The University of Hong Kong
Department of Computer Science

COMP3354 Statistical Learning

Project Report

Cancer Risk Analysis for Elderly

Group Member:

Sirui Chen 3035532882

Yuchen Liu 3035535145

Yutong Wu 3035331927

Yuqian Zhang 3035233565

Lei Zuo 3035331472

Supervisor:

Dr. Hao Luo

December 8, 2019

Table of Content

1 Introduction

1.1 Background and motivation

The growing development of computer technologies has stimulated its application in other areas, of which the cross-field of computer science and medicine is gaining increasing attention. One of the most important branches of this field is the prediction of severe diseases. Humans have been fighting cancer for generations but still could not eliminate it. With a better understanding of one's chance of having cancer, one would be able to take precautions to reduce the probability or have an early diagnosis, which could significantly enlarge the survival rate of a patient. Many aspects of one's life could have an influence on the probability of cancer, such as physiological features and family income. Using these features, it may become practical that cancer can be predicted in advance.

1.2 Purpose of the study

This project aims to use multiple statistical learning methods to predict the risk of the elderly is having cancer. This project will first perform feature selection on a dataset to determine the features that are closely related to cancer. After that, different regression and classification models will be applied to predict the result using the unselected or selected features. Various evaluation metrics will be used to test the accuracy of the models and an overall analysis will be given. The ultimate goal of this project is to realize high accuracy prediction of whether the elderly has the high risk of getting cancer using a suitable list of features.

1.3 Outline of the report

This report contains six major parts.
Chapter One introduces the topic of cancer risk analysis for elderly.
Chapter two reviews the dataset that is used in this project. Sources of the data and explanation of the questionnaires are introduced and discussed.
Chapter Three presents the methodology used in data preparation, model construction, and evaluation metrics. Details of the selection of features and models are discussed.

Chapter Four discusses the experimental results of different models in this project. How the models are applied and evaluated is explained in detail. The limitations of this project are examined afterward.
Chapter Five demonstrates potential future research in this field.
Chapter Six concludes the study.

2 Review of Dataset

2.1 Sources of the dataset

The Dataset from **CHARLS** (China Health and Retirement Longitudinal Survey) are used in this project.

CHARLS is a large-scale interdisciplinary investigation project hosted by multiple Faculties in Peking University and funded by the National Natural Science Foundation of China. It aims to collect high-quality data from families and individuals of middle-age (over 45 years old) by survey and questionnaire, in order to analyze the aging problem in China, promote interdisciplinary research, and provide a more scientific basis for the improvement of related policies in China.

Given its reliability, strong connection with our aging topic and sufficient data records, it is a desirable dataset for us to investigate. After the scrutinization of several waves of the dataset, we decide to use **Wave 4 2015 Data**, as it provides clear explanations of each variable and the data records are relatively new, which is believed to be a better proxy of the current status.

2.2 Explanation of the questionnaires and dataset

The CHARLS Wave 4 2015 Questionnaire ("The Questionnaire") are conducted in 6 dimensions: Family, Health status and functioning, Healthcare and Insurance, Work, retirement and pension, Income, expenditure and assets, and Housing characteristics. More than 20,000 over middle-aged individuals from a large number of communities in China participated in this field survey.

The Wave 4 2015 Dataset obtained from The Questionnaire consists of 17 sub-datasets including more than 20,000 records. The Dataset includes basic personal information, family structure and financial support, health status, physical measurement, medical service and health insurance, work, retirement and pension, income, consumption, assets, and community status.

Aiming to provide the risk index analysis for an elderly about the probability of getting cancer, our project uses the **"cancer" in Health Status section from question "ZDA007[x]" as the label, and other related information as features**. After our primary selection of

features, we obtained 122 related features of 8,402 records without NA. The detailed data processing method will be explained in chapter 3. Statistical models with additional feature selection process are trained with the selected features on the "cancer" label to reveal the risk of an elderly to get cancer.

3 Methodology

3.1 Data preprocessing

We investigated the 17 sub-datasets of around **7,000 columns of data** in total collected **from more than 20,000 individuals** of whom more than 10,000 also participated in the household questionnaire. However, a large number of columns contain too many NAs.

