**AVACADO PRICES PREDICTION**

**OBJECTIVE OF THE REPORT**

Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation:

Accurate forecast of Avocado prices.

**DESCRIPTION OF DATASET**

Brief description of the data set you chose and a summary of its attributes

The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g. greenskins) are not included in this table.

Some relevant columns in the dataset:

* Date - The date of the observation
* Average Price - the average price of a single avocado
* type - conventional or organic
* year - the year
* Region - the city or region of the observation
* Total Volume - Total number of avocados sold
* 4046 - Total number of avocados with PLU 4046 sold
* 4225 - Total number of avocados with PLU 4225 sold
* 4770 - Total number of avocados with PLU 4770 sold

**PRE-DATA PROCESSING**

Brief summary of data exploration and actions taken for data cleaning and feature engineering.

|  |  |
| --- | --- |
| Columns\_Name | No of msiing Values |
| Unnamed | 0 |
| Date | 0 |
| Average Price | 0 |
| Total Volume | 0 |
| 4046 | 0 |
| 4225 | 0 |
| 4770 | 0 |
| Total Bags | 0 |
| Small Bags | 0 |
| Large Bags | 0 |
| XLarge Bags | 0 |
| type | 0 |
| year | 0 |
| region | 0 |
| year | 0 |
| region | 0 |

* It was found that there no missing values.
* The whitespaces from the column names were removed.

Descriptive Statistics of the Variables



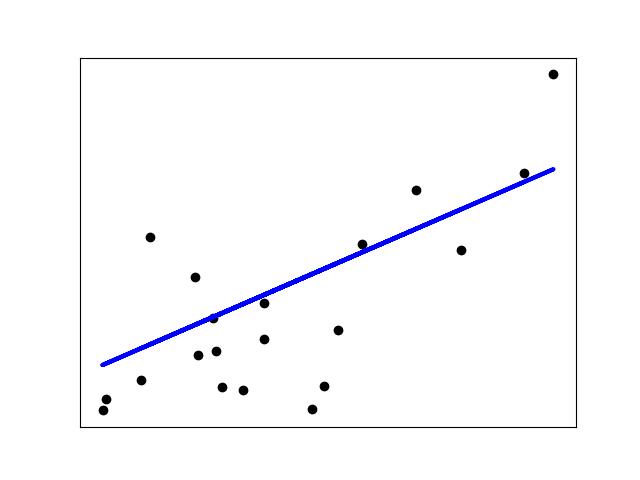
**MODEL FITTING**

Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method

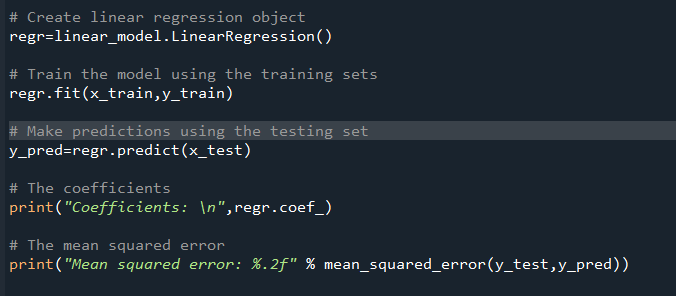
Linear Regression-Ordinary Least Squares

LinearRegression fits a linear model with coefficients

to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Mathematically it solves a problem of the form:

[](https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html)

Python code for Linear regression

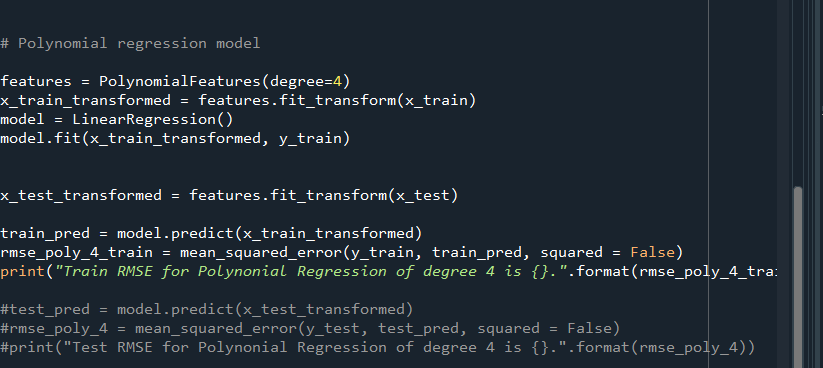


Polynomial Regression

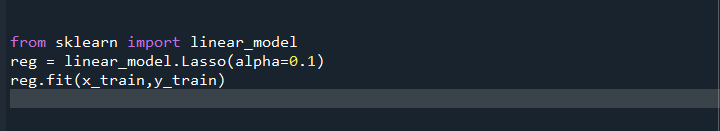
Polynomial regression: extending linear models with basis functions

One common pattern within machine learning is to use linear models trained on nonlinear functions of the data. This approach maintains the generally fast performance of linear methods, while allowing them to fit a much wider range of data.

For example, a simple linear regression can be extended by constructing polynomial features from the coefficients.



Regularization techniques in the Linear regression



**MODEL SELECTION**

A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

The RMSE of linear regression is 15 %.

The RMSE of Polynomial features is 37%.

The linear regression model was final chosen to be the model to be deployed.