Insurance Forecast

Aren Zita
Chamodi Basnayake
Weibin Huang

Introduction

Given patient information, can you accurately predict insurance costs?

1. Dataset
2. Descriptive Statistics
3. Variable Selection
4. LASSO
5. Multicollinearity
6. Introduce interaction term
7. Box Cox
8. Cross Validation
9. Final Models
10. Decision Tree
11. Conclusion

Dataset

1338 data points

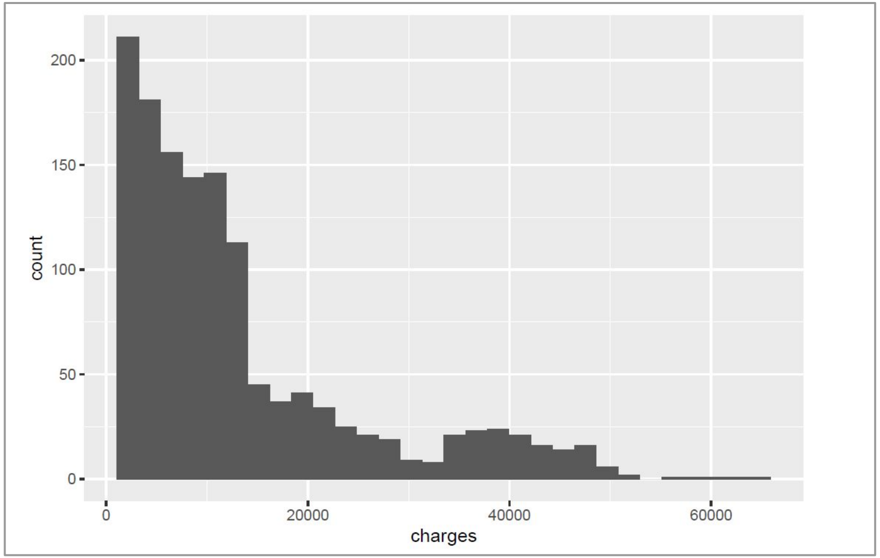
Response variable

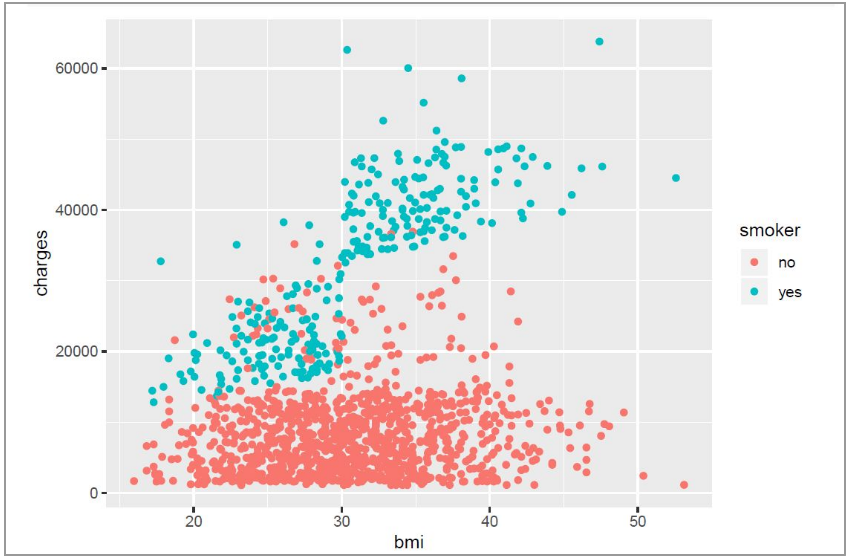
Charges	[$ 1.12K - 63.8K]