In order to **get a reasonable size of predictors** for our purpose of cancer risk analysis, we did the primary selection of features mainly from four aspects (i.e. Biomarkers, family members related features, individual and household features and pension and insurance features) based on their relation with cancer, and their number of valid data entries. To **deal with NA**, we eliminate most of high NA proportion columns, because some of NAs are neither reasonable to get filled with 0 nor with average level. Therefore, the proportion of NA in the column restricted the model features selection. To better **represent Yes or No question**, we change the original 1 and 2 representation into 0 and 1. To deal with columns that are **perfectly collinear** with each other, we eliminate one column, and to deal with the columns that are answering the same question, we merged it into one column. After all steps, we **inner join the individual data and household data** according to "ID", which means all data are household's data.

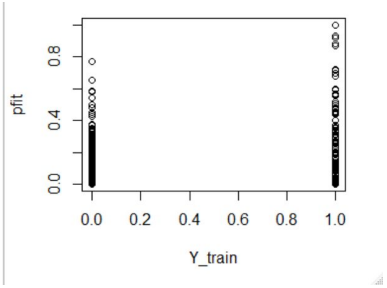
Ultimately, we get our **primary dataset of 122 features and 1 label as "cancer_true" of 8042 rows of records**.

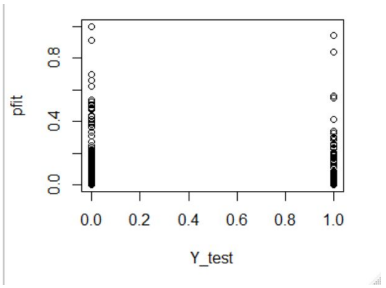
3.2 Models for feature selection and prediction

After the preprocessing of dataset, our dataset contains 122 features and 1 label. Hereinafter, unless further complement, all the random seed for sampling process and built-in cross validation is set to 2.

3.2.1 Forward and Backward stepwise selection

Before the further selection of features, the training and testing results of the logistic regression on the dataset containing 122 features and "cancer" label are shown in the following figures, which indicates an overfitting problem.





Therefore, forward and backward stepwise selection are applied in order to select a subset of the feature from the current 122 features to avoid overfitting problems. Combined with logistic regression, we can obtain a better model for risk index prediction.

We first run the full logistic model and empty logistic model on the training dataset. Afterwards, **use AIC as the selection criteria**, stepwise select features that contribute most to the model precision. After the forward and backward stepwise selection, we use the intersection of the features selected forward and backward and **19 features are selected**. The detailed selection result will be discussed in chapter 4.

3.2.2 Principal Component Analysis

Another way to shrink the predictor size is PCA. But in our case, PCA is only used for size shrinkage. After that, a supervised regression will be applied (not unsupervised learning), to be mentioned in chapter 4. The theoretical foundation of using PCR to do classification is that **principal components maximizes variances, and what an ideal classification expects is to maximize the differential across clusters and minimize variance within a cluster**. Although interpretability may be sacrificed to some extent, PCR will do a better job in prediction.

3.2.3 Lasso

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs **both variable selection and regularization** in order to enhance the prediction accuracy and interpretability of the statistical model it produces. In our project, we first carry out the cross validation to select the best lambda value and then find the coefficients of features with this parameter. Finally, we fit a lasso model based on the imported data.

