Predicting Avocado Prices with Different Models

Rebecca Gordon[1]

[1] University of Oregon

Author Note

Correspondence concerning this article should be addressed to Rebecca Gordon.
E-mail: rebeccag@uoregon.edu

Predicting Avocado Prices with Different Models

The outcome variable

Across the entire United States, avocados are being sold everyday in high volume despite their fluctuating price. Our goal for this project was to create a machine learning model capable of accurately predicting a given state's avocado price over time across states. The purpose of this project is to provide consumers, local grocers, and farmer's markets with a simple tool to predict avocado prices so that the largest profits are not taken by large-chain grocery stores.

We will be answering the research question: What is the strongest predictor of avocado prices in the United States? Thus, our goal is to find the feature in the data that most strongly predicts the price of avocados in the United States.

Description of the Data

Core features and descriptive statistics

To answer our question, we analyzed the avocado prices dataset retrieved from Kaggle.com and compiled by the Hass Avocado Board. The dataset consists of approximately 18,000 avocado sale records from 2015-2018. The dataset contains information about avocado prices by type (organic or conventional), region purchased in the United States, total volume sold, and date sold.

Frequencies and distribution of data

First, we subsetted the variables of interest from the dataset. From the histogram below, we can see that our outcome variable, average avocado price, is normally distributed. Mean price across data was $1.41 ($SD$ = $0.40).

Table 1

Frequencies of the data

	n	mean	sd	median	min	max	range	skew	kurtosis	se
type*	18249	1.50	0.50	1.00	1.00	2.00	1.00	0.00	-2.00	0.00
average_price	18249	1.41	0.40	1.37	0.44	3.25	2.81	0.58	0.32	0.00
total_volume	18249	850644.01	3453545.36	107376.76	84.56	62505646.52	62505561.96	9.01	92.07	25564.99
region*	18249	27.50	15.58	27.00	1.00	54.00	53.00	0.00	-1.20	0.12

We visually examined the distribution of average avocado price by type. From the figure below, we can see that organic avocados are on average more expensive than conventional avocados. The highest average price for organic avocados was in San Francisco, CA in 2016 for $3.25 and the lowest average price was in Cincinnati, OH in 2017 for $0.44.

Missing data check

```
##            x1           date average_price  total_volume         x4046
##             0              0             0             0             0
##         x4225          x4770    total_bags    small_bags    large_bags
##             0              0             0             0             0
##  x_large_bags           type          year        region
```

```
## 0 0 0 0
```

No missingness was found for the variables in the dataset.

Description of the models

Three different modeling approaches will be used to predict avocado price from sale features, including: Linear Regression, Decision Trees, and Random Forest. Since the purpose of this project is to provide consumers, local grocers, and farmer's markets with a tool to predict avocado prices, we want to examine the predictive power of several features that contribute to avocado price. Thus, we first examined the effect of all predictors in a linear regression model to compare with the more advanced models. Next, we added more complexity to the linear model by growing and pruning decision tree regression models to predict avocado price. Finally, we used a random forest regression model using the significant features from the analysis to reduce the variance to get a more accurate prediction.

Model Fits

Preparation

The dataset is split into training and test set with the following code. We used a 80-20 split for the data. The smaller test dataset will be used as a final hold-out set, and training dataset will be used to build the model. The training set has 14,599 observations, and the test set has 4,650 observations. We will evaluate model performance by examining fit features to predict avocado price (RMSE, MAE, and R^2).

```
require(recipes)
loc <- sample(1:nrow(df), round(nrow(df) * 0.8))
df_train  <- df[loc, ]
df_test  <- df[-loc, ]
```

