Airbnb Listing Price Analysis

Matthew Coleman, Austin Mac, Jeff Pittman, and Nick Reyes

10 March 2020

1 Abstract

The following project was created with the purpose of predicting prices of Airbnb rentals in New York City using various machine learning models. The dataset was downloaded from Kaggle, the original source being from insideairbnb.com, contains approximately 50,000 different Airbnb listings. Each listing had a variety of attributes including price, room type, latitude and longitude, along with information of the listing's host. Our models varied in complexity and accuracy, and no model was 100% correct, though many ended up being useful. As the old adage goes, there is no free lunch, and in the end, we concluded linear regression for log(price) prediction and logistic regression for price above median classification were the best models as it had the ideal combination of complexity, interpretability, and accuracy.

2 Introduction

There were two main goals for this project: prediction of New York City Airbnb listing prices and classification of a listing being above or below the median price. Through comparison of many parametric and non-parametric machine learning models, we hope to select the models which result in the lowest MSE and misclassification rates while also considering the complexity and interpretability of the model.

3 Methods

3.1 Data

The dataset we will be using for our analysis is the dataset New York City Airbnb Open Data from Kaggle. This dataset contains the listing activity and metrics for Airbnb in New York City, New York during 2019. There are 48895 observations and 16 attributes to the dataset. The main features we are going to use for our analysis include the following:

- Price: Our main response variable. The price, in dollars, of the listing per night. Log-transformed to normalize distribution.

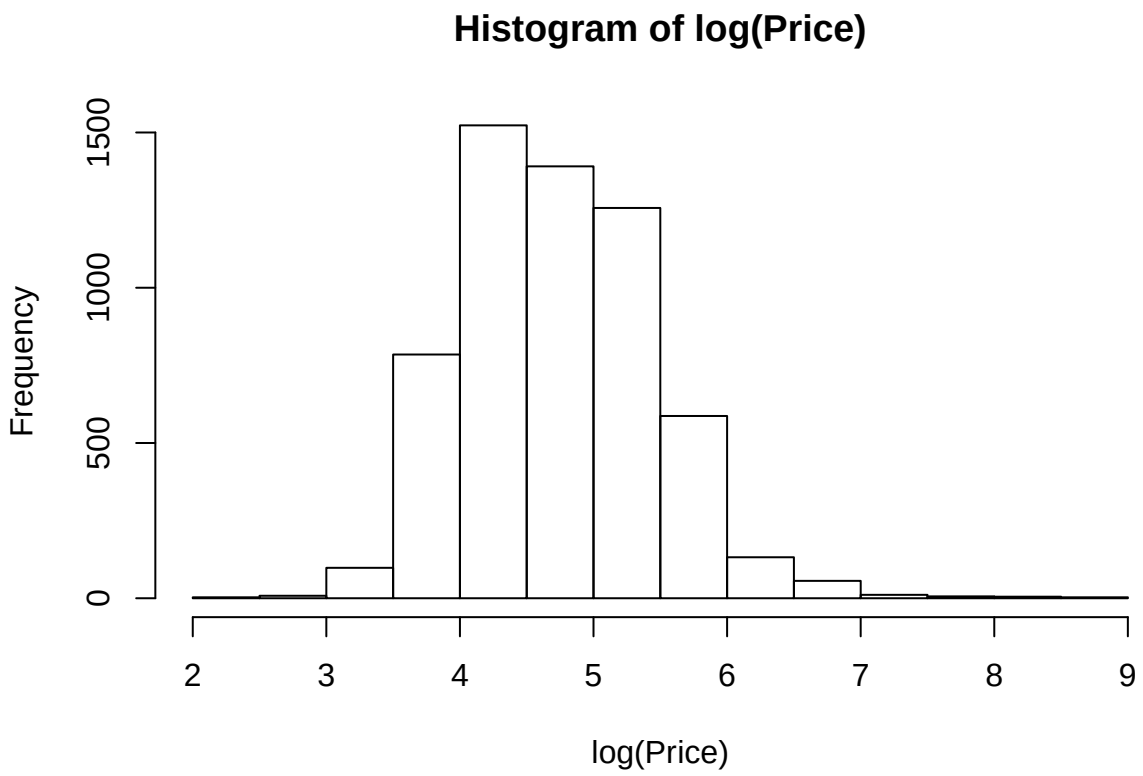


Figure 1: Log-transformed Price

- Price Above: Variable created from `price`, `price_above` is a binary variable of signaling whether a listings price is above the median listing price. 1 represents the price being above the median, and 0 represents the price being below the median.
- Neighbourhood: Categorical variable of the neighbourhood to which a listing belongs. This is a nested version of neigbhourhood group, with 221 unique neighbourhood groups.
- Neighbourhood Group: Factor variable of the neighbourhood group to which the listing belongs. There are 5 neighbourhood groups in the dataset.
  - Plots of both neigbourhood and neighbourhood group are shown below:

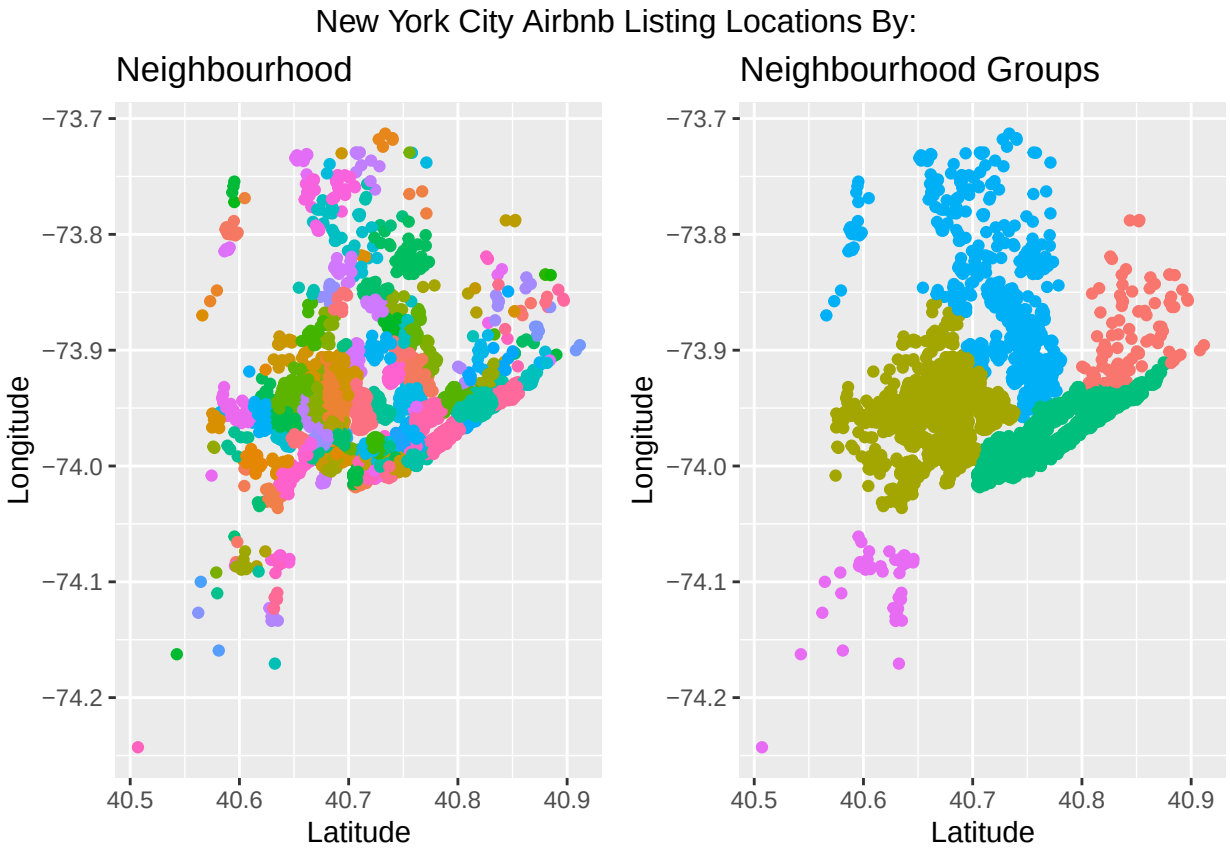


Figure 2: Neighbourhood and neighbourhood group

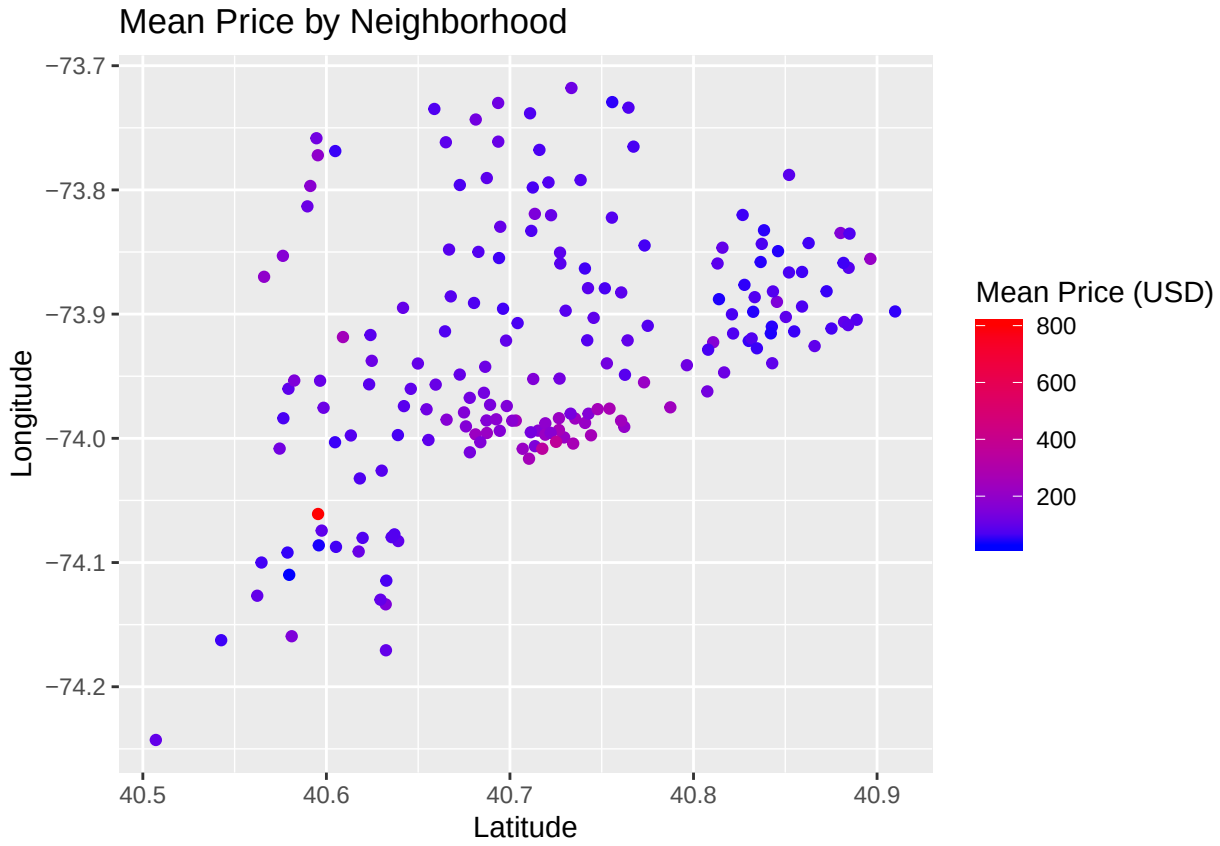


Figure 3: Justification for using neighbourhood group

- Latitude: Latitude coordinates of the listing.
- Longitude: Longitude coordinates of the listing.
- Room Type: The listing space type. Three types: *Entire home/apt*, *Private room*, *Shared room*.
- Minimum Nights: The mininum amount of nights someone can stay in the listing.
- Number of reviews: The number of reviews for the host.
- Reviews per Month: The number of reviews per month for the host. Formula: $\frac{Number\ of\ Reviews}{Months\ Listed}$.
- Calculated Host Listings Count: The number of listings per host.

All atributes were complete with the exception of `last_review`, which has the date of the last review, and `reviews_per_month`. Upon further exploration, the reviews per month feature was NA only when the host had no reviews. This resulted in us imputing 0's for NA values in the reviews per month column. Because the date of last review was unimportant to our analyses, we did not impute values for this column.

