Prediction of Medical Insurance Cost

Group A

PoAn Chen, Zhongwei Wang, Yibing Chen

Predictive Modeling

Professor： Lina Xu

04/26/2022

Distribution of Labor

Data Overview: Zhongwei Wang

Data Visualization: Yibing Chen

Best Select: Yibing Chen

Ridge & Lasso: Zhongwei Wang

Decision Tree: Po An Chen

Random Forest & Boosting: Po An Chen

Conclusion: Po An Chen&Zhongwei Wang

Data Overview

Target	Charges	How much health insurance premiums cost
Predictors	Age	Age of primary beneficiary
	Sex	Insurance contractor gender, female, male
	BMI	Body mass index
	Smoker	Smoking
	Children	Number of children covered by health insurance
	Region	The beneficiary's residential area in the U.S, northeast, southeast, southwest, northwest

Data Overview

Age&Sex

Medicare cost increases with age with some exceptions

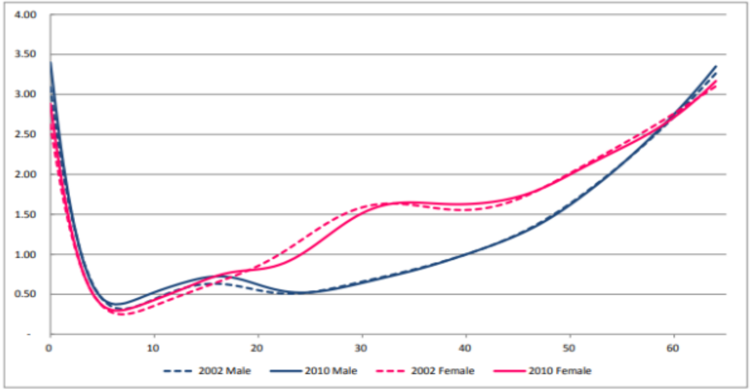
Relatively high and gradually decrease from 2 to 5

Male: Declines between 15 to 25 years old

Female: Stable between 30- to 40 years old

https://www.soa.org/globalassets/assets/files/research/projects/research-health-care-birth-death-report.pdf

Commercial Aging Curve (Chart 1)



Medicare Aging Curve (Chart 10)

Data Overview

BMI(Body mass index)

Weight divided by height squared (kg/m^2)

Normal range: 18.5 - 24.9

Too high or too low will lead to an increase in medicare expense

BMI, basic categories

Category	BMI (kg/m^2)[c]	BMI Prime[c]
Underweight (Severe thinness)	< 16.0	< 0.64
Underweight (Moderate thinness)	16.0 – 16.9	0.64 – 0.67
Underweight (Mild thinness)	17.0 – 18.4	0.68 – 0.73
Normal range	18.5 – 24.9	0.74 – 0.99
Overweight (Pre-obese)	25.0 – 29.9	1.00 – 1.19
Obese (Class I)	30.0 – 34.9	1.20 – 1.39
Obese (Class II)	35.0 – 39.9	1.40 – 1.59
Obese (Class III)	≥ 40.0	≥ 1.60

https://en.wikipedia.org/wiki/Body_mass_index

Data Overview

Medicare expense

Smoker:

Smoker > Non-Smoker

Number of children:

More children usually result in more Medicare cost

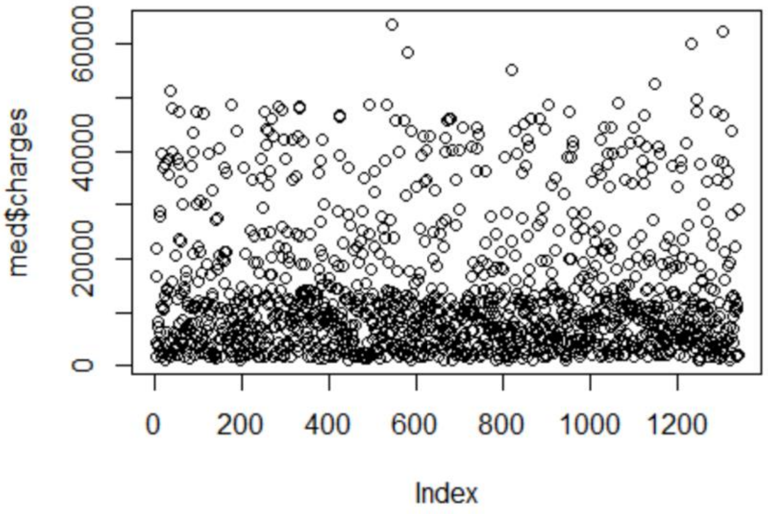
Region:

