Lung Cancer Data Analysis

Yan Cheng
Email:myan0613@tamu.edu

1. Statement of Research

Lung cancer is the leading cancer killer in both men and women in the U.S. In 1987, it surpassed breast cancer to become the leading cause of cancer deaths in women. An estimated 154,050 Americans are expected to die from lung cancer in 2018, accounting for approximately 25 percent of all cancer deaths. It is also the most common cancer worldwide, accounting for 2.1 million new cases and 1.8 million deaths in 2018. More than half of people with lung cancer die within one year of being diagnosed. If we can identify significant predictors and detect lung cancer at an earlier stage, we might be able to improve the lung cancer survival rate.

2. Study Design

2.1. Data Description

The data set we used is from Kaggle. It has 15 predictors and one binary response. Except age is a continuous variable, all other predictors are binary variables. There are 309 observations in total. In the original data, some binary values are coded as “yes” and “no”. We recoded it as 1 and 0. Some are coded as 2 and 1. Again, we recoded it as 1 and 0, so they are much easier to be analyzed with statistical programming language R. 1 denotes a patient who has the feature, and 0 denotes a patient without the feature. The binary response is recoded as 1 and 0 too. 1 represents a patient who has lung cancer and 0 represents a patient who doesn’t have lung cancer.

2.2. Exploratory Data Analysis

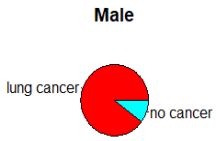
2.2.1.1. There is no missing value in this data set. We checked the correlation between the covariates and they are not highly correlated.

2.2.1.2. We also checked any outliers in age.

Figure 1. Boxplot of Age

There is only one 21 years old female with no lung cancer out of 309 observations. There are another two observations aged 38 and 39 respectively which is not far away from the majority of the data. Therefore, we are going to continue the analysis with these three observations.

2.2.1.3. Check differences between males and females, smokers and nonsmokers



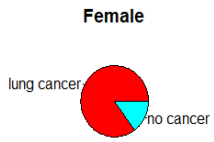
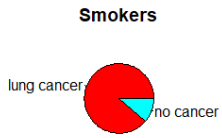


Figure 2: Gender Differences



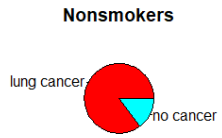


Figure 3: Smokers vs Nonsmokers

From the pie charts above, there is no obvious differences between male and female, smokers and nonsmokers.

2.2.1.4. Age

As we all know, age is a significant predictor of lung cancer. Here we plot a histogram of age.

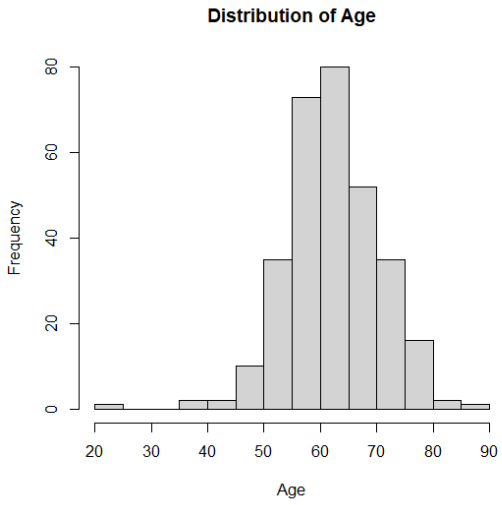


Figure 4. Histogram of Age

This plot tells us most people in our data set are older than 50. From Figure 5 below, we know that most people with lung cancer are older than 50.

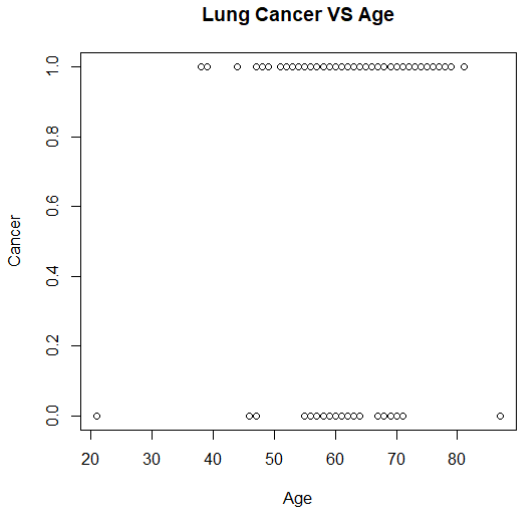


Figure 5. Age and Lung Cancer

3. Method

3.1. A logistic regression with age as the predictor

P=0.118 which is not significant. Therefore, age is not a significant predictor of lung cancer by itself.