Model 1: Linear Regression Model with Cross Validation

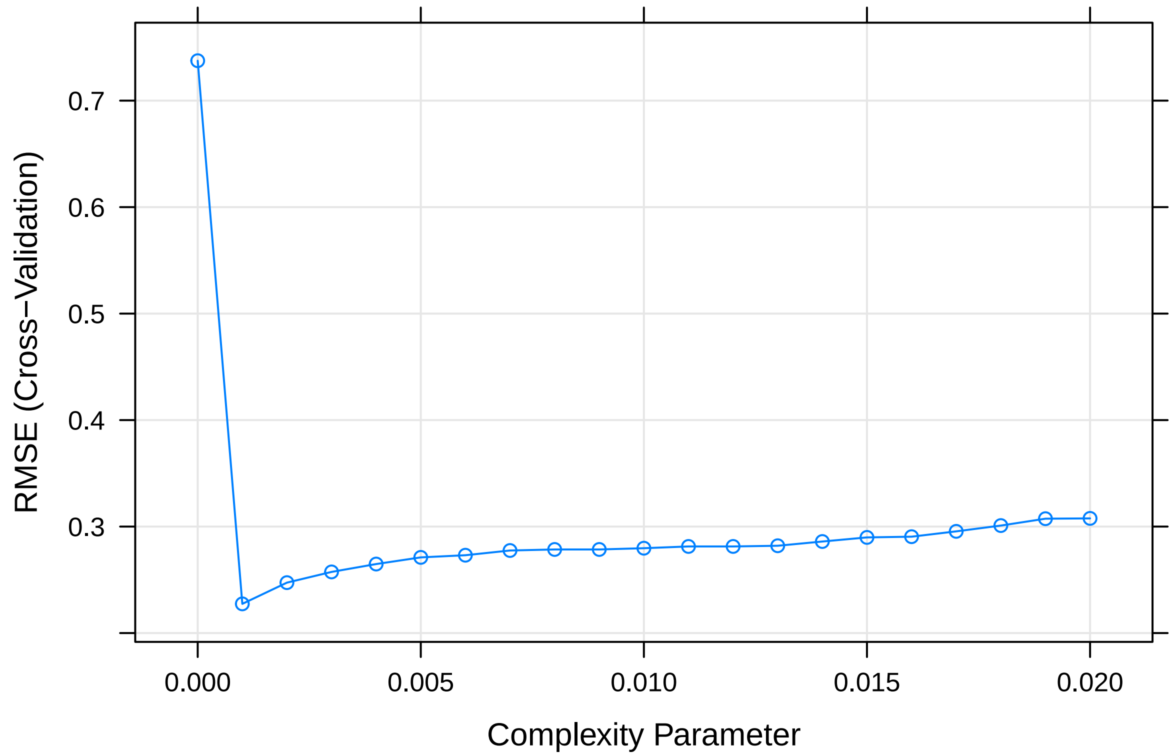
We first fitted a linear regression model without regularization. Since our outcome is continuous, we first want to examine if there is a correlation between the different variables. The equation generated by the linear model will then be applied to predict outcome of new unseen data. Our criteria for evaluation of model performance will be the root mean square error (RMSE) and R-sqaured (R^2). We used 10-fold cross validation to train and test classifiers.

