Predicting AirBnB Rental Prices

Group 11: Trevor Isaacson, Jonathan Olavarria, Jasmine DeMeyer

12/10/2021

Motivation

- You are looking for some additional income and decide renting on AirBnB is the best option
- How much should you rent your extra space for?

Data

- In general, AirBnB data is very open and be easily accessed
- The original dataset is from a past Kaggle competition
  - Contained over 74,000 individual listings between 2011-2018
- For sake of time and processing power, we took a random sample of 17,500 from those 74,000 listings
- They also provided a testing file
- Since the competition is over, we will compile our final predictions on that file using our best model

Data

- Original data consists of 30 variables
- Variables are about the property, property location, the host and host reviews
- After cleaning and eliminating variables, our data consisted of 22 variables
- Property:
  - property_type, room_type, accommodates, bedrooms, beds, bed_type, bathrooms
- Location:
  - latitude, longitude, city
- Host:
  - cancellation_policy, cleaning_fee, host_has_profile_pic, host_identify_verified, etc

Baseline Regression

```r
linear = lm(price ~ ., data = training)
```

```
## [1] "MSE of Testing Set:  0.165"
```

Regression Splines/Generalized Additive Models

- 20 Fold Cross-Validation was performed for different degrees of freedom ranging usually between 3 and 6
- Cross-Validation MSE used to pick degrees of freedom for splines

Splines

- Splines fit to variables Accommodates, review_scores_rating, bathrooms, and bedrooms
- Best performing spline based on Cross-Validation MSE was the spline on review_scores_rating with degrees of freedom = 4
- Use these splines with their optimal degrees of freedom in my general additive model

Spline Degrees of Freedom

```
##   degfree        cv
## 1       2 0.6076361
## 2       3 0.6077699
## 3       4 0.6083597
## 4       5 0.6084167
```

```
##   degfree        cv
## 1       3 0.4417728
## 2       4 0.4418893
## 3       5 0.4422561
```

```
##   degfree        cv
## 1       3 0.4939648
## 2       4 0.4934629
## 3       5 0.4937314
```

```
##   degfree        cv
## 1       3 0.5792754
## 2       4 0.5940714
```

GAM Model

- Performed the GAM on the training data set using all of the predictors plus splines on Accommodates, review_scores_rating, bathrooms, and bedrooms with their optimal degrees of freedom
- Not a great fitting model, $R^2 = 0.6388
- Decent MSE when fit on the test data set

```
## [1] "Test MSE of GAM:  0.1612"
```

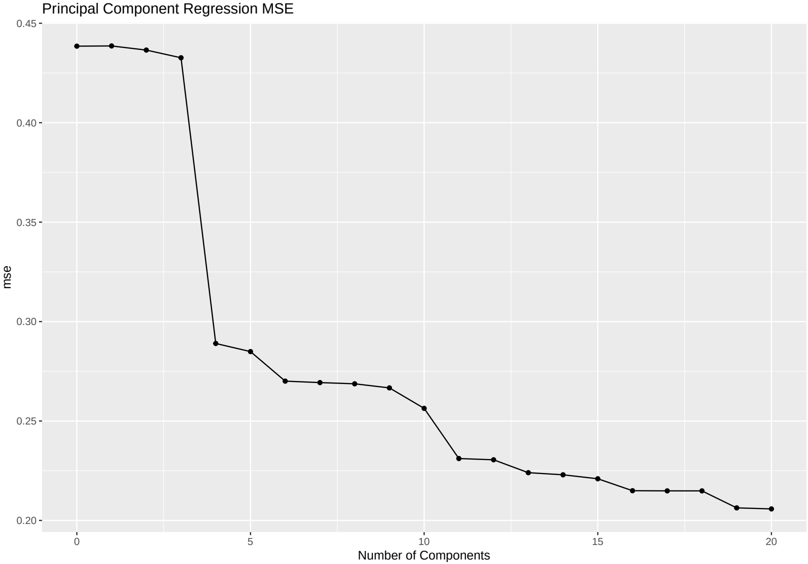
Future Modeling with Splines

- Received errors when using degrees of freedom larger than 6 or so
- Want to look into these errors and figure out if I could try larger degrees of freedom in my splines to get a better model.

