Forecasting Avocado Prices

Avocado King can use data to optimize profits

Renel Chesak, MSc Data Science
Winter 2020

The problem

Company	Context	Problem statement
Avocado King , a major distributor of avocados across the USA	Avocado **prices fluctuate** based on a variety of factors Avocado King stores historical **sales and price data** Google provides historical **search data**	There is a need to **forecast avocado prices** to enable company-wide financial planning

Challenges deep-dive

Understand data

Through data visualization, we can better **see relationships** in the data and develop analytical approach

Prepare data

The data given has a unique key: **[date, region, type]**

- Data must be **cleaned** and **normalized** carefully with key in mind

Model prices

Utilize sound statistical analyses and **machine learning** to model prices

- Random forest regression
- Linear regression

Solution

Linear Regression

On average, the selected model predicts avocado prices **within 11 cents of the true value**.

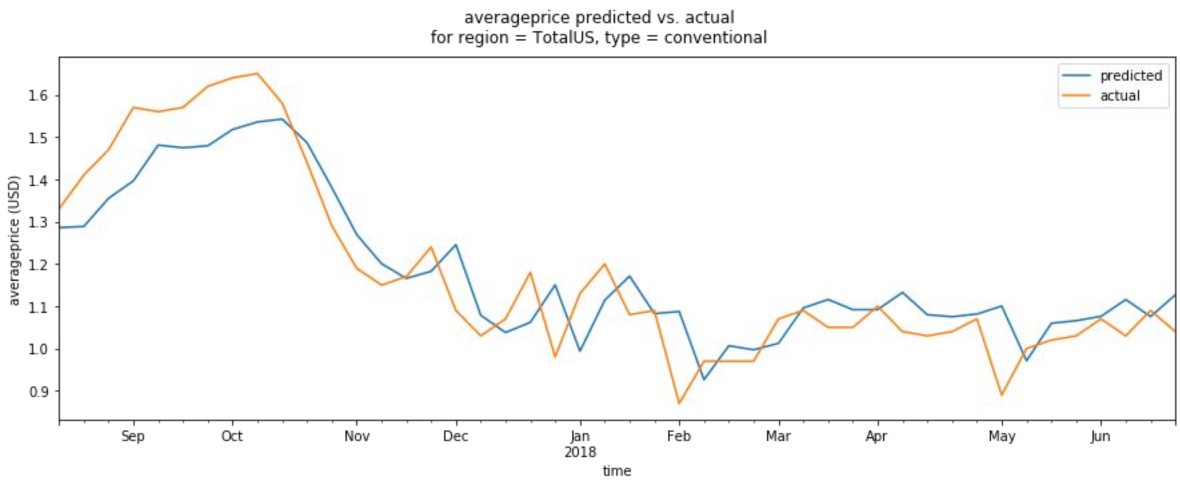
Forecasting 1 week ahead
Mean absolute error = 0.11

84%

of variation seen in price is explained by the selected model

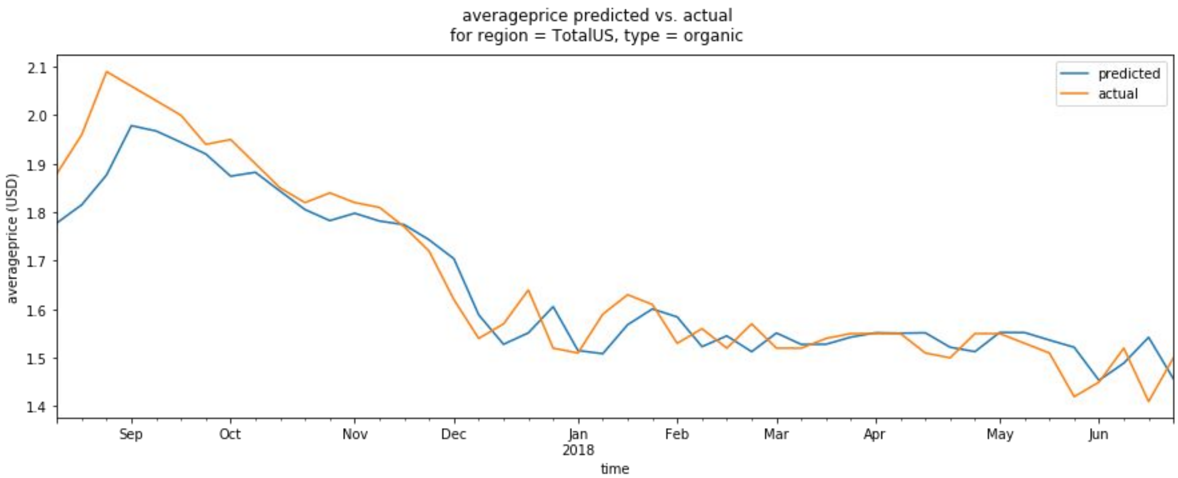
R-squared = 83.72% (coefficient of determination)