Here is the fitted line and the confidence band for the fit.

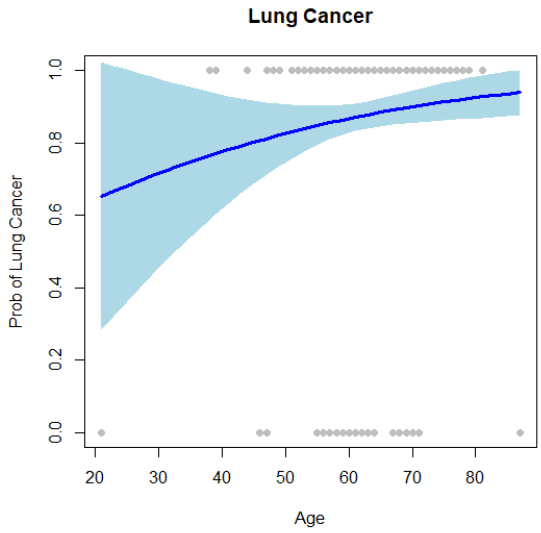


Figure 6. Probability of lung cancer based on age

Now, we are going to test whether the fit is linear or quadratic versus the need to do a semiparametric fit.

Analysis of Deviance Table
Model 1: LUNG_CANCER ~ AGE
Model 2: LUNG_CANCER ~ AGE + age2
Model 3: LUNG_CANCER ~ s(AGE, k = 27)

	Resid.	Df Resid. Dev	Df	Deviance	Pr(>Chi)
1	307.00	231.88			
2	306.00	231.43	1.000	0.453	0.5011431
3	291.17	189.01	14.826	42.422	0.0001746 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

From the Analysis of Deviance Table, we know that there is no significant differences between a linear fit and quadratic fit, but a semiparametric fit is necessary.

3.2. Generalized Additive Models

Next, we fit two GAMs with all the predictors, with age entering the model linearly and as a spline respectively.

Here is the summary of GAM1:

Family: binomial
Link function: logit
Formula:

```
LUNG_CANCER ~ GENDER + AGE + SMOKING + YELLOW_FINGERS + ANXIETY +
    PEER_PRESSURE + CHRONIC.DISEASE + FATIGUE + ALLERGY + WHEEZING +
    ALCOHOL.CONSUMING + COUGHING + SHORTNESS.OF.BREATH + SWALLOWING.DIFFICULTY +
    CHEST.PAIN
Parametric coefficients:
                          Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)               -8.33760    2.50692   -3.326  0.000882 ***
GENDERM                   -0.52611    0.70898   -0.742  0.458050
AGE                        0.02181    0.03394    0.643  0.520477
SMOKING                    1.77601    0.70190    2.530  0.011396 *
YELLOW_FINGERS             1.37642    0.74251    1.854  0.063777 .
ANXIETY                    0.88775    0.81268    1.092  0.274672
PEER_PRESSURE              1.73122    0.66025    2.622  0.008740 **
CHRONIC.DISEASE            3.19156    0.88829    3.593  0.000327 ***
FATIGUE                    3.07043    0.82521    3.721  0.000199 ***
ALLERGY                    1.64614    0.76895    2.141  0.032292 *
WHEEZING                   0.96625    0.83419    1.158  0.246737
ALCOHOL.CONSUMING          1.40981    0.79890    1.765  0.077616 .
COUGHING                   3.31128    1.07166    3.090  0.002003 **
SHORTNESS.OF.BREATH       -0.72889    0.76004   -0.959  0.337551
SWALLOWING.DIFFICULTY      3.12209    1.12984    2.763  0.005722 **
CHEST.PAIN                 0.55907    0.68913    0.811  0.417213
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) =  0.566   Deviance explained = 60.8%
UBRE = -0.599  Scale est. = 1         n = 309
```

From the summary above, we can see that smoking, peer pressure, chronic disease, fatigue, allergy, coughing and swallowing difficulty are all significant predictors of lung cancer. 60.8% of variance in the response is explained by the predictors in this model.

