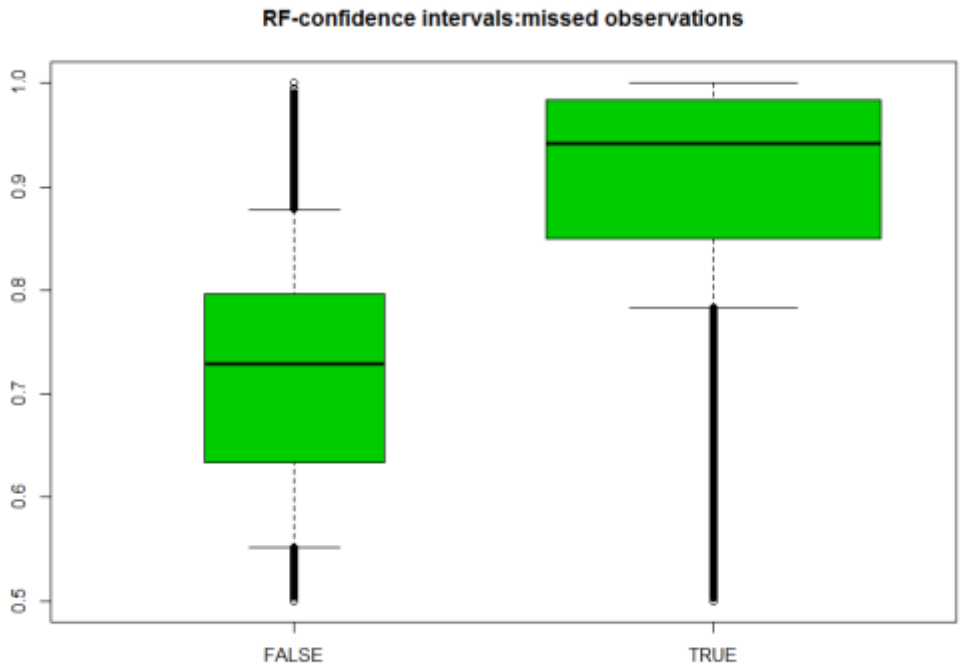
chronic*		*calculated 1(presence of one or more chronic health conditions such as diabetes, heart problems) or 0 (absence thereof)
agea	Age group	"14-25","25-40","40-60","60-80","80+"
fltdpr	Felt depressed how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
flteeff	Felt everything did as effort, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
slprl	Sleep was restless, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
wrhpp	Were happy, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
fltlnl	Felt lonely, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
enjlf	Enjoyed life, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
fltsd	Felt sad, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
cldgng	Could not get going, how often past week	1 (Never/Almost never) through 4 (All/Almost all of the time)
cnfpplh	Serious conflict between people in household when growing up, how often	"1Alwys_ofn","2Smtims","3No-Rarely"
fnsdfml	Severe financial difficulties in family when growing up, how often	"1Alwys_ofn","2Smtims","3No-Rarely"
Numpplhouse*	Number of people living regularly as member of household	"1","3","4-6","7-9","10-13"
female	Gender of respondent	1= Female, 0=Male
maritalb	Legal Marital status	"1Married_cvlu","2sep_divor","3widowd","4Notmarried"
chldhm	Children staying in the household	1=Yes, 0=No
eisced	Education level	<=lowerSec","2upperSec","3advVoc", "4BAandhigher","5other")) h$eisced <- h$edu
mnactic	Main activity of respondent in the past week	"1Paidwork","2education","3umemployed", "4Sick_disabled","5retired","6service","7housework_childcare","8other"
hincfel	Feeling about household income nowadays	"1comfortable","2coping","3difficult", "4v.difficult"
weights*	Population weight x Design weight	*calculated

Appendix: 2


rlgdgr
stfhlth
stfedu
trstplc
trstlgl
trstprl
pplhlp
ppltair
ppltrst
stflife
FAITH indicators
0
1
2
3
4
5
6
7
8
9
10
>--- Low to High --->
Gender
Male
Female


KNN tuning for Flat and Inverse error rates
Error Rates
0.15
0.25
0.35
0
10
20
30
40
50
kseq
Flat
Inverse

Appendix: 3



R Code and dataset: submitted as separate files