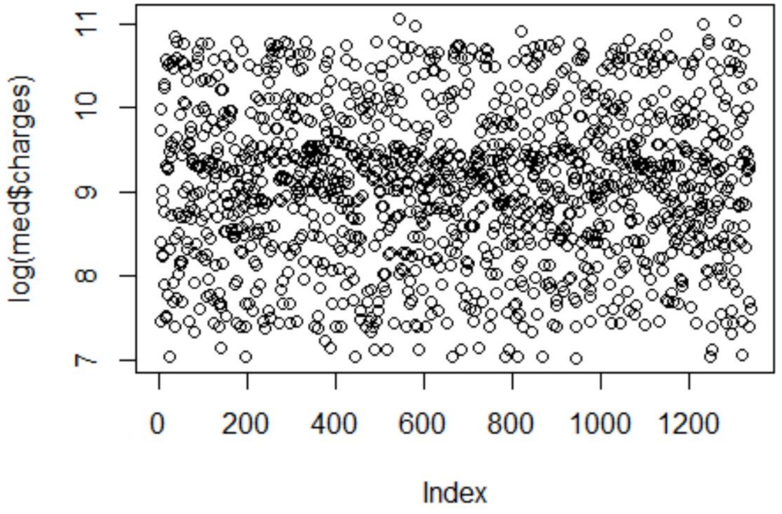
Regional differences may lead to differences in medical spending

Summary of Data

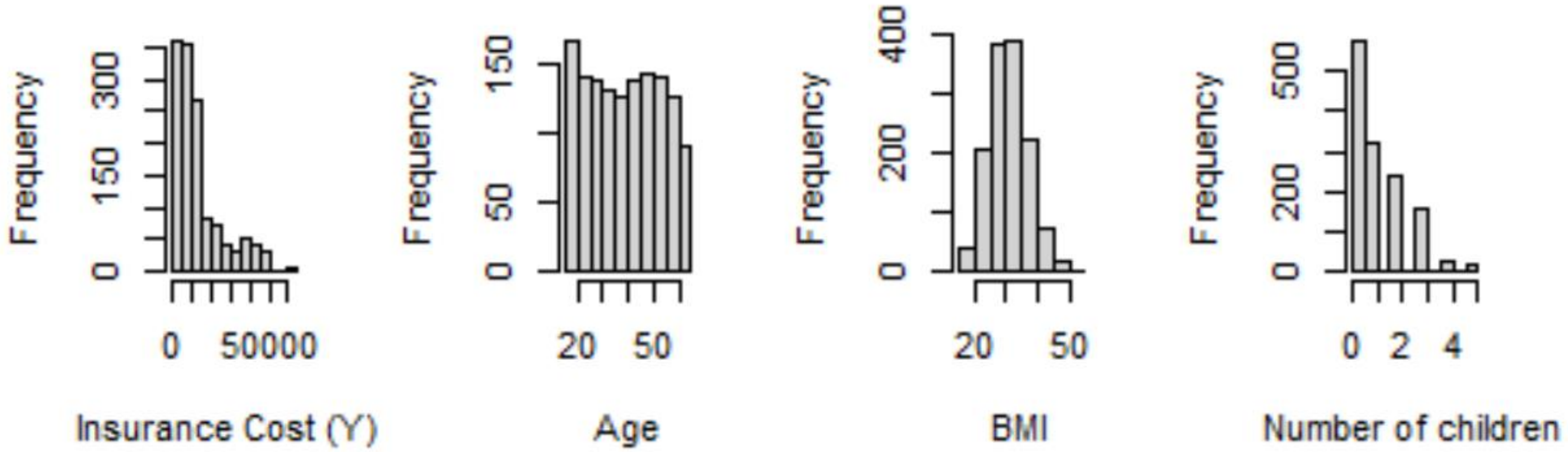
```
summary(med)

##       age            sex           bmi           children       smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   No :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   Yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southwest:364   