3.1.1 Assumptions.

Many of our machine learning methods are very computationally intensive, so we sampled 15% of the entire dataset, and then train-test split the 15% sample into 80% training 20% test dataset. To verify this was a viable practice, we plotted the distribution of our response variable, price, and verified the distribution is similar to the distribution of the overall dataset. The histogram is very similar, and even contains some of the outliers we can see in the overall dataset, so we assumed our smaller dataset was representative of the population.

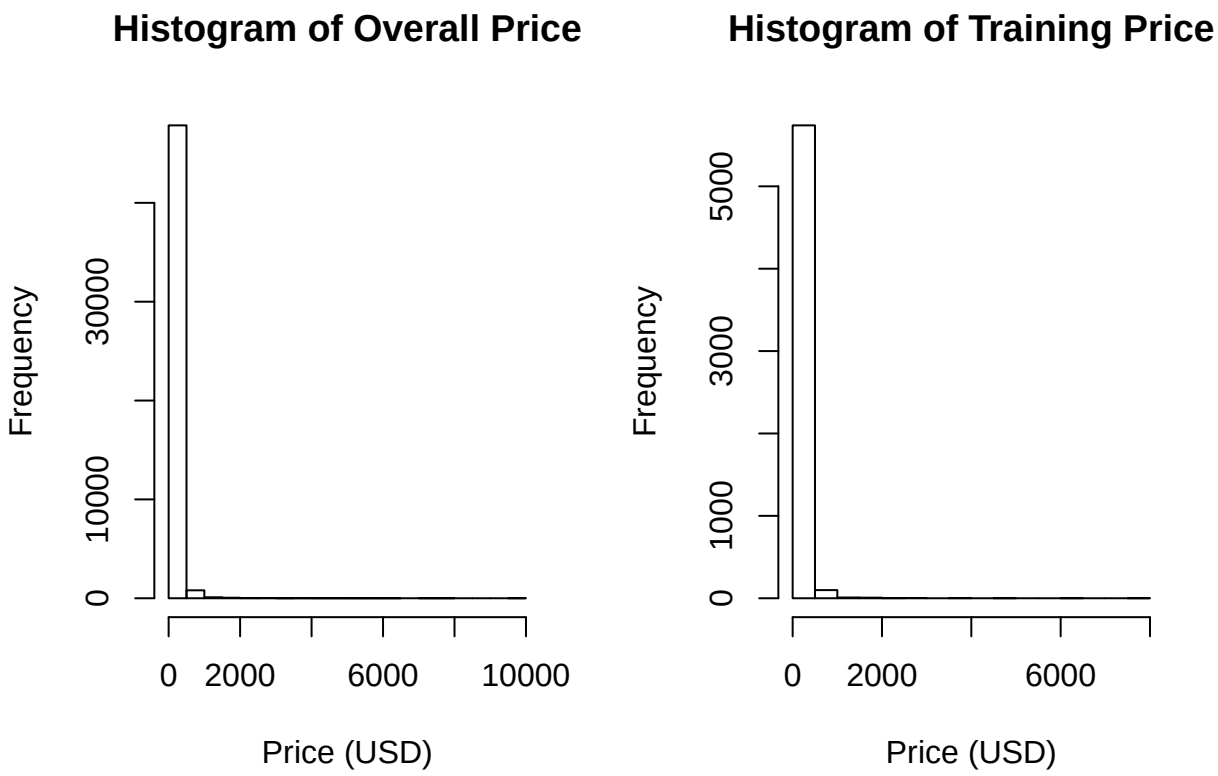


Figure 4: Training Data Histogram

We assessed the correlation between our variables with a correlation heatmap:

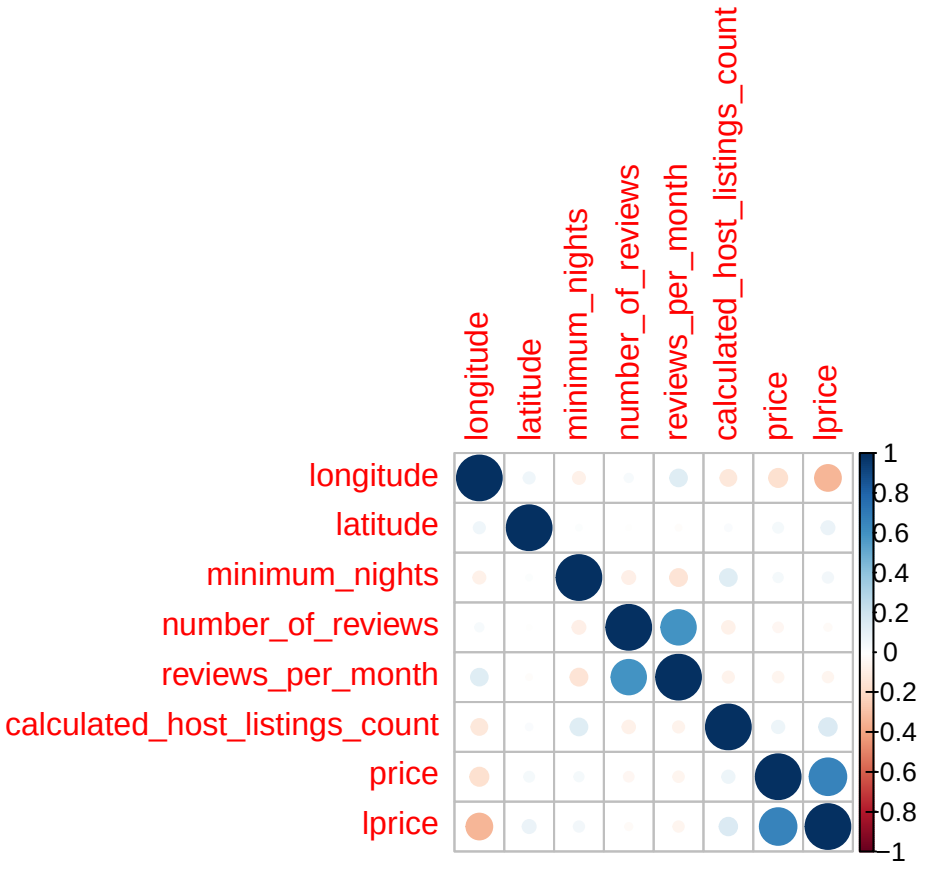


Figure 5: Feature Correlations

Reviews per month and number of reviews were highly correlated, so we decided to remove number of reviews to account for collinearity.

3.1.2 Sample Sizes

Our overall dataset is 48895. Taking the proposed 15% split on the data left us with an overall dataset of 7332 observations. The 80/20 train-test split left us with 5865 training samples and 1467 test samples.

3.2 Machine Learning Methods

3.2.1 Regression Methods

Methods used to predict the price of a listing:

- Ordinary Least Squares: Multiple Linear Regression with log(price) as the response variable.
- Tree Methods
  - Individual Trees: To compare the efficacy of ensemble tree methods, we will fit an individual regression tree on all variables of interest.
  - Bagging: We will fit an ensemble tree method which will grow large trees on bootstrapped data, resulting in high variance and low bias. All of these trees predictions will be averaged to give the final prediction.

  - Random Forest: We will create multiple decision trees similar to bagging, but try to decorrelate each of the bootstrap trees through selecting $m = \frac{p}{3}$ variables.
  - Boosting: We will fit multiple (weak) trees sequentially, grown on information from the previously grown tree. Final prediction is a weighted prediction of the weak learners.
- Ridge Regression:
  - Constraint optimization on the least squares criterion:

$$\hat{\beta}_{ridge} = \underset{\beta}{argmin}[||Y - XB||^2 + \lambda \sum_{j=1}^{p} \beta_j^2]$$

- Lasso Regression:
  - Constraint optimization and model selection on the least squares criterion:

$$\hat{\beta}_{ridge} = \underset{\beta}{argmin}[\ ||Y - XB||^2 + \lambda \sum_{j=1}^{p} |\beta_j|\ ]$$

By using these two methods, we can try to reduce our estimates for the linear model by imposing some Bias on our estimates for β. Another benefit of using Lasso regression is that we can also perform model selection, making a simpler model.

3.2.2 Classification Methods

Methods used to predict whether a listings price is above the median:

- Logistic Regression: We will fit a logistic regression model on all variables of interest,using a binary classification output to predict whether a listing's price is above or below the median.
- LDA: We will fit an LDA model on all variables of interest (assuming a Gaussian distribution), using a binary classification output to determine whether a listing's price is above or below the median. Additionally, LDA assumes equal variances in each group.
- QDA: We will fit a QDA model on all variables using the same binary classification and compare effectiveness to the LDA model.
- Tree Methods
  - Individual Trees: To compare the efficacy of ensemble tree methods, we will fit an individual classification tree on all variables of interest.
  - Bagging: We will fit an ensemble tree method which will grow large trees on bootstrapped data, resulting in high variance and low bias. All of these trees predictions will be chosen by majority voting for the final prediction.
  - Random Forest: We will create multiple decision trees similar to bagging, but try to decorrelate each of the bootstrap trees through selecting $m = \sqrt{p}$ variables. Final predictions will be through majority voting.
  - Boosting: We will fit multiple (weak) trees sequentially, grown on information from the previously grown tree. Final prediction is a weighted of the weak learners
- SVM: We will fit support vector machines with different kernels (Linear, Polynomial, Radial). In order to select the best possible support vector machines, we will use k-fold cross validation to tune the cost parameter to obtain the lowest misclassification rate.
- KNN: We will fit a K-Nearest Neighbours model with optimal K selected by cross validation.