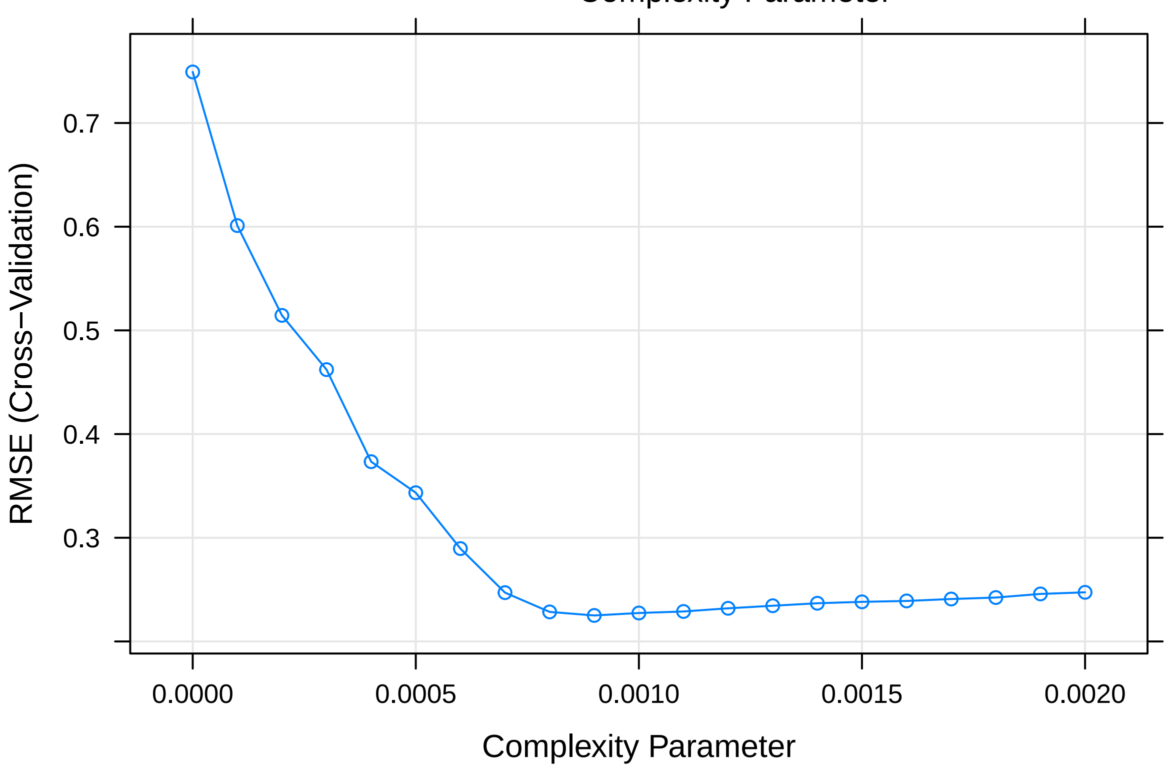
```
##   intercept      RMSE  Rsquared      MAE      RMSESD  RsquaredSD       MAESD
## 1      TRUE 0.2653203 0.5681895 0.199722 0.007260751 0.007208039 0.004739862
```

Model 2: Decision Trees

Next, we fitted decision tree regression model with cross validation to get a better estimate of the generalization error on unseen data using the split test data. We manually tuned the hyper-parameter grid as well as maximum depth and minimum number of observations to optimize the model fit, as shown in the figures below.