3.3 Tree Based Methods for prediction

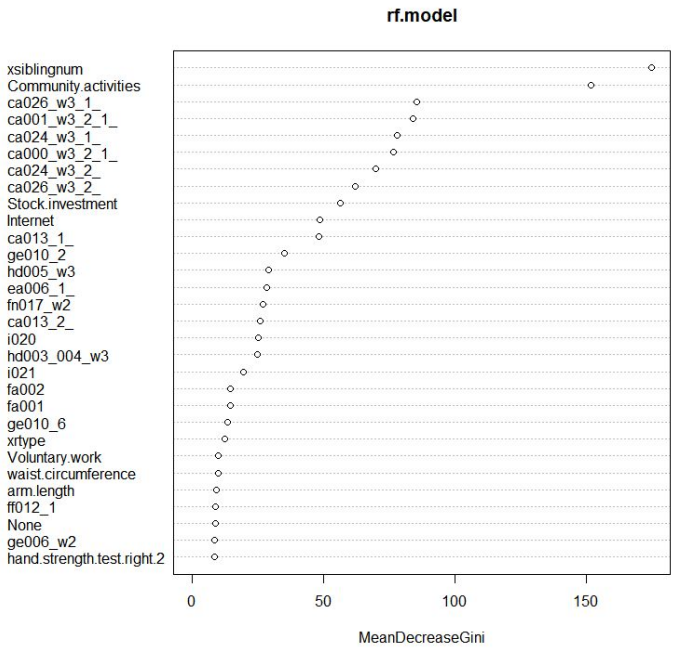
3.3.1 Random Forest: Classification

We applied random forest regression algorithm with 1000 trees on training data only (training : test : validation = 6:2:2) and then applied on training data combined with simulated data from oversampling method SMOTE. Then we thresholded the data at 0.026 with is the sample proportion. The performance seems to be quite similar in AUC with SMOTE or without SMOTE,but in terms of undiagnosed rate the model trained using SMOTE tend to

perform better. Moreover we use validation set method to get optimal feature size: 37. Detailed performance results is shown in the next chapter.

Importance data of different features: The table explained features in enumerated order of importance.

1. Number of Siblings	2. Do Community activity when free	3. Father can take care himself
4. Mother Still living	5.which Ethnic Minorities father is	6. Father Still living
7.Father is not ethnic minority	8. Father cannot take care of himself	9. Investing stocks or not
10. Surfing internet in free time or not	11. Whether Your father health condition is “very good”	12. Family travel expenditure
13. Yes or not you have being cheated	14.Amount of insurance purchase	15.Included durance of social pension program
16. Whether you father health condition is “good”	17.Does your residence have heating	18. Credit card loan amount
19. Type of heating energy source	20.Did you work for more than an hour last week?	21. Whether the participant did in agricultural work within a year
22. Family education expenditure	23. Whether the participant has participated in the survey before or not	24. Whether do voluntary work during free time
25. Waist Circumference	26. Length of arm	27.Monthly raw income
28. Whether the participant didn’t participate in any of the activities during free time.	29. Family weekly food expenditure	30. The result of right hand strength test.



3.3.2 Gradient Boosting Decision Tree

We applied Gradient Boosting Decision Tree with validation to selected hyperparameter max depth = 1, which is decision stump. since the Boosting method can adjusted weight of different sample according to previous trained trees, we don't need oversampling to increase its performance. We finalize the trees number to be 5000 without shrinkage parameter. As a result boosting tree behaves better than random forest in terms of AUC and the ratio of missed diagnoses. Detailed illustration can be seen in chapter 4.

3.4 Other models for prediction

3.4.1 Support Vector Machines

Support vector machines could be used to perform classification. Both radial and polynomial kernels will be investigated and fine-tuned to find suitable parameters. The results of the two methods will be compared and discussed. Since the computation for support vector machines tends to be more expensive, the experiments will be done on the basis of the selected features discussed in section 3.1. The result will be values of true or false. A numerical probability will not be given.

3.4.2 Neural network

Since the Logistic Regression is based on strong linear assumption, to test whether apply an extremely non-linear model will fit data better we tested Neural network with Sigmoid Output and with the architecture of below, since training neural network need more balanced data we applied SMOTE data augmentation algorithm, but the outcome has not improved:

Layers	
Input layers	23 Units Linea
Hidden Layer 1	25 Units Sigmoid Activation With Bias
Hidden Layer 2	25 Units Sigmoid Activation With Bias
Hidden Layer 3	15 Units Sigmoid Activation With Bias
Hidden Layer 4	5 Units Sigmoid Activation With Bias
Output Layer	1 Unit Sigmoid Positive Probability Output

3.5 Evaluation Metrics

The training of all the models will be result-oriented, and our optimal goal is to get as accurate as possible on whether an elderly has cancer. Taking a step back, we aim to improve the True Positive Rate, which means the elderly who gets the cancer should be predicted to

have cancer. In our evaluation, we use confusion matrix and ROC and AUC to evaluate our models, and detailed results will be shown in section 4.