6 Regressor variables

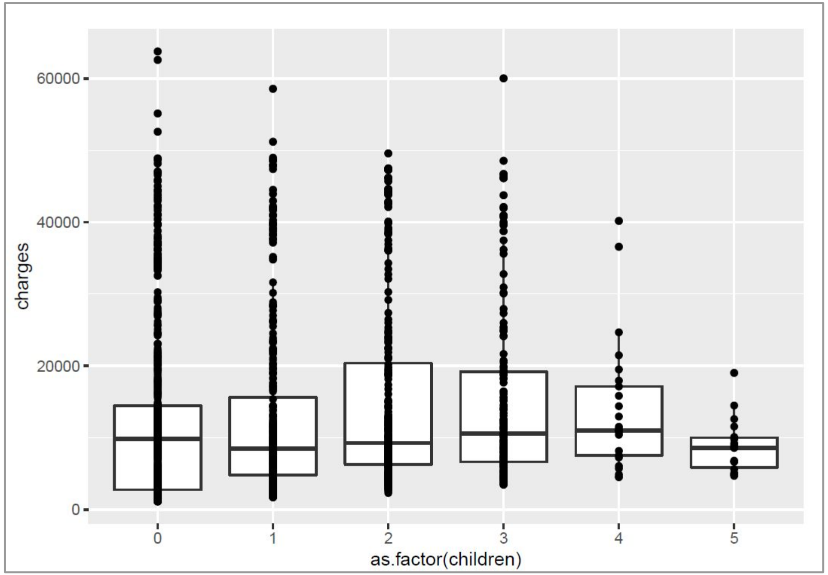
Age	[18 - 64]
Sex	[Female, Male]
Bmi	[16 - 50]
Children	[0, 1, 2, 3, 4, 5]
Smoker	[true, false]
Region	[SE, SW, NE, NW]

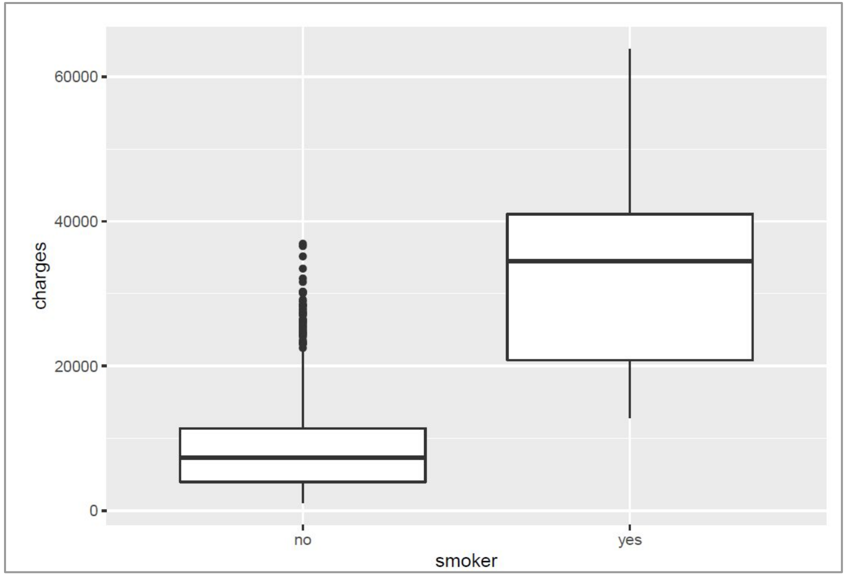
Descriptive Statistics





Descriptive Statistics





Variable Selection

Forward Selection

Backward Selection

Stepwise Regression

```
## Step:  AIC=23314.58
## charges ~ smoker + age + bmi + children + region
##
##              Df  Sum of Sq        RSS   AIC
## <none>                     4.8845e+10 23315
## - region     3 2.3320e+08 4.9078e+10 23315
## + sex        1 5.7164e+06 4.8840e+10 23316
## - children   1 4.3596e+08 4.9281e+10 23324
## - bmi        1 5.1645e+09 5.4010e+10 23447
## - age        1 1.7151e+10 6.5996e+10 23715
## - smoker     1 1.2301e+11 1.7186e+11 24996
```

- Select all variables except Sex
- R^2 = 74%
- RSE = 6060

charges ~ a (age) + b (bmi) + c (children) + s (smoker) + r (region)

LASSO(least absolute shrinkage and selection operator)