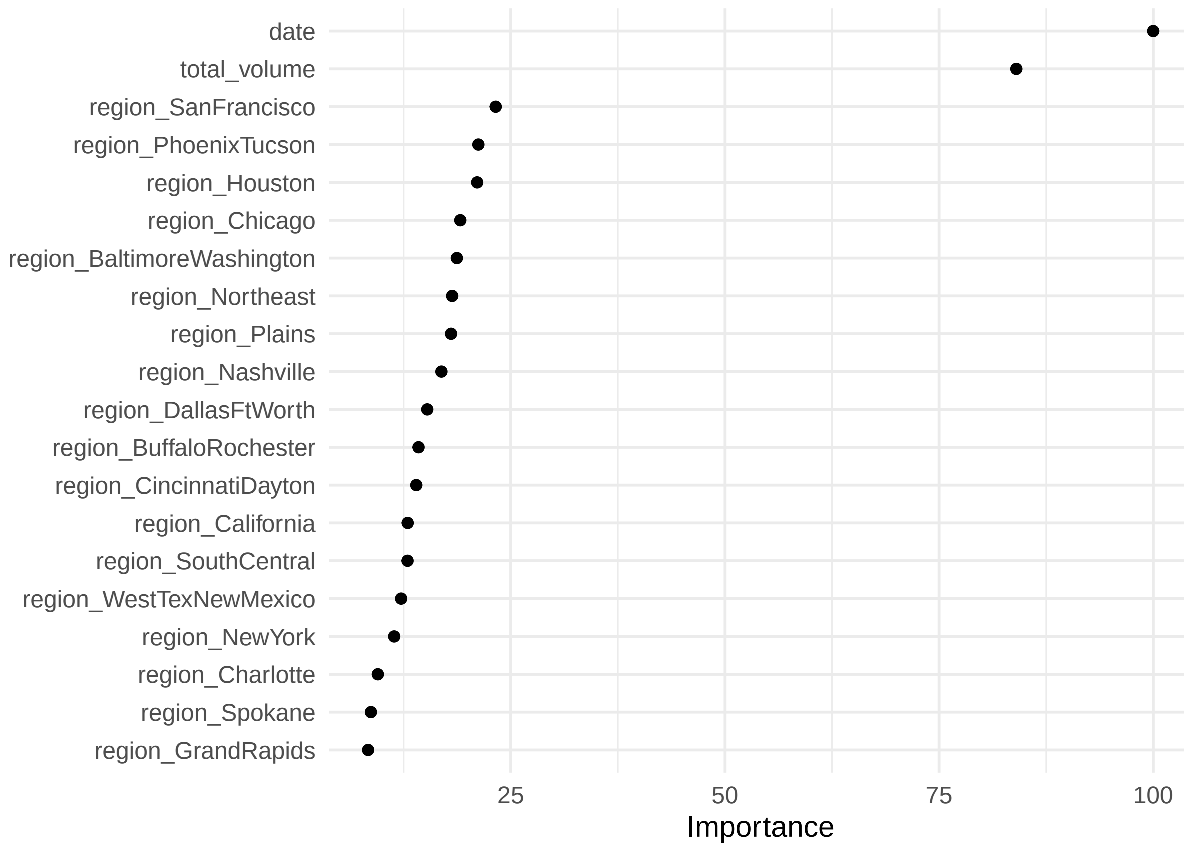
```
##   parameter   class                label
## 1        cp numeric Complexity Parameter
```


RMSE (Cross−Validation)
0.7
0.6
0.5
0.4
0.3
0.000
0.005
0.010
0.015
0.020
Complexity Parameter


RMSE (Cross−Validation)
0.7
0.6
0.5
0.4
0.3
0.0000
0.0005
0.0010
0.0015
0.0020
Complexity Parameter

```
##       cp
## 2 0.001
```