4 Experiments and Results

4.1 Models based on linear assumption

4.1.1 Logistic Regression with Stepwise feature selection

The features obtained with Forward and Backward selection and their coefficient with logistic regression are shown in the following table:

Feature	Coefficient	Description	Feature	Coefficient	Description
(Intercept)	-6.0400154		Water.cigarettes	-13.4325617	if smoke water cigarettes
xsiblingnum	-7.4648404	number of siblings	knee.height	50.7807839	average knee height
ca000_w3_2_1_	0.8896064	if father alive	ge004	-23.7232990	how many guests ate at home this week
fa001	1.6661591	if engaged in agriculture	hd005_w3	0.5139286	if get swindled
cc012_w3_1_	2.8158552	if siblings are healthy	i011	0.5651318	number of stairs to home
ge010_6	5.0869601	medical expenditure	Systolic.2	2.5163201	systolic value
Stock.investment	1.0259044	if have stock investment	cg003_w2_1_	-0.8428383	if need help to finish survey
ca001_w3_2_1_	0.6524237	if mother alive	fa006	2.0168594	if receive salaries
hc039_w3	0.4897536	if others owe you money	breath.test.1	-1.1265443	vital capacity
hand.strength.test.right.2	3.4311835	right hand strength	waist.circumference	-10.0490569	waist circumference

After stepwise selecting the features, we run **logistic regression** with maximum likelihood method on the selected features. Then calculate the probability of response to be 1. Since the original dataset response is skewed to 0, which means most of the participants are not affected by cancer, we need to decide a good threshold aka logistic boundary to classify the result. A good proxy for the **threshold is the natural cancer incidence** in our training dataset, which is:

$$\frac{\#\{classified\ to\ 1 \mid boundary = x\}}{\#\{validation\ set\}} = natrual\ incidence = \frac{\#\{patients\}}{\#\{subjects\}}$$

With the threshold set as the natural incidence of catching cancer (i.e. 0.025), the training error and test error are shown in the following figures.

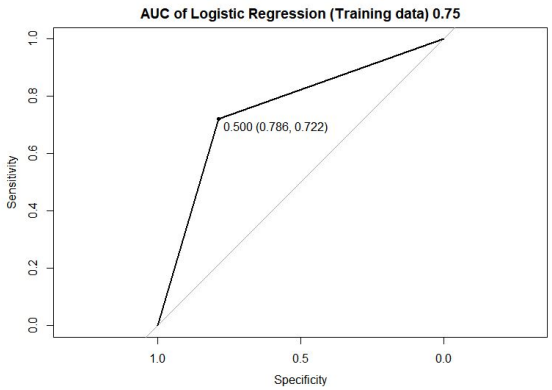
logit.train.pred	FALSE	TRUE
FALSE	4110	40
TRUE	1122	104

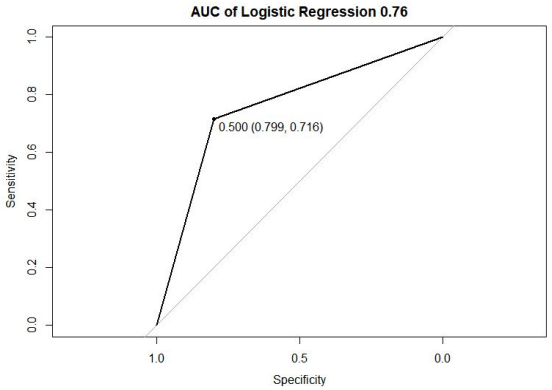
training confusion matrix

logit.pred	FALSE	TRUE
FALSE	2353	23
TRUE	592	58

test confusion matrix

The training and testing ROC and AUC results are shown in the following figures.

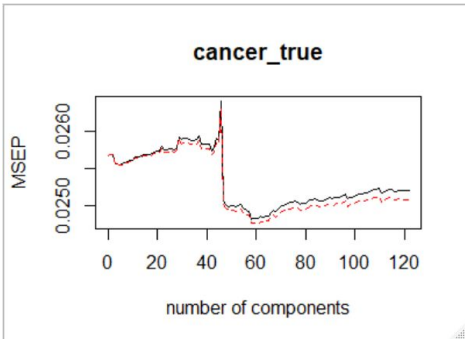


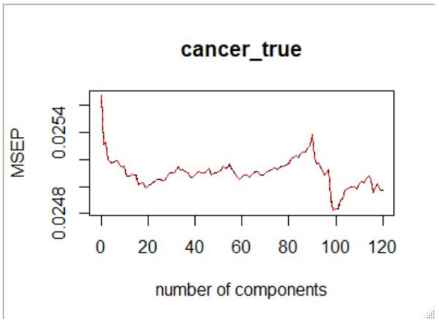


Since we are more concerned with type 1 error, which means when the individual is indeed affected by cancer, the prediction should return true, the prediction result is better than random guess on whether a person is affected with cancer or not.