One of the Shrinkage Method

- Shrinks coefficients estimates to zero
- Minimize the criterion

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

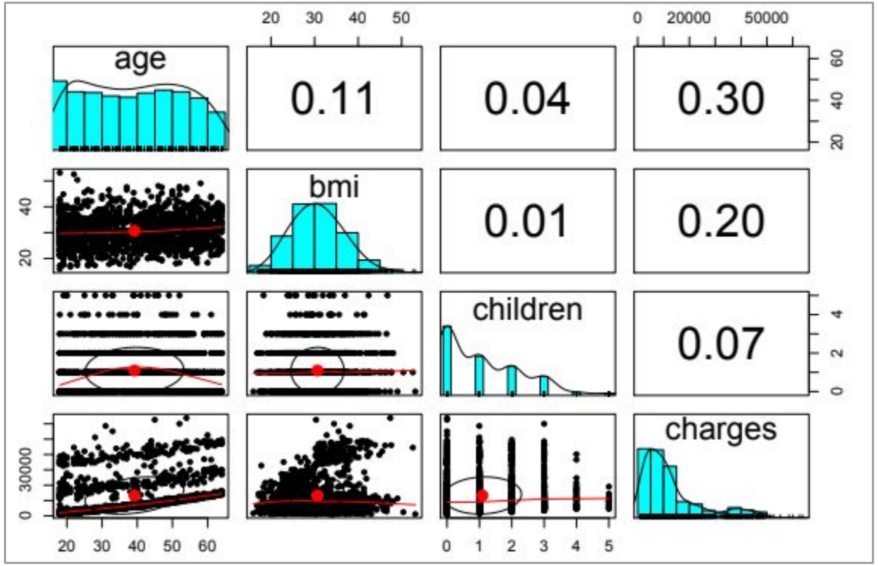
- Lasso coefficients

```
(Intercept)                age          sexmale              bmi         children
-11270.2740           250.9591           0.0000         314.5540         392.1943
  smokeryes  regionnorthwest  regionsoutheast  regionsouthwest
 23567.3485           0.0000        -393.8077        -380.5802
```

Multicollinearity

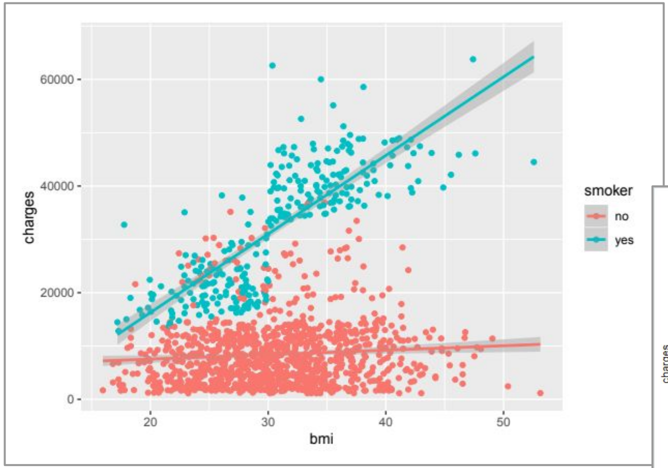
- Multicollinearity - High intercorrelations among the independent variables
  - High standard errors
  - Impacts significance

	GVIF	Df	GVIF^(1/(2*Df))
age	1.016188	1	1.008061
bmi	1.104197	1	1.050808
children	1.003714	1	1.001855
smoker	1.006369	1	1.003179
region	1.098870	3	1.015838

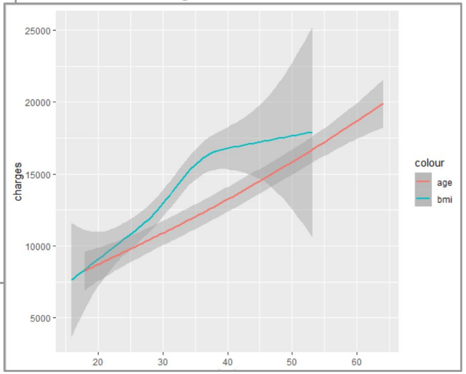


Introduce Interaction Terms

smoker::bmi



age::bmi



- Interaction effects exist?
- BMI, Age and Smoker
- Analyze slopes and significance

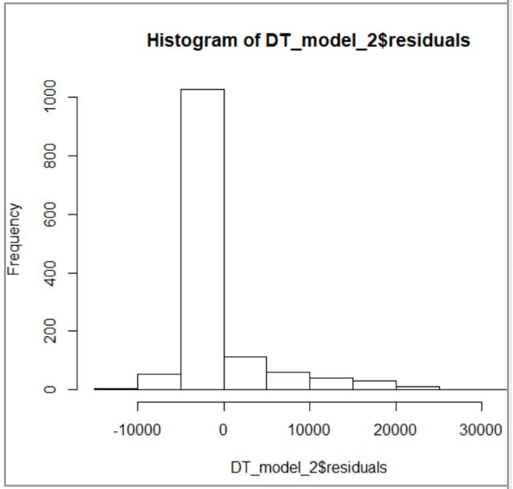
- R^2 = 84%
- RSE = 4851

charges ~ a (age) + b (bmi) + c (children) + s (smoker) + r (region) + sb (smoker::bmi)