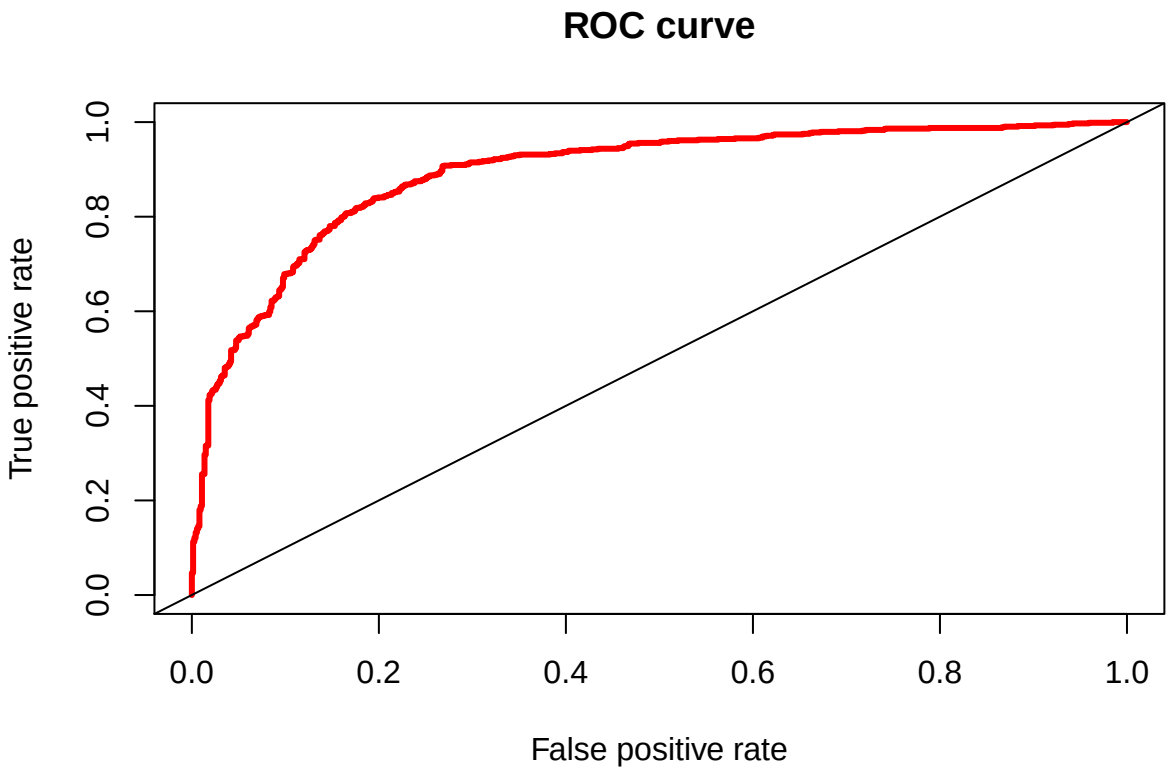
4 Analysis and Discussion

4.1 Linear Models

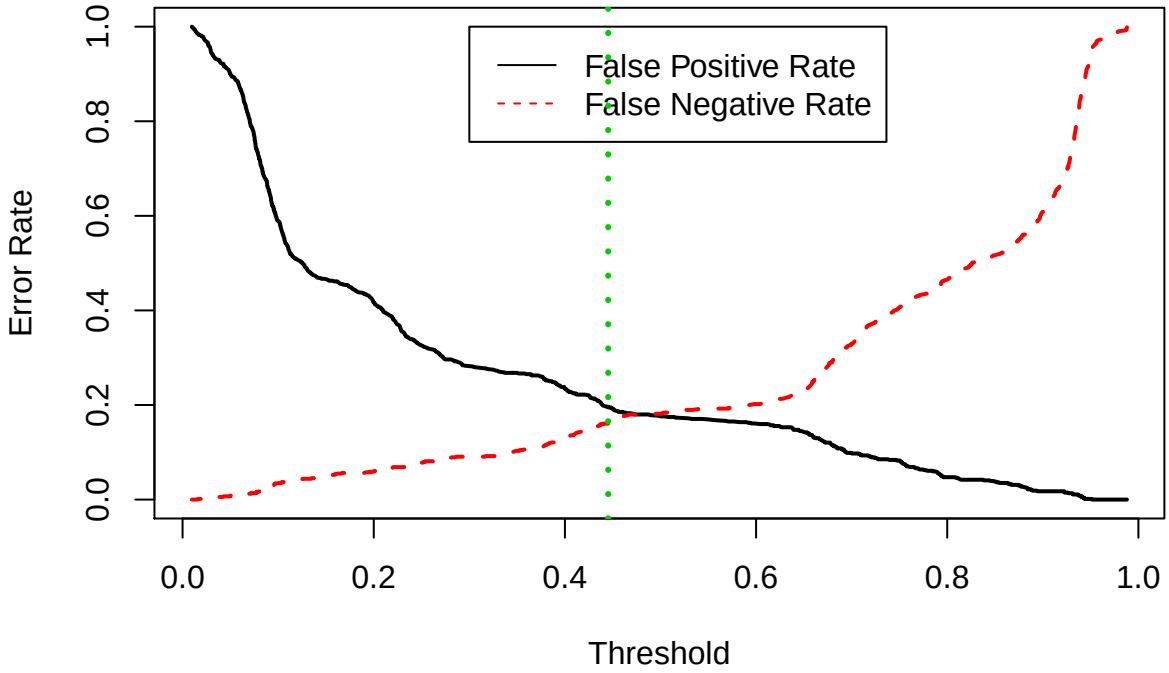
4.1.1 Ordinary Least Squares

The MSPE for out ordinary least squares model was 0.258828. This is slightly higher than some of the methods we explore. Despite this, the model is simpler.

4.1.2 Logistic Regression



The ROC curve above shows how the true positive rate and false positive rate change based on the threshold we choose. The AUC for our logistic classifier is 0.891641511397935, demonstrating our classifier is very good. Now, we can find the best threshold value for our logistic model:



Using the ROC curve we were able to find out the best threshold value to reduce the distance between the FPR and FNR. This threshold value was 0.4452299.

The confusion matrix outputted below gives an idea as to the distribution of predictions.

```
##           True
## Predicted   0   1
##         0 595 118
##         1 144 610
```

4.2 Discriminant Analysis

The linear discriminant analysis misclassification rate is 0.1833674 and the quadratic discriminant analysis MSPE is 0.1923274. These are very similar to the other classification methods we explore in our analysis.

4.3 Tree-based Methods

4.3.1 Classification and Regression Trees

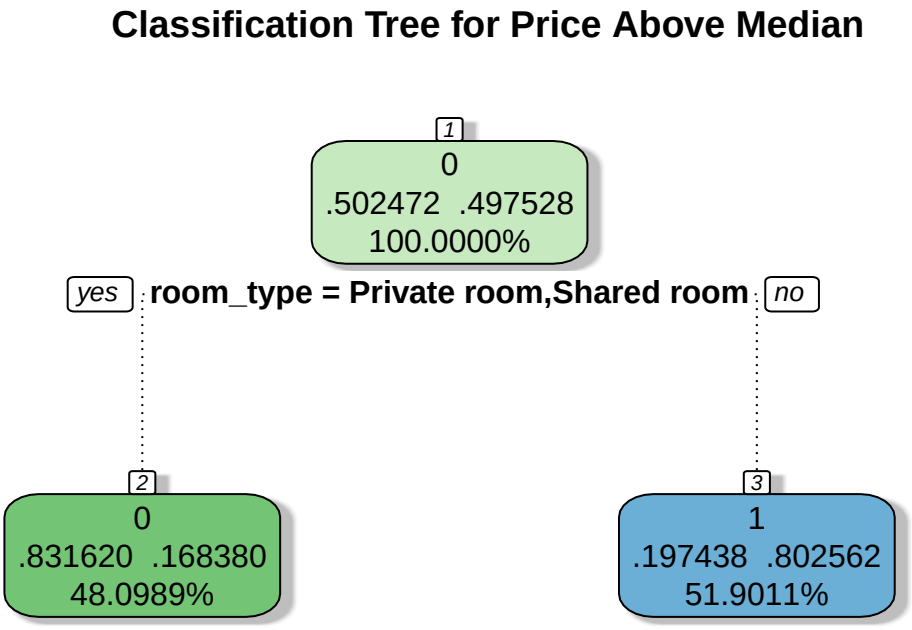


Figure 6: Classification Tree

The classification tree on all predictors split only on the type of room. We can see from the dendrogram that if a room is a private room or a shared room, the listing would be classified as "below the median price," and if it is a whole apartment or home then it would be classified as "below the median price."

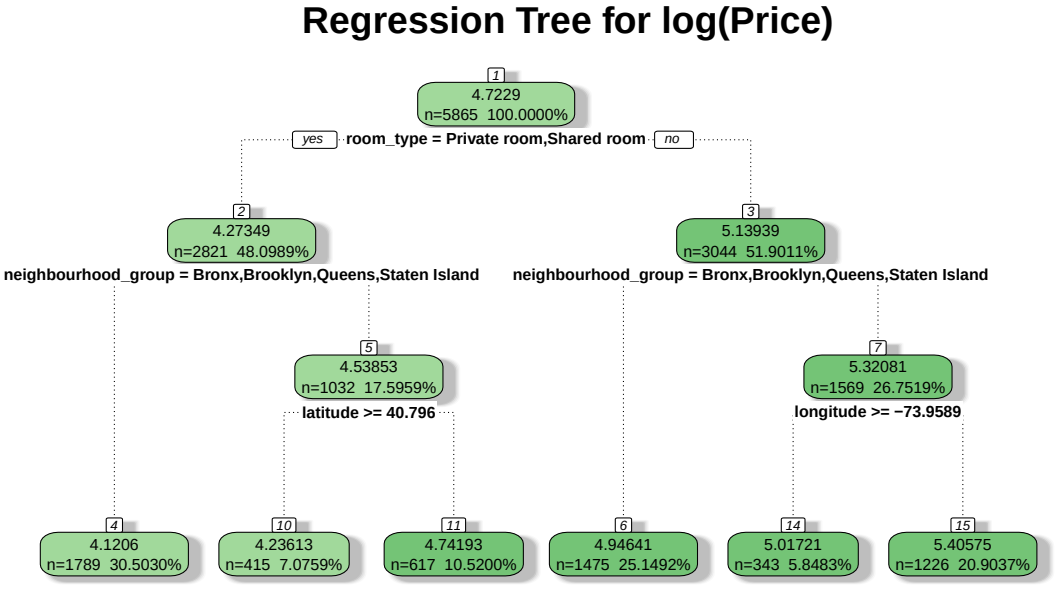


Figure 7: Regression Tree

The regression tree on is more intricate than the classification tree. The main split is on the room type of the

listing, and then the next two splits are made on the neighbourhood group. The last splits made are on the location of the the listing.

4.3.2 Random Forests

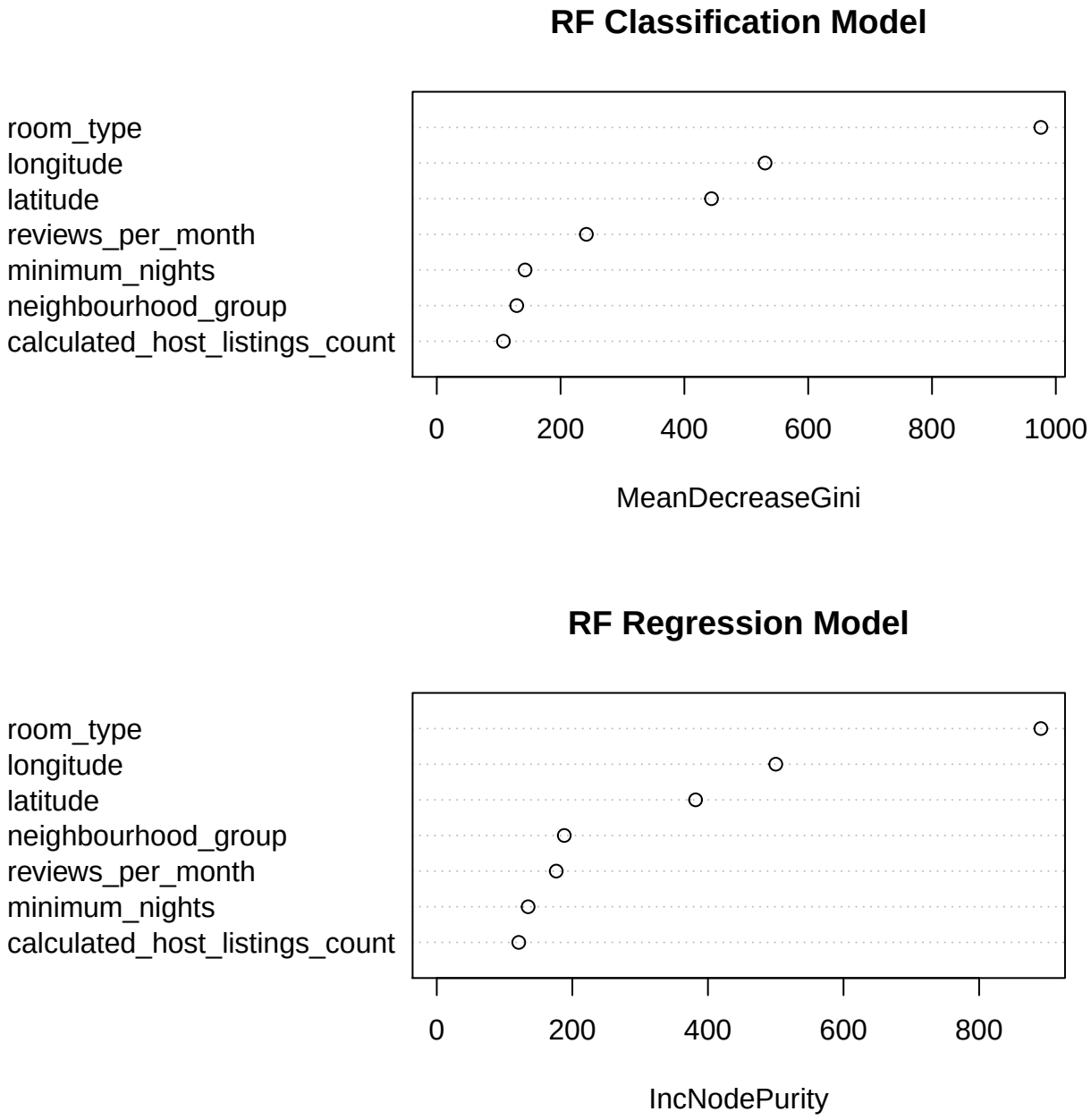


Figure 8: Variable Importances for RF Models

The random forest model imporances showed that room type, longitude, latitude, and reviews per month were the most important variables for the decrease in gini impurity for classification forests. For the regression model, room type, longitude, latitude, and neighbourhood group were the most important variables for the increase in node purity.

