Machine Learning 1 Project

Avocado Pricing Dataset

Presented by Clifer Fernandez – April 2020 cohort

Agenda

- Introduction
- Executive Summary
- Data Profiling and Analysis
- Machine Learning Model
- Conclusions
- Recommendations

Introduction

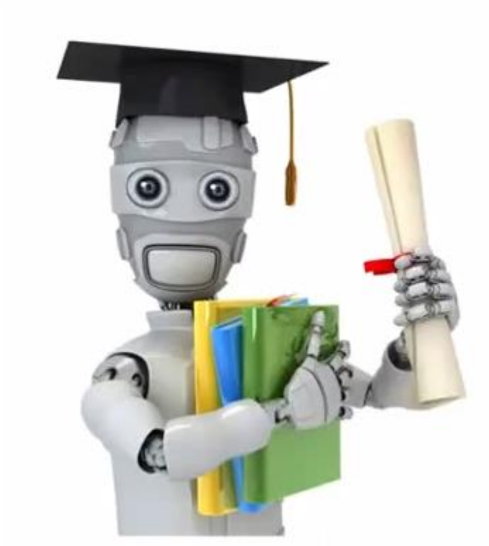
- Dataset - Average Pricing for single Avocado
  - 10 numerical – price, sales volume, etc.
  - 3 categorical – type, region, year

Our Task

- Select and Train a ML model to help predict the average price of a single avocado

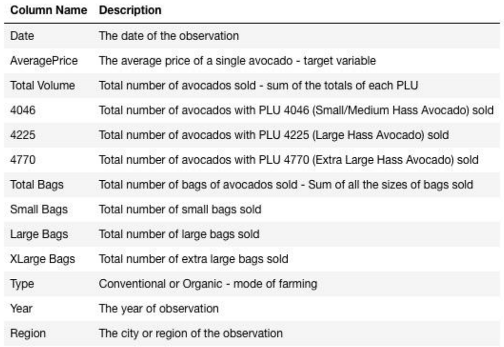
Executive Summary

- Original data set – 18249 records, no missing / null
- Analysis
  - Covers years 2015 – Q1 2018
  - Extremely high correlation between majority of variables
  - ***Small / medium size of conventionally grown, packed in small bags are preferred***
  - Increasing sales volume over the years
- Machine Learning Model
  - Regression models preferred
  - Random Forest Regressor model gives best result
- Recommendations
  - Other feature engineering methods
  - More powerful algorithms
    - Support Vector Machines (SVM)
    - Artificial Neural Networks (ANN)

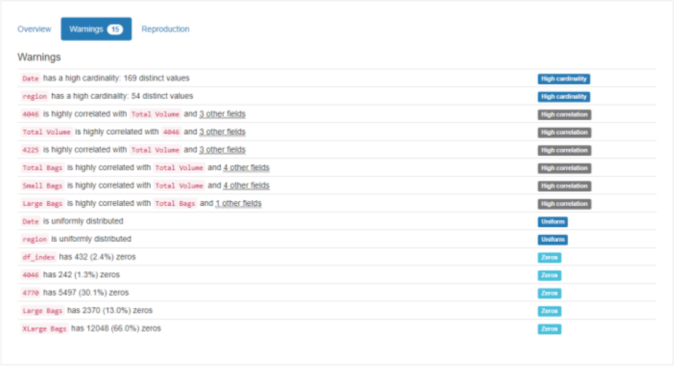


Data Profiling

- Data set
  - Import data
  - View records
  - Basic stats
  - Profiling



Column Name	Description
Date	The date of the observation
AveragePrice	The average price of a single avocado - target variable
Total Volume	Total number of avocados sold - sum of the totals of each PLU
4046	Total number of avocados with PLU 4046 (Small/Medium Hass Avocado) sold
4225	Total number of avocados with PLU 4225 (Large Hass Avocado) sold
4770	Total number of avocados with PLU 4770 (Extra Large Hass Avocado) sold
Total Bags	Total number of bags of avocados sold - Sum of all the sizes of bags sold
Small Bags	Total number of small bags sold
Large Bags	Total number of large bags sold
XLarge Bags	Total number of extra large bags sold
Type	Conventional or Organic - mode of farming
Year	The year of observation
Region	The city or region of the observation



Warnings

Date has a high cardinality: 169 distinct values — High cardinality

region has a high cardinality: 54 distinct values — High cardinality

4046 is highly correlated with Total Volume and 3 other fields — High correlation

Total Volume is highly correlated with 4046 and 3 other fields — High correlation

4225 is highly correlated with Total Volume and 3 other fields — High correlation

Total Bags is highly correlated with Total Volume and 4 other fields — High correlation

Small Bags is highly correlated with Total Volume and 4 other fields — High correlation

Large Bags is highly correlated with Total Bags and 1 other fields — High correlation

