**Table of Contents**

[**1.0** **Introduction** 4](#_Toc111991652)

[**1.1.** **What are the different types of Machine Learning?** 4](#_Toc111991653)

[**1.2.** **Benefits of Using Machine Learning in Avocado Marketing Industry** 5](#_Toc111991654)

[**1.3.** **About Industry** 5](#_Toc111991655)

[**1.3.1** **AI / ML Role in Avocado Marketing Industry** 6](#_Toc111991656)

[**2.0** **Avocado Marketing Industry** 7](#_Toc111991657)

[**2.1.** **Main Drivers for AI Auto Quote Analysis** 7](#_Toc111991658)

[**2.2.** **Internship Project - Data Link** 8](#_Toc111991659)

[**3.0** **AI / ML Modelling and Results** 9](#_Toc111991660)

[**3.1.** **Your Problem of Statement** 9](#_Toc111991661)

[**3.2.** **Data Science Project Life Cycle** 9](#_Toc111991662)

[**3.2.1** **Data Exploratory Analysis** 10](#_Toc111991663)

[**3.2.2** **Data Pre-processing** 10](#_Toc111991664)

[**3.2.2.1.** **Check the Duplicate and low variation data** 10](#_Toc111991665)

[**3.2.2.2.** **Identify and address the missing variables** 10](#_Toc111991666)

[**3.2.2.3.** **Handling of Outliers** 11](#_Toc111991667)

[**3.2.2.4.** **Categorical data and Encoding Techniques** 11](#_Toc111991668)

[**3.2.2.5.** **Feature Scaling** 12](#_Toc111991669)

[**3.2.3** **Selection of Dependent and Independent variables** 12](#_Toc111991670)

[**3.2.4** **Data Sampling Methods** 13](#_Toc111991671)

[**3.2.4.1.** **Stratified sampling** 13](#_Toc111991672)

[**3.2.4.2.** **Simple random sampling** 13](#_Toc111991673)

[**3.2.5** **Models Used for Development** 13](#_Toc111991674)

[**3.2.5.1.** **Model 01** 14](#_Toc111991675)

[**3.2.5.2.** **Model 02** 14](#_Toc111991676)

[**3.2.5.3.** **Model 03** 14](#_Toc111991677)

[**3.2.5.4.** **Model 04** 14](#_Toc111991678)

[**3.2.5.5.** **Model 05** 14](#_Toc111991679)

[**3.2.5.6.** **Model 06** 15](#_Toc111991680)

[**3.2.5.7.** **Model 07** 15](#_Toc111991681)

[**3.2.5.8.** **Model 08** 15](#_Toc111991682)

[**3.3.** **AI / ML Models Analysis and Final Results** 15](#_Toc111991683)

[**3.3.1** **Different Model codes** 15](#_Toc111991684)

[**3.3.2** **Linear Regression Python Code** 16](#_Toc111991685)

[**3.3.3** **Random Forest Python Code** 16](#_Toc111991686)

[**3.3.4** **Decision Tree Python Code** 17](#_Toc111991687)

[**3.3.5** **SVR Python Code** 17](#_Toc111991688)

[**3.3.6** **LGBMRegressor Python Code** 18](#_Toc111991689)

[**3.3.7** **ExtraTreesRegressor Python Code** 18](#_Toc111991690)

[**3.3.8** **XGBRegressor Python Code** 18](#_Toc111991691)

[3.3.9 **GradientBoostingRegressor Python Code** 19](#_Toc111991692)

[**4.0** **Conclusions and Future work** 20](#_Toc111991693)

[**5.0 References** 21](#_Toc111991694)

[**6.0** **Appendices** 22](#_Toc111991695)

[**6.1**  **Python code Results** 22](#_Toc111991696)

[**6.2** **List of Charts** 22](#_Toc111991697)

[**6.2.1 Chart 01: Price distribution graph** 22](#_Toc111991698)

[**6.2.2 Chart 02: Change of average price per calendar year** 22](#_Toc111991699)

[**6.2.3 Chart 03: Average Price in Each Region** 24](#_Toc111991700)

[**6.2.4 Chart 04: Number of conventional and organic type** 25](#_Toc111991701)

**Abstract**

Persea americana, commonly known as avocado, is becoming increasingly important in global agriculture. There are dozens of avocado varieties, but more than 85% of the avocados harvested and sold in the world are of the Hass one. Furthermore, information on the market of agricultural products is valuable for decision-making; this has made researchers try to determine the behaviour of the avocado market, based on data that might affect it one way or another. Here, a machine learning approach for estimating the total sales of avocados, using data based on total no. of avocado’s sold, total volume available and type, is presented. For that purpose, seven algorithms were evaluated: Linear Regression, Random Forest Regression, Support Vector Machine for Regression and Decision Tree Regression, Extra Trees Regression, XGBRegressor, LGBMRegressor Prediction Model.