Model performance



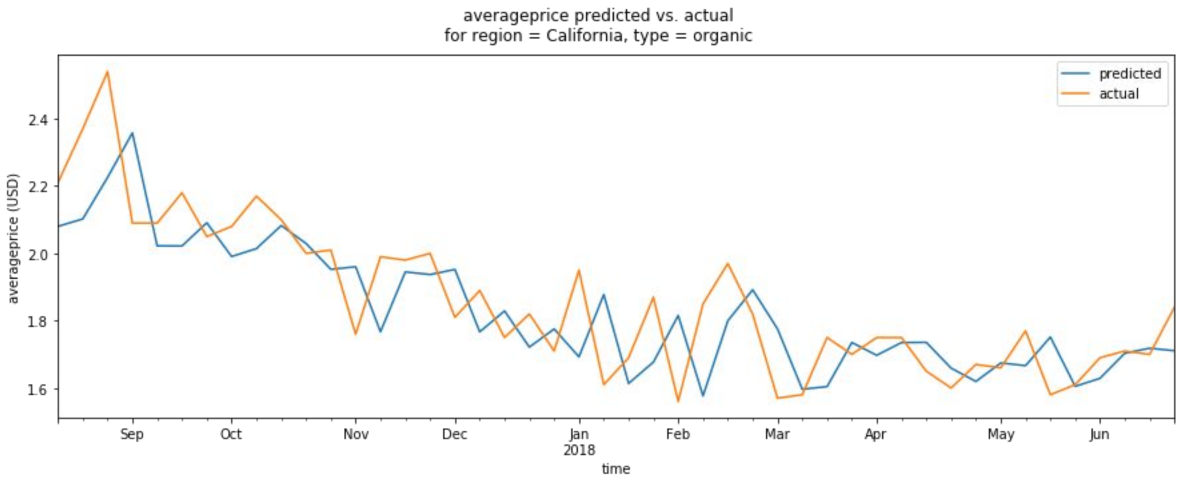
Validation set

Model performance



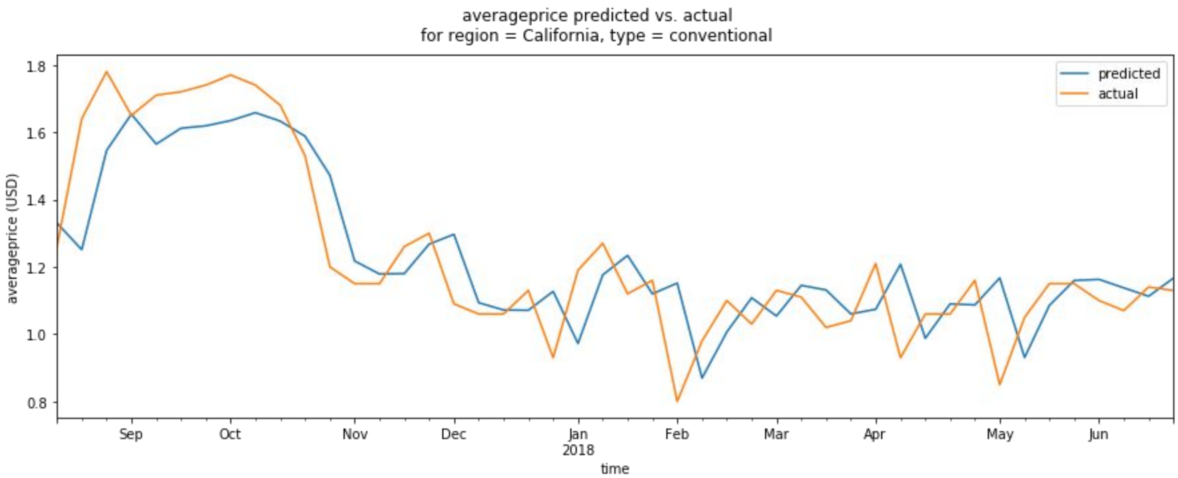
Validation set

Model performance



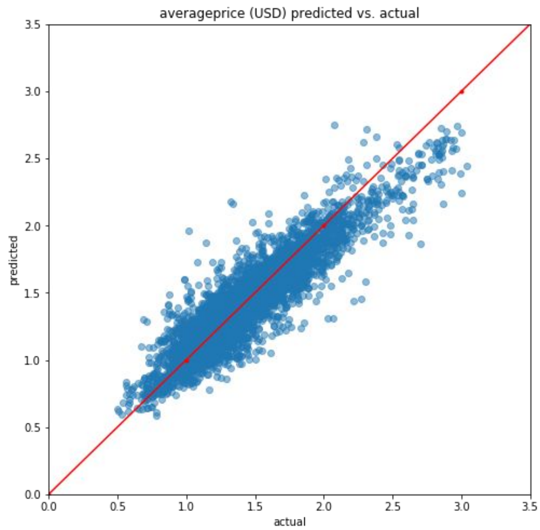
Validation set

Model performance



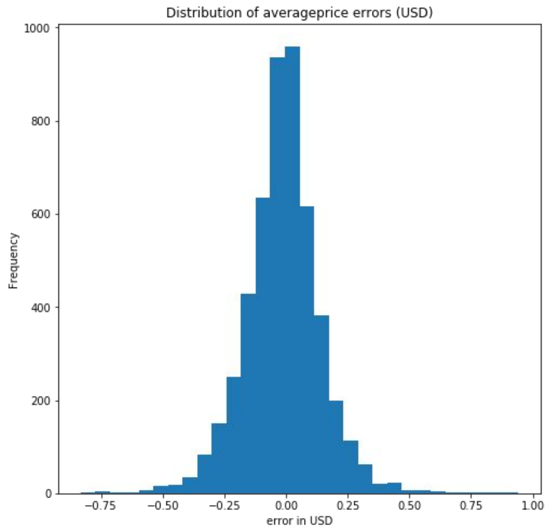
Validation set

Model performance



Validation set

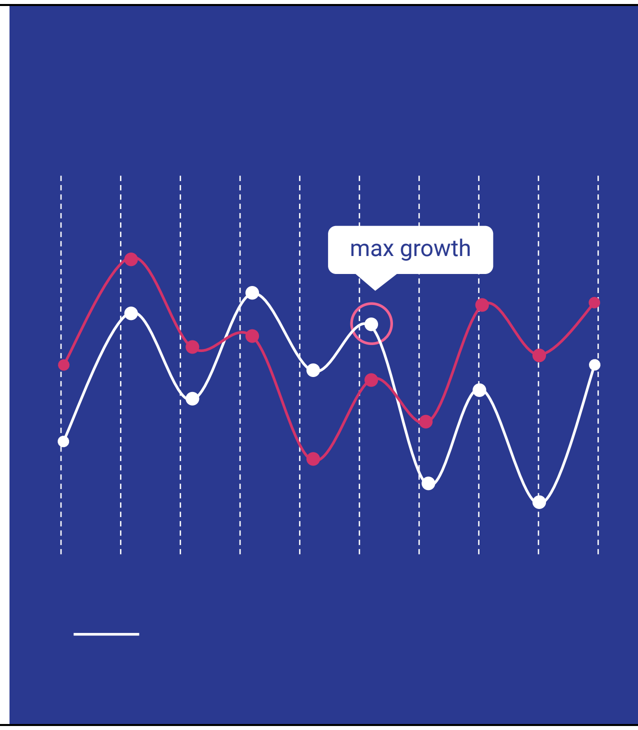
Model performance



The errors are approximately normally distributed, meaning the model is low in bias.
Validation set

Impact

Forecast prices => **Plan ahead**
=> **Optimize profits**
=> **Maximize growth**



Next steps

This proof of concept can be expanded upon:

- Higher accuracy
- Forecasting weekly, monthly, quarterly, yearly
- Forecasting volume sold