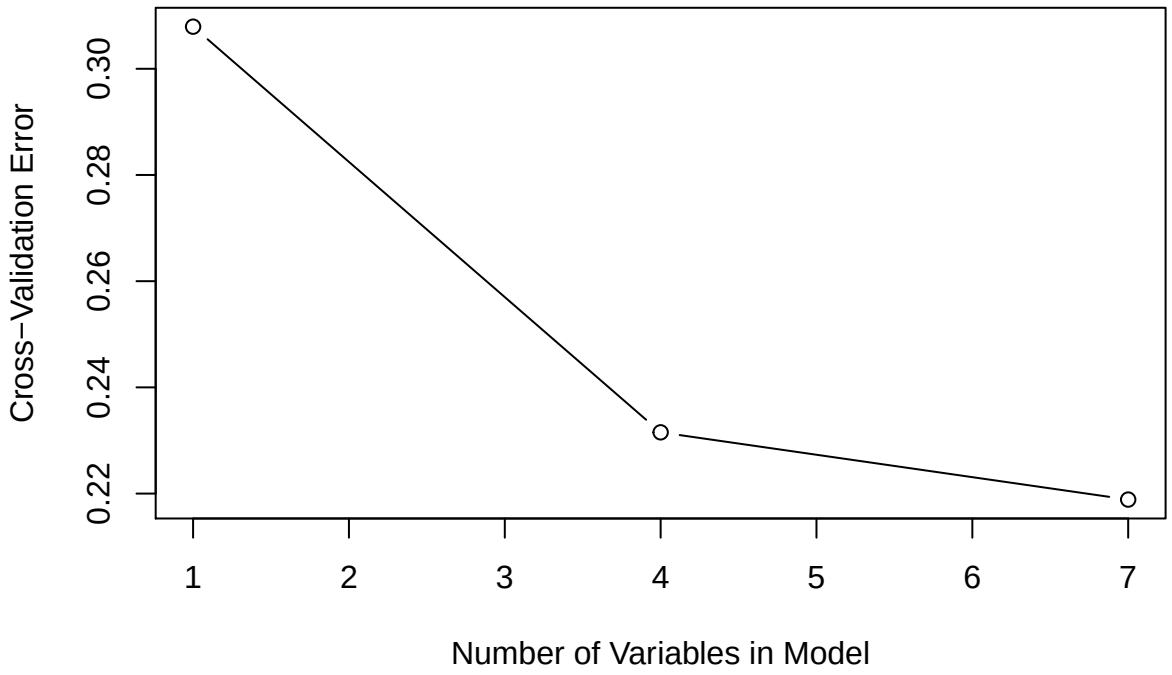


Figure 9: Cross Validation Error by # of Predictors

Through cross validation, we were able to determine 4 variables in the model leads to the greatest decrease in the cross-validation error while accounting for model complexity. To choose the 4 variables we would use in the reduced random forest model, I used the criteria of greatest variable importance from above. This means we chose room type, longitude, latitude, and reviews per month for the classification model and room type, longitude, latitude, and neighbourhood group of a regression model.

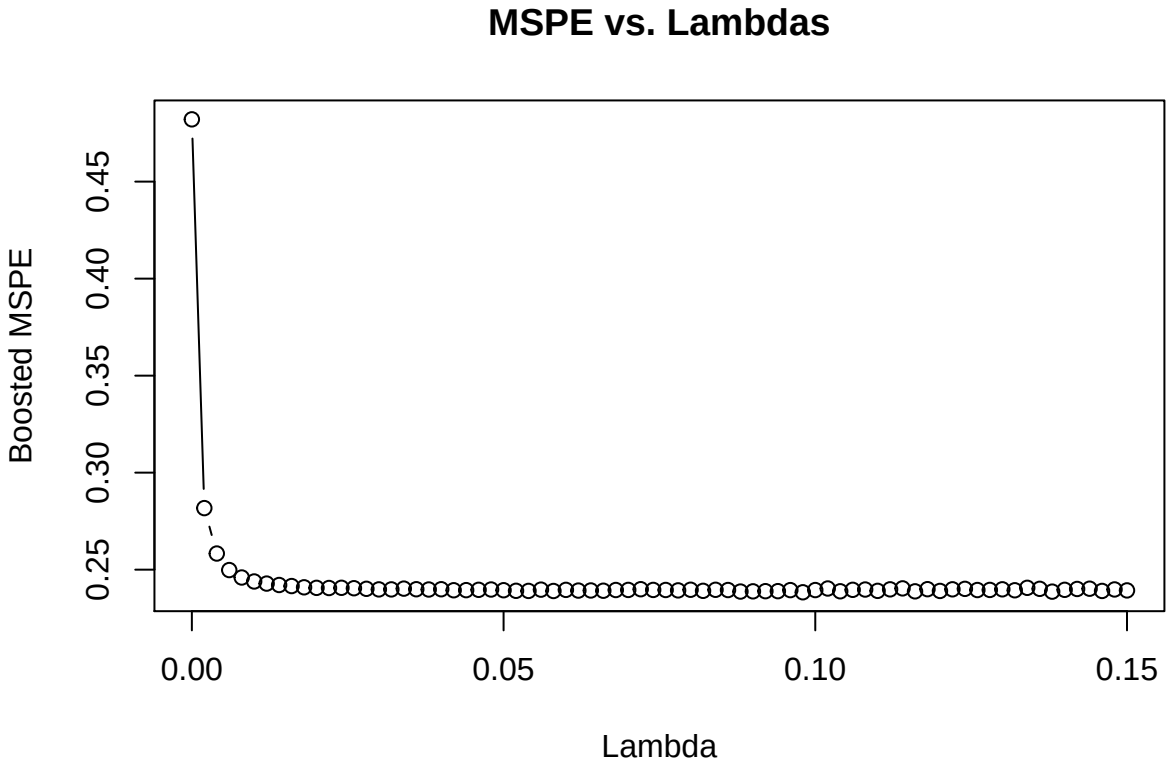
The misclassification rate for the random forest model with all the predictors included was 0.1731425, as opposed to the 0.1799591 misclassification rate of the reduced model. While the RF model with all the predictors is more accurate than the smaller model, the smaller model is simpler and more likely to be scalable in different scenarios. The mean squared prediction error of the full model, 0.222965, is also lower than the 0.2435506 MSPE of the smaller model. As with the classification forest, the simpler model is more scalable at the cost of prediction error. Another downside of the larger model is the possibility of overfitting to the training dataset.

4.3.3 Boostrap-Aggregating (Bagging)

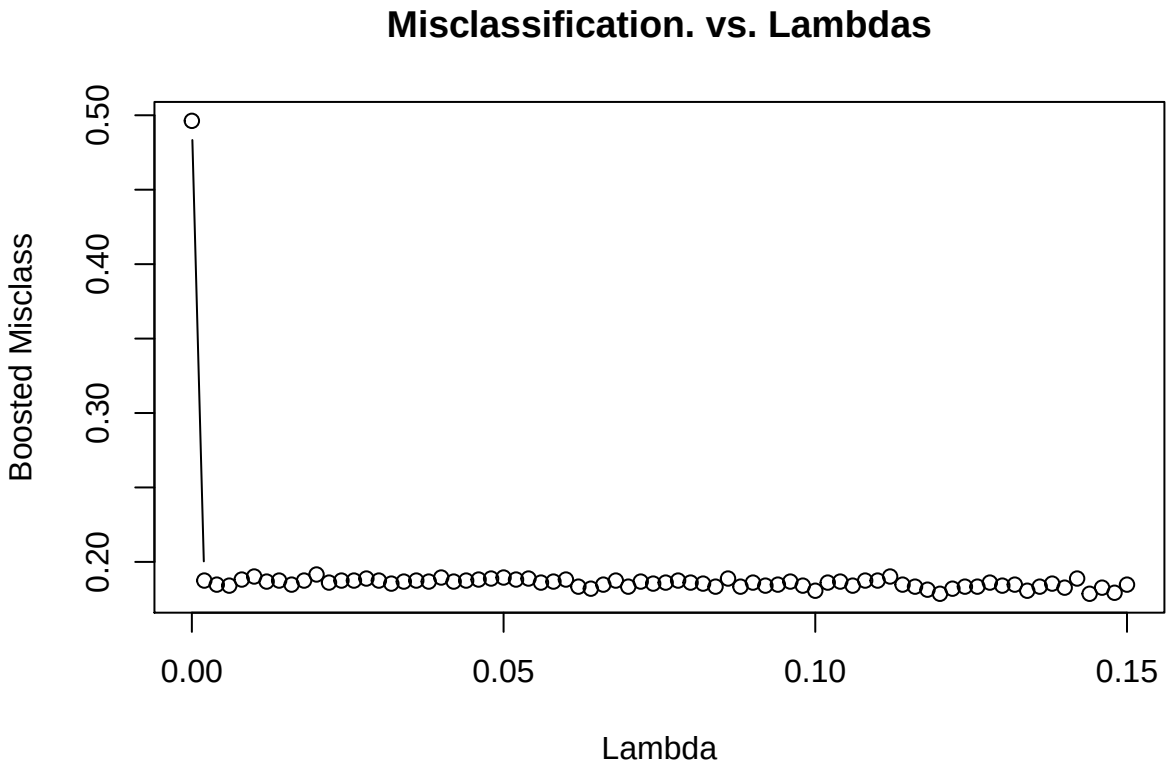
The bagging model was similar to the random forest model with a misclassification rate of 0.1820041 and a MSPE of 0.2426593.

4.3.4 Boosting

The first parameter we tuned in the boosting model was the shrinkage parameter, λ. The lambda which results in the lowest MSPE will be the lambda used in the final boosting model.

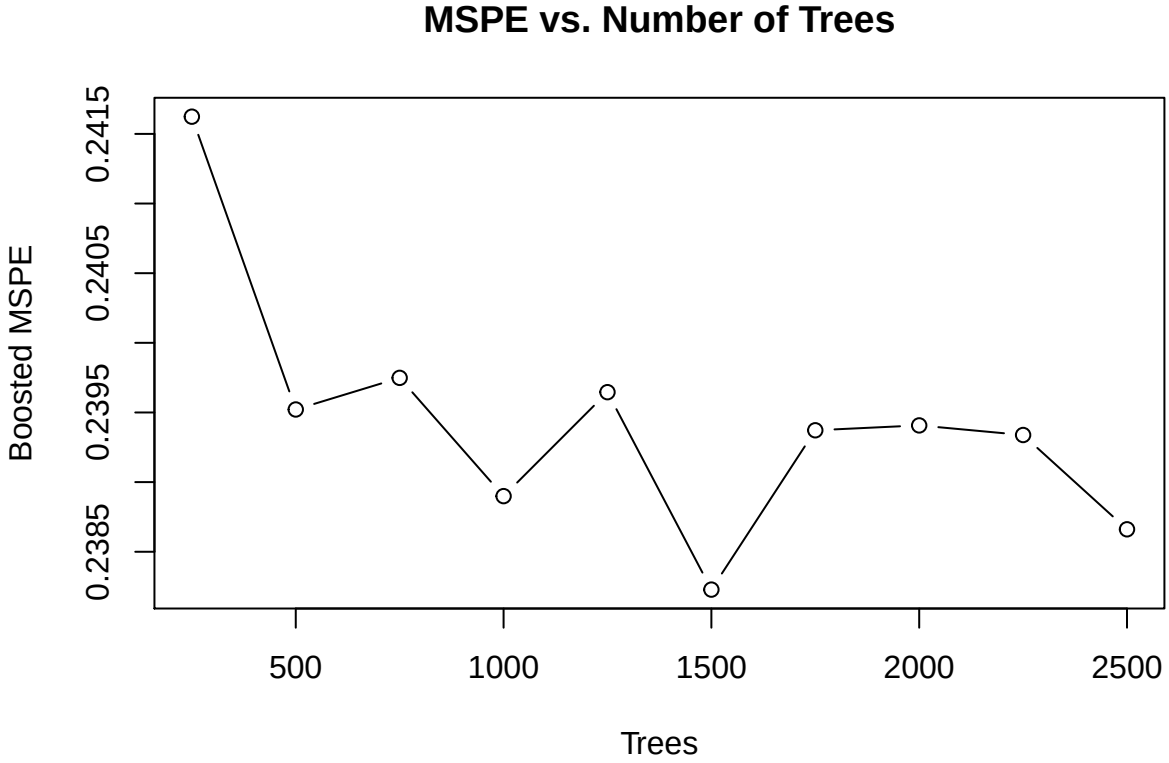


There was not a discernable optimal lambda for test MSE, so we decided to use the default value of $\lambda = 0.1$. This made model parameter selection the easiest.



As with the regression boosting model, the misclassification error rate does not seem to be at a minimum for any value of lambda, so we will also choose the default value for λ.

The other parameter tuned for the boosting model was the number of trees used.



The MSPE was the lowest for the model with 1500 trees, so we implemented this into the final boosting model.