Median : 9382
##  southeast:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Data Visualization

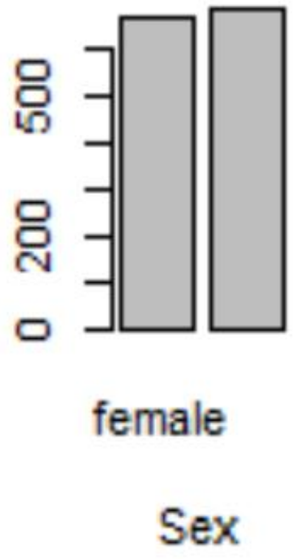


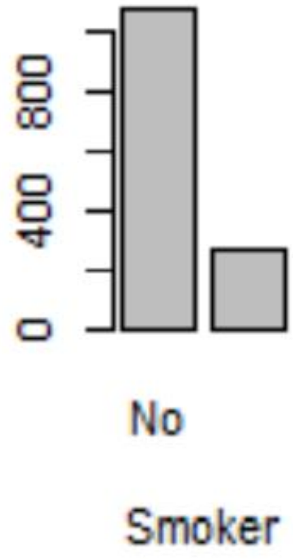


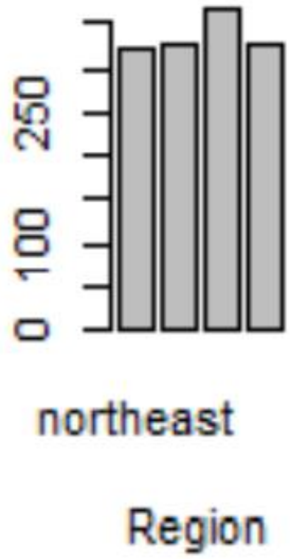
Data Visualization



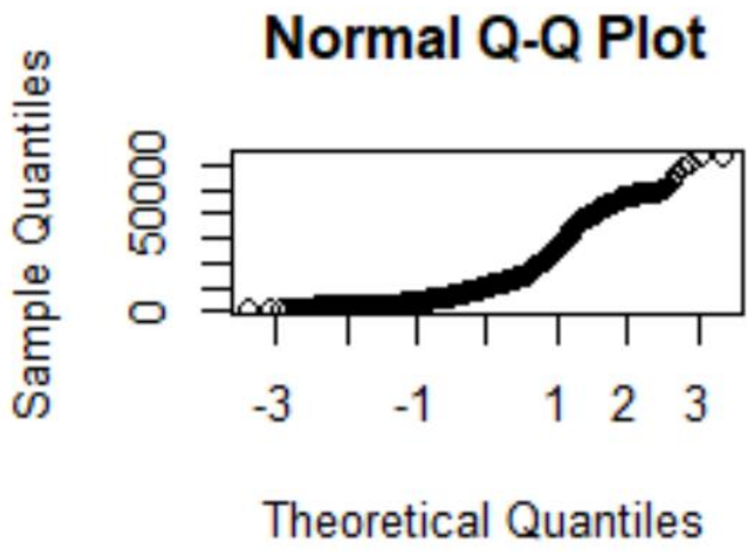
Data Visualization

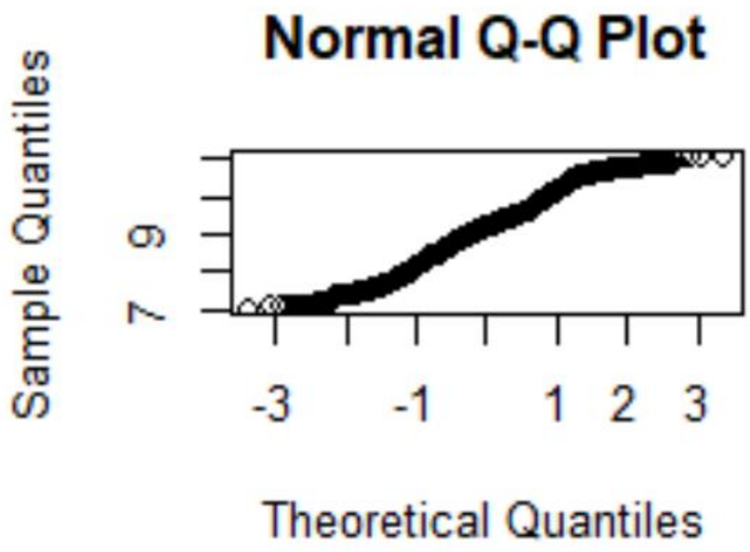






Data Visualization





Fitting Model