Keywords: avocado; regression model; machine learning;

# **Introduction**

MACHINE LEARNING PROJECT 3

CSCI 4525 Project IV: Machine Learning Project

Introduction:

Machine learning is a sub-domain of computer science which evolved from the study of

pattern recognition in data, and also from the computational learning theory in artificial

intelligence. It is the first-class ticket to most interesting careers in data analytics today[1]. As

data sources proliferate along with the computing power to process them, going straight to the

data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision

making, clustering, classification, forecasting, deep-learning, inductive logic programming,

support vector machines, reinforcement learning, similarity and metric learning, genetic

algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine

learning task of inferring a function from a labeled data [2]. In Supervised learning, we have a

training set, and a test set. The training and test set consists of a set of examples consisting of

input and output vectors, and the goal of the supervised learning algorithm is to infer a function

that maps the input vector to the output vector with minimal error. In an optimal scenario, a

model trained on a set of examples will classify an unseen example in a correct fashion, which

requires the model to generalize from the training set in a reasonable way. In layman’s terms,

supervised learning can be termed as the process of concept learning, where a brain is exposed to

a set of inputs and result vectors and the brain learns the concept that relates said inputs to

outputs. A wide array of supervised machine learning algorithms are available to the machine

learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines,

Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier[4,7,8,9] etc.,

MACHINE LEARNING PROJECT 3

CSCI 4525 Project IV: Machine Learning Project

Introduction:

Machine learning is a sub-domain of computer science which evolved from the study of

pattern recognition in data, and also from the computational learning theory in artificial

intelligence. It is the first-class ticket to most interesting careers in data analytics today[1]. As

data sources proliferate along with the computing power to process them, going straight to the

data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the stud

Machine learning is a sub-domain of computer science which evolved from the study of pattern recognition in data, and also from the computational learning theory in artificial intelligence. It is the first-class ticket to most interesting careers in data analytics today. As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions. Machine Learning can be thought of as the study of a list of sub-problems, viz: decision making, clustering, classification, forecasting, deep-learning, inductive logic programming, support vector machines, reinforcement learning, similarity and metric learning, genetic algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine learning task of inferring a function from a labelled data. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman’s terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms is available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier etc.

## **What are the different types of Machine Learning?**

There are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement.

Supervised learning

In supervised learning, the machine is taught by example. The operator provides the machine learning algorithm with a known dataset that includes desired inputs and outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. While the operator knows the correct answers to the problem, the algorithm identifies patterns in data, learns from observations and makes predictions. The algorithm makes predictions and is corrected by the operator – and this process continues until the algorithm achieves a high level of accuracy/performance.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to

what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.

Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.

Forecasting: Forecasting is the process of making predictions about the future based on the past and present data, and is commonly used to analyse trends.

Semi-supervised learning

Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this

combination, machine learning algorithms can learn to label unlabelled data.

Unsupervised learning

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analysing available data. In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organise that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organised.

As it assesses more data, its ability to make decisions on that data gradually improves and becomes more refined.

Under the umbrella of unsupervised learning, fall:

Clustering: Clustering involves grouping sets of similar data (based on defined criteria). It’s useful for segmenting data into several groups and performing analysis on each data set to find patterns.

Dimension reduction: Dimension reduction reduces the number of variables being considered to find the exact information required.

Reinforcement learning

Reinforcement learning focuses on regimented learning processes, where a machine learning algorithm is provided with a set of actions, parameters and end values. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal. Reinforcement learning teaches the machine trial and error. It learns from past experiences and begins to adapt its approach in response to the situation to achieve the best possible result.