4.3.5 Tree Method Summary

All The tables for tree methods put together. This will allow us to evaluate the efficacy of our tree models and determine which is the best.

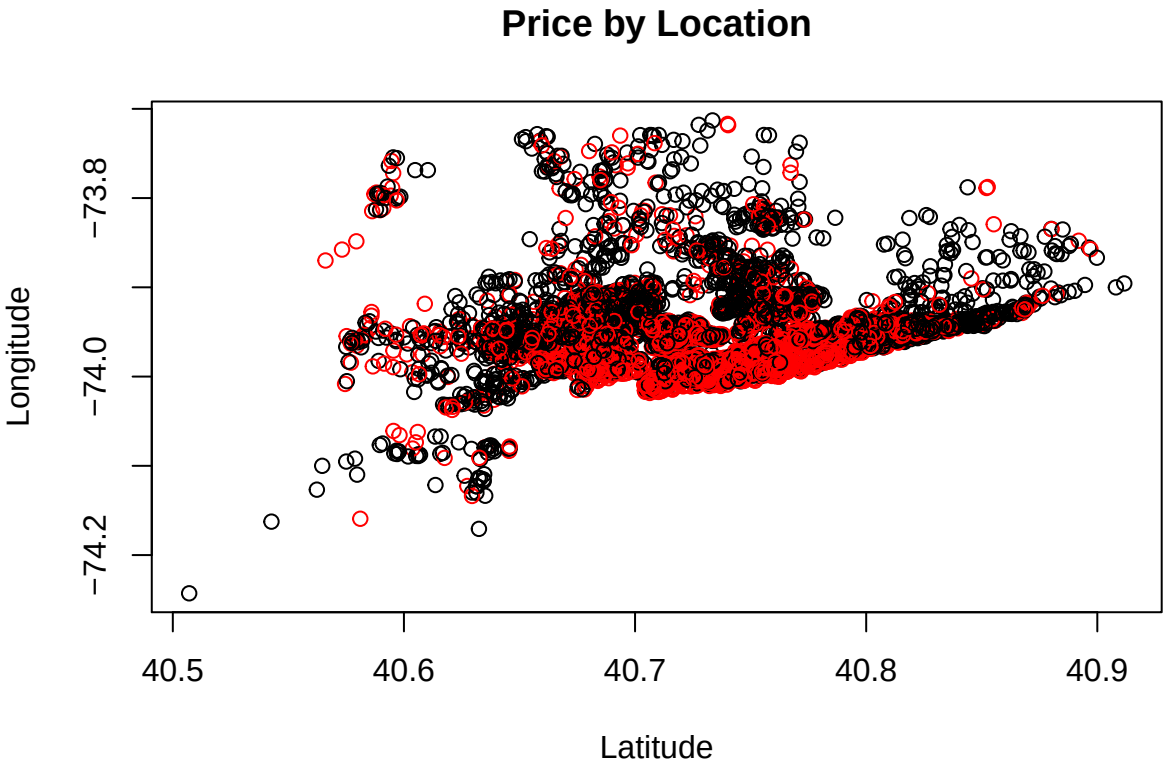
Table 1: Error Rates for Tree Models

Methods	MSPE	Methods	Misclassification
Tree	0.25993	Tree	0.187457
Bagging	0.242659	Bagging	0.182004
Boosting	0.240589	Boosting	0.184731
Random Forest	0.222965	Random Forest	0.173142
Reduced Random Forest	0.243551	Reduced Random Forest	0.179959

The table with all the tree method error rate shows there is not a clear "best model" for the data. The random forest with all predictors has the lowest test MSE and misclassification error rate, but suffers from the possibility of overfitting the data and being too complex of a model. Because all of the methods are approximately equal in terms of predictive performance, a simpler model such as a simple tree or the reduced random forest may be best. Both of these models are simpler, and would therefore be more scalable in larger-data environments.

4.4 SVM

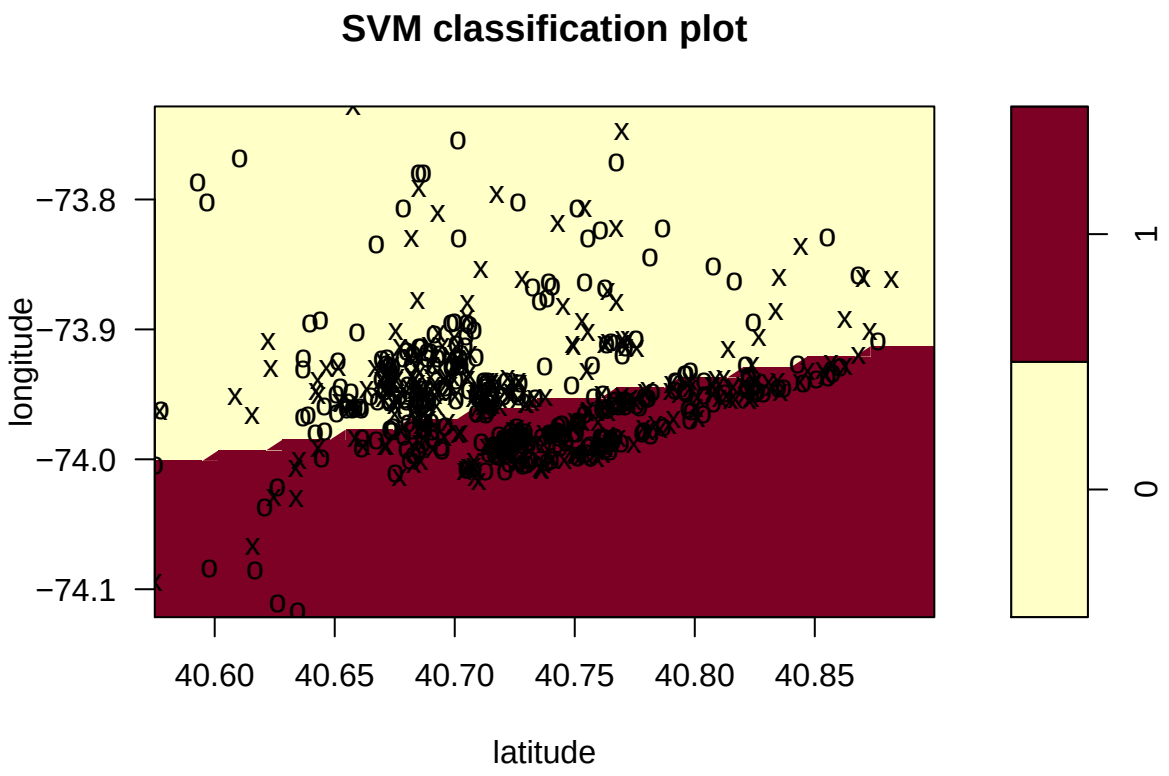
In order to gain a rudimentary understanding of our data, we plotted a scatterplot of the listings coded by price above or below the median. It appears that the price of properties is noticably higher along waterfront properties.



4.4.1 Best Linear Kernel SVM

- Misclass. Rate: 0.1867757
- Cost Parameter: 0.06
- Support Vectors: 4598

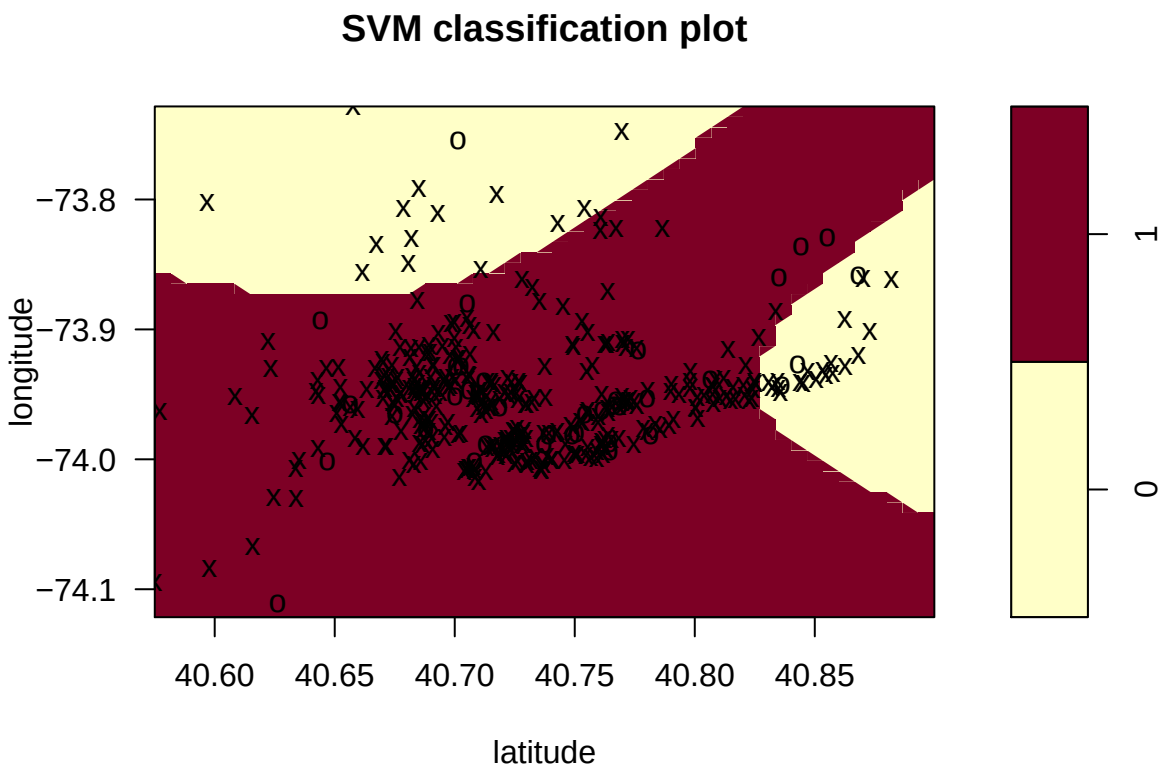
Below is a preliminary plot of a linear kernel SVM model fit off of latitude and longitude. After including other predictors including `neighbourhood_group`, `minimum_nights`, `room_type`, `number_of_reviews`, `calculated_host_listings_count`, `availability_365`, and using 10-fold cross validation to select the best cost parameter, we obtain a cost parameter with value 0.06. This gives us a test misclassification rate of 0.1867757.



4.4.2 Best Polynomial Kernel SVM

- Misclass. Rate: 0.1724608
- Cost Parameter: 100
- Degree: 3
- Support Vectors: 5737

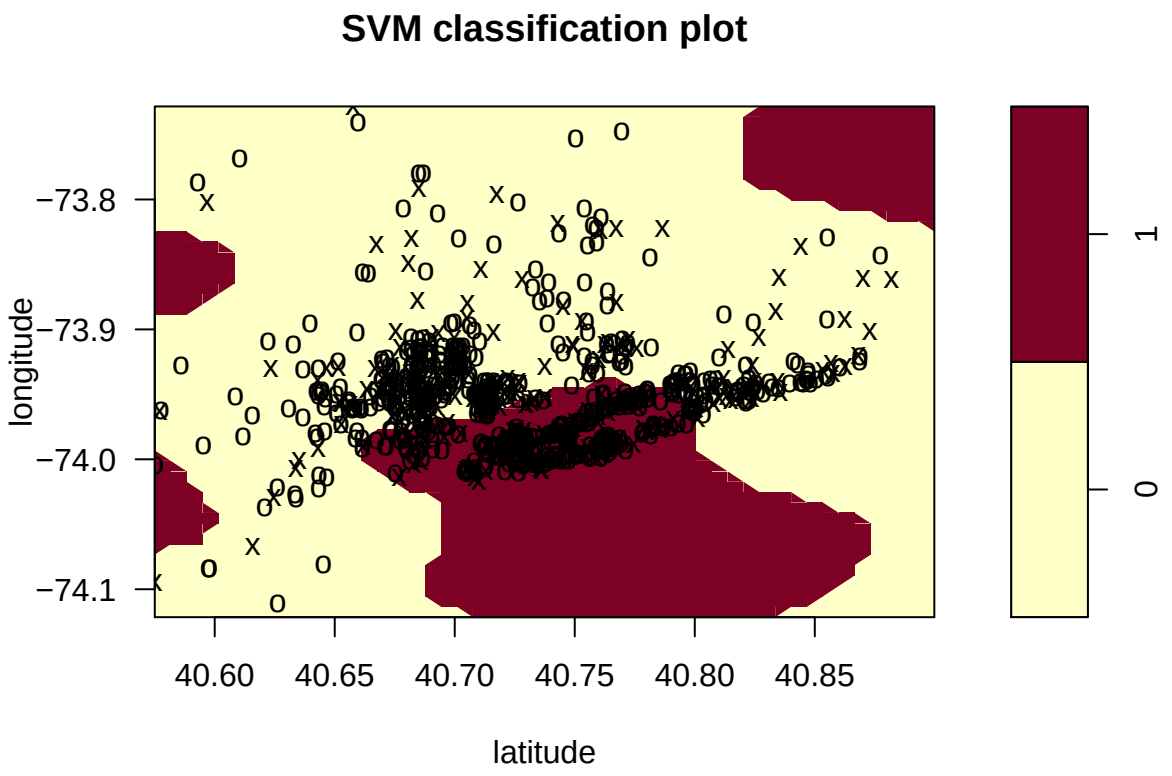
Below is a plot of a polynomial kernel SVM decision boundary fit off of `longitude` and `latitude`. Fitting the model with the same predictors as our linear kernel and using 10-fold cross validation, we obtain an optimal cost parameter value of 100. Our test misclassification rate is 0.1724608.