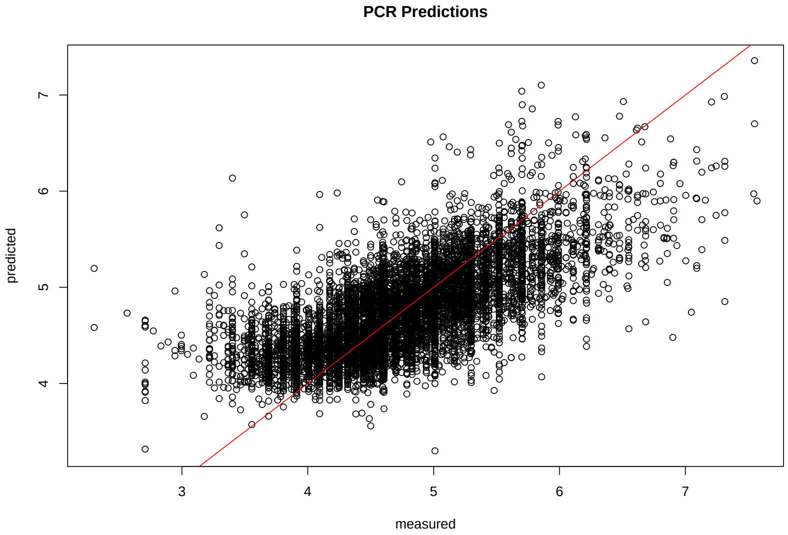
PCR and PLS

- 10 Fold Cross-Validation was performed for number of components ranging from 1 to 20.
- The Cross-Validation MSE was used to pick optimal number of components for both models.

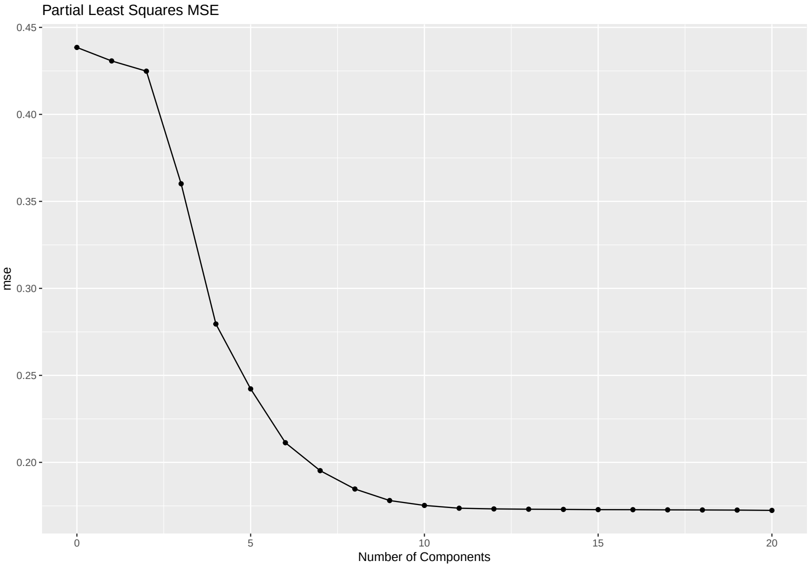
PCR


Principal Component Regression MSE
mse
0.45
0.40
0.35
0.30
0.25
0.20
0
5
10
15
20
Number of Components

PCR Predictions

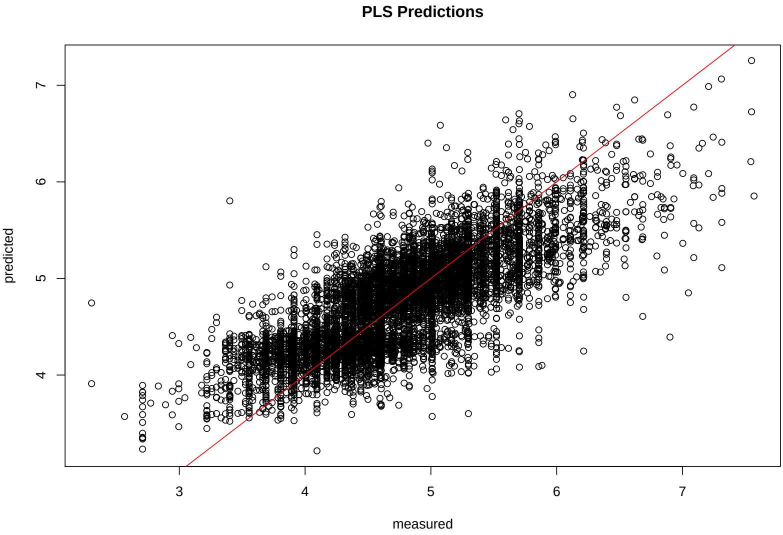


PLS



Partial Least Squares MSE
0.45
0.40
0.35
0.30
0.25
0.20
mse
0
5
10
15
20
Number of Components