## **Benefits of Using Machine Learning in Avocado Marketing Industry**

Collecting information on the market and on better practices concerning avocado cultivation would be of great help to producers, vendors, associations, and companies. This could be used to choose the right places to sell avocados, to carry out successful marketing campaigns or to develop innovations for the production and sales of such product. In this study, machine learning techniques were used to estimate the number of units sold and the total sales of avocados. This will allow avocado producers to plan. With this innovative solution, producers, vendors, associations and companies can get to know the sales expected to be registered in advance. It could be an essential input for making rational decisions regarding the avocado market, such as encouraging consumption or shifting supplies to markets of high demand for the product.

## **About Industry**

The [global avocado market](https://www.persistencemarketresearch.com/mediarelease/avocado-market.asp) is anticipated to show significant market share and a high growth rate during the period of forecast 2017-2027. The global avocado market has seen an upward trend since 2012. The global avocado market is projected to grow at a CAGR of 6.2% throughout the period of assessment to reach a valuation of about US$ 23 Bn by the end of the assessment year (2027) from a value a bit lower than US$ 13 Bn in 2017.This significant growth can be attributed to growing awareness among consumers about health concerns coupled with increasing focus on healthy lifestyle promoting intake of healthy food items, [increasing use of avocados](https://www.persistencemarketresearch.com/market-research/avocado-market.asp) across various end use industries and growing sales channel for distribution of avocados across the globe. Since 2010, investment in the avocado industry has significantly increased to meet the surging global import demand; however, avocado production remains the smallest of the major tropical fruits (Altendof, 2019). Continuous growth in avocado import demand and consumption places pressure in finding potential sustainable avocado 3 markets (FreshPlaza, 2018). For example, South African and Peruvian avocado industries have spent U$ 2.5 million on market and production promotions for Hass avocado cultivar in Europe (FreshPlaza, 2018). This budget will not decline anytime soon. The growing demand in the United Arab Emirates, Saudi Arabia, and China— due to increasing middle income and population (ITC,2019) presents potential market opportunities. Therefore, tapping into these markets is essential for South Africa (Sihlobo, 2018), If the global demand trend continues to grow. Growing world demand opens more opportunities for the upcoming producing regions (i.e. Western Cape) except Mexico and other well established exporters.

### **AI / ML Role in Avocado Marketing Industry**

Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data. Machine learning (ML) is getting more and more attention and is becoming increasingly popular in many other industries. Within the Avocado Market, there is more application of ML regarding the claims.

# **Avocado Marketing Industry**

Aiming at observing the fluctuation of the avocado market based on few factors, several machine learning techniques were evaluated to estimate the number of avocados sold and the total sales of this agricultural product.

The main factors are:

Date: The date of the observation

AveragePrice: The average price of a single avocado

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

Total Bags

Small Bags

Large Bags

XLarge Bags

type: conventional or organic

year: The year

region: The city or region of the observation

## **Main Drivers for AI Auto Quote Analysis**

Predictive modelling allows for simultaneous consideration of many variables and quantification of their overall effect. When a large number of claims are analysed, patterns regarding the characteristics of the claims that drive loss development begin to emerge.

The following are the main drivers which influencing the Claims Analytics:

|  |  |
| --- | --- |
| * **Policy Characteristics** * Exposures * Limits and Deductibles * Coverages and Perils * **Insured Characteristics** * Credit Information * Prior loss experience * Payment history * **Geography based on insured locations** * Auto Repair Costs * Jurisdictional Orientation * Demographics * Crime * **Agency Characteristics** * Exclusive Agents * Independent Agents | * **Claim information** * FNOL * Claimant data (Credit info, geography, social data, etc.) * Other participants (insured, doctors, lawyers, witnesses, etc.) * Cause, type of Injury/Damage * Injury or damaged object * Coverage * Loss Location * Date and time of Loss and Report * Weather at time & location of loss * **Details from Prior Claims** * from same insured * from same claimant * from same location * **Household Characteristics** |