Here is the summary of GAM2 with age entering the model as a spline:

```
Family: binomial
Link function: logit
Formula:
LUNG_CANCER ~ GENDER + s(AGE, k = 27) + SMOKING + YELLOW_FINGERS +
    ANXIETY + PEER_PRESSURE + CHRONIC.DISEASE + FATIGUE + ALLERGY +
    WHEEZING + ALCOHOL.CONSUMING + COUGHING + SHORTNESS.OF.BREATH +
    SWALLOWING.DIFFICULTY + CHEST.PAIN

Parametric coefficients:
                   Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)        -7.2269     1.4285    -5.059  4.21e-07 ***
GENDERM            -0.5924     0.7115    -0.833  0.405087
SMOKING             1.8433     0.7180     2.567  0.010248 *
YELLOW_FINGERS      1.3442     0.7403     1.816  0.069418 .
ANXIETY             0.8295     0.8152     1.017  0.308916
PEER_PRESSURE       1.8408     0.6824     2.697  0.006988 **
CHRONIC.DISEASE     3.4125     0.9232     3.696  0.000219 ***
FATIGUE             3.2938     0.8666     3.801  0.000144 ***
ALLERGY             1.8439     0.8055     2.289  0.022073 *
WHEEZING            0.9381     0.8453     1.110  0.267094
ALCOHOL.CONSUMING   1.3798     0.7948     1.736  0.082537 .
```

```
COUGHING                  3.4987      1.1125      3.145      0.001661 **
SHORTNESS.OF.BREATH      -0.8320      0.7738     -1.075      0.282300
SWALLOWING.DIFFICULTY     3.2953      1.1764      2.801      0.005090 **
CHEST.PAIN                0.4620      0.7013      0.659      0.510029
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Approximate significance of smooth terms:
            edf  Ref.df    Chi.sq    p-value
s(AGE)    2.024   2.568     2.243       0.35
R-sq.(adj) =  0.585   Deviance explained = 62.4%
UBRE = -0.60435     Scale est. = 1         n = 309
```

From the summary, we can see that smoking, peer pressure, chronic disease, fatigue, allergy, coughing and swallowing difficulty are all significant predictors of lung cancer. This is the same as GAM1. Age is still not significant even if it enters the model as a spline. 62.4% of variance in the response is explained by the predictors in this model. It's improved slightly.

We used an ANOVA test to compare the two GAMs to see if the spline is needed. We got p=0.11. So the spline is not necessary for age.

Next, we did stepwise regression to select variables using gam::step.Gam function from R package gam. The predictors selected here are: yellow fingers, peer pressure, chronic disease, fatigue, allergy, alcohol consuming, coughing, swallowing difficulty. Smoking enters the model as a factor.

Here is the summary from our final model:

```
Family: binomial
Link function: logit
Formula:
LUNG_CANCER ~ YELLOW_FINGERS + PEER_PRESSURE + CHRONIC.DISEASE +
    FATIGUE + ALLERGY + ALCOHOL.CONSUMING + COUGHING + SWALLOWING.DIFFICULTY +
    as.factor(SMOKING)

Parametric coefficients:
                          Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)               -6.6110      1.2677       -5.215     1.84e-07 ***
YELLOW_FINGERS             1.7409      0.6397        2.722      0.006497 **
PEER_PRESSURE              1.8743      0.6372        2.942      0.003265 **
CHRONIC.DISEASE            2.6949      0.7619        3.537      0.000404 ***
FATIGUE                    2.8705       0.6719       4.272      1.94e-05 ***
ALLERGY                    1.8342       0.7238       2.534      0.011267 *
ALCOHOL.CONSUMING          1.7514       0.7117       2.461      0.013861 *
COUGHING                   3.0653       0.8369       3.663      0.000250 ***
SWALLOWING.DIFFICULTY      3.4267       0.9797       3.498      0.000469 ***
as.factor(SMOKING)1        1.4536       0.6535       2.224      0.026120 *
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) =  0.552      Deviance explained = 58.8%
UBRE = -0.62255       Scale est. = 1         n = 309
```

In our final model, all the predictors are significant. With all other variables fixed, the odds for a patient with yellow fingers to have lung cancer is increased by $e^{1.74} = 5.7$; the odds for a person with peer pressure to have lung cancer is increased by $e^{1.87} = 6.49$; the odds for a person with chronic disease to have lung cancer is increased by $e^{2.69} = 14.73$; the odds for a person experiencing fatigue to have lung cancer is increased by $e^{2.87} = 17.64$; the odds for a person with allergy to have lung cancer is increased by $e^{1.83} = 6.23$; the odds for a person consuming alcohol to have lung cancer is increased by $e^{1.75} = 5.75$; the odds for a person who coughs to have lung cancer is increased by $e^{3.07} = 21.54$; the odds for a person with swallowing difficulty to have lung cancer is increased by $e^{3.43} = 30.88$; the odds for a smoker to have lung cancer is increased by $e^{1.45} = 4.26$.