```
## (Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
## age              0.0345816  0.0008721  39.655  < 2e-16 ***
## sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
## bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
## children         0.1018568  0.0100995  10.085  < 2e-16 ***
## smokerYes        1.5543228  0.0302795  51.333  < 2e-16 ***
## regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
## regionsouthwest -0.1571967  0.0350828  -4.481 8.08e-06 ***
## regionsoutheast -0.1289522  0.0350271  -3.681 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Model Selection

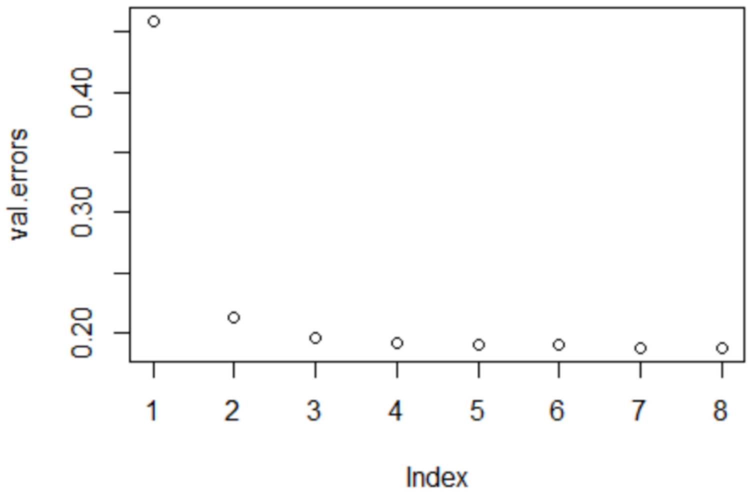
- Full dataset

```
##   model        r2     adjr2         cp       bic
## 1     1 0.4428978 0.4424809 1856.61244  -768.341
## 2     2 0.7395465 0.7391564  159.65828 -1778.456
## 3     3 0.7572654 0.7567195   60.17950 -1865.527
## 4     4 0.7621566 0.7614429   34.16713 -1885.564
## 5     5 0.7639274 0.7630413   26.02496 -1888.365
## 6     6 0.7657049 0.7646487   17.84507 -1891.278
## 7     7 0.7673647 0.7661403   10.33949 -1893.591
## 8     8 0.7679478 0.7665509    9.00000 -1889.750
```