## **Internship Project - Data Link**

The internship project data has taken from Kaggle and the link is:

https://www.kaggle.com/datasets/alanluo418/avocado-prices-20152019

# **AI / ML Modelling and Results**

## **Your Problem of Statement**

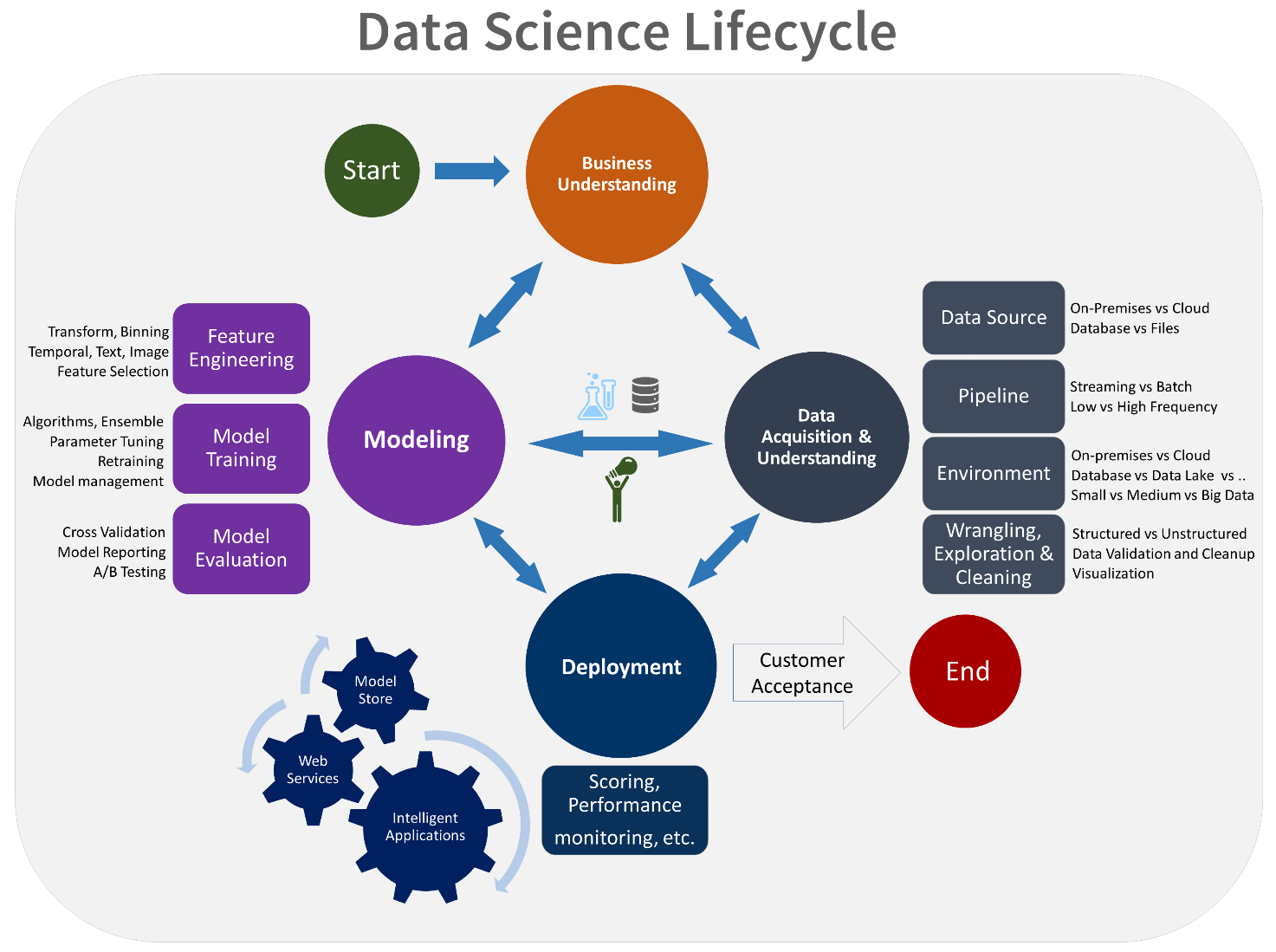
Predictive models are most effective when they are constructed using a company’s own historical claims data since this allows the model to recognize the specific nature of a company’s exposure as well as its claims practices. The construction of the model also involves input from the company throughout the process, as well as consideration of industry leading claims practices and benchmarks.

Predictive modelling can be used to quantify the impact to the claims department resulting from the failure to meet or exceed claim service leading practices. It can also be used to identify the root cause of claim leakage. Proper use of predictive modelling will allow for potential savings across two dimensions:

* Early identification of claims with the potential for high leakage, thereby allowing for the proactive management of the claim
* Recognition of practices that are unnecessarily increasing claims settlement payments

## **Data Science Project Life Cycle**

Data Science is a multidisciplinary field of study that combines programming skills, domain expertise and knowledge of statistics and mathematics to extract useful insights and knowledge from data.



