AVOCADO AVERAGE PRICE PREDICTION MODEL

Submitted by:

UJJWAL PRATIK

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Data Trained Faculty for the guidance. They have covered the topics like Machine Language, Python & SQL. Under their guidance I learned a lot about this project. their suggestions and directions have helped in the completion of this project. I had also taken help from YouTube & online videos.

INTRODUCTION

ABOUT AVOCADO

Avocado is a tree. The fruit, a popular food, is a good source of potassium and healthy fats. The fruit, leaves, and seeds are sometimes used to make medicine. The oil from the fruit is also used as a medicine and for cooking, it is used for high cholesterol, psoriasis, arthritis, sexual desire, obesity, and many other conditions, but there is no good scientific evidence to support these uses.

PROBLEM STATEMENT

This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. 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. 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.

DATA ANALYSIS

COLUMNS IN THE DATAFRAME

- 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

	Unnamed: 0	Date	AveragePrice	Total Volume	PLU4046	PLU4225	PLU4770	Total Bags	Small Bags	Large Bags	XLarge Bags	type	year	region
0	0	27-12-2015	1.33	64236.62	1036.74	54454.85	48.16	8696.87	8603.62	93.25	0.0	conventional	2015	Albany
1	1	20-12-2015	1.35	54876.98	674.28	44638.81	58.33	9505.56	9408.07	97.49	0.0	conventional	2015	Albany
2	2	13-12-2015	0.93	118220.22	794.70	109149.67	130.50	8145.35	8042.21	103.14	0.0	conventional	2015	Albany
3	3	06-12-2015	1.08	78992.15	1132.00	71976.41	72.58	5811.16	5677.40	133.76	0.0	conventional	2015	Albany
4	4	29-11-2015	1.28	51039.60	941.48	43838.39	75.78	6183.95	5986.26	197.69	0.0	conventional	2015	Albany

OBJECTIVE

- Our main objective is to predict the Average Price of avocado at using machine learning algorithms.
- All the parameters will be analysed through Machine Learning algorithms like Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression etc which will help to predict the Average Price.

DATA DESCRIPTION

- The dataset contains the detailed study of Date, type, year, region, Total Volume, 4025(PLU), 4046(PLU), 4770(PLU).
- The source of data is taken from GitHub. (https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/avocado.csv)

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the sales because it gives the best outcome.
- The dependent variable is Average Price, whereas other variables i.e., Date, type, region, year, total volume, 4025(PLU), 4046(PLU), 4770(PLU), are the independent variables.

METRIC USAGE

a. Linear Regression.
b. Lasso Regression.
c. Ridge Regression.
d. Elastic Net Regression.
e. KNeighbors Regressor.
f. Random Forest Regressor.

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score : is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error : is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Zscore: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- Random Forest Regressor: A Random Forest is an ensemble technique capable of performing both regression and classification tasks.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- K Nearest Neighbors Regressor: The K Nearest Neighbors Regressor instead computes the mean of the nearest neighbor labels.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regression
d. Ridge Regression
e. Elastic Net Regression
f. Grid Search CV
g. Cross Validation

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values and converted categorical columns into numerical.
- Analysed data for distribution.
- Checked and removed outliers through zcore method.
- Removed skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r2 score from overfitting.
- Hyper Parameter tuning.

CONCLUSION

As our conclusion we proclaim that, after checking r2 score, cross validation, Ensemble Techniques, we declare Linear Regression best suited model for our purpose of predicting average avocado price.

```
loaded_model=pickle.load(open('avocado_avg_price.pkl','rb'))
result=loaded_model.score(x_test, y_test)
print(result)
```

```
0.9996826528035346
```

```
conclusion=pd.DataFrame([loaded_model.predict(x_test)[:],pred_decision[:]],index=['predicted','original'])
conclusion
```

	0	1	2	3	4	5	6	7	8	9	...	306	307	308	309
predicted	0.953348	0.755797	1.10977	1.00858	0.740487	0.968115	0.833161	1.258965	0.711312	1.219915	...	1.508126	1.360946	1.00858	1.039949
original	0.953348	0.755797	1.10977	1.00858	0.740487	0.968115	0.833161	1.258965	0.711312	1.219915	...	1.508126	1.360946	1.00858	1.039949

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc.

It also helped me to learn how to apply various model techniques on data and enable predications.

REFERENCES

- Data trained course videos.
- Google Search.
- YouTube.
- GitHub.
- UCI Machine learning repository.