EFFECTS OF LIFESTYLE ON AGING

Akshay Naik (02)
Anirudh Kuruwada (04)
Ramakrishna Ronanki (34)
Rohit Musle (37)

Introduction

- There have been many studies focused on understanding physical and cognitive changes that define aging.
- These studies spread their effort in identifying genetic, physical, behavioral, and environmental factors.
- This are primary factors that affect the aging process and understand the interrelationship between aging and various diseases.

Objective

- Explore the data to uncover insights around impacts of lifestyle on aging.
- Specifically, We need to predict three types of diseases like ‘Arthritis’, ‘Angina’ and ‘Chronic Lung’ conditions.

Work flow

Step 1 Title

- Raw Data
- Analyze data
- Removing NA values
- Consulting Domain Expert

Step 2 Title

- Dim Red. of categorical variables (Decision trees)
- Dim. Red. Of continuous variables(PCA)
- Segregate train and validate data

Step 3 Title

- Implementation of Machine Learning Algo.
- Random Forest,
- Max Entropy
- Naïve Bayes

Step 4 Title

- Selection of best technique.

Data Explanation:

- Unique features : 31 features consist of Economic condition, Personal health, Residence, Smoking/Alcohol etc.
- The total features in our data set were about 269
    1. 3 dependent variables(Angina, Lung Cancer, Arthritis)
    2. 265 explanatory variables
- Number of observations around 13K.
- Each set of unique features consist of mixed data type for e.g. ‘Physical activity’ consist of categorical as well as continuous data.
- Some of the Unique features set also consist of ordinal data like Self care, Memory.
- Initially most of raw data was also filled with NA values.

This is how our data looks like!!

```
> levels(data_dic$Category)
 [1] "Anxiety"                          "Blood pressure (Diastolic)"     "Blood pressure (Systolic)"
 [4] "Chest pain"                       "Childhood"                      "Conflict"
 [7] "Daily Functionality Discomfort"   "Demographic"                    "Dependent 1"
[10] "Dependent 2"                      "Dependent 3"                    "Diet"
[13] "Disease conditions"               "Economic"                       "Education"
[16] "Energy Level"                     "Family and Household"           "General visible health issue"
[19] "Health Care"                      "Learning"                       "Memory"
[22] "Mobility"                         "Parents"                        "Personal Health"
[25] "Personal relationship"            "Physical Activity"              "Physical Strength"
[28] "Pulse rate"                       "Residence"                      "Self-care"
[31] "Sleeping"                         "Smoking/ Alcohol"               "Vigorous Activity"
[34] "Visibility"
>
>
```


WHAT'S MISSING???

Dealing with missing values in the data

- The first hurdle of our data analysis was to predict and fill the missing values in our data.
- We had two ways to deal with missing values :

1. Eliminating variables which had more than 30% missing values.
2. Predicting missing values.

The second option is the best choice as it enables us to give better results in prediction in terms of accuracy. There are several packages in R that provide built in functions to make this task easier.

MissForest

- We have used missForest package to find the missing data values.
- The package missForest is used to impute missing values. It uses a random forest trained on observed values of data matrix to predict missing values. It can be used to predict both continuous and categorical variables.
- In missForest, primarily the N.A. values are replaced with mean values for continuous variables and mode values for the categorical values. After each iteration the difference between the previous and the new imputed data matrix is assessed for the continuous and categorical parts. The stopping criterion is defined such that the imputation process is stopped as soon as both differences have become larger once.

$$\sqrt{\frac{mean((X_{true} - X_{imp})^2)}{var(X_{true})}}$$

  - where Xtrue the complete data matrix, Ximp the imputed data matrix

Code snippet

- install.packages("missForest")
  library("missForest", lib.loc="C:/Program Files/R/R-3.2.2/library")
  test_fill<-missForest(test2,ntree=50,mtry=8,maxiter = 10)

- Ntree : No of trees to grow in each forest.
- Mtry : No of variables that are randomly selected at each split.
- Maxiter : maximum number of iteration to be performed given the stopping criterion is not met.