4.4.3 Best Radial Kernel SVM

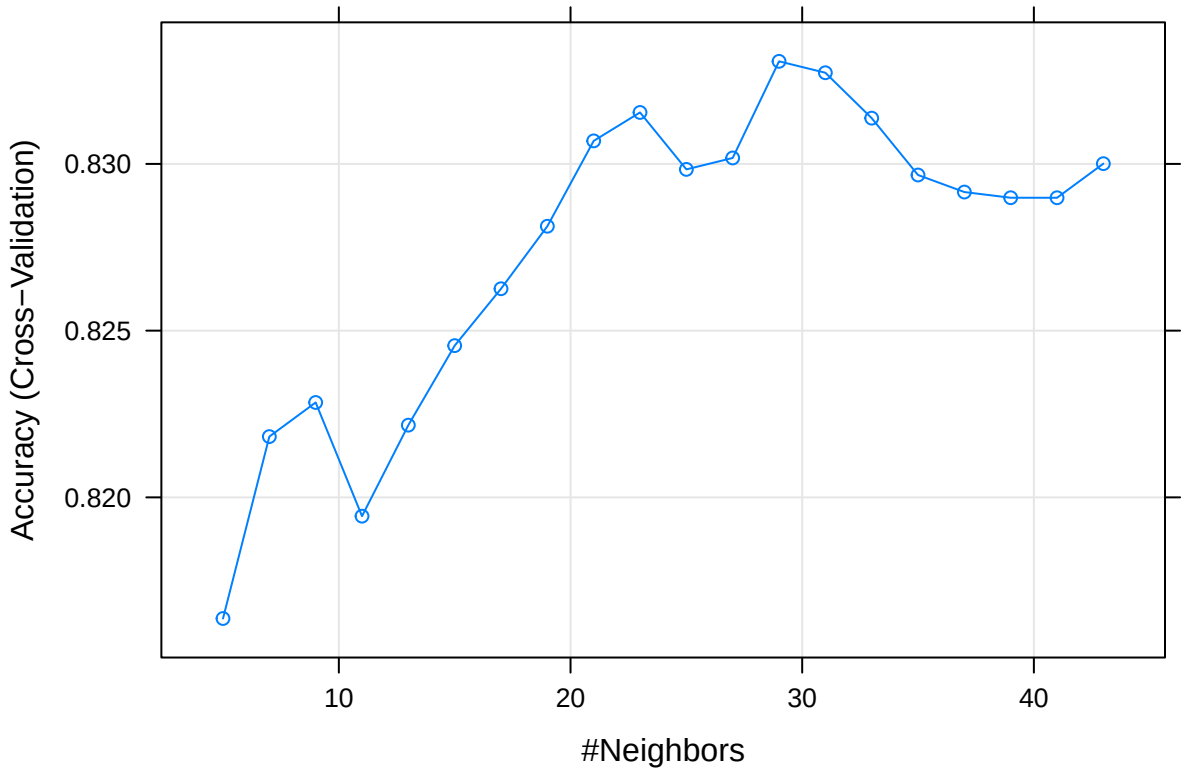
- Misclass. Rate: 0.1785958
- Cost Parameter: 8
- Support Vectors: 3615

Below is a plot of the decision boundary of a radial kernel SVM fit on `latitude` and `longitude`. Fitting a model with the same parameters as the previous two models, we obtain a cost parameter value of 8 through 10-fold cross validation. Our test misclassification rate is 0.1785958.



4.5 KNN

Using cross validation to find K such that the misclassification rate is minimized, we find that k = 29 gives us the minimum misclassification rate of 0.1656442. In the plot below, accuracy is maximized at k = 29.



4.6 Regularization

4.6.1 Ridge and Lasso Regression

The test MSE for ridge and lasso regression were both approximately 0.42. It is reasonable these models would have worse test MSE values than our other models since both the ridge and lasso models did not contain non-numeric predictors. When we compare with the previous results, we see that some of the non-numeric variables (such as room-type and neighborhood group) were important in predicting in other models.

5 Conclusion

There were many aspects to take into account when selecting our final model. First, we had to consider the accuracy of the model. If our test error rate was too high, there would be no reason for to implement a poorly-constructed model. Next was model complexity: we could use a model with multiple predictors and parameters, decreasing the error rate of our model, but increasing the variance, or we could contstruct a simpler model with more bias, but lower variability. Choosing a simpler model would allow us to use our model in a larger setting and achieve similar test error rates. We can compare the test error rates of our different models and determine a best model.

5.1 Error Rates

5.1.1 Linear Models

Table 2: Error Rates for Linear Models

Method	MSPE	Method	Misclass.
OLS	0.258828	Logistic	0.1785958

5.1.2 Tree Error Rates

Table 3: Error Rates for Tree Models

Methods	MSPE	Methods	Misclassification
Tree	0.25993	Tree	0.187457
Bagging	0.242659	Bagging	0.182004
Boosting	0.240589	Boosting	0.184731
Random Forest	0.222965	Random Forest	0.173142
Reduced Random Forest	0.243551	Reduced Random Forest	0.179959

5.1.3 SVM and KNN Error Rates

Table 4: Misclassification Rates for SVM and KNN

Methods	Misclassification
Linear Kernel SVM	0.187457
Polynomial Kernel SVM	0.173824
Radial Kernel SVM	0.179277
KNN	0.165644

5.1.4 Regulatization Methods

Table 5: MSPE for Regularization Methods

Methods	MSPE
Ridge	0.417889226551229
Lasso	0.418745748401298

5.2 Final Models and Overall Dataset Error

5.2.1 Regression

5.2.1.1 Ordinary Least Squares Regression

The final mean squared prediction error, 0.2572045, of the model is slightly higher than some of the more complex models. This choice was made for our final model because the model is simpler and more efficient with larger datasets. While the accuracy is slightly decreased, it is not a large enough difference where we would consider another, more accurate model. Another reason for choosing the linear model is because of the ability to interpret the coefficients. Some of the more significant coefficients can be interpreted below:

Note: All interpretations are given all other predictors are held constant.

- **Latitude**: A one unit increase in latitude results in a -66.9781261% decrease in the price of the listing.
- **Longitude**: A one unit increase in longitude results in a -92.3387881% decrease in the price of the listing.
- **Neighbourhood Group**:
  - Being in the *Brooklyn* neighbourhood group results in the price of the listing being -7.1133265% lower than a listing in the Bronx.
  - Being in the *Manhattan* neighbourhood group results in the price of the listing being 37.0807524% higher than a listing in the Bronx.
  - Being in the *Queens* neighbourhood group results in the price of the listing being 5.9100539% higher than a listing in the Bronx.
  - Being in the *Stanten Island* neighbourhood group results in the price of the listing being -50.2967568% lower than a listing in the Bronx.
- **Room Type**
  - Being a *Private Room* results in the price of the listing being -52.7208126% lower than a *Whole Apartment or House* listing.
  - Being a *Shared Room* results in the price of the listing being -68.8389105% lower than a *Whole Apartment or House* listing.
- **Calculated Host Listing Count**: A one unit increase in longitude results in a 0.050903% decrease in the price of the listing.

5.2.2 Classification

5.2.2.1 Logistic Regression

```
##           True
## Predicted     0     1
##         0 19536  3954
##         1  4925 20469
```

The logistic regression model ended up being the best choice for our classification model. As with the regression model, the misclassification error rate, 0.1816341, was slightly higher than some of the more complex models, but we decided to choose the simpler model at a small penalty to misclassification rate. By searching for the best threshold, it seems 0.4452299 is the best choice. There seemed to be many more false positives than false negatives in our overall dataset predictions. As we can see from the confusion matrix of the logistic predictions, our False Postitive Rate was 0.2013409 and our False Negative Rate was 0.1618966.

We can interpret the most important coefficients as follows:

Note: All interpretations are given all other predictors are held constant.

- **Latitude**: A one unit increase in latitude results in a -4.699 decrease in the log-odds of the listing price being above the median.
- **Longitude**: A one unit increase in longitude results in a -13.08 decrease in the log-odds of the listing price being above the median.
- **Minimum Nights**: A one unit increase in the minimum amount of nights results in a decrease of -.002933 of the listing price being above the median.
- **Neighbourhood Group**:
  - Being in the *Brooklyn* neighbourhood group results in the log-odds of the price being below the median to be reduced by -.1586, compared to being a listing in the Bronx.
  - Being in the *Manhattan* neighbourhood group results in the log-odds of the price being below the median to be increased by 1.411, compared to being a listing in the Bronx.
  - Being in the *Queens* neighbourhood group results in the log-odds of the price being below the median to be increased by .5048, compared to being a listing in the Bronx.
  - Being in the *Stanten Island* neighbourhood group results in the log-odds of the price being below the median to be reduced by -3.371, compared to being a listing in the Bronx.
- **Room Type**
  - Being a *Private Room* results in the log-odds of the price being below the median to be reduced by -3.078, compared to a whole apartment or house listing.
  - Being a *Shared Room* results in the log-odds of the price being below the median to be reduced by -4.019, compared to a whole apartment or house listing.
- **Calculated Host Listing Count**: A one unit increase in longitude results in a .00375 increase in the price of the listing.

5.2.3 Future Research

One of the limiations of our study was the computing power of our machines. Train-test splitting the entire dataset was not feasible because of the computing power and resulted in a train-test split of the subsetted data. This resulted in some of the more expensive methods such as Ensemble Tree methods, Support Vector Machines, and K-Nearest-Neighbours. One possible direction is to try and run some of the more expensive methods on a stronger machine, which could make some of the more complex ML methods more useful. Another possible direction for the dataset would be to using the Host Name to classify a host as male or female (or a combination of the two). One could use this data to explore whether the host's gender played a role in predicting the amount of business the listing receives. Gender could be combined with the price to make inferences on differences in net income from listings across different genders.

6 Appendix

```
# Anything below here before the abstract has to be here so we can run code in the analysis and have ou
library(tidyverse)
library(dplyr)
library(randomForest)
library(class)
library(tree)
library(gbm)
library(caret)
library(rpart.plot)
library(rattle)
library(knitr)
library(fastAdaboost)
library(ggpubr)
library(MASS)
library(kableExtra)
library(glmnet)
library(faraway)
library(ROCR)
library(e1071)
#library(train)
```

Read in the CSV and check the dimensions of the data.

```
bnb <- read.csv('AB_NYC_2019.csv')

bnb <- as_tibble(bnb)

dims <- dim(bnb)

#sprintf('Our dataset has %d observations and %d attributes', dims[1],dims[2])
# [1] "Our dataset has 48895 observations and 16 attributes"
```

Since observations which have a price of 0 will not be useful to our analysis, and are likely to be representative of a bad data point, we will remove these observations.

```
bnb <- bnb[(bnb$price!=0),]
bnb[is.na(bnb$reviews_per_month),'reviews_per_month'] <- 0
bnb$price_above <- ifelse(bnb$price> median(bnb$price),1,0)
```

Train-test split the data

```
set.seed(123)
train.ind <- sample(1:nrow(bnb),size = .8*nrow(bnb))
small.ind <- sample(1:nrow(bnb), size = .15*nrow(bnb))


train.big <- bnb[train.ind,]
test.big <- bnb[-train.ind,]

small <- bnb[small.ind,]
train.small <- sample(1:nrow(small), size = .8*nrow(small))
train <- small[train.small,]
test <- small[-train.small,]
```

We can get the column names of the columns which contain NA's with the following code:

```
#colnames(train)[apply(train, 2, anyNA)]
```

We can see that there are NA reviews in the `reviews_per_month`, the number of reviews per month. We can also see upon visual inspection `last_review`, the date of the last review the host received, also contains empty values. We will not be using last_review in our analysis, so we will not worry about imputing values here.