Model Selection

- Train-test splitting methods

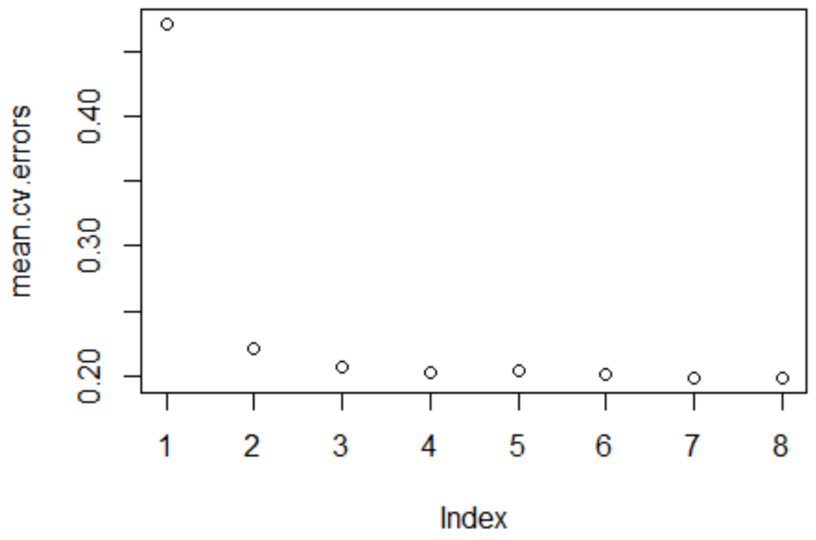
```
##   model val.errors
## 1     1  0.4586110
## 2     2  0.2126700
## 3     3  0.1962058
## 4     4  0.1916911
## 5     5  0.1911752
## 6     6  0.1905866
## 7     7  0.1886521
## 8     8  0.1881355
```



Model Selection

- Cross validation method
- 10-fold

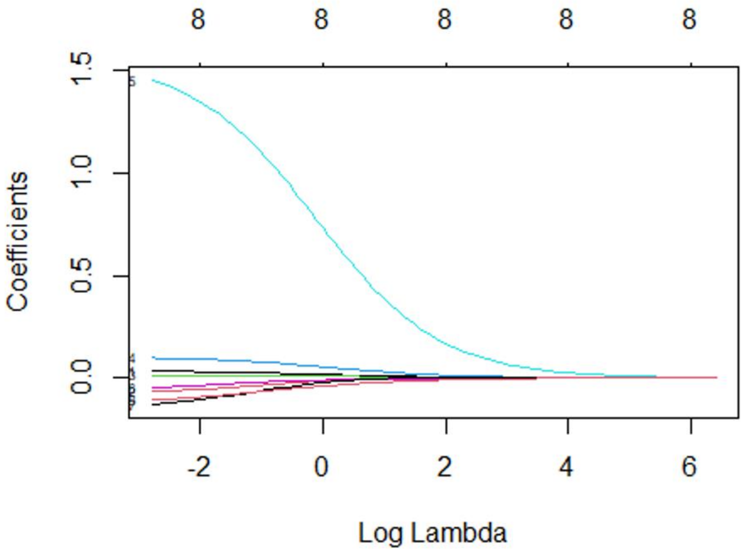
```
##   model mean.cv.errors
## 1     1      0.4711180
## 2     2      0.2209773
## 3     3      0.2061846
## 4     4      0.2021111
## 5     5      0.2033770
## 6     6      0.2013673
## 7     7      0.1986568
## 8     8      0.1982912
```



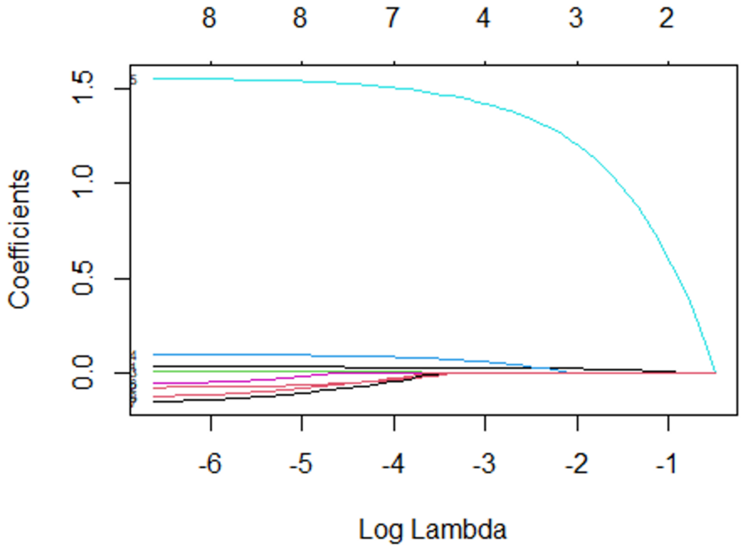
Model Selection

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.2877234  0.0387040 188.294   <2e-16 ***
## age         0.0352849  0.0008839  39.919   <2e-16 ***
## children    0.1016311  0.0102990   9.868   <2e-16 ***
## smokerYes   1.5442724  0.0307364  50.242   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4535 on 1334 degrees of freedom
## Multiple R-squared:  0.7573, Adjusted R-squared:  0.7567
## F-statistic:  1387 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Ridge & Lasso