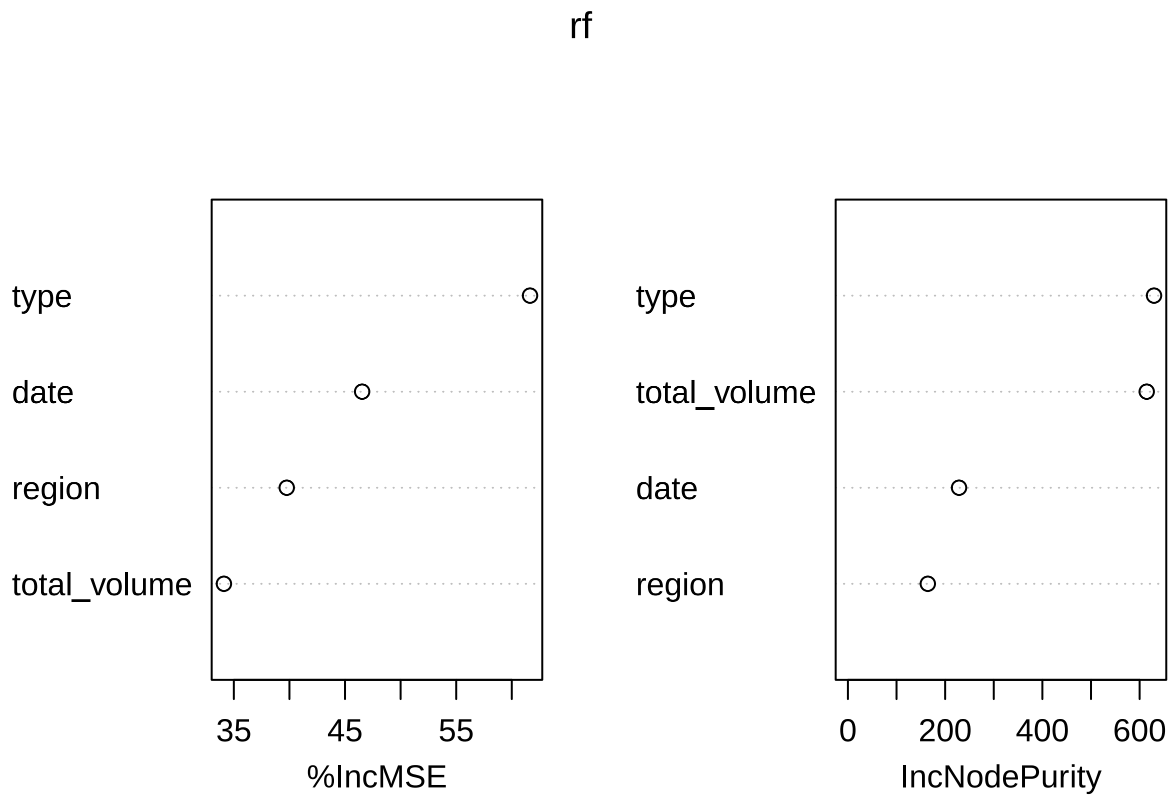
Next, we examined the complexity parameters and importance for the model and found that date sold and total volume sold were the most important factors as predictors for average avocado price.



Model 3: Random Forest

Finally, we fitted a random forest regression model. We chose an `mtry` value of 5 as it is the total number of variables. We left the number of trees and node size as the standard values. According to the random forest regression, the top predictor of avocado prices is type (i.e. whether the avocado is organic or conventional). This result aligned with our expectations, as our preliminary data analyses depicted differences in distributions between organic and conventional avocado prices.

Next, we extracted the importance variables from the random forest model. The plot below shows that `type` has the strongest impact on average avocado price with the highest percentage increase in MSE and in node purity.




```
##                %IncMSE IncNodePurity
## date         46.53848      228.5543
## type         61.63902      629.6293
## total_volume 34.10729      614.6171
## region       39.75783      164.3508
```

Comparing Models

The linear model showed that most variables in the data were predictive of avocado price, thus further testing was necessary to develop and fine tune our tool. In the decision tree and random forest models we found an increase in variance predicted and reduction of error from the original linear model. Examining the predictions of each model, we can see that the random forest model outperformed the linear and decision tree models. This is because it has the highest R^2 and the least error. Thus, we can assume that random forest models can be trusted to predict avocado prices.

Model	Rsquare	RMSE	MAE
Linear Regression	0.5543246	0.2657232	0.2009898
Decision Trees	0.6932190	0.2205053	0.1697523
Random Forest	0.7723804	0.1974473	0.1473851

Data Visualization

Figure 1: Avocado Prices and total volume sold by type with regression lines

We examined our variables of interest visually with several plots. First we log-transformed the total volume sold to examine its relationship with average price. We can see that more conventional type avocados were sold at a lower price than organic avocados.