Date is uniformly distributed — Uniform

region is uniformly distributed — Uniform

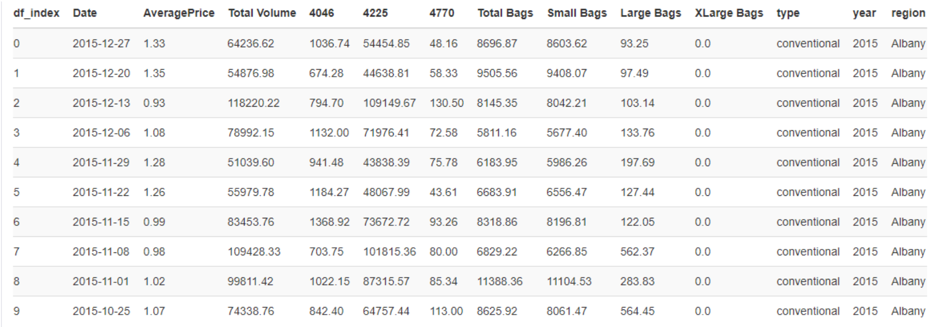
df_index has 432 (2.4%) zeros — Zeros

4046 has 242 (1.3%) zeros — Zeros

4770 has 5497 (30.1%) zeros — Zeros

Large Bags has 2370 (13.0%) zeros — Zeros

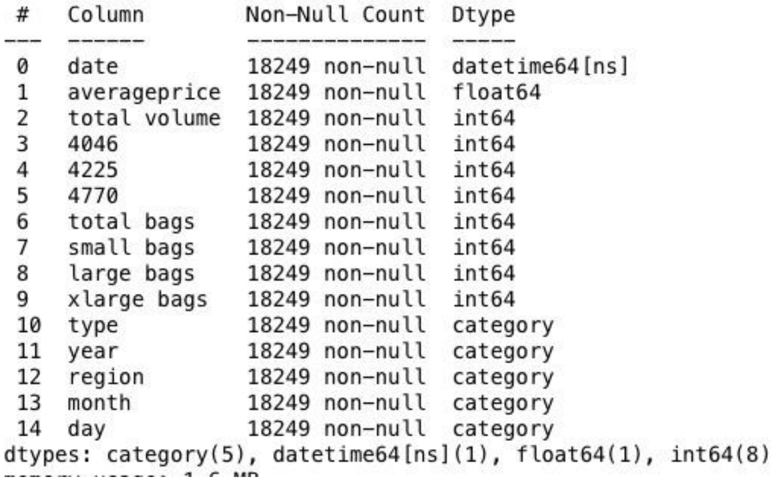
XLarge Bags has 12048 (66.0%) zeros — Zeros



df_index	Date	AveragePrice	Total Volume	4046	4225	4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	2015-12-27	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	2015-12-20	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2015-12-13	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	2015-12-06	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	2015-11-29	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany
5	2015-11-22	1.26	55979.78	1184.27	48067.99	43.61	6683.91	6556.47	127.44	0.0	conventional	2015	Albany
6	2015-11-15	0.99	83453.76	1368.92	73672.72	93.26	8318.86	8196.81	122.05	0.0	conventional	2015	Albany
7	2015-11-08	0.98	109428.33	703.75	101815.36	80.00	6829.22	6266.85	562.37	0.0	conventional	2015	Albany
8	2015-11-01	1.02	99811.42	1022.15	87315.57	85.34	11388.36	11104.53	283.83	0.0	conventional	2015	Albany
9	2015-10-25	1.07	74338.76	842.40	64757.44	113.00	8625.92	8061.47	564.45	0.0	conventional	2015	Albany

Data Profiling - Pre-processing

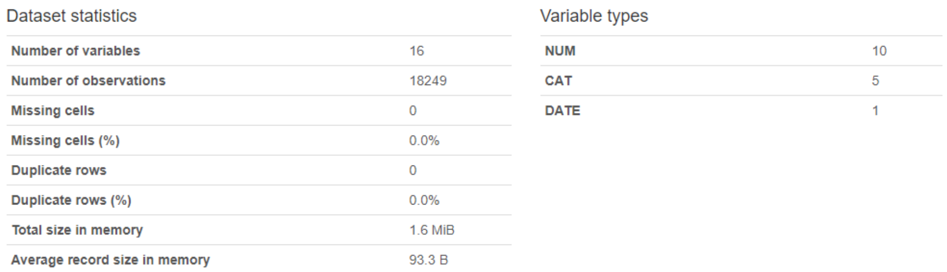
- Pre-processing
  - Check data distribution
  - Review and convert to relevant data type
  - Review perceived outliers



```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   date          18249 non-null  datetime64[ns]
 1   averageprice  18249 non-null  float64
 2   total volume  18249 non-null  int64
 3   4046          18249 non-null  int64
 4   4225          18249 non-null  int64
 5   4770          18249 non-null  int64
 6   total bags    18249 non-null  int64
 7   small bags    18249 non-null  int64
 8   large bags    18249 non-null  int64
 9   xlarge bags   18249 non-null  int64
 10  type          18249 non-null  category
 11  year          18249 non-null  category
 12  region        18249 non-null  category
 13  month         18249 non-null  category
 14  day           18249 non-null  category
dtypes: category(5), datetime64[ns](1), float64(1), int64(8)
```