PLS Predictions



PCR and PLS Summary

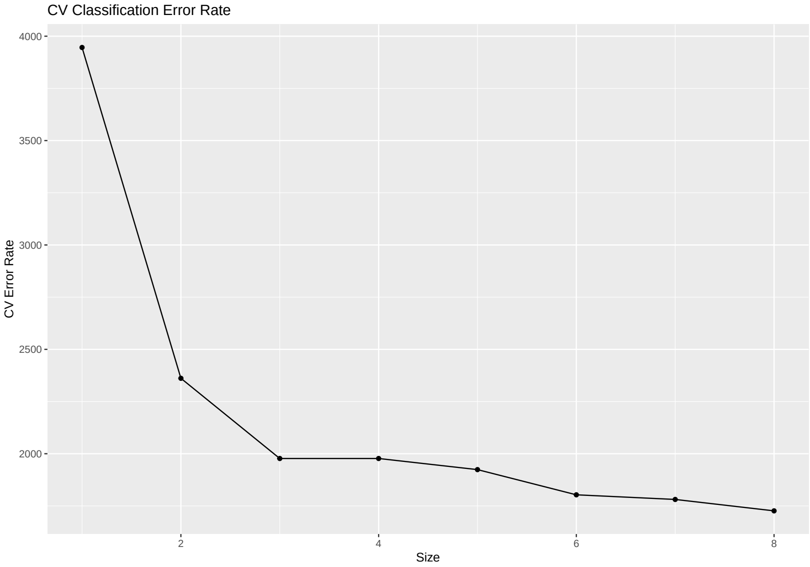
```
##                          PCR     PLS
## Components           15.0000 10.0000
## Test MSE              0.1765  0.2192
## % Variance Explained 99.7000 99.9000
```

Regression Trees

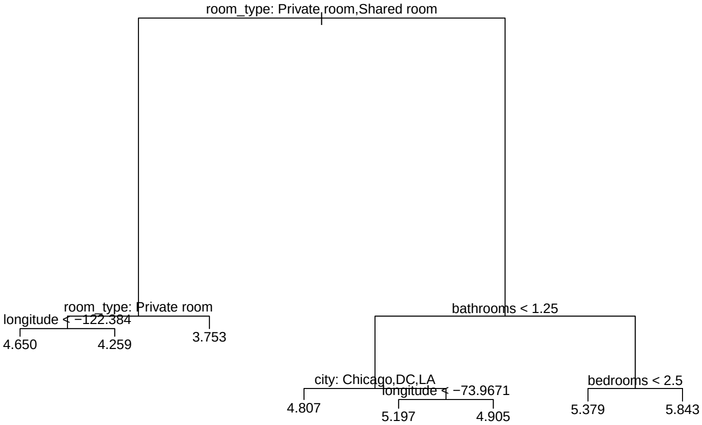
```
##
## Regression tree:
## tree(formula = price ~ ., data = training)
## Variables actually used in tree construction:
## [1] "room_type" "longitude" "bathrooms" "city"       "bed
## Number of terminal nodes:  8
## Residual mean deviance:  0.1885 = 1695 / 8992
## Distribution of residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.5050 -0.2999 -0.0196  0.0000  0.2558  2.8310

## [1] "Test MSE of Initial Tree:  0.1926"
```

Regression Trees



Regression Trees

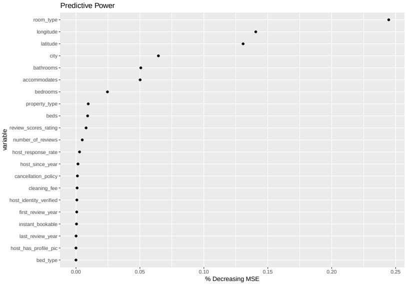


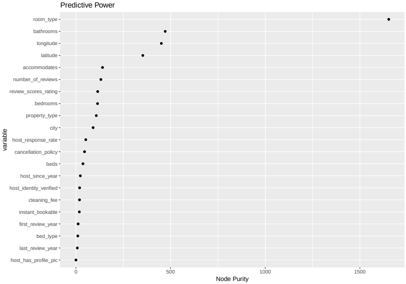
Bagging

```
bag_fit <- randomForest(price ~ ., data = training, mtry =
bag_predict = predict(bag_fit, testing, type = "response")
bag_MSE = round(mean((testing$price - bag_predict)^2), 4)
print(paste("Test MSE of Bagging: ", bag_MSE))
```

```
## [1] "Test MSE of Bagging:  0.1294"
```

Bagging


Predictive Power
variable
room_type
longitude
latitude
city
bathrooms
accommodates
bedrooms
property_type
beds
review_scores_rating
number_of_reviews
host_response_rate
host_since_year
cancellation_policy
cleaning_fee
host_identity_verified
first_review_year
instant_bookable
last_review_year
host_has_profile_pic
bed_type
0.00
0.05
0.10
0.15
0.20
0.25
% Decreasing MSE