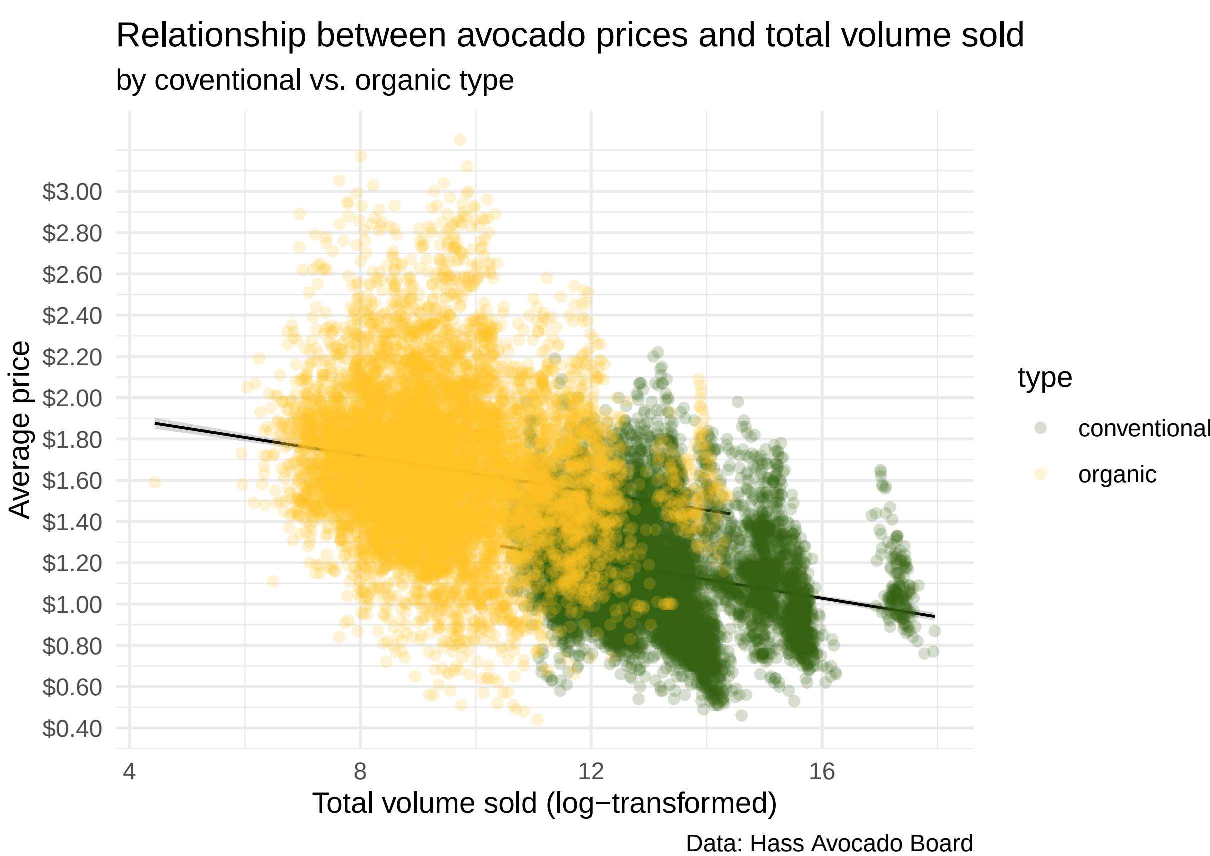


Figure 2: Distribution of avocado prices by type

We examined the difference in average price distribution across data by type and found that there is a clear difference between organic and conventional prices, such that organic avocados are more expensive.