Data Profiling – Post Profiling

- Post Profiling
  - Rerun stats
    - profiling



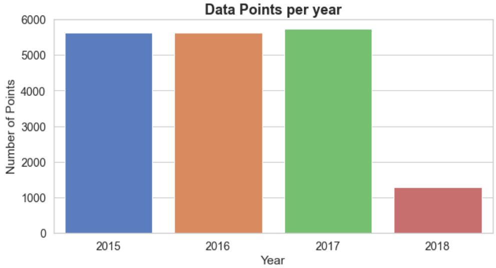
Dataset statistics

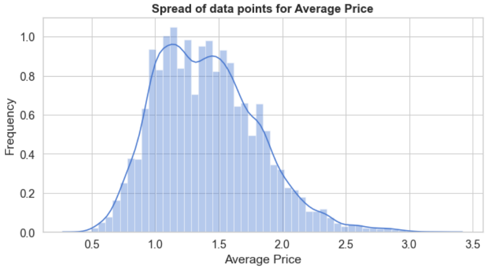
Number of variables	16
Number of observations	18249
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	1.6 MiB
Average record size in memory	93.3 B

Variable types

NUM	10
CAT	5
DATE	1

Data Analysis

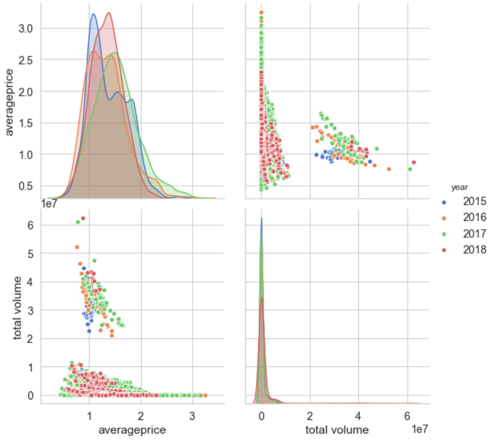




Data spread

- 3+ years worth of data

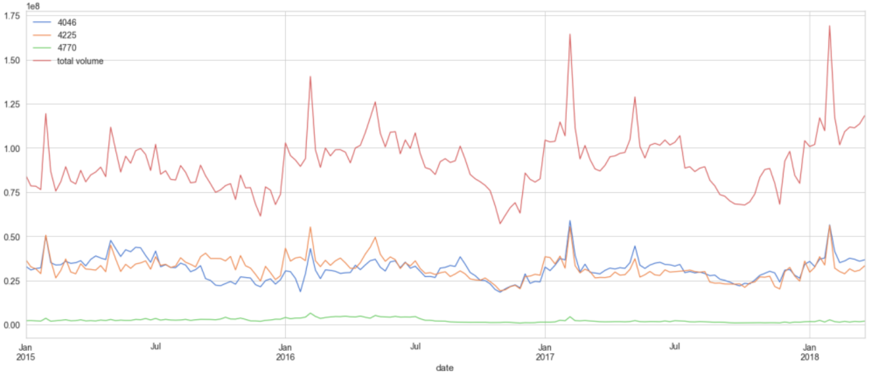
- Average Price
  - Normal distribution
  - Target variable for prediction
- Total Volume
  - Right skew in data
  - Includes data points for ‘Total US’

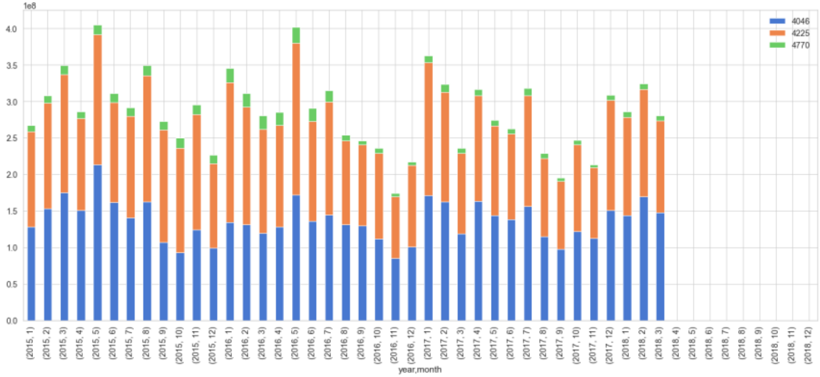


Data Analysis

Volume Sale by Avocado Grade

- Small and medium size preferred over large
- X large is a niche product
- Spike in sale around Feb every year