### **Data Exploratory Analysis**

Exploratory data analysis has been done on the data to look for relationship and correlation between different variables and to understand how they impact or target variable.

The exploratory analysis is done for Auto Quote / Policy Conversion with different parameters

### **Data Pre-processing**

We removed variables which does not affect our target variable (Claimed Target) as they may add noise and also increase our computation time, we checked the data for anomalous data points and outliers. We did principal component analysis on the data set to filter out unnecessary variables and to select only the important variables which have greater correlation with our target variable.

### **Check the Duplicate and low variation data**

These can be of two types: Duplicate Values: When two features have the same set of values. Duplicate Index: When the value of two features is different, but they occur at the same index.

There are two ways you can remove duplicates. One is deleting the entire rows and other is removing the column with the most duplicates. Method 1: Removing the entire duplicates rows values. For removing the entire rows that have the same values using the method drop duplicates ().

Low variance means there is a small variation in the prediction of the target function with changes in the training data set. At the same time, High variance shows a large variation in the prediction of the target function with changes in the training dataset.



### **Identify and address the missing variables**

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

7 ways to handle missing values in the dataset:

Deleting Rows with missing values

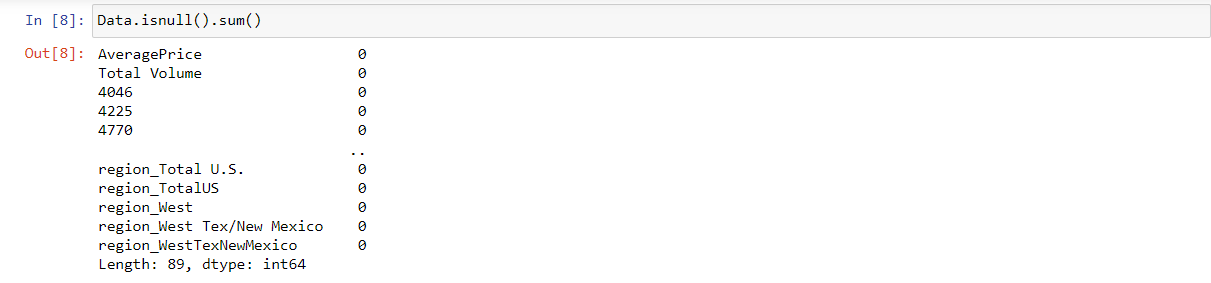
Impute missing values for continuous variable

Impute missing values for categorical variable

Other Imputation Methods

Using Algorithms that support missing values

Prediction of missing values

Imputation using Deep Learning Library

### **Handling of Outliers**

In statistics, we call the data points significantly different from the rest of the dataset outliers. In other words, an outlier contains a value that is inconsistent or doesn’t comply with the general behavior.

You can choose from four main ways to detect outliers:

Sorting your values from low to high and checking minimum and maximum values.

Visualizing your data with a box plot and looking for outliers.

Using the interquartile range to create fences for your data.

Using statistical procedures to identify extreme values.

**There are some techniques used to deal with outliers.**

1. Deleting observations.
2. Transforming values.
3. Imputation.
4. Separately treating

### **Categorical data and Encoding Techniques**

Categorical Data is the data that generally takes a limited number of possible values. Also, the data in the category need not be numerical, it can be textual in nature. All machine learning models are some kinds of mathematical model that need numbers to work with.

Why do we need encoding?

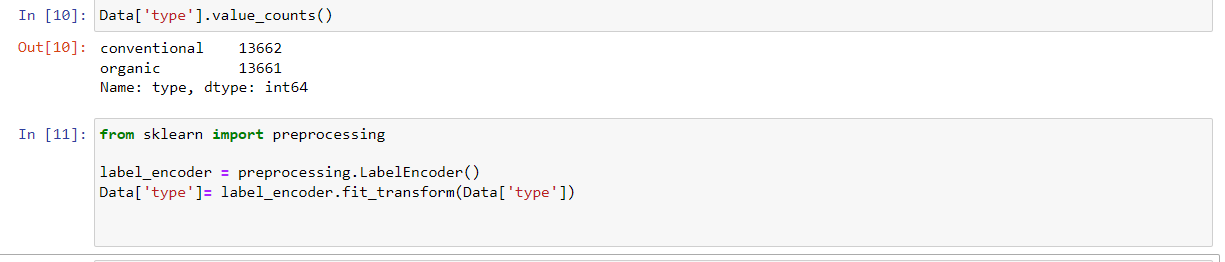
Most machine learning algorithms cannot handle categorical variables unless we convert them to numerical values

Many algorithm’s performances even vary based upon how the categorical variables are encoded

Categorical variables can be divided into two categories:

Nominal: no particular order

Ordinal: there is some order between values

 The two most popular techniques are an **Ordinal Encoding** and **One-Hot Encoding**.