In the first two parts, we compared two GAMs and used stepwise regression to identify significant predictors. The generalized additive model we fitted gives us estimates of different coefficients of all the significant predictors. So we are able to understand how the predictors are related to the response and how much they contribute to the variation in the response.

In the next part, we are going to fit a few machine learning models to do predictions. We will identify the best model for prediction by comparing their prediction accuracy.

3.3. Machine Learning Models

There are 309 observations in the original data set. We divided the data set into training set and testing set. We randomly sampled 250 observations from the original data and used them as training set. We used the training set to train the models, and then used the testing set (the rest 59 observations) to predict.

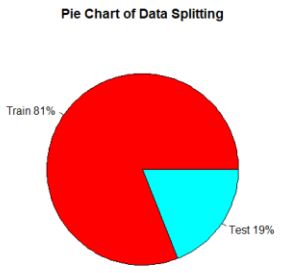


Figure 7. Data Splitting

3.3.1.1. Multiple Logistic Regression

In a Multiple Logistic Regression, we predict whether a patient have lung cancer or not using multiple predictors.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{15} X_{15}$$

where $X = (X_1, \ldots, X_{15})$ are 15 predictors.

We used maximum likelihood methods to estimate regression coefficients $\beta_0, \beta_1, \ldots, \beta_{15}$.

Here are the estimates of the coefficients, their standard errors, corresponding z values and p-values.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.96975	3.54114	-1.403	0.160487
GENDER	-1.03739	0.79561	-1.304	0.192266
AGE	-0.07513	0.05815	-1.292	0.196340
SMOKING	2.05633	0.87180	2.359	0.018338 *
YELLOW_FINGERS	1.06430	0.87318	1.219	0.222893
ANXIETY	1.08366	0.91301	1.187	0.235263
PEER_PRESSURE	2.81336	1.13690	2.475	0.013339 *
CHRONIC.DISEASE	3.82559	1.28256	2.983	0.002856 **
FATIGUE	6.02530	1.79832	3.351	0.000807 ***
ALLERGY	3.16495	1.45773	2.171	0.029920 *
WHEEZING	-0.89006	1.25771	-0.708	0.479146
ALCOHOL.CONSUMING	2.46425	0.98041	2.513	0.011955 *
COUGHING	5.76199	1.93755	2.974	0.002941 **
SHORTNESS.OF.BREATH	-1.85157	1.07470	-1.723	0.084913
SWALLOWING.DIFFICULTY	6.55397	2.45120	2.674	0.007500 **
CHEST.PAIN	-0.66835	1.05012	-0.636	0.524486

Table 1. Estimated coefficients of Logistic Regression

From the table above, we know that *smoking*, *peer pressure*, *chronic disease*, *fatigue*, *allergy*, *alcohol consuming*, *coughing*, *swallowing difficult* are significant predictors. This result is similar to the significant predictors identified by the stepwise regression. The only difference is that the stepwise regression also identified yellow fingers as a significant predictor. We know yellow fingers are caused by tobacco stain on fingers. It is a clinical sign of smoking. It can be treated as a marker of tobacco-related disease.

The logistic regression coefficient β associated with a predictor X is the expected change in log odds of having the outcome per unit change in X. So increasing the predictor by 1 unit (or going from 1 level to the next) multiplies the odds of having the outcome by e^{β}.

From the estimates of the coefficients of the Logistic Regression model, we know that if we have all other predictors fixed, the odds for a smoker to have lung cancer is increased by $e^{2.06} = 7.85$; the odds for a patient with peer pressure to have lung cancer is increased by $e^{2.81} = 16.61$; the odds for a patient with chronic disease to have lung cancer is increased by $e^{3.83} = 46.1$; the odds for a patient experiencing fatigue to have lung cancer is increased by $e^{6.03} = 415.72$; the odds for a patient with allergy to have lung cancer is increased by $e^{3.16} = 23.57$; the odds for a patient consuming alcohol to have lung cancer is increased by $e^{2.46} = 11.7$; the odds for a patient who coughs to have lung cancer is increased by $e^{5.76} = 317.35$; the odds for a patient with swallowing difficulty to have lung cancer is increased by $e^{6.55} = 699.24$. Swallowing difficulty, fatigue and coughing are the three most significant predictors of lung cancer.

The following table shows the prediction result we got from a logistic regression model:

	Predicted No	Predicted Yes
Actual No	5	5
Actual Yes	1	48

Table 2. Confusion table of Logistic Regression

The prediction error rate of logistic regression is: $\frac{1+5}{5+48+1+5} = 0.102$

3.3.1.2. Tree Based Methods

We used a classification tree to predict a qualitative response: having lung cancer vs no lung cancer. For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. To grow a classification tree, we use recursive binary splitting. We use classification error rate to evaluate the prediction accuracy of a classification tree. The classification error rate is the fraction of the training observations in that region that do not belong to the most common class. However, since classification error rate is not sufficiently sensitive for tree-growing, so we use other two measures: Gini index and entropy.