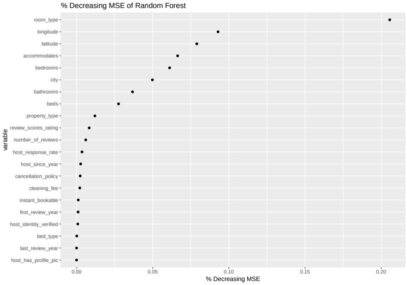

Predictive Power
variable
room_type
bathrooms
longitude
latitude
accommodates
number_of_reviews
review_scores_rating
bedrooms
property_type
city
host_response_rate
cancellation_policy
beds
host_since_year
host_identity_verified
cleaning_fee
instant_bookable
first_review_year
bed_type
last_review_year
host_has_profile_pic
0
500
1000
1500
Node Purity

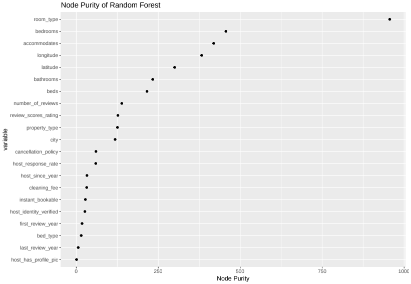
Random Forests

```r
rf_fit <- randomForest(price ~ ., data = training, mtry = s
```

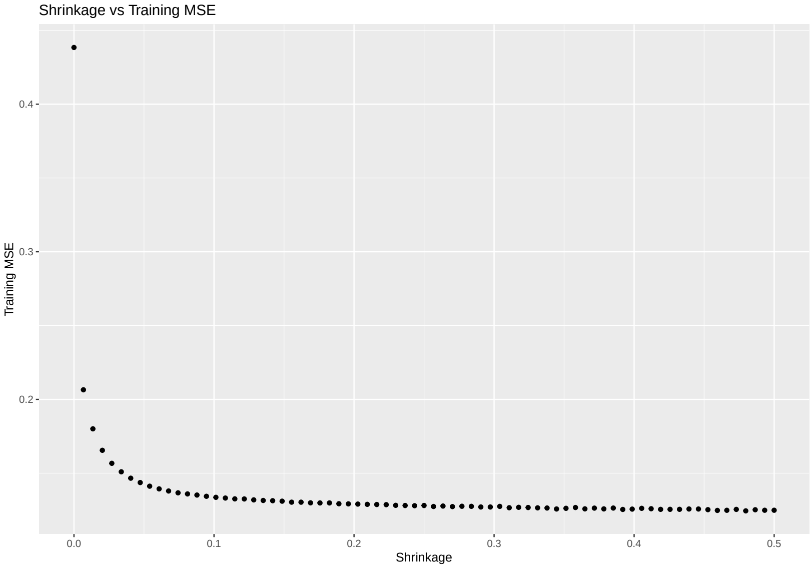
```
## [1] "Test MSE of Random Forest:  0.1301"
```

Random Forests





Boosting


Shrinkage vs Training MSE
Training MSE
0.2
0.3
0.4
Shrinkage
0.0
0.1
0.2
0.3
0.4
0.5

Boosting

```
## Using 2500 trees...

## [1] "Testing MSE for Boosted Model: 0.131"

## [1] "Best Variables: property_type, room_type, bathrooms

##                                             var     rel.i
## property_type                   property_type 22.9265735
## room_type                           room_type 19.4884646
## bedrooms                             bedrooms 14.0558039
## bathrooms                           bathrooms  9.5079769
## accommodates                     accommodates  7.6994146
## longitude                           longitude  7.6563738
## latitude                             latitude  6.7749833
## beds                                     beds  3.9324401
## review_scores_rating     review_scores_rating  2.6240559
## city                                     city  1.9246036
## number_of_reviews           number_of_reviews  0.8176040
## bed_type                             bed_type  0.8083937
```

MSE Table

```
##             Methods    MSE MSE_Dollars
## 1 Linear Regression 0.1652        1.18
## 2               PCR 0.2192        1.25
## 3               PLS 0.1765        1.19
## 4           Splines 0.4423        1.56
## 5               GAM 0.1612        1.17
## 6             Trees 0.1926        1.21
## 7           Bagging 0.1294        1.14
## 8     Random Forest 0.1301        1.14
## 9          Boosting 0.1310        1.14
```

Going Forward

- Our data has listings from multiple cities across the country
- Can we apply this to a certain city and see similar results?
- Is this accurate enough to help AirBnB hosts in selected cities?
  - Using current data, can this model help hosts correctly adjust their rates?

Questions?