- Training data after filling up the missing values.

```
> summary(train_fill_10_1)
     CAX_ID         Arthritis          Angina          Chronic_Lung          V1        
 Min.   :    1   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :1.000  
 1st Qu.: 3241   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.000  
 Median : 6464   Median :0.0000   Median :0.00000   Median :0.00000   Median :2.000  
 Mean   : 6472   Mean   :0.2202   Mean   :0.08877   Mean   :0.08695   Mean   :1.508  
 3rd Qu.: 9712   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:2.000  
 Max.   :12957   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :2.000  
       V2              V3              V4              V5              V6        
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :2.000   Median :2.000   Median :2.000   Median :2.000   Median :2.000  
 Mean   :1.932   Mean   :1.919   Mean   :1.996   Mean   :1.977   Mean   :1.966  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000  
       V7              V8              V9              V10             V11       
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :2.000   Median :2.000   Median :2.000   Median :2.000   Median :2.000  
 Mean   :1.997   Mean   :1.995   Mean   :1.991   Mean   :1.997   Mean   :1.909  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000  
 Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000  
      V12             V13             V14             V15             V16       
 Min.   :1.000   Min.   :1.000   Min.   :50.00   Min.   :1.000   Min.   : 1.000  
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:55.00   1st Qu.:2.000   1st Qu.: 1.000  
 Median :2.000   Median :2.000   Median :61.00   Median :2.000   Median : 1.000  
 Mean   :1.995   Mean   :1.533   Mean   :63.02   Mean   :2.449   Mean   : 1.044  
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:70.00   3rd Qu.:2.000   3rd Qu.: 1.000  
 Max.   :2.000   Max.   :2.000   Max.   :99.00   Max.   :8.000   Max.   :87.000  
      V17             V18             V19             V20             V21       
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   :1.000   Min.   : 0.000  
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.: 1.000  
 Median : 1.000   Median : 2.000   Median : 1.000   Median :2.000   Median : 2.000  
 Mean   : 2.224   Mean   : 3.898   Mean   : 3.387   Mean   :2.065   Mean   : 2.631  
 3rd Qu.: 2.000   3rd Qu.: 7.000   3rd Qu.: 6.000   3rd Qu.:3.000   3rd Qu.: 3.000  
 Max.   :87.000   Max.   :87.000   Max.   :87.000   Max.   :7.000   Max.   :20.000  
```

How we use Decision trees….

- The data was really complicated. It consist of categorical, continuous and ordinal as well.
- Also, the data dictionary suggests that one parameters may consist of multiple features. For eg. Parameter -Economic conditions consist of several features from V17-V45 (28 features!!!)
- To reduce data from 28 features to few features and decision trees prove its worth.
- We applied decision trees to every set of unique features for dimension reduction.
- After using decision tree we were able to reduce the dimensions.
- For Angina : 64
- For Arthritis : 58
- For chronic_ lung : 41

Contd….

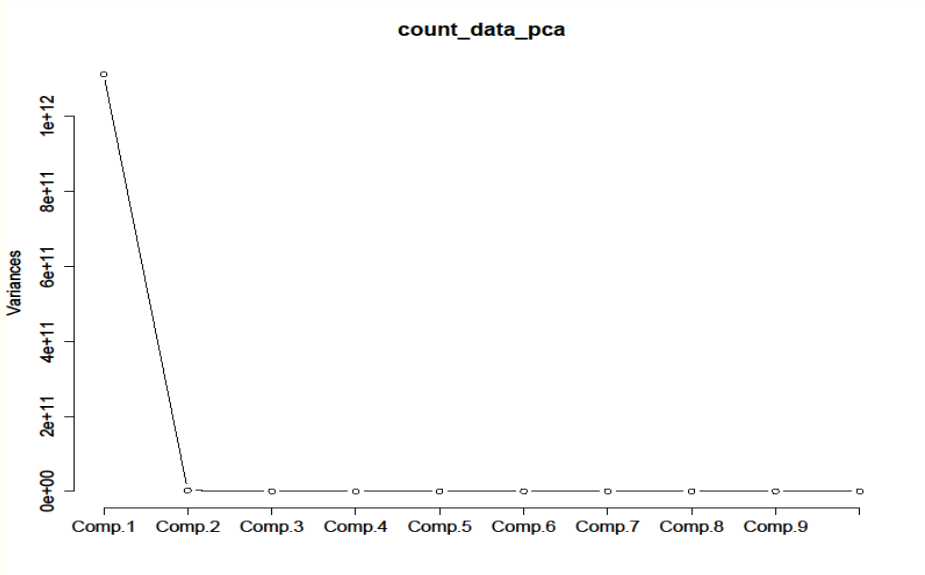
- We first check the interdependency of response variables. Going by statistical way and domain knowledge, we consulted doctor to know about this.
- Conclusion:
    - Lung cancer is totally independent of other two.
    - Arthritis can be linked with Angina with probability of 0.05 that too in rare cases.
    - As a result, we can safely assumed that all response variables are independent of each other.