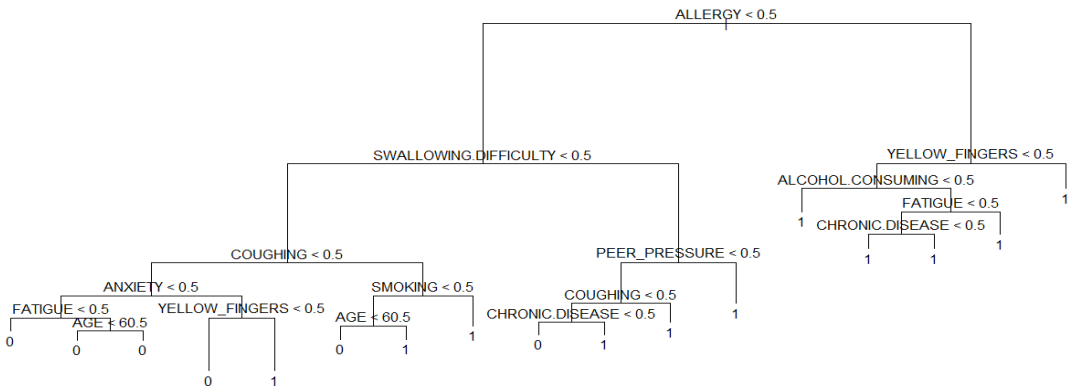


Figure 8. Unpruned Classification Tree

By inspecting Figure 7, we found a surprising characteristic: some of the splits yield two terminal nodes that have the same predicted value. For instance, consider the split *yellow fingers < 0.5* on the right of the unpruned tree. Regardless of the value of *yellow fingers*, a response value of yes is predicted for those observations. The split is performed only because it leads to increased node purity. Even though the split *yellow fingers < 0.5* doesn't reduce the classification error, it improves the Gini index and the entropy, which are more sensitive to node purity.

The prediction error rate of the classification tree is 0.119. This tree was grown to full depth, and might be too variable. In order to reduce the tree complexity, we used cross validation to identify the tree size that produces lowest misclassification rate.

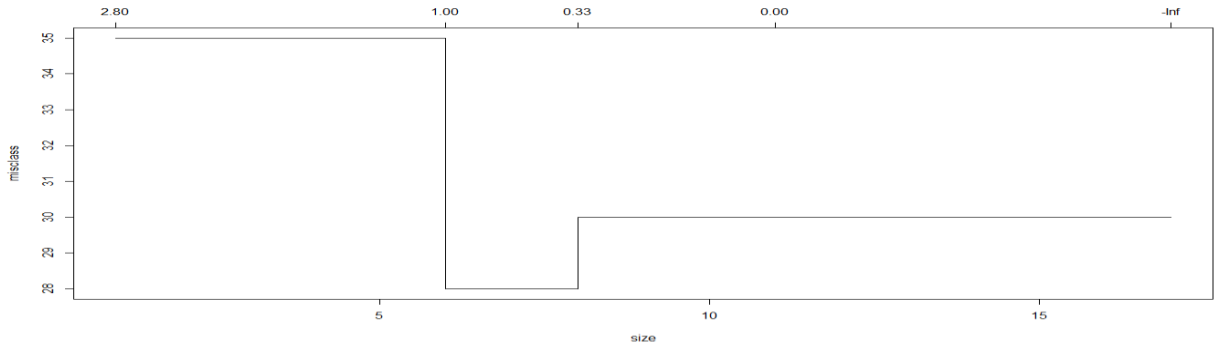


Figure 9. Misclassification rate against size of tree

From Figure 8, we can see that a tree with 6, 7 or 8 terminal nodes produces lowest misclassification rate. Therefore, we chose the classification with 6 terminal nodes.