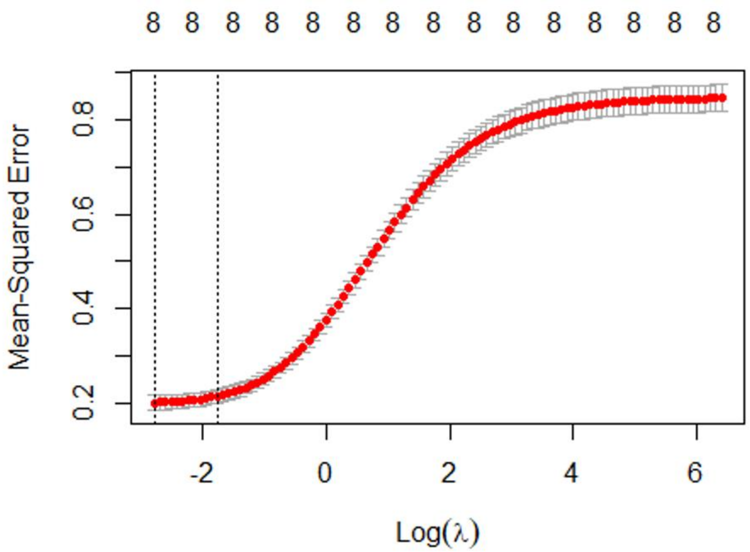


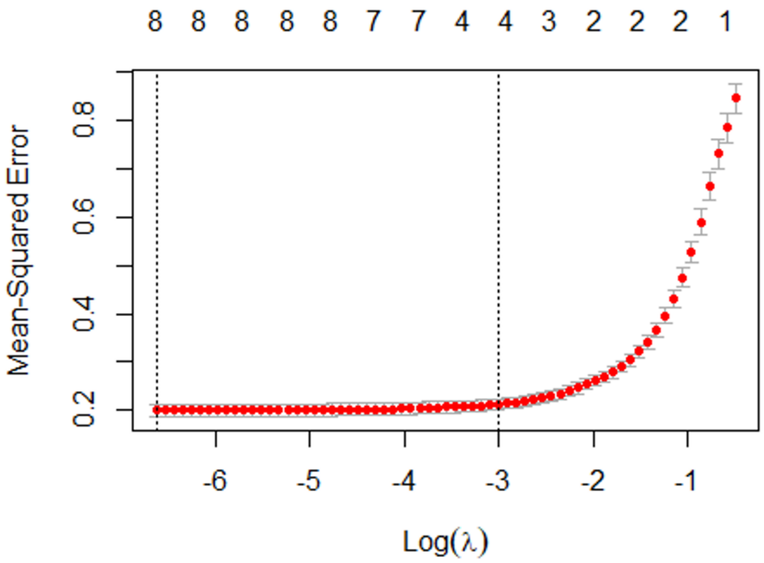
$$SSE + \lambda \sum_{j=1}^{p} b_j^2$$



$$SSE + \lambda \sum_{j=1}^{p} |b_j|$$

Ridge & Lasso


8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
Mean-Squared Error
0.2
0.4
0.6
0.8
-2
0
2
4
6
Log(λ)

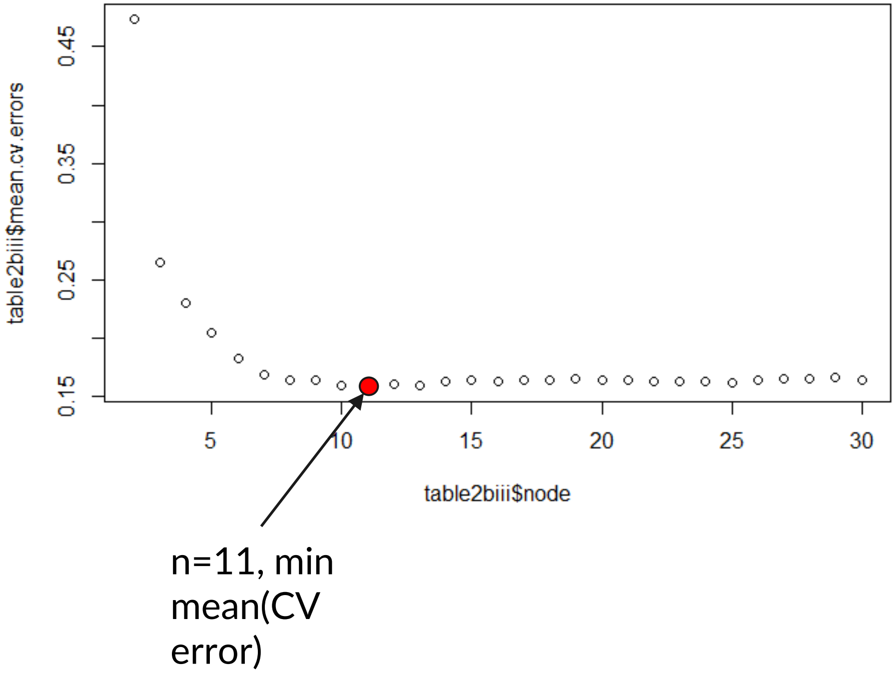

8 8 8 8 8 7 7 4 4 3 2 2 2 1
Mean-Squared Error
0.2
0.4
0.6
0.8
-6
-5
-4
-3
-2
-1
Log(λ)