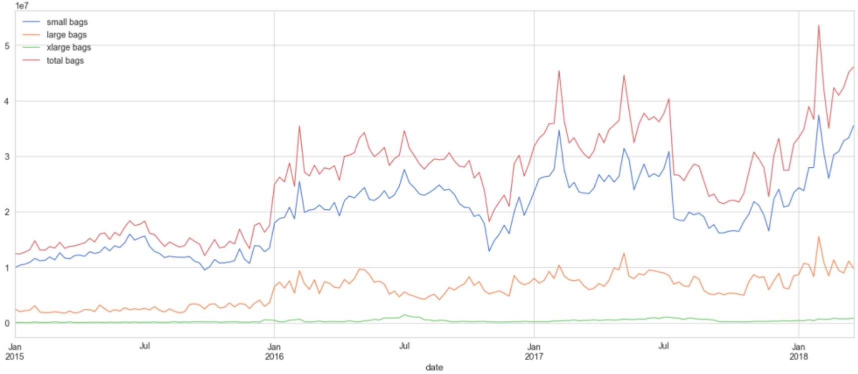


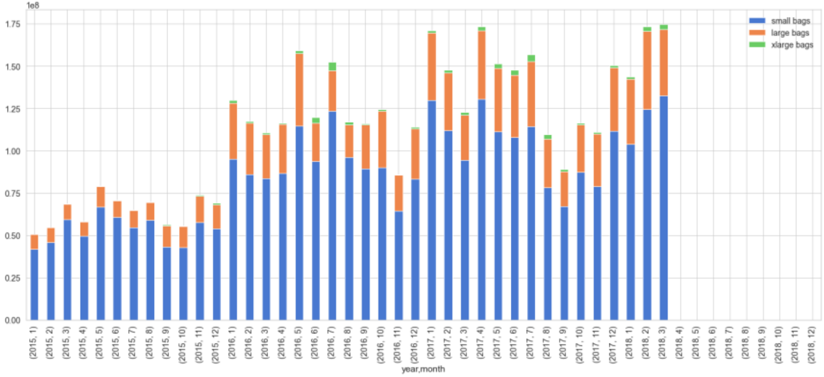


Data Analysis

Bag size preference

- Small bag size preferred
- XL size sold is very small numbers

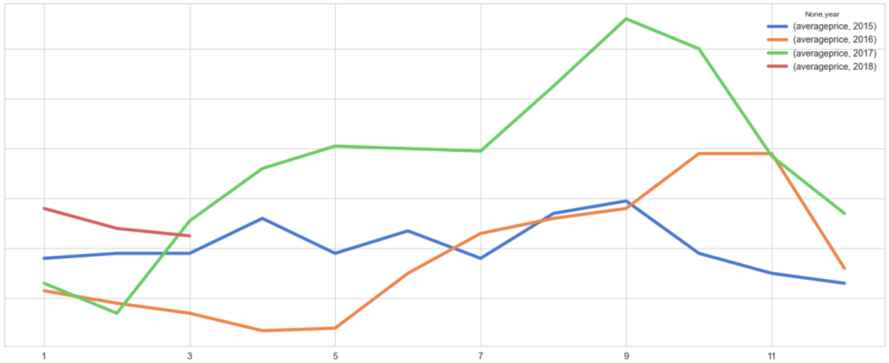


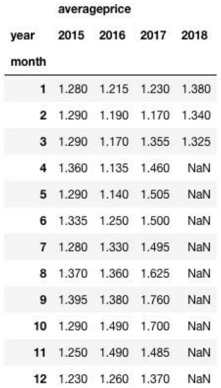


Data Analysis

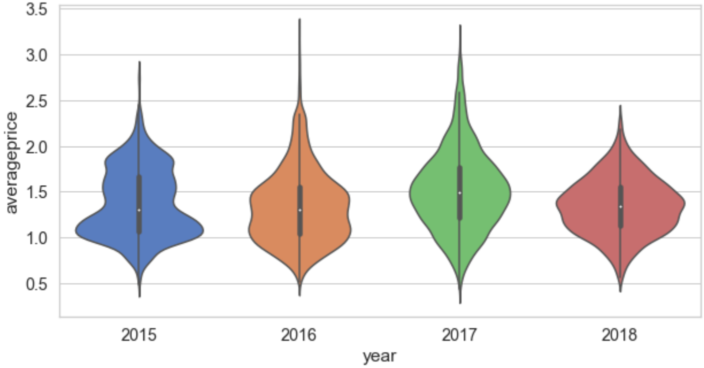
Annual variations in Average Price

- range is between 1.2 – 1.5
- peaks during Sep – Oct every year
- 2017 prices were generally higher
  - attributed to a poor harvest leading to shortage in supply





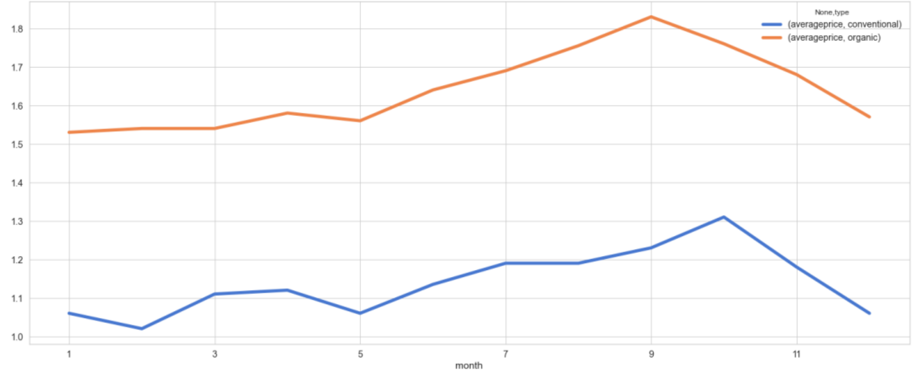
	averageprice			
year	2015	2016	2017	2018
month				
1	1.280	1.215	1.230	1.380
2	1.290	1.190	1.170	1.340
3	1.290	1.170	1.355	1.325
4	1.360	1.135	1.460	NaN
5	1.290	1.140	1.505	NaN
6	1.335	1.250	1.500	NaN
7	1.280	1.330	1.495	NaN
8	1.370	1.360	1.625	NaN
9	1.395	1.380	1.760	NaN
10	1.290	1.490	1.700	NaN
11	1.250	1.490	1.485	NaN
12	1.230	1.260	1.370	NaN