Important features extracted:-

	lung	arthritis	angina
general health	V9,	V2,V6,V11	V2,V3,V5,V6,v11,
anxiety	none	V170	V170,
economic	V11,V17,V19,V34,V36,V38,	V19,30,33	V18,V38,V43,V24,V25,V36,V39,V13,V17,V20,
education	V46,V47,	V46,V47	V46,V47,
demographic	V13,V14,	V13,V14,V15	V13,V14,
parents	none	V59,V60	V59,V60,
residence	V51	V51	V51,V53,
smoking	V67,V68,V69,	V67,V68,V69	V67,V68,V69,V79,
diet	none	none	none
disease	V263,V265,	V265,V263	V265,V263,V263,
physical	V70,	V70,V73	V70,V73,V77,
personal_health	V102,V106,V114,V124,V256,	V139,V105,V148,V111,V141,V121,V135,V117,V149,V102,V124	V105,V121,V154,V138,V135,V118,V106,V102,V110,V256,
health care	V88,V89,V90,V92,V99,V100.	V88,V89,V90,V92,V100	V88,V89,V90,V96,V100,
chest pain	V261,V258,	V258	V258,V261,
mobility	V155,	V155,157,159	V155,
vigorous	V160,	V160	V160,V165,
conflict	none	none	none
relationship	none	none	V77,V79,
daily discomfort	V239,V242,V246,V249,	V240,V238,V246,V244,V244,V249,V235	V239,240,246,249,
energy level	none	none	V200,
sleeping	none	V198	none
visiblity	none	V205,V211,V212	none

How we handled Continuous data

- We have about 35 attributes which are continuous and we used PCA to reduce the dimension of the data.
- Using the train data set, we found the principal components and their respective contributions towards variances.
- As first PC contributes around 99.7 % is selected.
- We found loadings of that particular PC and calculated scores for train set and test set.



Code Snippet

```
count_data_pca = princomp(cont_data, cor="False")
summary(count_data_pca)
screeplot(count_data_pca, type="lines")
cont_data_pc <- cont_data_pca$loadings[,1]
cont_mat <- as.matrix(cont_data)
cont_train_scores <- cont_mat %*% cont_data_pc
cont_test_scores <-  as.matrix(filled_test) %*% cont_data_pc
```

Prediction Method:

- We have used 3 methods to predict the effect of lifestyle on aging in fact to predict whether a person will be affected with Arthritis, Angina and lung cancer.

  1. Random Forest
  2. Naive bias
  3. Maxent (Maximum Entropy method)

Random Forest

- **Random forest** (or **random forests**) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.
- Code snippet :

```
lung_fit <-randomForest(Chronic_Lung~V9+V11+V17+V19+V34+V36+V38+
                        V46+V47+V13+V14+
                        V51+V67+V68+V69+
                        V263+V265+
                        V70+V102+V106+V114+V124+V256+
                        V88+V89+V90+V92+V99+V100+
                        V261+V258+V155+
                        V160+V239+V242+V246+V249+
                        V50+V115+V116, data = train_lung, ntree = 300, mtry=10)

lung_pred <- predict(lung_fit, train_dat_validate)

prop.table(lung_pred, train_dat_validate$lung)
```

Contd…

- For the random forest, we used below parameters :

mtry = 10

Ntree = 300

- We used the most important variables after performing dimension reduction on the individual variables for growing the random forest.

- Prediction accuracy :

Disease	Prediction Accuracy
Chronic lung cancer	91
Artthritis	93
Angina	90

Maximum Entropy Model :

- Features are often added during model development to target errors
- Then, for any given feature weights, we want to be able to calculate: • Data conditional likelihood • Derivative of the likelihood wrt each feature weight • Uses expectations of each feature according to the model n then find the optimum feature weights .

  Maxent behaviour:

  max_lung<-maxent(trainlungmat,lungtrain$Chronic_Lung)

  Accuracy : 89.5 %

Output of Maxent

labels	0	1
0	0.948254802010323	0.0517451979896773
0	0.914058847718694	0.0859411522813062
0	0.948637927322902	0.0513620726770981
0	0.972864850005791	0.0271351499942094
0	0.949015334262956	0.0509846657370443
0	0.958008663299277	0.0419913367007234
0	0.970703077944098	0.029296922055902
0	0.922098295330272	0.0779017046697279
0	0.960074889611562	0.0399251103884381
0	0.957223840650902	0.0427761593490985
0	0.864028479076695	0.135971520923305
0	0.957036719400233	0.0429632805997672
0	0.93621154528532	0.0637884547146802
0	0.950437115268224	0.0495628847317761
0	0.913232480748307	0.0867675192516935
0	0.953314729716039	0.0466852702839605
0	0.892625770196265	0.107374229803735
0	0.932312591566497	0.0676874084335032

Naive Bayes :

Naive-Bayes is for classification:

- We have a bunch of random variables (data features) which we would like to use to predict another variable (the class)
- **naive Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

- Navlung=naiveBayes(Chronic_Lung~V9+V11+V17+V19+V34+V36+V38+V46+V47+V13+V14+V51+V67+V68+V69+V263+V265+V70+V102+V106+V114+V124+V256+V88+V89+V90+V92+V99+V100+V261+V258+V155+V160+V239+V242+V246+V249+V50+V115+V116, data = train_lung)

- navpredlung<-predict(navlung,train_lung)
- Accuracy : 88 %

Refrences

- https://www.crowdanalytix.com
- www.stackoverflow.com
- Dr. Nakod and Dr. Sangle
- https://cran.r-project.org
- GOOGLE!!

THANKYOU