Decision Tree

- Recursive binary splitting process that splits data into a finite set of non-overlapping regions.
- Pro: easy to interpret and display; no probability distribution assumption
- Con: vulnerable to overfitting.
- test MSE: 0.2464

Pruned Decision Tree

- Pruning is necessary to reduce the size of a tree and remove less valuable splits.
- Pro: reduces overfitting and can lead to a simpler, more interpretable tree; automatically performs variable selection.
- Using Cross Validation to find the best pruned tree, n=11

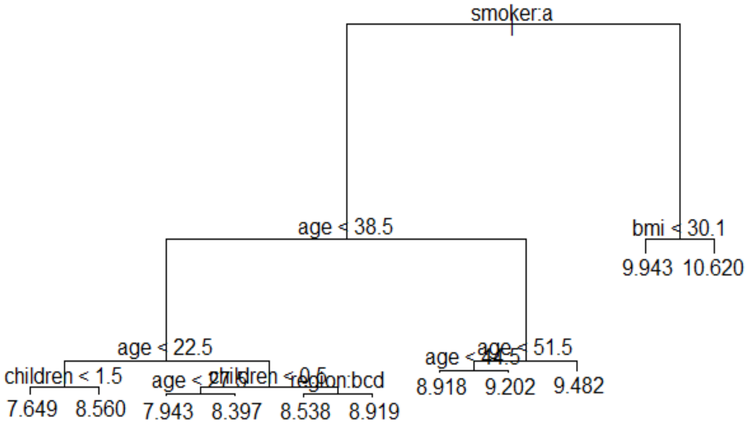


Pruned Decision Tree

- Using Cross Validation to find the best

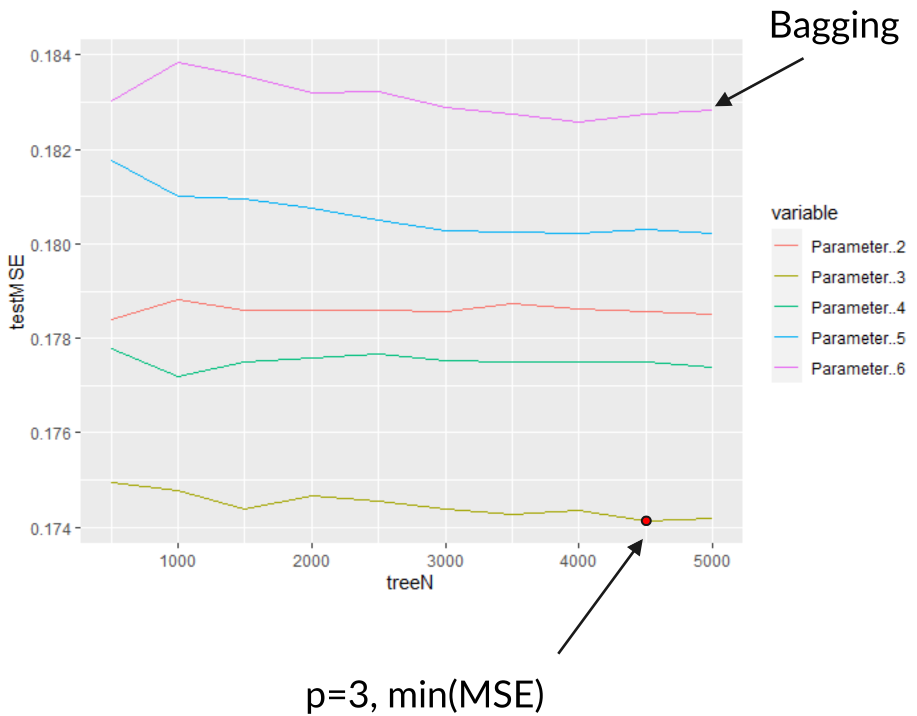
pruned tree, n=11

- test MSE: 0.1890



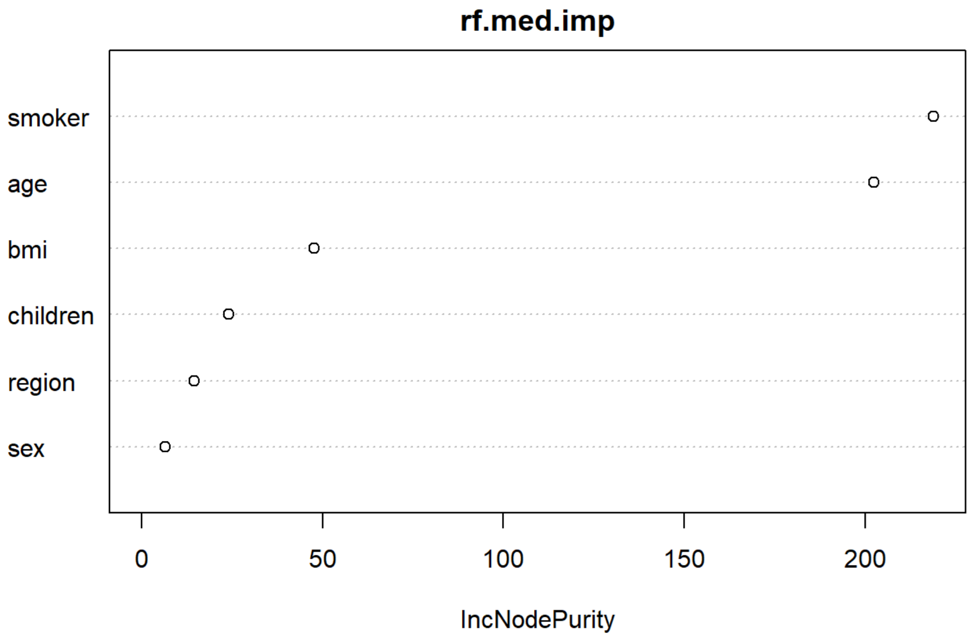
Random Forest

- Combines the results of a set of decision trees fitted to a different bootstrapped sample of the training data, then using the average to make a final prediction.
- Pro: reduces overfitting and variance of the base tree, leading to higher prediction accuracy.
- Con: loses the interpretability of decision trees and is computationally intensive.
- Randomly select 3 parameter and build 4500 trees -> minimum MSE



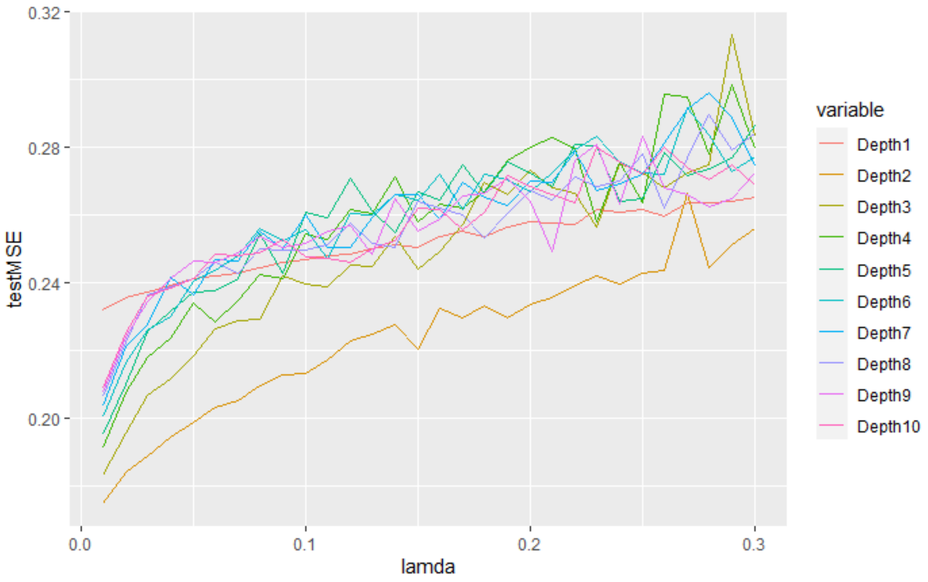
Random Forest

- Variable importance
- smoker, age
- The result is same as the best selection in the linear regression model with p=2.