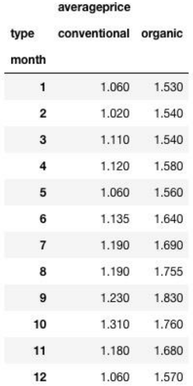


Data Analysis

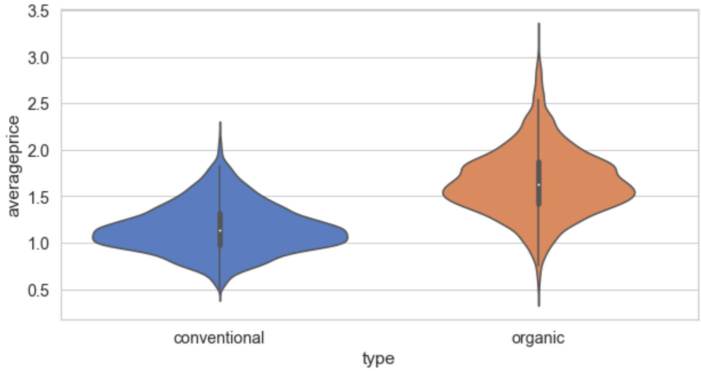
Annual variations in Average Price by Type

- Organic avocados are higher priced
  - Can be attributed to higher cost of production





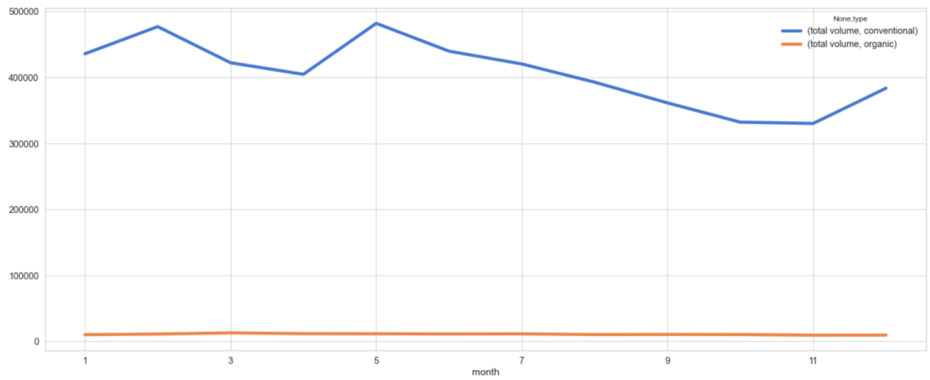
	averageprice	
type	conventional	organic
month		
1	1.060	1.530
2	1.020	1.540
3	1.110	1.540
4	1.120	1.580
5	1.060	1.560
6	1.135	1.640
7	1.190	1.690
8	1.190	1.755
9	1.230	1.830
10	1.310	1.760
11	1.180	1.680
12	1.060	1.570



Data Analysis

Annual variations in Total Volume by Type

- Conventionally farmed avocados are by far the most in demand
- Organic avocados are a very ‘niche’ sale item