We believe the reason there are NA's in the `reviews_per_month` column is because the hosts have 0 reviews overall. We further explore this claim the below:

```
#with(train, sum((is.na(reviews_per_month)) & (number_of_reviews!=0)) )
# [1] 0
```

We can see there are no cases where the `number_of_reviews` and `reviews_per_month`. As a result, we will impute 0 where `reviews_per_month` is NA.

```
train[is.na(train$reviews_per_month),'reviews_per_month'] <- 0
test[is.na(test$reviews_per_month),'reviews_per_month'] <- 0

#sum(is.na(train$reviews_per_month))
# [1] 0
#sum(is.na(train$reviews_per_month))
# [1] 0
```

We can assess the correlation between numeric features with a correlation heatmap:

```
num.feat <- train %>% dplyr::select(longitude, latitude, minimum_nights, number_of_reviews,
                                    reviews_per_month, calculated_host_listings_count, price)
num.feat$lprice <- log(num.feat$price)
feat.corr <- cor(num.feat)
corrplot::corrplot(feat.corr)
```

```
pairs(num.feat)
```

```
# levels(bnb$room_type)
#
# levels(bnb$neighbourhood_group)
#
# n_distinct(bnb$neighbourhood)
#
# n_distinct(bnb$neighbourhood_group)

# [1] "Entire home/apt" "Private room"    "Shared room"
# [1] "Bronx"         "Brooklyn"      "Manhattan"     "Queens"        "Staten Island"
# [1] 221
# [1] 5
```

There are 221 neighborhoods covered in the overall data, but only 5 neighbourhood groups. We will further investigate whether we need to use the neighbourhood, or whether we would like to use the negihbourhood groups for simplicity of our model.

We will determine whether we should use the neigbourhood by seeing if there is a large disparity in mean price by calculating the mean price for the neighbourhood. If there seems to be large disparities within the neighbourhood group for mean pricing, we will attempt to use neighbourhood itself.

```
hist(log(train$price), main='Histogram of log(Price)', xlab = 'log(Price)')
```

```
n <- ggplot(data = train, aes(x = latitude, y = longitude, color = neighbourhood)) +
  geom_point() + theme(legend.position="none") + xlab('Latitude') + ylab('Longitude') +
  ggtitle('Neighbourhood')

ng <- ggplot(data = train, aes(x = latitude, y = longitude, color = neighbourhood_group)) +
  geom_point() + theme(legend.position="none") + xlab('Latitude') + ylab('Longitude') +
  ggtitle('Neighbourhood Groups')

fig <- ggarrange(n, ng, nrow = 1, ncol = 2)

annotate_figure(fig, top = 'New York City Airbnb Listing Locations By:')
```

```
mean_price <- train %>% group_by(neighbourhood) %>% summarise(mean_price = mean(price),
                                                              latitude = median(latitude), longitude =
                                                              median(longitude))

#plot(mean_price$latitude, mean_price$longitude, col = mean_price$mean_price)

ggplot(data = mean_price, aes(x = latitude, y = longitude, color = mean_price)) +
  geom_point() + scale_color_gradient(low="blue", high="red", name = 'Mean Price (USD)') + xlab('Latitu
  ylab('Longitude') + ggtitle('Mean Price by Neighborhood')
```

It does not seem there are any large disparities in pricing, and all the neighbourhood groups seems to be similar to their nearby neighbours. To reduce the complexity of our model, we will use the neighbourhood group.

```
par(mfrow = c(1,2))
hist(bnb$price, main='Histogram of Overall Price', xlab = 'Price (USD)')
hist(train$price, main='Histogram of Training Price', xlab = 'Price (USD)')
```

```
num.feat <- train %>% dplyr::select(longitude, latitude, minimum_nights, number_of_reviews,
                                    reviews_per_month, calculated_host_listings_count, price)
num.feat$lprice <- log(num.feat$price)
feat.corr <- cor(num.feat)
corrplot::corrplot(feat.corr)
```

```
set.seed(123)
ols.price <- lm(log(price) ~ latitude + longitude + minimum_nights + reviews_per_month +
                        neighbourhood_group +  room_type + calculated_host_listings_count, data = train)
ols.pred <- predict(ols.price, test)

ols.mspe <- mean((log(test$price)-ols.pred)^2)
#ols.mspe

#summary(ols.price)
```

```
#initial LR model
set.seed(123)
logit.fit <- glm(price_above ~longitude + latitude+ minimum_nights+ calculated_host_listings_count+
                    reviews_per_month + room_type + neighbourhood_group, data=train,
                 family=binomial('logit'))

fit.pred <- predict(logit.fit, test, type="response")

#Finding the best threshold value
```

```r
# thresh <- seq(0,1,.01)
# miss <- NULL
#
# for(t in thresh){
#   fit.pred.lab <- ifelse(fit.pred>t, 1, 0)
#   err <- mean(fit.pred.lab!=test$price_above)
#   miss  <- append(miss, err)
# }
#
# best.thresh  <- thresh[which.min(miss)]

#table(fit.pred,test$price_above)
pred <- prediction(fit.pred, test$price_above)
perf <- performance(pred, "tpr","fpr")
plot(perf, col=2, lwd=3, main="ROC curve")
abline(0,1)

auc = performance(pred, "auc")@y.values
```

```r
set.seed(123)
fpr = performance(pred, "fpr")@y.values[[1]]
cutoff = performance(pred, "fpr")@x.values[[1]]
fnr = performance(pred,"fnr")@y.values[[1]]

rate = as.data.frame(cbind(Cutoff=cutoff, FPR=fpr, FNR=fnr))
rate$distance = sqrt((rate[,2])^2+(rate[,3])^2)

index = which.min(rate$distance)
best.thresh = rate$Cutoff[index]

matplot(cutoff, cbind(fpr,fnr), type="l",lwd=2, xlab="Threshold",ylab="Error Rate")
legend(0.3, 1, legend=c("False Positive Rate","False Negative Rate"),
col=c(1,2), lty=c(1,2))
abline(v=best.thresh, col=3, lty=3, lwd=3)
```

```r
set.seed(123)
logit.fit <- glm(price_above ~longitude + latitude+ minimum_nights+ calculated_host_listings_count+
                   reviews_per_month + room_type + neighbourhood_group, data=train,
                 family=binomial('logit'))

fit.pred <- predict(logit.fit, test, type="response")

logit.pred.best <- ifelse(fit.pred>best.thresh, 1, 0)
logit.miss <- mean(logit.pred.best!=test$price_above)

table(Predicted = logit.pred.best, True = test$price_above)
```

```r
set.seed(123)
class.tree <- rpart(as.factor(price_above)~latitude + longitude + minimum_nights + reviews_per_month +
                      neighbourhood_group +  room_type + calculated_host_listings_count, data = train)

tree.class.prediction <- predict(class.tree, test, type = 'class')
tree.class.misclass <- mean(test$price_above !=tree.class.prediction)
```

```
fancyRpartPlot(class.tree, digits = 6, main = 'Classification Tree for Price Above Median', sub = '')
```

```
set.seed(123)

#Pruning tree did not improve tree

# prune <- prune.tree(loc.tree, best = 4, newdata = test)
# plot(prune)
# text(prune)
# tree.class.misclass
```

```
set.seed(123)
tree.reg <- rpart(log(price) ~ latitude + longitude + minimum_nights + reviews_per_month +
                        neighbourhood_group +  room_type + calculated_host_listings_count, data = train)

reg.prediction <- predict(tree.reg, test)
tree.mspe <- mean((log(test$price)-reg.prediction)^2)

fancyRpartPlot(tree.reg, digits = 6,  sub = '', main = 'Regression Tree for log(Price)')
```

```
set.seed(123)
rf.class <- randomForest(as.factor(price_above)  ~ latitude + longitude + minimum_nights +  reviews_per_
                        neighbourhood_group +  room_type + calculated_host_listings_count,
                      data = train, mportance = TRUE)

#randomForest::importance(rf.class)

rf.class.pred <- predict(rf.class,test)
rf.misclass <- mean(test$price_above!=rf.class.pred)

#sprintf('The misclassification rate for the classification random forest is %f', rf.misclass)
#varImpPlot(rf.class)
```

```
set.seed(123)
rf.reg <- randomForest(log(price) ~ latitude + longitude + minimum_nights +  reviews_per_month +
                        neighbourhood_group +  room_type + calculated_host_listings_count,
                      data = train, mportance = TRUE)

rf.reg.pred <- predict(rf.reg,test)
rf.mspe <- mean((log(test$price)-rf.reg.pred)^2)

#sprintf('The mean squared prediction error for the regression random forest is %f', rf.mspe)
```

There still does not seem to be much of an improvement over the regression tree. We can try to re-evaluate the random forest model through cross validation and seeing if we can select important features.

The variable importance plot shows us that `room_type`, `longitude`, `latitude`, and `reviews_per_month` are the most important variables.

```
par(mfrow = c(2,1))
varImpPlot(rf.class, main = 'RF Classification Model')
varImpPlot(rf.reg, main = 'RF Regression Model')
```

```
set.seed(123)
rf.cv.trainx <- train %>% dplyr::select(latitude, longitude, minimum_nights, reviews_per_month,
                                neighbourhood_group, room_type, calculated_host_listings_count)
```

```
rf.cv.trainy <- log(train$price)
cv.rf <- rfcv(rf.cv.trainx,rf.cv.trainy, cv.fold = 5)
plot(cv.rf$n.var, cv.rf$error.cv, type = 'b', xlab = 'Number of Variables in Model', ylab = 'Cross-Vali
```

As we can see, the cross validation error is the lowest when we use the most predictors. Despite this, There does not seem to be much of a decrease after there are 4 variables in the model, so we will try to fit a model with 4 variables.

We will try fitting the 4 most important variables from the regression random forest, and seeing whether this model is better, or the same, as out more complex model.

```
set.seed(123)
rf.reg.reduced <- randomForest(log(price) ~ latitude + longitude + neighbourhood_group + room_type,
                               data = train, importance = TRUE)

rfr.reg.pred <- predict(rf.reg.reduced,test)
rfr.reg.mspe <- mean((log(test$price)-rfr.reg.pred)^2)

#sprintf('The mean squared prediction error for the reduced regression random forest is %f', rf.red.msp
```

By reducing the number of predictors, we were able to slightly increase the MSE, while creating a much simpler model.