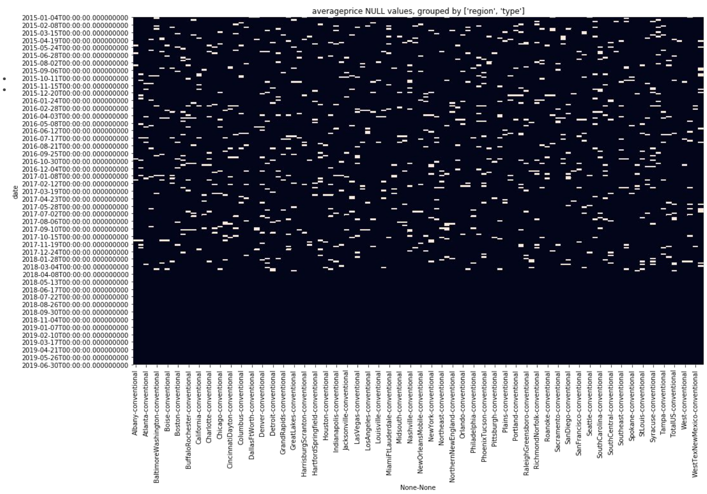
Technical details

Data cleaning

There is a unique key for this dataset: [date, region, type]. We ensured no data was missing:

1. Check for **extra dates** (out of expected weekly sequence)
2. Check for **missing dates** (based on expected weekly sequence)
3. Use **forwardfill** to fill nulls
4. Use **backwardfill** to fill remaining nulls

Each unique key combination was visualized:

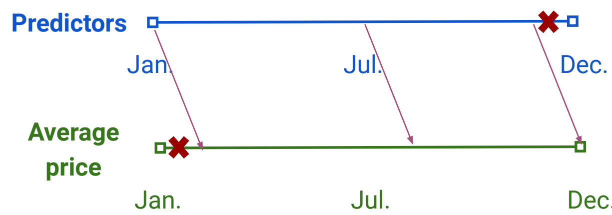


Technical details

Time shifting

In forecasting, we use values from this week predict next week's values. So we shifted the data accordingly.

1. Drop the latest date in x
   a. (since it doesn't have an associated future y at *t*+1)
2. Add 7 days to each date in x
3. Drop the earliest date in y
   a. (since it doesn't have a corresponding past x at *t*-1)
4. Join the x (*t*-1) and y (*t*) dfs on the unique key [date, region, type]

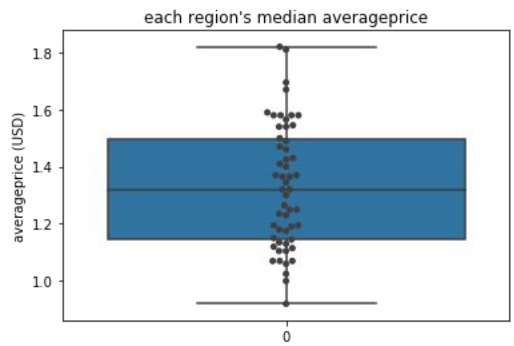


Technical details

The data

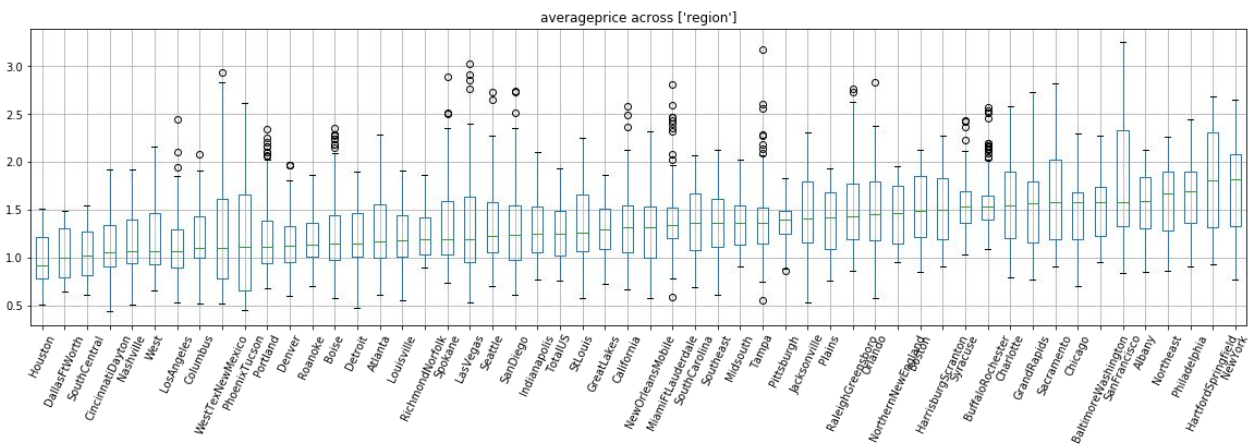
All predictive variables were shifted 1 week back to enable forecasting 1 week ahead.