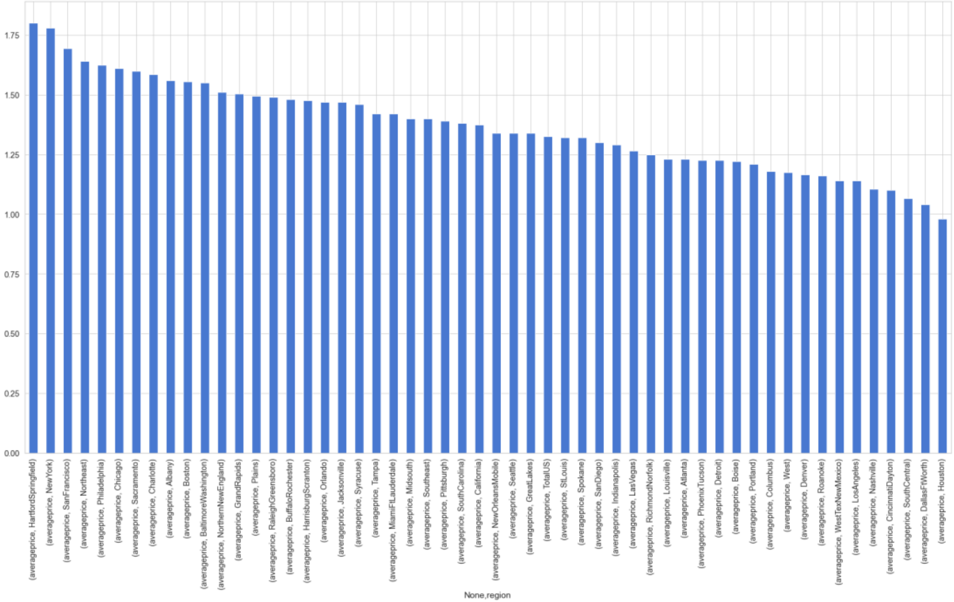


Data Analysis

Top 5 cities with highest average price

- Hartford Springfield
- New York
- San Francisco
- Philadelphia
- Chicago

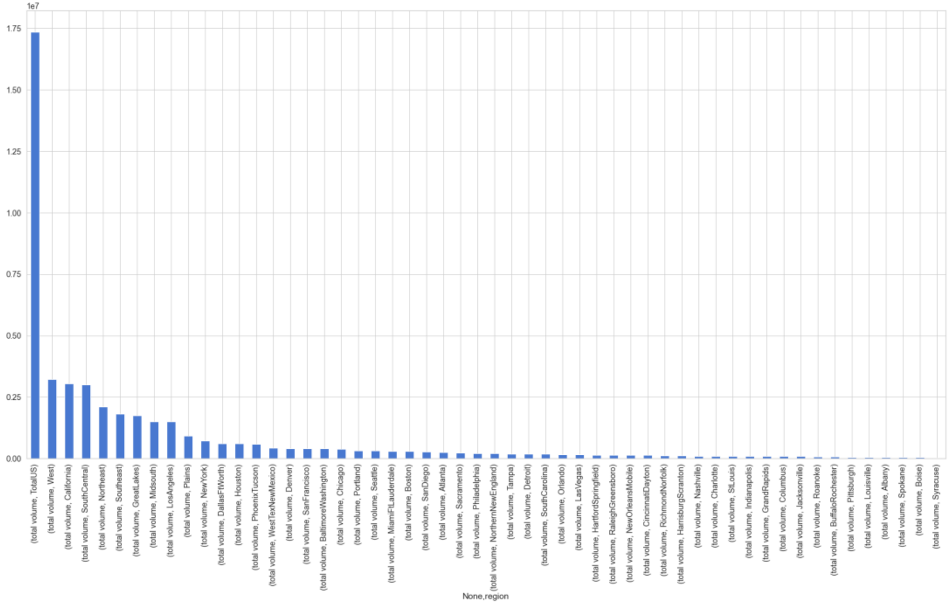
- Note: The ‘regional’ areas were not selected for the listing although they are plotted in the graph



Data Analysis

Top 5 cities with highest total volumes

- Los Angeles
- New York
- Dallas
- Houston
- Phoenix

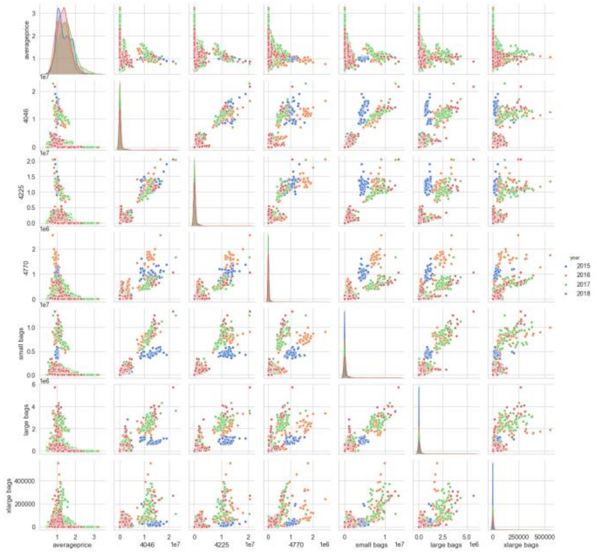


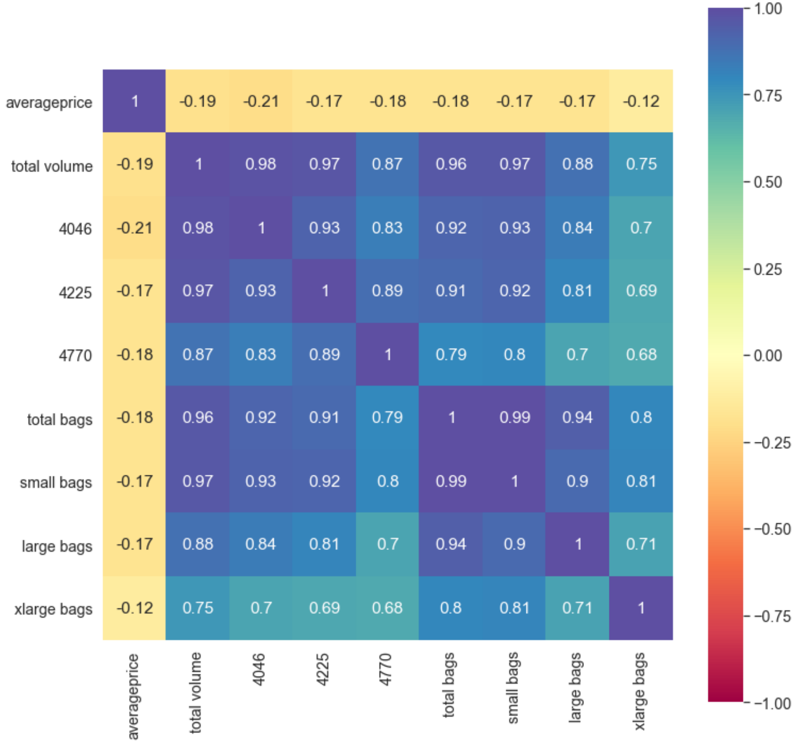
- Note: The ‘regional’ areas were not selected for the listing although they are plotted in the graph. The ‘high’ tower is the Total US volume data point