### **Feature Scaling**

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

There are several ways to do feature scaling. The top 5 of the most commonly used feature scaling techniques.

Absolute Maximum Scaling

Min-Max Scaling

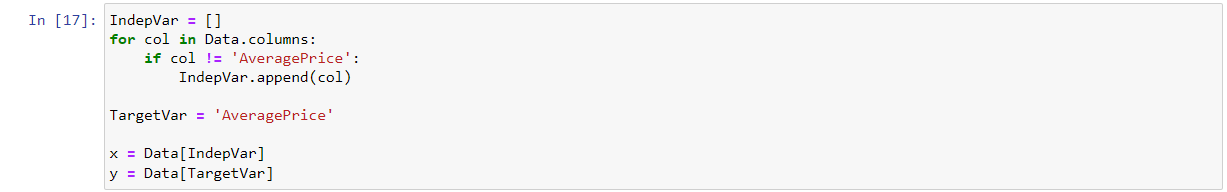
Normalization

Standardization

Robust Scaling

### **Selection of Dependent and Independent variables**

The dependent or target variable here is Claimed Target which tells us a particular policy holder has filed a claim or not the target variable is selected based on our business problem and what we are trying to predict.

The independent variables are selected after doing exploratory data analysis and we used Boruta to select which variables are most affecting our target variable.

### **Data Sampling Methods**

The data we have is highly unbalanced data so we used some sampling methods which are used to balance the target variable so we our model will be developed with good accuracy and precision. We used three Sampling methods

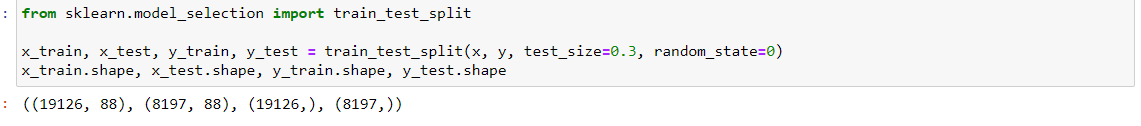
### **Stratified sampling**

Stratified sampling randomly selects data points from majority class so they will be equal to the data points in the minority class. So, after the sampling both the class will have same no of observations.

It can be performed using strata function from the library sampling.

### **Simple random sampling**

Simple random sampling is a sampling technique where a set percentage of the data is selected randomly. It is generally done to reduce bias in the dataset which can occur if data is selected manually without randomizing the dataset.

We used this method to split the dataset into train dataset which contains 70% of the total data and test dataset with the remaining 30% of the data.

### **Models Used for Development**

We built our predictive models by using the following eight algorithms

### **Model 01**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

### **Model 02**

Random forest is an algorithm that consists of many decision trees. It was first developed by Leo Breiman and Adele Cutler. The idea behind it is to build several trees, to have the instance classified by each tree, and to give a "vote" at each class. The model uses a "bagging" approach and the random selection of features to build a collection of decision trees with controlled variance. The instance's class is to the class with the highest number of votes, the class that occurs the most within the leaf in which the instance is placed.

The error of the forest depends on:

* Trees correlation: the higher the correlation, the higher the forest error rate.
* The strength of each tree in the forest. A strong tree is a tree with low error. By

using trees that classify the instances with low error the error rate of the forest

decreases.

### **Model 03**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

### **Model 04**

Support Vector Machine (SVM) is a very popular Machine Learning algorithm that is used in both Regression and Classification. Support Vector Regression is similar to Linear Regression in that the equation of the line is y= wx+b. In SVR, this straight line is referred to as **hyperplane**. The data points on either side of the hyperplane that are closest to the hyperplane are called **Support Vectors** which is used to plot the boundary line. Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value (Distance between hyperplane and boundary line), **a**. Thus, we can say that SVR model tries satisfy the condition -a < y-wx+b < a. It used the points with this boundary to predict value.