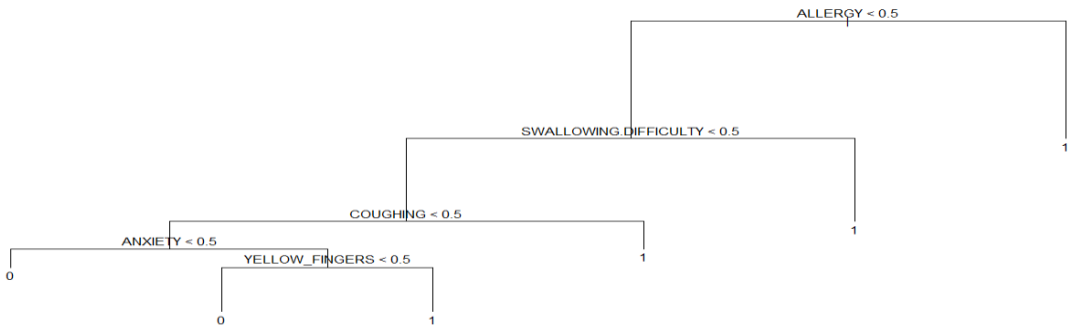


Figure 10. The Pruned Classification Tree

A patient with no *allergy*, no *swallowing difficulty*, no *coughing* and no *anxiety* is predicted to not have lung cancer. A patient with no *allergy*, no *swallowing difficulty*, no *coughing*, but has *anxiety* and *yellow fingers* is predicted to have lung cancer. A patient with *allergy* is predicted to have lung cancer. A patient with *allergy* and *swallowing difficulty* is predicted to have lung cancer. A patient with *allergy*, *swallowing difficulty* and *coughing* is predicted to have lung cancer.

The prediction error rate of the pruned tree is 0.102 which is improved significantly.

3.3.1.3. Random Forest

Decision trees suffer from high variance. It means if we split the training data into two parts at random and fit decision trees to both halves, the results that we get could be quite different. So here we are going to use a statistical learning method with low variance: random forest.

We build a number of decision trees on bootstrapped training samples. When building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and we choose $m \approx \sqrt{p}$, that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (3 out of 15 for our lung cancer data).

```
> print(rf.lcc)
Call:
 randomForest(formula = LUNG_CANCER ~ ., data = lung_cc, mtry = 3,  ntree = 10000, importance = TRUE, subset = train)
               Type of random forest: classification
                     Number of trees: 10000
No. of variables tried at each split: 3
        OOB estimate of error rate: 10%
```

Confusion matrix:

	0	1	class.error
0	17	16	0.48484848
1	9	208	0.04147465

The confusion matrix is also known as the error matrix that shows the visualization of the performance of the classification model. The out of bag estimate of prediction error rate is 0.1.

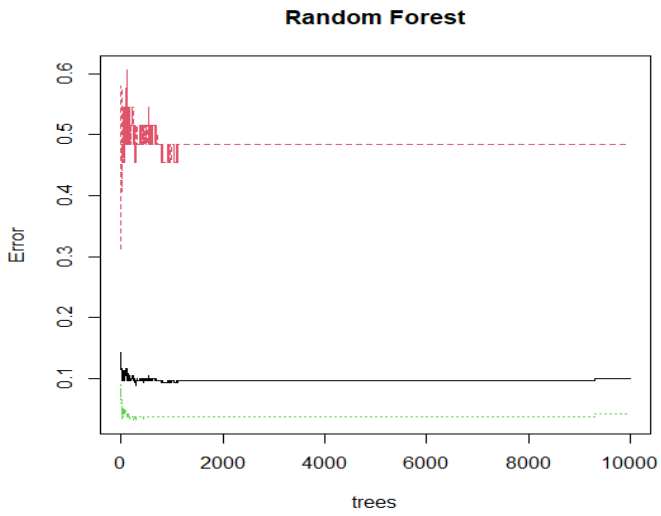


Figure 11. Classification error against number of trees

Figure 10 shows the number of decision trees developed using the classification model for random forest algorithms, i.e. 10000 decision trees. As we can see the classification errors become stable after 2000 trees.

The following R output shows the importance of all predictors in the model.

> importance(rf.lcc)

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
GENDER	12.115474	17.263108	21.7647338	1.693064
AGE	-9.386666	7.754418	0.9007504	6.945716
SMOKING	7.814522	11.221880	13.9340507	1.483324
YELLOW_FINGERS	40.565845	57.729060	67.6345506	3.632339
ANXIETY	31.006388	51.066970	58.8804945	2.986496
PEER_PRESSURE	34.654262	51.209980	59.5169287	3.121805
CHRONIC.DISEASE	13.757288	30.644608	32.9380042	2.322190
FATIGUE	31.231114	29.307926	40.4655019	2.676079
ALLERGY	78.129840	32.614039	72.2055762	5.047228
WHEEZING	36.851370	10.780785	31.8715231	2.331312
ALCOHOL.CONSUMING	66.485084	57.210526	77.7258862	4.563822
COUGHING	47.716167	24.649163	48.0378352	3.270282
SHORTNESS.OF.BREATH	25.060193	15.363812	27.0042660	1.932821
SWALLOWING.DIFFICULTY	45.630689	35.552073	54.1376866	3.349959
CHEST.PAIN	21.259105	11.780724	23.0631446	1.774113