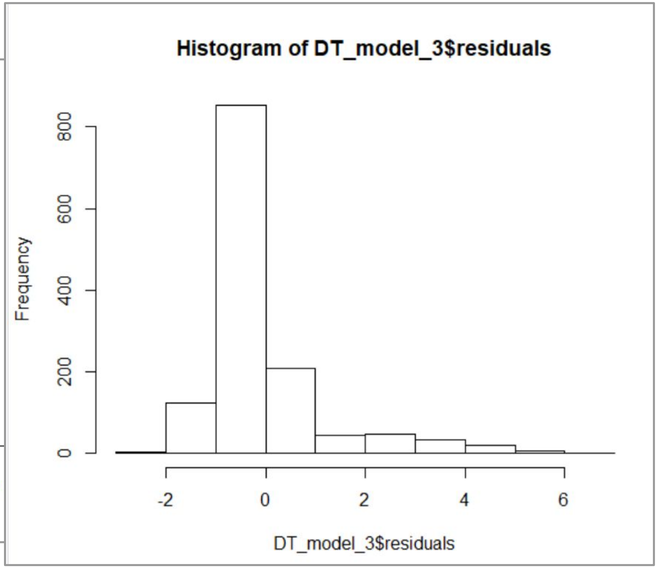
Box Cox Transformation

- Transform y variable
- Meet normality assumption
- λ = 0.262623

Before



After



- R^2 = 80%
- RSE = 1.181

charges ^(0.262623) ~ a (age) + b (bmi) + c (children) + s (smoker) + r (region) + sb (smoker::bmi)

k-fold Cross Validation

img: http://karlrosaen.com/ml/learning-log/2016-06-20/

- Original data is randomly split into k partitions
- Use one subsample as test data and other remaining subsamples as training data. Calculate the residual sum of squared.
- Repeat it k times, picking different test data each iteration.
- Calculate the mean of residual sum of squared to evaluate the model

10-fold Cross Validation of the Models

Model without the interaction term:

charges^0.1414 ~ smoker + age + bmi + children + region

```
1338 samples
   5 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1205, 1204, 1203, 1204, 1206, 1205, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.2246241  0.776145  0.143058
```

Model with the interaction term:

charges^0.2626 ~ smoker + age + bmi + children + smoker:bmi

```
1338 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1204, 1205, 1203, 1203, 1206, 1204, ...
Resampling results:

  RMSE      Rsquared  MAE
  1.182119  0.806899  0.7138265
```

Final Model

charges^0.2626 ~ smoker + age + bmi + children

+ smoker:bmi

R^2 Adj Value = 0.8074

Residual Standard Error = 1.181

```
Call:
lm(formula = charges^0.262626 ~ age + bmi + children + region +
    sex + smoker:bmi, data = data.insurance)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3805 -0.4977 -0.2375  0.0359  6.4309

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.319609   0.191614  32.981  < 2e-16 ***
age              0.090696   0.002319  39.113  < 2e-16 ***
bmi              0.013902   0.005593   2.486 0.013054 *
children         0.239781   0.026851   8.930  < 2e-16 ***
regionnorthwest -0.185733   0.092795  -2.002 0.045537 *
regionsoutheast -0.411757   0.093286  -4.414  1.1e-05 ***
regionsouthwest -0.363111   0.093117  -3.900 0.000101 ***
sexmale         -0.200442   0.064921  -3.088 0.002060 **
bmi:smokeryes    0.160602   0.002571  62.457  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.181 on 1329 degrees of freedom
Multiple R-squared:  0.8085,    Adjusted R-squared:  0.8074
F-statistic: 701.6 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Regression Trees

Why Use Regression Tree?

- Effective when there are different clusters of observations.

- It is easier to visualize the effectiveness of each regressor.

- It can be helpful when making a rational decision.

Procedure

1. Pick a model

2. Construct a Large Decision Tree

3. Apply Pruning to the Tree (Cross Validation)

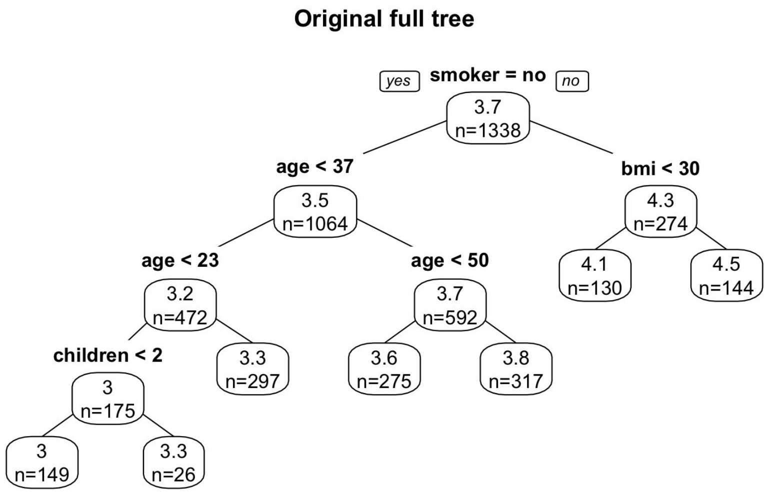
4. Decide a Final Tree

First Decision Tree

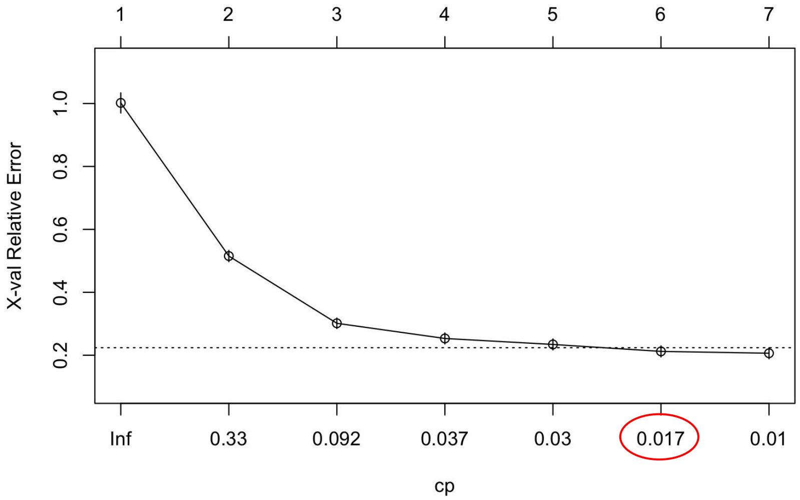
Model Used:

charges^0.1414~smoker+age+bmi+children+region

