Random Forest Ensemble (Regressor and Classifier) Preprocessing Steps

Dataset

We trained Random Forest Ensemble model on prices and production dataset.

- **y-variable** (target value) is the average price of the avocados in the US cities.
- **X-variables** are week, month, year, units sold (3 different types), bags sold (3 different types), production data for avocados in California, Chile, Peru, Mexico, Columbia and the ratio between total volume sold and total volume produced. This step was taken to reflect the ratio per each city, since the production data was repetitive per city. Using ratio in ML model showed improvement on model performance.

Preliminary data preprocessing

From our EDA we recognized seasonality in dataset, where we saw patterns in weeks and months. We approached to this problem, by splitting the date into weeks, months and years and use those features in ML model as separate X-variables.

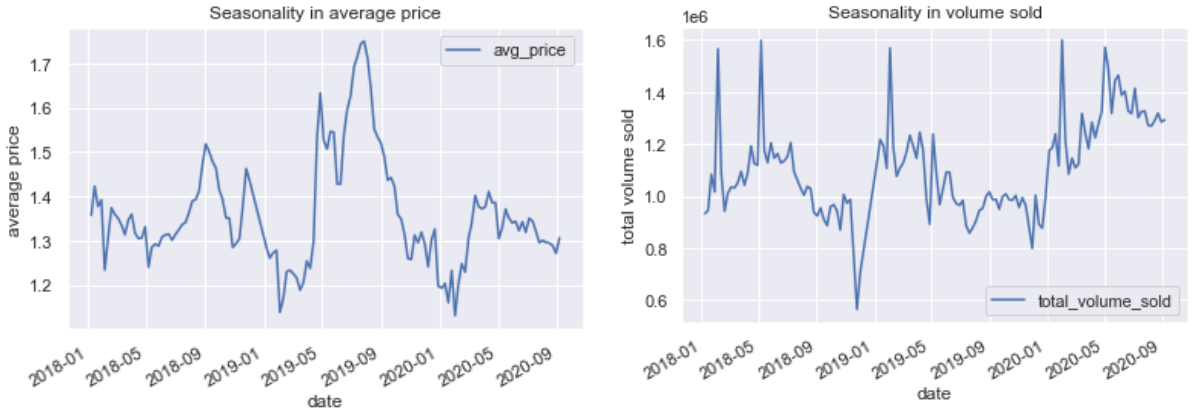

Figure 1 & 2: Seasonality in average price and volume sold.

Additionally, our dataset contained some outliers. These datapoints were combined regions that contained sums of selected cities. We recognized those as a noise in the dataset and decided to remove them from the dataset. Our data preprocessing includes:

- Converting date datatype from object to datetime and extracting week, month and year.
- Making necessary calculations (ratios).
- Drop NaN values.
- Handling outliers and noise in the dataset.
- Cleaning column “type” (some datapoints “conventional” had space at the end, so the value appeared twice “conventional” and “conventional ”).
- Encoding strings and dropping non-beneficial columns.
- Normalizing data set, using Standard Scaler.

Preliminary feature engineering and preliminary feature selection. Splitting data into training and testing sets, decision making and description of how model was trained.

Our y-variable, that is avocado average price is continuous variable. We decided to use both approaches Regression and Classification. In order to use classification model y-variable has to be categorical value. We used **qcut function** to evenly distribute values into 3 and 4 categories and checked the range of the values. One advantage of qcut function is that values are evenly distributed; yet on the other hand the range between categories might not be even. We kept this in mind when analyzing final results in confusion matrix.

We split data into training and testing at 75% and 25% respectively. Next, we run the models on various combinations of the X variables. For example:

- Using ratio between total volume sold and total volume produced and without.
- Using only one type of the avocado at the time (conventional or organic) and combined.
- Removing noise (larger regions vs only cities and vice versa).
- Using only weeks or months and all three weeks, months and years.

When deciding about final model we looked at model performance scores, feature importances, and our final decision what we would like to predict. For example, model performed better when using only regions vs. cities; however, we decided to use the model with the cities since prediction price in selected city is important for our project. Ratios contributed to improved ML model, so we used ratios as one of our X-variable. y-variable was split into 3 and 4 categories. Model performed better when split into 3 categories vs 4 categories.