Oliver Brooker, Christain Muresan, Kayelin Santa Elena, Nina Valdez

Professor Doosti

MGSC 310

May 22, 2020

Introduction: Avocado Dataset

The data set used for this project is an avocado price data set from kaggle.com. This data set consists of historical data on the price and sales volume of avocados across the United States, or US. From a business management standpoint, our goal is to analyze the price trends of avocados to observe the variables that allow avocado farmers to reach optimal revenues. Our objective is to compare variables such as organic versus conventional, or non-organic avocados, along with finding the optimal season for selling avocados. In addition, we want to identify the regions that have the largest impact on price. To determine this, we have generated 4 models, Linear Regression, Lasso Regression, K-Mean Clustering and a Random Forest model. Through these models, we will determine which variables are significant along with comparing the models' accuracy .

Dataset Overview

This dataset collected data for avocados, recorded every Sunday between January 4th 2015 and March 25th, 2018. It includes data on average avocado prices from 54 different regions in the U.S. Before performing any data analysis, the variables we predicted to be most important and relevant to our project were: Organic vs. Conventional avocados, Avocado Size, Bag Size, Location (Sold), Quarters (3 month periods allocated in each quarter), Total Volume (of Avocado's sold) and Region.

The avocados in the dataset were sold through different channels such as grocery stores, warehouse clubs, the drug industry, and the military. Additionally, the variables Region and

Type shown in the summary statistics table did not reveal much information. Prior to creating any models, the data set contained a number of 18,249 observations. As per common practice, we divided this into 25% testing and 75% training subsets in order to build our models. Lastly, we decided to create a new variable for assessing revenues. This was accomplished by multiplying the average price by the total volume of avocados. In performing this data transformation, we are able to simultaneously visualize the growth of volumes and price as it relates to Region, Type, and other important variables.

We found that the Extra Large Bags of Avocados tended to sell more volumes of conventional avocados compared to organic. Whereas, the Small and Large Bags of Avocados were mostly comprised of organic avocados. The small volume of organic avocados is offset by it's higher price average. Organic avocados also saw far less sporadic movements over time in regard to total volumes. There was greater volatility within the average prices for organic avocados compared to conventional.

Analysis: Linear Regression Model

Our linear model used all variables, with the expectation of dates, as predictors, and we used average price as our outcome variable. Due to omission of dates, we did not regress this variable, and we decided to use the Average Price variable as our outcome because we wanted to see how all the other variables would impact the price. Primarily, we are interested in the effect that region and type (organic or conventional) has on the average price. According to the coefficient plot, sorted by magnitude of largest to smallest, avocado type has the largest impact on the average price. In addition, the regions with the largest impact are all major cities include, San Francisco, Hartford, New York, Philadelphia, and Sacramento. There was a presence of heteroskedasticity in our train predictions, mainly among the organic datapoints, where the residual variance was marginally greater in organic than compared to conventional. We only saw a slight overfitting issue for this model, since the test RMSE of 0.2737 was very close to the train RMSE of 0.2659. However, the R^2 value was very low at a score of 0.5604. Unfortunately, this means that the model is not very accurate for predicting new inputs, which compelled us to try different and more sophisticated models.

Analysis: Lasso Regression Model

Our lasso model used all variables, with the exception of dates, and with average prices as our outcome. We used a cross-validated lasso model to find the best lambda. Lambda min was 0.0003347769 and removed 6 variables, and lambda 1se was 0.004127286 which removed 15 variables. We chose lambda min as our best lambda because it is smaller and also received a slightly higher R^2 score. It was interesting to discover that it had removed Total.Volume and Type Organic, which are important variables in our dataset. The model scores were very close to our linear regression model with a train MSE of 0.2659, a test RMSE of 0.2735, and an R^2 0.5493. Although the test RMSEs are similar, lasso's was a slightly lower, meaning that linear regression provided superior results when predicting prices on unseen data. Overall this model was quite disappointing because usually using a form of dimensionality reduction provided a better model.

Analysis: K-Means Clustering Model

Our K-Means modeling cluster was made of a 1,000 variable subset of the original dataset. The model could not compute the original amount of variables so we reduced it to 6%. The best number of clusters for our dataset was 2 because it had the highest silhouette score however, we decided to proceed with 3 clusters as we believed it could reveal more patterns and trends.

In the markdown report, Cluster 1 is displayed as dark green, cluster 2 as light green and cluster 3 as purple. Cluster 2 is the least dense and cluster 1 is the most dense. Cluster 2 also had the highest average total volume (31,606,403.80 avocados) and revenue ($34,647,017.20). Cluster 3 had the lowest average price at $1.05 and cluster 1 has the highest at $1.45.