4.1.2 Principal Component Regression

Due to the small correlation across each feature, the number of principal components is not small enough to an optimal level. (Note that the ratio of test:train = 7:3)





In the unnormalized model, MSE minimises when there are 60 components.

In the normalized model, MSE minimises when there are 100 components.

Using above two models with the best components and 0.5 as the threshold, results are:

Raw data (unnormalized)

pcr.result	FALSE	TRUE
FALSE	642	6
TRUE	1815	58

Normalized data

pcr.result2	FALSE	TRUE
FALSE	2400	63
TRUE	57	1

4.1.3 The Lasso

After cross validation, we find that the best lambda in feature selection process is 0.003914072. Then we use this parameter to find the most significant features. With information of coefficients, we could derive the following feature chart with top 6.

Significant features	Meaning	Coefficient
Stock.investment	if have stock investment	0.104306313
ca024_w3_2_	Father is not ethnic minority	0.012082346
fa001	if engaged in agriculture	0.008946915
ca024_w3_1_	which Ethnic Minorities father is	0.003851552
knee.height	average knee height	0.002472166
ca026_w3_2_	Father cannot take care of himself	0.002090989

Then we carry out the prediction experiment. The ratio of training set and testing set is 7:3. We also set a threshold 0.025 for the final result.

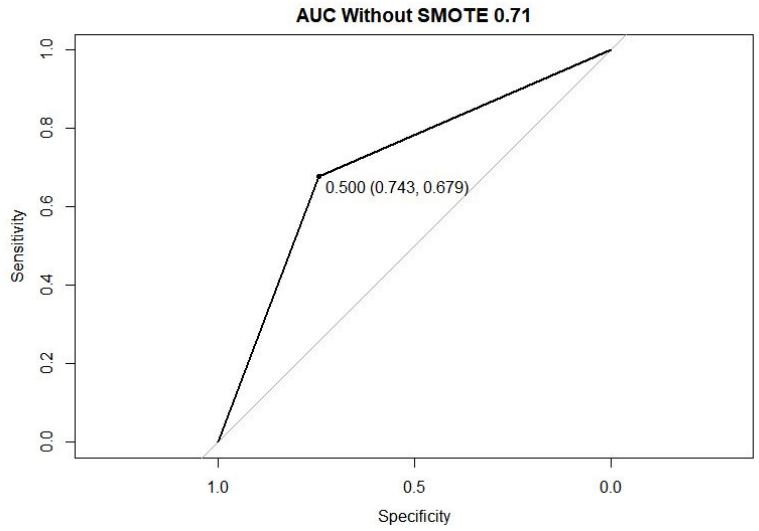
```
> table(test.y,pred)
      pred
test.y    0    1
     0 1366 1075
     1   13   66
```

4.2 Tree Based Methods

4.2.1 Random Forest

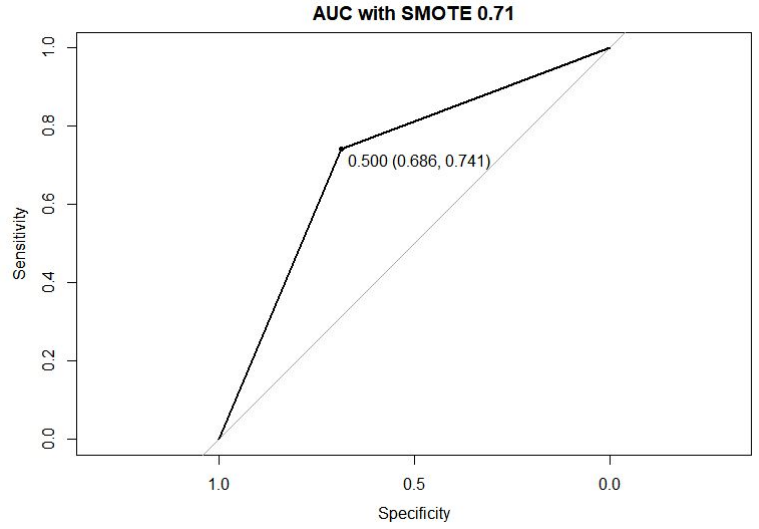
We use **AUC** and **Undiagnosed** rate as metrics for two models we obtained one with oversampling(SMOTE) one without oversampling. We applied threshold to the regression random forest to be 0.026 with is the approximately the positive sample rate in the population. Below is the confusion matrix as well as AUC plot.

```
> rf.table = table(rf.thresh,cancer_true[test])
> rf.table

rf.thresh    0    1
    FALSE 2189   26
    TRUE   756   55
```