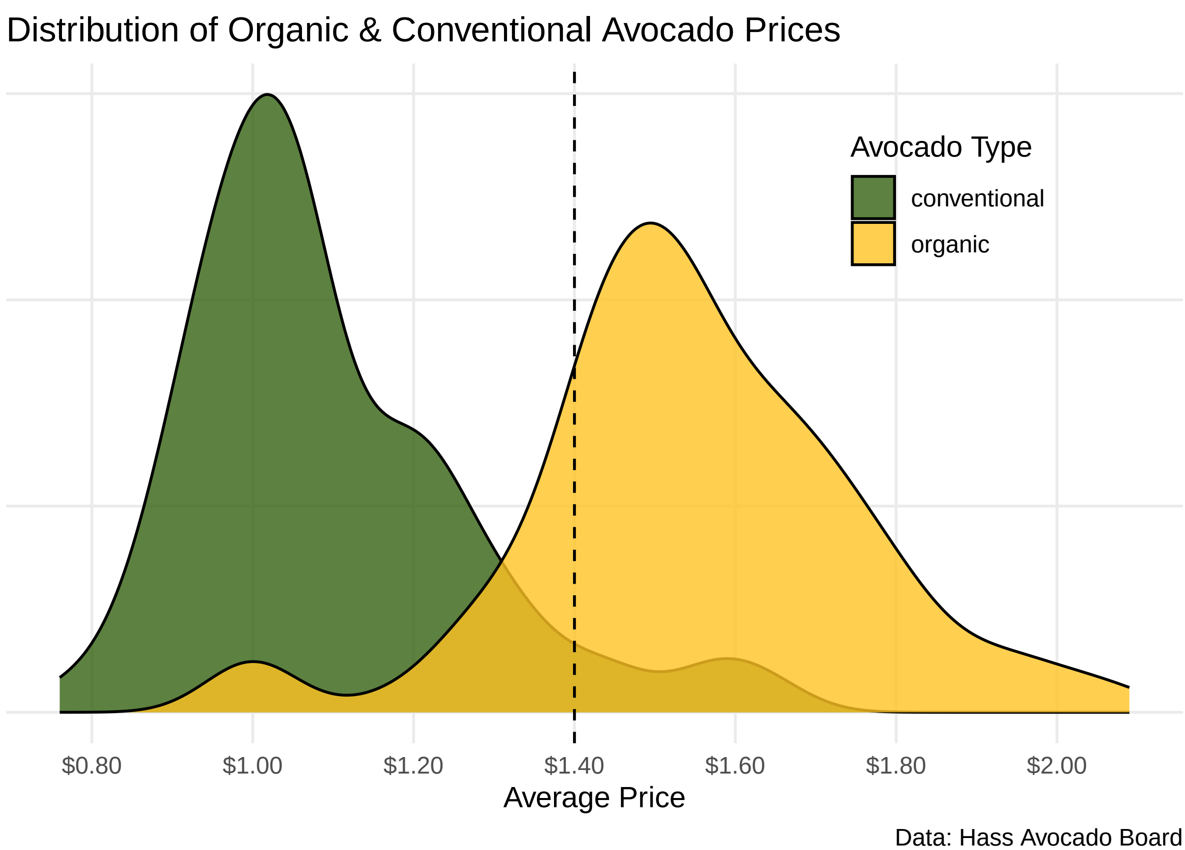
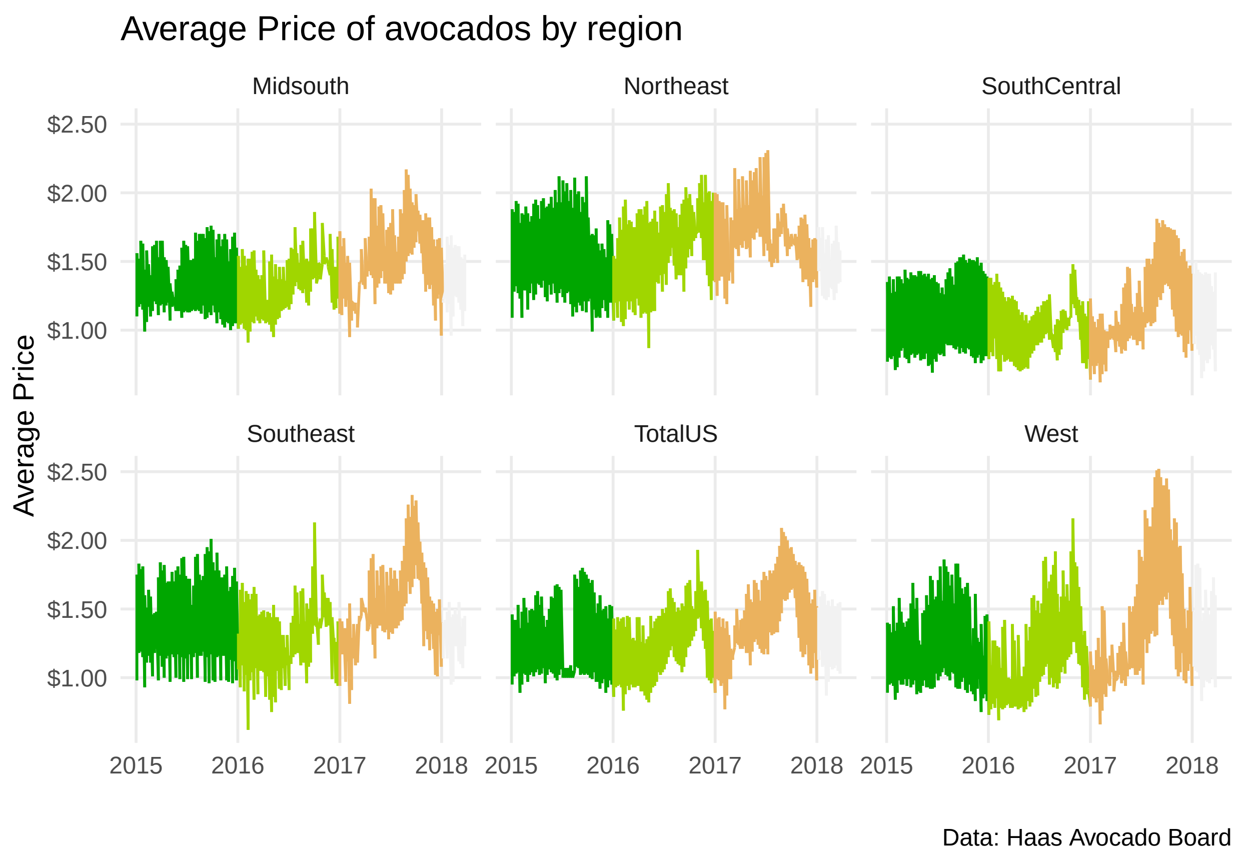