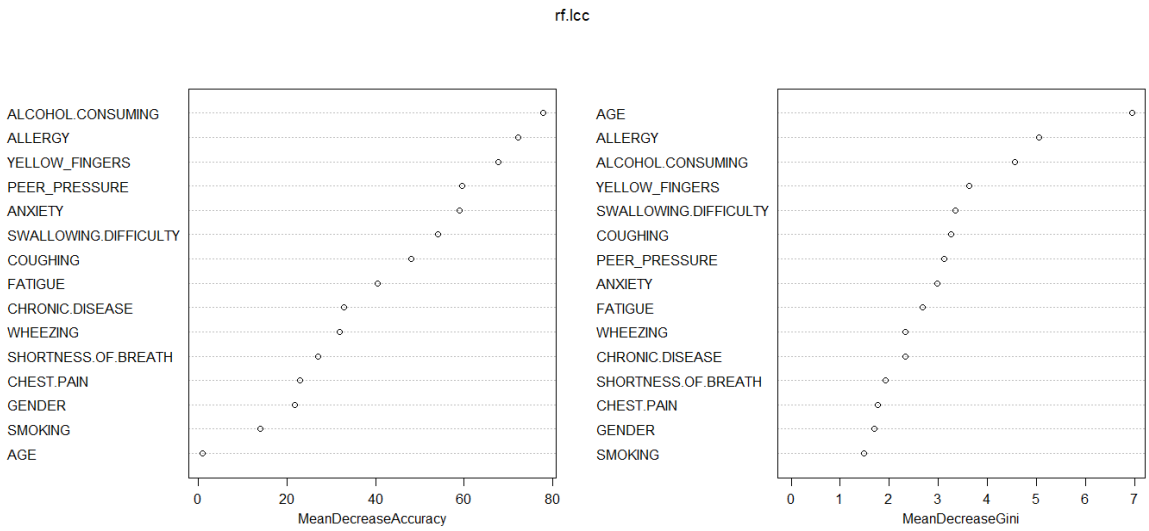


Figure 12. Influence of Predictors

In Figure 12, the x-axis displays the average decrease in prediction accuracy and Gini index of the classification trees based on splitting on the various predictors displayed on the y-axis. From the plot we can see that *age* is the most important predictor variable, followed closely by *allergy* and *alcohol consuming*. Table 3 shows the prediction result of the random forest:

	Predicted No	Predicted Yes
Actual No	4	2
Actual Yes	2	51

Table 3. Confusion Table of Random Forest

The margin of a data point is defined as the proportion of votes for the correct class minus maximum proportion of votes for the other classes. Thus under majority votes, positive margin means correct classification, and vice versa.

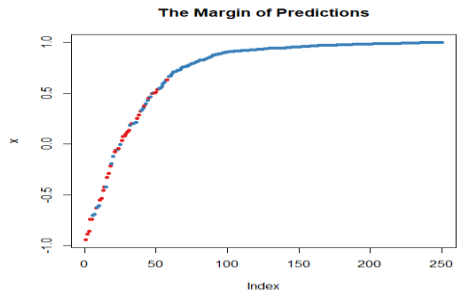


Figure 13. Margins of Random Forest Predictions

We can also tune the hyper parameter mtry.

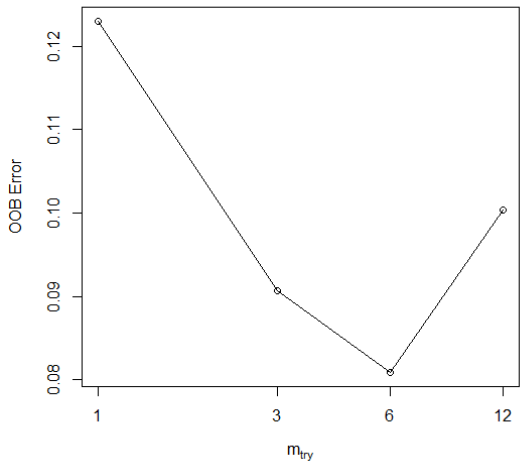


Figure 14. OOB values against various values of mtry

```
> print(tune.rf)
    mtry         OOBError
1.     OOB  1 0.12621359
3.     OOB  3 0.08737864
6.     OOB  6 0.08090615
12.    OOB 12 0.09061489
```

We can see that mtry=6 produces lowest out of bag error rate, so we fit another random forest with mtry=6, but it didn't reduce our prediction error rate in our lung cancer data. Therefore, we will stay with mtry=3.