Lets try bagging with the smaller subset of variables

```
set.seed(123)
bag.reg <- randomForest(log(price)  ~ latitude + longitude + reviews_per_month +  room_type,
                    data = train, mtry = 4 , importance = TRUE)

bag.reg.pred <- predict(bag.reg,test)
bag.mspe <- mean((log(test$price)-bag.reg.pred)^2)

# sprintf('The mean squared prediction error for bagging is %f', bag.mspe)
```

The prediction error is about the same as it is for a random forest.

```
set.seed(123)
bag.class <- randomForest(as.factor(price_above) ~ latitude + longitude + reviews_per_month +  room_type
                    data = train, mtry = 4 , importance = TRUE)

bag.class.pred <- predict(bag.class,test)
bag.misclass <- mean(test$price_above!=bag.class.pred)

# sprintf('The misclassification rate for the bagged classification model is %f', bag.misclass)
```

```
set.seed(123)
boost.mod <- gbm(log(price) ~ latitude + longitude +   reviews_per_month +  room_type, data = train,
                 n.trees = 1000, cv.folds = 5, distribution = 'gaussian')
boost.pred <- predict(boost.mod, test, n.trees = 1000)

boost.mspe <- mean((log(test$price)-boost.pred)^2)

#sprintf('The mean squared prediction error for boosting is %f', boost.mspe)
```

As with all our other models, this one is about the same. We can try different values of the shrinkage and see if we can find a best model for cross validation error

```
set.seed(123)
lambdas <- seq(0,.15, .002)
b.mspe.list <- NULL

for(lambda in lambdas){
  boost.l.mod <- gbm(log(price) ~ latitude + longitude +   reviews_per_month +  room_type, data = train
                n.trees = 1000, shrinkage = lambda, distribution = 'gaussian')
  boost.l.pred <- predict(boost.l.mod, test, n.trees = 1000)

  b.mspe.list <- append(b.mspe.list,mean((log(test$price)-boost.l.pred)^2))
}
plot(lambdas,b.mspe.list, type = 'b', ylab = 'Boosted MSPE', xlab = 'Lambda',
     main = 'MSPE vs. Lambdas')
```

There does not seem to be a discernable lambda from the plot.

```
set.seed(123)
best.lambda <- lambdas[which.min(b.mspe.list)]

best.boost <- gbm(log(price) ~ latitude + longitude +   reviews_per_month +  room_type, data = train,
                n.trees = 1000, distribution = 'gaussian')
best.boost <- predict(boost.mod, test, n.trees = 1000)

b.boost.mspe <- mean((log(test$price)-best.boost)^2)

#sprintf('The mean squared prediction error for bagging with the optimal lambda is is %f', b.boost.mspe
```

```
set.seed(123)
tree.err <- NULL

ntrees <- seq(250,2500,250)

for(ntree in ntrees){
  boost.t.mod <- gbm(log(price) ~ latitude + longitude +
                       reviews_per_month +  room_type, data = train, n.trees = ntree,
                     shrinkage = best.lambda, distribution = 'gaussian')
  boost.t.pred <- predict(boost.t.mod, test,n.trees = ntree)

  tree.err <- append(tree.err,mean((log(test$price)-boost.t.pred)^2))
}
#tree.err

plot(ntrees, tree.err, type = 'b', ylab = 'Boosted MSPE', xlab = 'Trees',
     main = 'MSPE vs. Number of Trees')
best.tree <- ntrees[which.min(tree.err)]
```

```
set.seed(123)
lambdas <- seq(0,.15, .002)
b.mis.list <- NULL

for(lambda in lambdas){
  boost.m.mod <- gbm(as.character(price_above) ~ latitude + longitude +
                       reviews_per_month +  room_type, data = train,
                     n.trees = 1000, shrinkage = lambda, distribution = 'bernoulli')
  boost.m.pred <- predict(boost.m.mod, test, n.trees = 1000, type = 'response')
```

```
  boost.m.pred <- ifelse(boost.m.pred>=.51, 1,0)

  b.mis.list <- append(b.mis.list, mean(test$price_above!=boost.m.pred))
}
plot(lambdas, b.mis.list, type = 'b', ylab = 'Boosted Misclass', xlab = 'Lambda',
     main = 'Misclassification. vs. Lambdas')

best.class.lambda = 0.1
```

```
set.seed(123)

# class.boost <- train(as.factor(price_above) ~ latitude + longitude + reviews_per_month +  room_type,
#                      method = 'gbm', data = train, verbose = FALSE)

class.boost <- gbm(as.character(price_above) ~ latitude + longitude + reviews_per_month +  room_type,
                   n.trees = best.tree, data = train, distribution = 'bernoulli')


boost.class.pred<- predict(class.boost, test, n.trees = 1000, type = 'response')
boost.class.pred <- ifelse(boost.class.pred>=.51, 1,0)


boost.misclass <- mean(test$price_above!=boost.class.pred)
#boost.misclass
```

```
set.seed(123)
boost.reg.final <- gbm(log(price) ~ latitude + longitude +   reviews_per_month +  room_type, data = tra
                 n.trees = 1000,  distribution = 'gaussian')
boost.pred <- predict(boost.reg.final, test, n.trees = 1000)

boost.mspe <- mean((log(test$price)-boost.pred)^2)

#sprintf('The mean squared prediction error for boosting is %f', boost.mspe)
```

```
train$price_above <- as.factor(train$price_above)

plot(rbind(train, test)$latitude,
     rbind(train,test)$longitude,
     col = rbind(train, test)$price_above,
     main = "Price by Location",
     xlab = "Latitude",
     ylab = "Longitude")
```

```
# determine approximate best cost parameter
# tune.linear <- tune(svm, price_above ~ latitude+longitude, data = train, kernel = "linear", ranges =
# increase precision of cost parameter
# tune.linear <- tune(svm, price_above ~ longitude
#                     +latitude
#                     +neighbourhood
#                     +minimum_nights
#                     +room_type
#                     +minimum_nights
#                     +number_of_reviews
#                     +calculated_host_listings_count
#                     +availability_365, data = train, kernel = "linear", ranges = list(cost = seq(.05,
```

```r
plot(svm(price_above ~ longitude+latitude, data = train, kernel = "linear", cost = 0.06), test[,c("pric
best.linear <- svm(price_above ~ longitude
                   +latitude
                   +neighbourhood_group
                   +minimum_nights
                   +room_type
                   +number_of_reviews
                   +calculated_host_listings_count
                   +availability_365, data = train, kernel = "linear", cost = 0.06)
pred.linear <- predict(best.linear, test)

# confusion matrix
conf.linear <- table(obs = test$price_above, pred = pred.linear)
acc.linear <- 1 - sum(diag(conf.linear)/sum(conf.linear))
```

```r
# tune.poly <- tune(svm, price_above ~ longitude
#                     +latitude
#                     +neighbourhood_group
#                     +minimum_nights
#                     +room_type
#                     +minimum_nights
#                     +number_of_reviews
#                     +calculated_host_listings_count
#                     +availability_365, data = train, kernel = "polynomial", ranges = list(cost = c(10

plot(svm(price_above ~ longitude + latitude, data = train, kernel = "polynomial", cost = 100),
     test[,c("price_above", "longitude", "latitude")])
best.poly <- svm(price_above ~ longitude
                   +latitude
                   +neighbourhood_group
                   +minimum_nights
                   +room_type
                   +number_of_reviews
                   +calculated_host_listings_count
                   +availability_365, data = train, kernel = "polynomial", cost = 100)
pred.poly <- predict(best.poly, test)
# (MSE.poly <- mean((as.numeric(test$price_above) - as.numeric(pred.poly))^2))

# confusion matrix
conf.poly <- table(obs = test$price_above, pred = as.factor(pred.poly))
acc.poly <- 1 - sum(diag(conf.poly)/sum(conf.poly))
```

```r
# tune.rad <- tune(svm, price_above ~ longitude+latitude, data = train, kernel = "radial", ranges = lis
# tune.rad <- tune(svm, price_above ~ longitude
#                     +latitude
#                     +neighbourhood_group
#                     +minimum_nights
#                     +room_type
#                     +number_of_reviews
#                     +calculated_host_listings_count
#                     +availability_365, data = train, kernel = "radial", ranges = list(cost = seq(8, 1
plot(svm(price_above ~ longitude+latitude, data = train, kernel = "radial", cost = 8), test[,c("price_a
best.rad <- svm(price_above ~ longitude
                   +latitude
```

```
                    +neighbourhood_group
                    +minimum_nights
                    +room_type
                    +number_of_reviews
                    +calculated_host_listings_count
                    +availability_365, data = train, kernel = "radial", cost = 8)
pred.rad <- predict(best.rad, test)
conf.rad <- table(obs=test$price_above, pred=pred.rad)
acc.rad <- 1 - sum(diag(conf.rad)/sum(conf.rad))
```

```
set.seed(3)
knn.caret <- train(as.factor(price_above) ~ neighbourhood_group+latitude+longitude+room_type+minimum_nig
plot(knn.caret)
```

```
set.seed(3)
pred.knn <- predict(knn.caret, test)

conf.matrix <- table(pred = pred.knn, obs = test$price_above)
acc.knn <- 1 - sum(diag(conf.matrix))/sum(conf.matrix)
```

```
xTrain <- num.feat %>% dplyr::select(-c(price,lprice)) %>% as.matrix()
yTrain <- num.feat$lprice

X <- data.matrix(num.feat)
y <- log(train$price)
Ridge <- glmnet(x = xTrain, y = yTrain, alpha = 0)
RidgeCV <- cv.glmnet(x = xTrain, y = yTrain, alpha = 0, lambda = Ridge$lambda, nfolds = 10)
lambda.ind <- which.min(RidgeCV$cvm)
lambda.best <- Ridge$lambda[lambda.ind]
#lambda.best
#Ridge$beta[, lambda.ind]



proRidgeT <- glmnet(x = xTrain, y = yTrain, alpha = 0, lambda = lambda.best)
XTest <- test %>%  dplyr::select(longitude, latitude, minimum_nights, number_of_reviews,
                                 reviews_per_month, calculated_host_listings_count) %>% as.matrix()
yTest <- log(test$price)
yPredRidge <- proRidgeT$a0 + XTest%*%proRidgeT$beta
proMSPE.Ridge <- mean((yTest - yPredRidge@x)^2)
#proMSPE.Ridge
```

```
xTrain <- num.feat %>% dplyr::select(-c(price,lprice)) %>% as.matrix()
yTrain <- num.feat$lprice
Lasso <- glmnet(x = xTrain, y = yTrain, alpha = 1)


LassoCV <- cv.glmnet(x = xTrain, y = yTrain, alpha = 1, lambda = Lasso$lambda, nfolds = 10)
lambda.indL <- which.min(LassoCV$cvm)
lambda.bestL <- Lasso$lambda[lambda.indL]
#lambda.bestL
#Lasso$beta[, lambda.indL]
```

```
proLassoT <- glmnet(x = xTrain, y = yTrain, alpha = 1, lambda = lambda.best)
XTest <- test %>%  dplyr::select(longitude, latitude, minimum_nights, number_of_reviews,
                                 reviews_per_month, calculated_host_listings_count) %>% as.matrix()
yTest <- log(test$price)
yPredLasso <- proLassoT$a0 + XTest%*%proLassoT$beta
proMSPE.Lasso <- mean((yTest - yPredLasso@x)^2)
#proMSPE.Lasso
```

```
ols.tab <- data.frame(Method = 'OLS', MSPE = ols.mspe)
logit.tab <- data.frame(Method = 'Logistic', Misclass. = logit.miss)

kable(list(MSPE = ols.tab, Misclassification  = logit.tab),
      caption = 'Error Rates for Linear Models') %>% kable_styling(latex_options = "hold_position")
```

```
Methods <- c('Tree','Bagging','Boosting', 'Random Forest', 'Reduced Random Forest')
MSPE <- round(c(tree.mspe, bag.mspe, boost.mspe, rf.mspe, rfr.reg.mspe),6)
Misclassification <- round(c(tree.class.misclass, bag.misclass ,boost.misclass, rf.misclass, rfr.miscla

mspe.tab  <- cbind(Methods,MSPE)
misclass.tab <- cbind(Methods,Misclassification)

kable(list(`MSPE` = mspe.tab, Misclassification  = misclass.tab),
      caption = 'Error Rates for Tree Models') %>% kable_styling(latex_options = "hold_position")
```

```
Methods <- c("Linear Kernel SVM", "Polynomial Kernel SVM", "Radial Kernel SVM", "KNN")
Misclassification <- round(c(acc.linear, acc.poly, acc.rad, acc.knn), 6)

svm.misclass.tab <- cbind(Methods,Misclassification)

kable(list(`Misclassification`  = svm.misclass.tab),
      caption = 'Misclassification Rates for SVM and KNN') %>% kable_styling(latex_options = "hold_posi
```

```
Methods <- c('Ridge', 'Lasso')
MSPE <- c(proMSPE.Ridge, proMSPE.Lasso)

reg.tab <- cbind(Methods,MSPE)
kable(reg.tab,
      caption = 'MSPE for Regularization Methods') %>% kable_styling(latex_options = "hold_position")
```

```
set.seed(123)
ols.price <- lm(log(price) ~ latitude + longitude + minimum_nights + reviews_per_month +
                      neighbourhood_group +  room_type + calculated_host_listings_count, data = train)

final.ols.pred <- predict(ols.price, bnb)
final.ols.mspe <- mean((log(bnb$price)-final.ols.pred)^2)
```

```
set.seed(123)
final.logit <- glm(price_above ~longitude + latitude + minimum_nights +
                   reviews_per_month + calculated_host_listings_count + room_type +
                   neighbourhood_group, data=train, family=binomial('logit'))

final.logit.pred <- predict(final.logit, bnb, type="response")
final.logit.pred <- ifelse(final.logit.pred>best.thresh, 1, 0)
final.logit.misclass <- mean(final.logit.pred!=bnb$price_above)
```

```r
#kable(list(table = table(predicted = final.logit.pred,true = bnb$price_above)), caption = 'Confusion M
#  kable_styling(latex_options = "hold_position")

log.tab <- table(Predicted = final.logit.pred,True = bnb$price_above)
#kable(table(Predicted = final.logit.pred,True = bnb$price_above)) %>% kable_styling(latex_options = "h
log.tab
```