Figure 3: Regional avocado prices change over time

We examined avocado prices by data from a subset of regions to examine differences in avocado prices across regions compared with the total US. For this we included: Mid South, Northeast, South Central, Southeast, West regions, and the total US. From visually inspection, it is clear that Western regions have more expensive avocado prices, specifically in 2018, than the rest of the US.



Discussion

Conclusion

We optimized three models to predict average avocado prices and we found that the test scores for our predictive models were overall high. For the random forest regression model, the total variance explained was 91% and the decision tree model explained 69% of the variance.

We discovered some interesting findings from the models. The decision tree regression model predicted that date sold and total volume sold are the most important features for predicting avocado price. However, The random forest regression model predicted that type is the most important feature for predicting avocado price. This may be due to different parameters being used in the models. The random forest is more likely accurate since the prior data exploration showed such a strong difference between organic and conventional avocado prices. Further investigation is needed to examine this.

The region where the avocado was sold was an important feature in the pricing of avocados in the decision tree model. For instance, regions such as Baltimore/Washington and Houston were the third and fourth most important predictors of average avocado price. This tool can be used by small business owners to predict where the best time to buy and sell avocados is based on the predictions in this model. Overall, this tool can help the economy by allowing equity among food sellers in the United States.

References

Kiggins, J. (2018). *Avocado prices: Historical data on avocado prices and sales volume in multiple US markets.* Retrieved from https://www.kaggle.com/neuromusic/avocado-prices

Shahbandeh, M. (2019). *Average sales price of avocados in the u.s. 2012-2018.* Retrieved from https://www.statista.com/statistics/493487/average-sales-price-of-avocados-in-the-us/