- Average price (per avocado, USD)
- Region group
- Total volume (avocados sold)
- Total bags (bags sold)
- “Avocado recipe” (google search)
- Season (spring, summer, fall, winter)
- Type (conventional, organic)



- **Average_price** was included because at time *t*-1, it correlates strongly (r = 0.6) with average_price at time *t*
- **Region_group** was created to reduce the number of features in the model (reduce overfitting) while preserving the important relationship between geography and price.
- **Total_volume**, **total_bags**, and **“avocado recipe”** (google search) were chosen because when grouped by region_group, they correlated weakly (r = 0.2) to moderately (r = 0.5) with average_price.
- **Season** was added because avocados from Mexico come in season in the fall and winter. 70% of avocados sold in the USA are sourced from Mexico.
- **Type** was important to include because in the exploratory visualizations, organic can be seen as almost always more expensive

Technical details

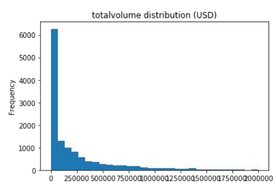


Technical details

Normalization

All numeric predictive variables were **tested** via several normalization methods, seeking to best preserve the original distribution

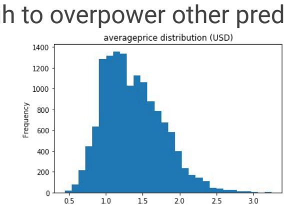
- Z-score
- Methods for preserving skewed distributions (e.g. Pareto):
    - MinMax
    - Log + MinMax
    - Cube Root + MinMax



Z-score normalized:

- Total volume
- Total bags
- “Avocado recipe” (google search)

Left average price on its original scale to preserve distribution. Original was scale not big enough to overpower other predictors.



- MinMax Normalization seems to preserve the shape of the normally distributed data best, but Z-score Standardization seems to work best in preserving the Pareto-distributed data. However using both will mean that there will be a mix of data centered at zero and data in the range [0, 1]. It is best to have all variables on the same scale. Since \`averageprice\` is the main target of interest, is normally distributed, and falls in the relatively small range [0.44, 3.25], we will leave it on its original scale and use Z-score normalization on the rest.
- Log + MinMax and Cube Root + MinMax normalization don't seem to improve over just MinMax scaling.

Technical details

Train / Validation / Test split

Since this is time series data, **we want full years (with all their seasonal variations) to exist** in the train, validation, and test set.

- training set:
  - Jan 2015 - July 2017
- validation set:
  - August 2017 - July 2018
- test set:
  - August 2018 - July 2019

→

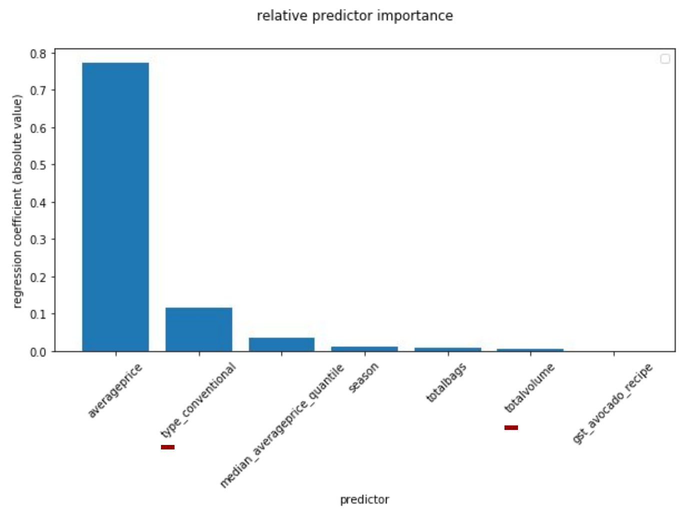
- training set:
  - 58%
- validation set:
  - 21%
- test set:
  - 21%

Technical details

Predictor importance

Based on the model, the rank of importance is:

- Average price (per avocado, USD)
- Type (conventional, organic)
- Region group (median_averageprice_quantile)
- Season (spring, summer, fall, winter)
- Total bags (bags sold)
- Total volume (avocados sold)
- "Avocado recipe" (google search)



Technical details

Error Metric

Mean absolute error was used when evaluating model performance:

- Enables us to intuitively understand how accurate our predictions will be, on average.
- MAE = 0.11
  - “On average, the model predicts avocado prices within 11 cents of the true value.”