Data Analysis

Parameter correlation

- Very high correlation among variables
  - Redundant information will be excluded for model building

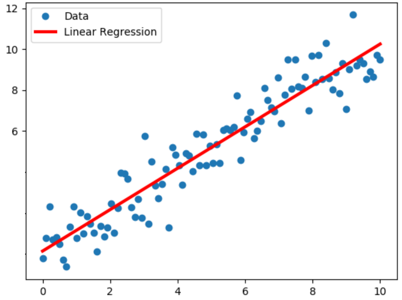


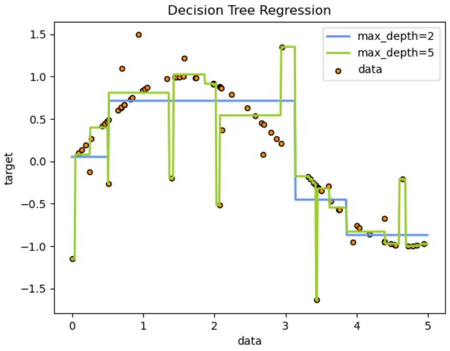


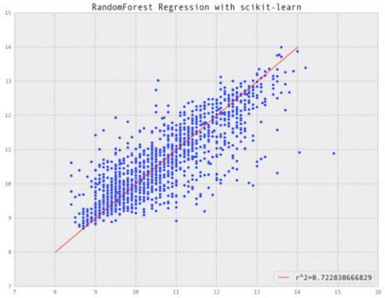
Machine Learning Model

Regression Models Selected

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor



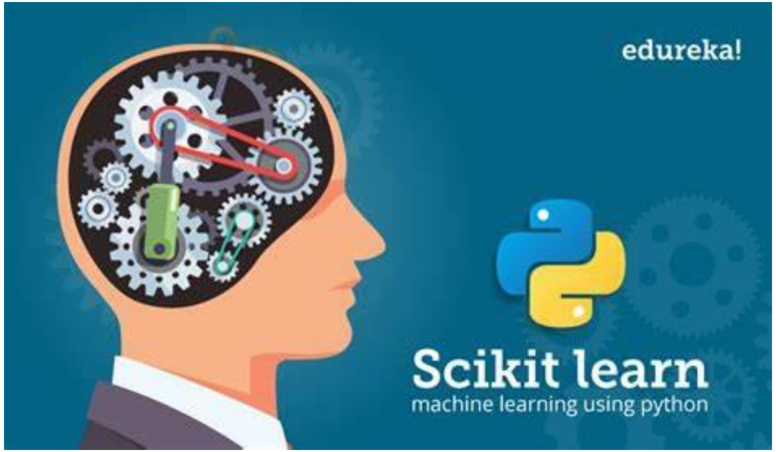




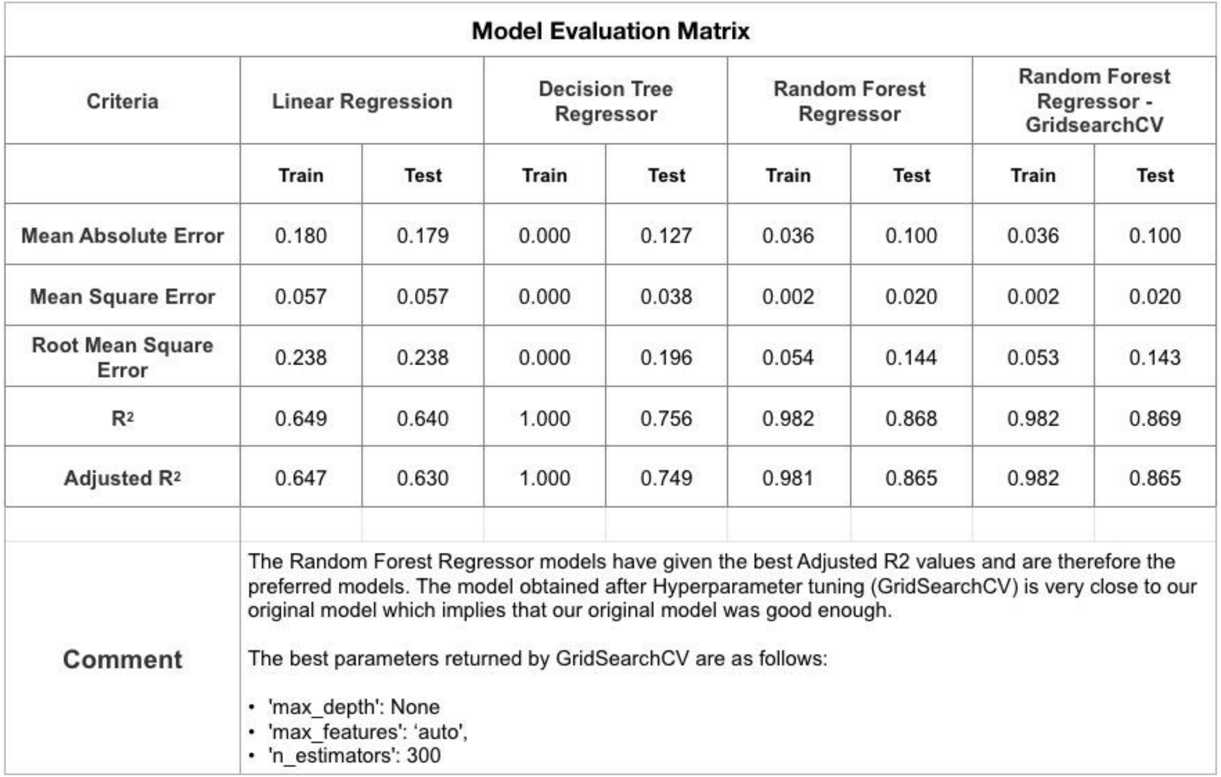
Machine Learning Model

Model building

- Train – Test Split dataset
- Scaling of the data
- Running the regression models
- Model evaluation



Machine Learning Model



Model Evaluation Matrix

Criteria	Linear Regression		Decision Tree Regressor		Random Forest Regressor		Random Forest Regressor - GridsearchCV	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Mean Absolute Error	0.180	0.179	0.000	0.127	0.036	0.100	0.036	0.100
Mean Square Error	0.057	0.057	0.000	0.038	0.002	0.020	0.002	0.020
Root Mean Square Error	0.238	0.238	0.000	0.196	0.054	0.144	0.053	0.143
R^2	0.649	0.640	1.000	0.756	0.982	0.868	0.982	0.869