Observing the “Average Price vs Revenue” and “Average Price vs Total Volume” graphs in the markdown report, it is clear that cluster 2 is superior compared to the other clusters in volume and revenue. This cluster sells avocados at a low price in large volumes which generates higher revenue, compared to cluster 1, who has the largest range in average price and sells the least in volume.

As the graphs “Average Price vs Total Volume by Quarter and Type” and “Total Volume vs Revenue) by Quarter and Type” demonstrate, clusters 2 and 3, sell only conventional avocados year round. They both have very high volumes and reasonable prices which generate good revenue. Cluster 3 demonstrates an increase in prices around quarter 3. In addition, cluster 1, which sells both organic and conventional year round, notices that generally prices and volume stay constant but around quarter 3 and quarter 4 prices increase for organic avocados. Perhaps this is due to the fact that prime avocado season is during Spring and Summer and in the colder months supply for organic avocados is limited.

Based on the graphs in this model, we came to a conclusion that Cluster 1, the least dense cluster, sells conventional avocados year round, sells low prices in high volume, and are most likely to consist of mass-market retailers and warehouse club stores like Costco and Sam’s Club where consumers can shop in bulk. Cluster 2, the most dense cluster, sells both organic and conventional avocados year round, has a large price range and the least volume, and increases their organic prices around quarter 3 and quarter 4. This cluster is most likely to consist of consumer grocery stores. Similarly to Cluster 1, Cluster 3 sells conventional avocados year round. They have low prices and sell a reasonable amount of volume. In addition, this cluster increases their prices around quarter 3. Just as Cluster 2, Cluster 3 is likely to consist of consumer grocery stores.

Our K-Means clustering model performed well as it received a silhouette score of about 0.85. This means that our clusters are cohesive (clusters members are similar) and separated (clusters are different from each other). To reiterate, the reason we decided to proceed with 3 clusters, despite having a lower silhouette score compared to 2 clusters, was because we believed the 3rd cluster would reveal more patterns and trends. The results proved to be interesting.

Analysis: Random Forest Regression Model

The random forest model was our most accurate model for determining average price. The model could not handle the original amount of variables so we reduced it to 6%, and for consistency and comparison. This model used the same subset reduction as the K-Means Clustering Model. We initially used a decision tree model, but after careful consideration, we decided to use a slightly more sophisticated model such as the random forest model. We based this change on a few main reasons; random forest is a supervised machine learning technique that forms many different decision trees, and tests each data point by running it through every tree and is trained using the bagging method. As a result, the model can use a large variety of combinations and tree models to accurately predict the outcome of an input. We decided to set the mtry parameter to 5, after manually testing multiple forests with varying mtry values. This way, it is more accurate which is evident by the low test RMSE of 0.2298 and low train RMSE of 0.2385. The R^2 value was higher than the other models at 0.7023. Evidently, this model significantly outperformed the other 3 models based on these measures. As predicted at the beginning of the project, Region and Type are among the most important variables in this model. In analyzing the importance charts, it shows that region and type had the largest increase in node purity for the random forest model, in addition to providing a high percent increase of MSE.

Conclusion

Our linear regression model demonstrated the importance of the Region and Type variables, as they had a substantial influence on the model's performance. Through this model, we were able to recognize that organic avocados had the largest impact on the average price. When comparing the linear and lasso regression model's RSME, the linear regression model proved to be more accurate. The linear model provided a marginally better test RMSE and R^2. Through the comparison of these two models, we discovered that the dimensionality reduction performed through lasso did not improve our model thus, it is unnecessary to use the lasso regression for this data set. Through the K-Means Clustering model we discovered that Clusters 2 and 3 demonstrated a constant sell of conventional avocados year-round. Both clusters displayed high volumes at reasonable prices which in turn provide high revenues for farmers. Cluster 1 during quarters 3 and quarter 4 showed to have an increase in prices for organic avocados. In addition, Cluster 3 during quarter 3 demonstrates an increase in prices overall. Based on the K-Means Clustering model, farmers should focus on conventional avocados. Farmers should also be cautious during quarters 3 and 4 as we see a change in prices and volume. We conclude that avocados peak seasons are during warmer quarters of the year, such as spring and summer. As a result, the supply for avocados decreases during the colder quarters of autumn and winter increasing demand thus increasing price. The model that proved to be the most accurate was our Random Forest regression model. In this model, we noticed that the most important variables were again region and type.

Overall, the models suggest that farmers should focus on conventional avocados. Farmers should also be cautious during quarters 3 and 4. The optimal season to sell avocados are quarter 1 and quarter 2. In addition, the regions that have the largest impact on price include San Francisco, Hartford, New York, Philadelphia, and Sacramento.