```
insurance.tree <- rpart(data = data.insurance,
                        final.model,
                        method = "anova")
```



Pruning (Cross-Validation of Regression Tree)



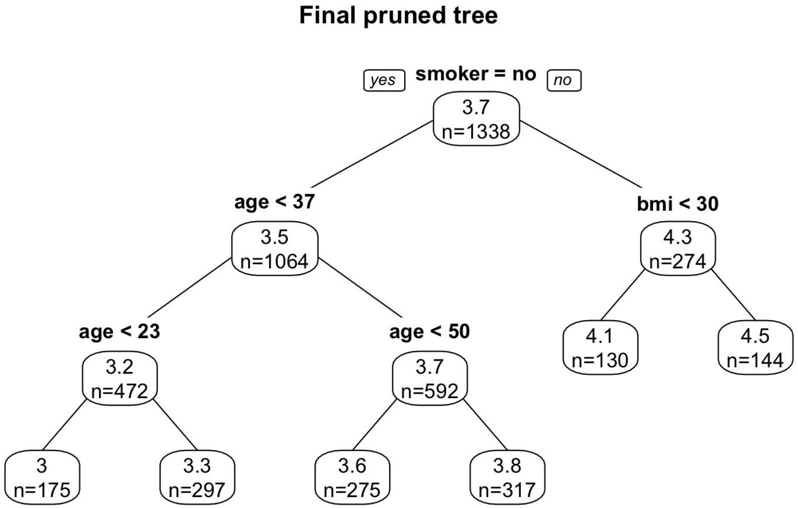
- The graph shows the Relative Error vs cp (Complexity Parameter)

$$\sum_{Leaves} (\text{RSS at each leaf}) + \lambda S$$

- The horizontal line represents the highest cross-validated error + 1 standard deviation of the error at the tree.

- Pick the cp at 0.017

Final Decision Tree



-People who smoke and are obese tend to have higher charges on insurance .

Height	Weight Range	BMI	Considered
5′ 9″	124 lbs or less	Below 18.5	Underweight
	125 lbs to 168 lbs	18.5 to 24.9	Healthy weight
	169 lbs to 202 lbs	25.0 to 29.9	Overweight
	203 lbs or more	30 or higher	Obese
	271 lbs or more	40 or higher	Class 3 Obese

ref: https://www.cdc.gov/obesity/adult/defining.html

Conclusion

- More complex transformations to better meet normality assumption
- Consider a Random Forest model
- Support Vector Machine might work better since there were several data apart on each other

THANK YOU

QUESTIONS?