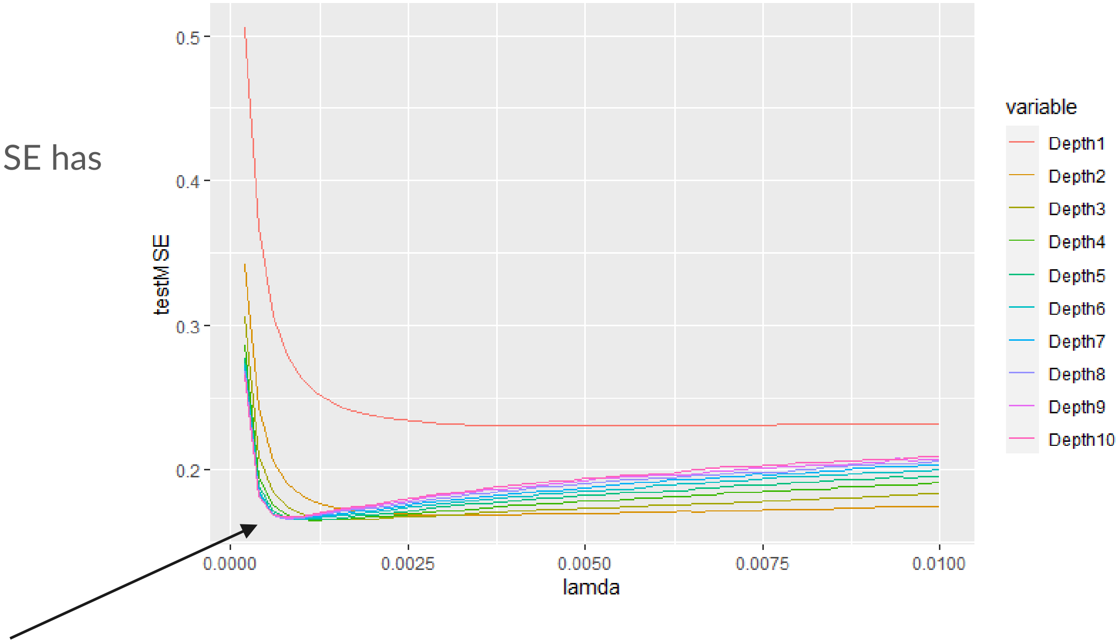
Gradient Boost Trees

- The boosting approach learns slowly controlled by shrinkage parameter λ.
- Given the current model, boosting fit a decision tree to the residuals from the model.
- Lamda=0.01 has the smallest test MSE in each Depth.



Gradient Boost Trees

- Lambda: Around 0.0012, the test MSE has the minimum.



Model Comparison & Observations

- Accuracy
- Interpret ability

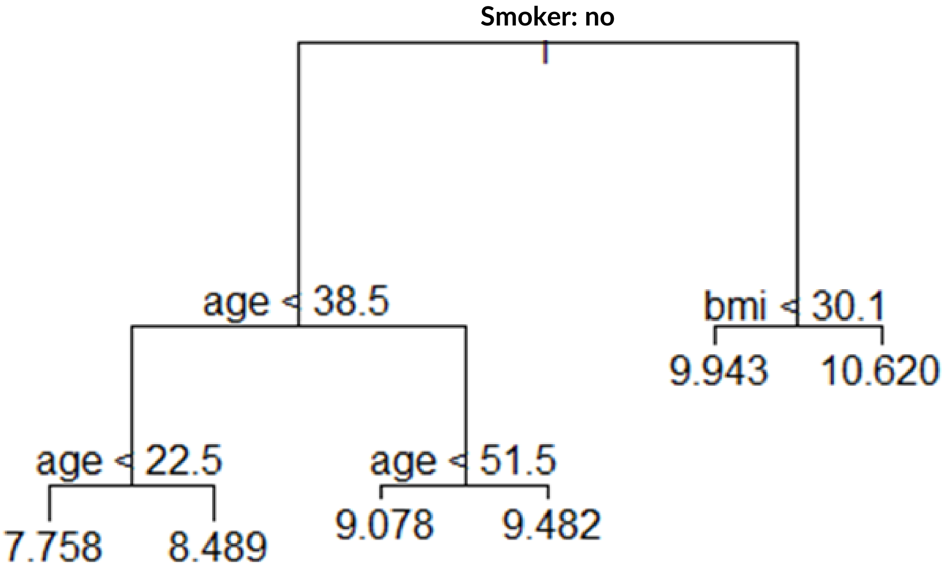
model	MSE.lm	branchs	MSE.tr	treeN	MSE.rf_p3	Depth	MSE.bo_l0.0012
<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
1	0.4586110	5	0.2207118	1500	0.1743677	Depth3	0.1673605
2	0.2126700	6	0.2029762	2000	0.1746719	Depth4	0.1652118
3	0.1962058	7	0.2043064	2500	0.1745387	Depth5	0.1656875
4	0.1916911	8	0.1908290	3000	0.1743732	Depth6	0.1667413
5	0.1911752	9	0.1904854	3500	0.1742815	Depth7	0.1675478
6	0.1905866	10	0.1907478	4000	0.1743549	Depth8	0.1684804
7	0.1886521	11	0.1890266	4500	0.1741359	Depth9	0.1692687
8	0.1881355	12	0.1915374	5000	0.1741971	Depth10	0.1702054

Final Model

- Interpret ability
- Pruned tree with 6 branches
- Test MSE=0.2030
- Question:

Man, smoker, age 37, children 0 , BMI 32,

Northeast



Final Model

Age & Sex: Significantly affects premiums

Long term: Different rates for each age and gender

Short term: In same level of age, premium are the same regardless of sex

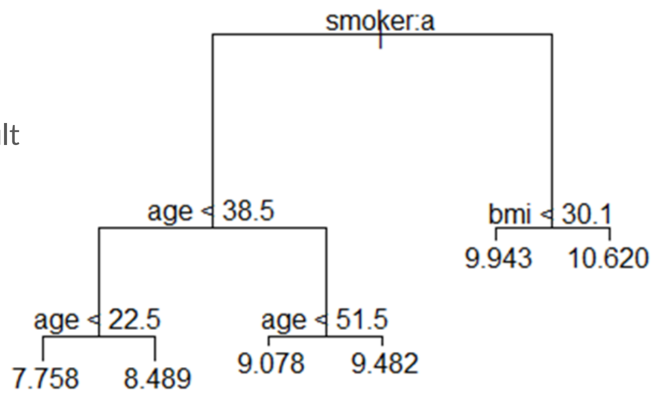
BMI: May affect the underwriting results of long-term insurance, and may result in more premium or decline

Smoker: Usually affects premiums, but does not result in a decline

Children: Not that significant as other factors

Region: The partition is too broad, the difference is decreased

Smoker: no



Thank you !