Above is the result without oversampling.

```
> rf.table2 = table(rf.thresh2,On_stage_data2$cancer_true[test])
> rf.table2

rf.thresh2 FALSE TRUE
     FALSE  2021   21
     TRUE    924   60
```



Above is the AUC and undiagnosed rate with oversampling we can see that result improved in terms of undiagnosed rate.

4.2.2 Gradient Boosting decision tree

We applied **threshold of 0.026 after normalizing the result**, which reflected as 5 in non-standardized form. Still we use AUC as well as misdiagnosed rate on test set as metrics.

```
> table(boost.thre,On_stage_data2$cancer_true[test])

boost.thre FALSE TRUE
     FALSE  2132   23
     TRUE    813   58
```

4.3 Advanced Non-linear classifiers

4.3.1 Support Vector Machines

We experimented with two kernels, polynomial and radial. We first used misclassification error to fine tune the model.

With polynomial kernel, 10-fold cross validation was performed on cost range (0.001, 0.01, 0.1, 1, 10, 100) and degree range (2, 3). The best parameter obtained is cost 1 and degree 2. The confusion matrix obtained using the best model is as follows:.

true	pred FALSE	TRUE
FALSE	2446	5
TRUE	68	2

For radial kernel, 10-fold cross validation gives cost 1000 and gamma 0.001, and the errors are:

true	pred FALSE	TRUE
FALSE	2446	5
TRUE	67	3

Since our project aims to provide early warning to the elderly who are more susceptible to cancer, our focus is on minimizing false negative predictions. Therefore, we modified the error function and again fine-tuned both of the models.

For svm model with polynomial kernel, the modified error function gives cost 10000 and degree 3 as best parameters. The error rate is as follows:

true	pred FALSE	TRUE
FALSE	1554	78
TRUE	36	12

For model with radial kernel, the best parameters are cost 50000 and gamma 0.005. The resulting table is:

true	pred FALSE	TRUE
FALSE	1588	44
TRUE	38	10

From the above we can see that modifying the error function is better suited to our needs. The successful predictions of cancer are significantly higher. We can see that the polynomial kernel gives best results.

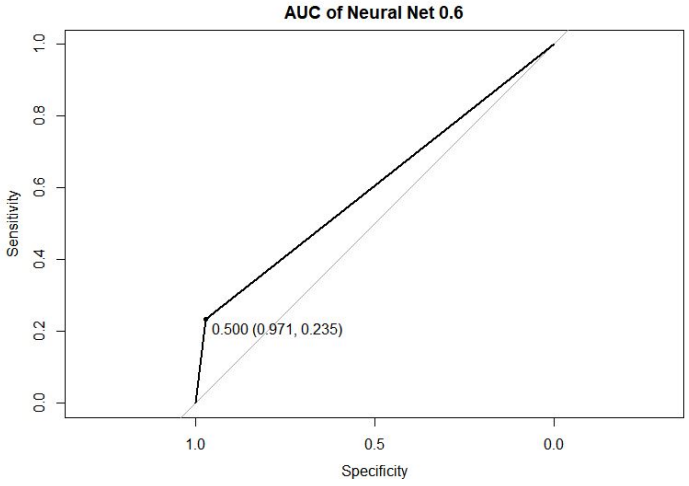
4.3.2 Neural Network

Apply classification neural network with illustrated structure and 0.026 as threshold.