3.3.1.4. Summary

Table 4 and Figure 15 show us the prediction error rates of the four classifiers we trained for our lung cancer data.

Classifiers	Logistic	Tree	Pruned Tree	Random Forest
Test Error Rate	0.102	0.119	0.102	0.068

Table 4. Test error rates of four classifiers

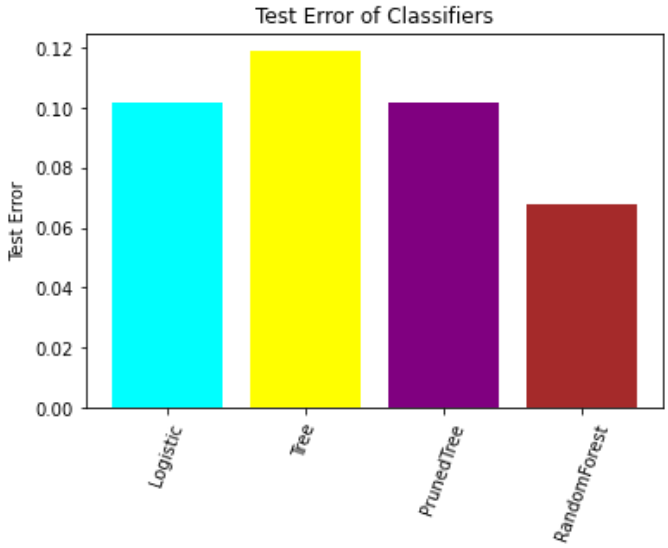


Figure 15. Bar graph of test error rates

From the bar graph above, we can see that Random Forest is the best performer with the lowest prediction error rates of 0.068, so it has the most prediction power. However, Random Forest is a complex black box model with low interpretability. It cannot help us identify the significant predictors. Logistic Regression model and the pruned tree give us the same prediction error of 0.102.

4. Results

We use Akaike information criterion (AIC) and Bayesian Information Criterion (BIC) to select the best model for our lung cancer data analysis. The following table shows us the AIC and BIC values of the Generalized Additive Models and the Logistic Regression model.

Models	GAM1	GAM2	Stepwise Regression	Logistic Regression
AIC	124	122	117	97
BIC	184	186	154	153

Table 5. AIC & BIC

Both AIC and BIC favor the Logistic Regression model. Therefore, the Logistic Regression model is the best model with good interpretability and high prediction accuracy.

5. Conclusions

In this analysis, we fitted several regression models and three tree based models. By comparing their prediction error rates, AIC and BIC, we think Logistic Regression Model is the best model for our lung cancer data. By fitting a Logistic Regression model, we identified all the significant predictors for whether a person have lung cancer.

However, as I dug deeper and deeper in this data analysis process. I realized that the data is flawed due to a bad experiment design. For example, there are 15 predictors in the data set to predict if a patient has lung cancer or not, such as *"smoking", "anxiety", "wheezing", "coughing", "shortness of breath", "chest pain", "alcohol consuming", "swallowing difficulty"* and so on. All of these predictors are binary data except *age*. If we have a better experiment design and get more accurate data, we will be able to detect more accurate relationships between predictors and the response. We can improve the experiment design by asking more specific questions, such as how frequent and how much does a patient smoke? How much chest pain does a patient feel? How often and how bad does a patient cough? How difficult does a patient swallow? If all these variables can be measured on different degrees, maybe on a 1-5 scale, we will be able to get more accurate information to predict if a person has lung cancer or not.

There is also sampling bias in this data set. 99% of people are older than 40. 95% of people are older than 50. This is not a good representation of the whole population. It's biased towards older people. So this data set couldn't reveal the true relationship between age and lung cancer.

In addition, we are facing an imbalanced classification problem here. 270 out 309 people have lung cancer, which is 87% of the whole data set. So we have a classification predictive modeling problem where the distribution of examples across the classes is not equal. This is caused by the biased sampling. Therefore, the imbalance can be corrected by improved sampling methods. Most machine learning algorithms work best when the number of instances of each classes are roughly equal. When the number of instances of one class far exceeds the other, the machine learning classifier tends to be more biased towards the majority class, causing bad classification of the minority class.

This is an example that tells us data professionals should get involved in a project right from the beginning. Statistical consulting should start from experiment design and data collecting. If data is collected from a bad experiment design, the following analysis will not be accurate and meaningful.