Adjusted R^2	0.647	0.630	1.000	0.749	0.981	0.865	0.982	0.865
Comment	The Random Forest Regressor models have given the best Adjusted R2 values and are therefore the preferred models. The model obtained after Hyperparameter tuning (GridSearchCV) is very close to our original model which implies that our original model was good enough. The best parameters returned by GridSearchCV are as follows: • 'max_depth': None • 'max_features': 'auto', • 'n_estimators': 300							

Conclusions

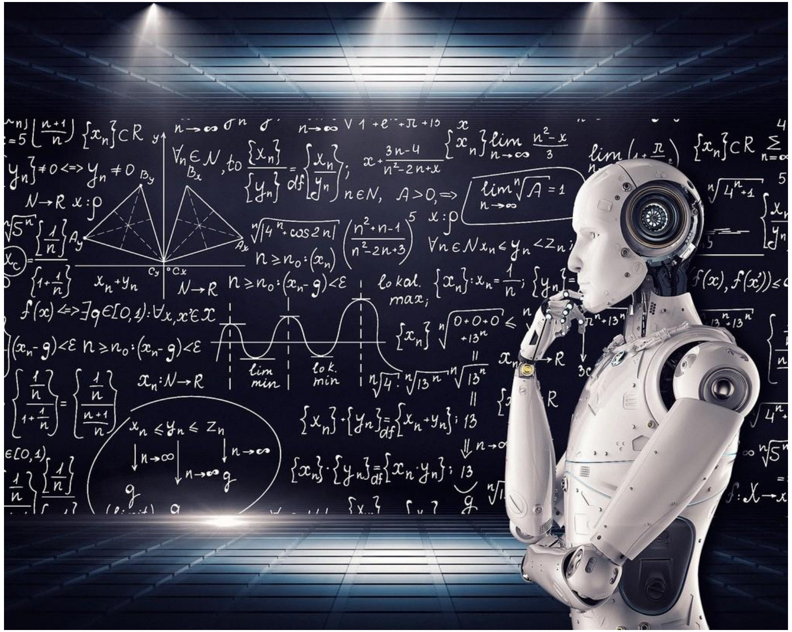
- Dataset for 3 full years and Q1 of the fourth
- Average Price is normally distributed
- Small / medium sized avocados in small sized bags are preferred
- Annual variation in price, peaking in Sept – Oct
- Conventionally farmed option preferred
  - Perhaps due to lower costs

Conclusions

- Pricing can be predicted using Machine Learning Models
- Regression Models selected
  - Random Forest Regressor provided the best model

Recommendations

- Model Improvement
  - More powerful Algorithms
    - Support Vector Machines (SVM)
    - Artificial Neural Networks (ANN)



Thank You!

Bon Appetite!