The confusion matrix:

ann.logit	FALSE	TRUE
FALSE	2860	62
TRUE	85	19

The AUC Diagram:



4.4 Limitations

4.4.1 Preprocessing

Since the original dataset contains a lot of columns which has a high proportion of NAs, when selecting the features, we may need better ways to deal with NA, instead of currently eliminate high NA columns, which restricted the further feature selection.

4.4.1.1 Data Balance

Since our objective is investigating risk factors that lead to cancer, in our dataset the positive sample of cancer is too limited so that general characteristic of cancer patients cannot be fully represented by them. Although we applied data augmentation algorithms like SMOTE the improvement is quite limited.

4.4.2 Logistic Regression with Stepwise selected features

We tried both scaled and unscaled features for the prediction, though it doesn't have much difference when applied to the test set. When running the logistic regression, there are some features are extremely skewed, which may make it seem more significant. Moreover, when selecting the suitable logistic boundary to map probability to 0 and 1, we use the natural cancer incidence. If we are more concerned about the TPR, we may use a lower threshold than we are currently using.

4.4.3 Support Vector Machines

The data contain many dimensions, making it difficult to visually test their separability. If the data are not highly separable, the support vector machines may not perform very well on them. For example, if most of the points are near the hyperplane, the results obtained on test set can be much worse than that on training set. The kernels that can be examined are limited, and thus may not be able to correctly classify the data. As a result, the true positive rate for svm is relatively low compared with other models.

4.4.4 Principal Component Regression

First, because the pcr() function cannot automatically apply normalization to such a big dataset, features with large values get more weights, which may be unnatural. Second, alone with PCR, the interpretability of the predictors are low. It is hard to tell the correlation between features, and the direct influence of each predictor on cancer diagnosis. Third, even if we manually normalize the data and apply PCA, the result is still unsatisfying. The reason may be the missing of significant features related to cancer and the high irreducible error in the data.

4.4.5 Tree Based Method

One of the main shortage of tree based method is lack of probability estimation, the current outcome of the tree is only the classification result, when applying ensemble learning with other learning algorithms it is very hard to perform soft vote which believes to deliver better performance than directly hard vote. Moreover ,although random forest and GBDT tend to be more robust, the lack of direct interpretability is still a problem, if we choose one singular decision tree, then we sacrifice the robustness of the algorithm.

5 Future Work

5.1 Design more thoughtful survey and change the target population.

Through all the models we have trained the accuracy performance seems to be quite limited although we can conservatively evaluated the risk of getting cancer at a relatively high accuracy. The reason is due to the extremely unbalanced data, since our focus is evaluating cancer risk we should intentionally survey more cancer patients to get more sufficient data and more representative characteristic of them.

5.2 Use more powerful computer for fine tune

Due to the computational limitations some hyper parameters including number of trees and cost of svm cannot be investigated in detail and cross validation is not feasible. In the future we will use more computational capable devices to train and evaluate model so that more fine tune to the model will be made possible.

5.3 Further investigate into feature correlation and selection.

Due to the lack of social science related background knowledge we cannot analytically select or screen out some interrelated data from the dataset, thus the dataset we obtained and put to data selection has quite redundant features.

5.4 Improve the procedure and structure of survey avoid NAs.

In the data we obtained from CHARLS WAVE 4, some previously related assumption and some repetitive and hierarchical questions generate many not available data. To increase the response rate and avoid NA we should use more parallel and formatted designs to conduct the survey.

5.5 Try to analyse characteristic of distribution of data from different classes.

If we hope to apply more suitable method for extremely unbalanced data including LDA and Anomaly detection, we need to know the form of the distribution and comparison to normal distribution to estimate the probable utility. Moreover, the more we know the distribution of the data, more oriented anomaly detection model can be built.

6 Conclusion

There are around 20 million people in the world get cancer every year, and most of them are elderlies. We used statistical learning to predict cancer risk of elderlies.

In our project, we tried to analyze the CHARLS dataset with various statistical models including Linear models, Tree based models, and advanced Non-linear Models. Currently, the best model obtained using AUC and confusion matrix as criteria, is the Logistic Regression with test AUC=0.76.

From this project, we see the cancer risk is not only related to personal physical measurement, but also has a close relation with the family background, financial status, etc. In the future, we hope to arouse public attention for elderly cancer prevention.