### **Model 05**

**LightGBM** is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage.   
It uses two novel techniques: **Gradient-based One Side Sampling** and **Exclusive Feature Bundling (EFB)** which fulfils the limitations of histogram-based algorithm that is primarily used in all GBDT (Gradient Boosting Decision Tree) frameworks. The two techniques of GOSS and EFB described below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks   
**Gradient-based One Side Sampling Technique for LightGBM:**  
Different data instances have varied roles in the computation of information gain. The instances with larger gradients (i.e., under-trained instances) will contribute more to the information gain. GOSS keeps those instances with large gradients (e.g., larger than a predefined threshold, or among the top percentiles), and only randomly drop those instances with small gradients to retain the accuracy of information gain estimation. This treatment can lead to a more accurate gain estimation than uniformly random sampling, with the same target sampling rate, especially when the value of information gain has a large range.

### **Model 06**

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. **XGBoost** can be used directly for **regression predictive**

### **Model 07**

Extra Trees is an ensemble machine learning algorithm that combines the predictions from many decision trees. It is related to the widely used random forest algorithm. It can often achieve as-good or better performance than the random forest algorithm, although it uses a simpler algorithm to construct the decision trees used as members of the ensemble. It is also easy to use given that it has few key hyperparameters and sensible heuristics for configuring these hyperparameters.

### **Model 08**

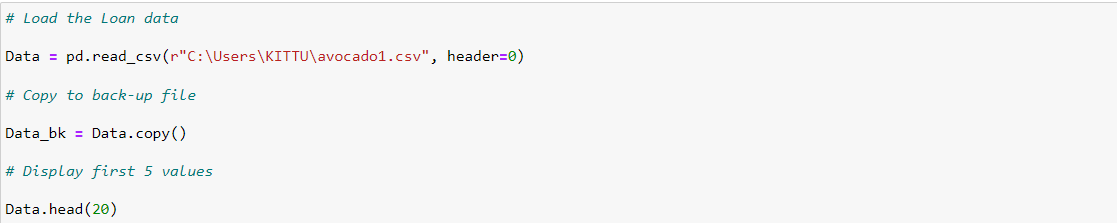
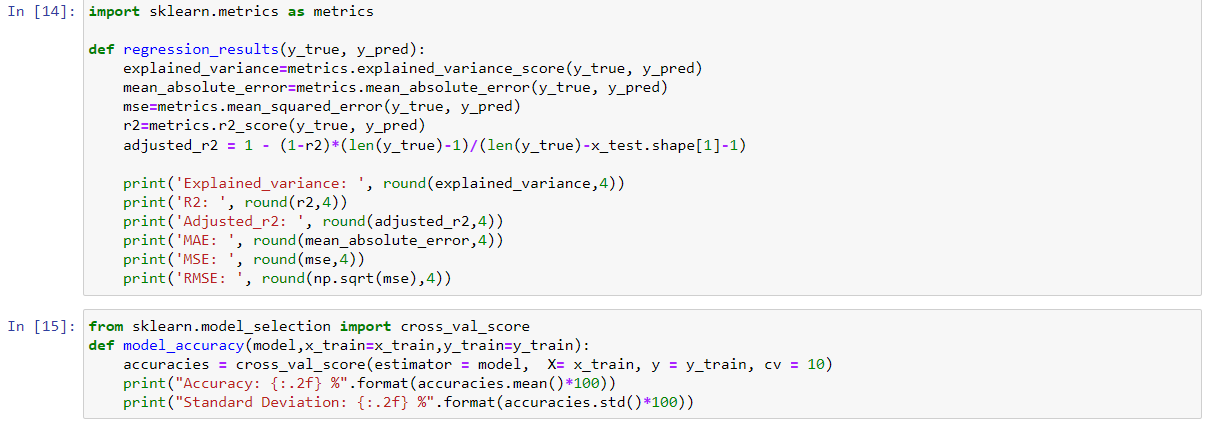
Gradient Boosting Regression is an analytical technique that is designed to explore the relationship between two or more variables (X, and Y). Its analytical output identifies important factors (Xi) impacting the dependent variable (y) and the nature of the relationship between each of these factors and the dependent variable. Gradient Boosting Regression is limited to predicting numeric output so the dependent variable has to be numeric in nature. The minimum sample size is 20 cases per independent variable.

## **AI / ML Models Analysis and Final Results**

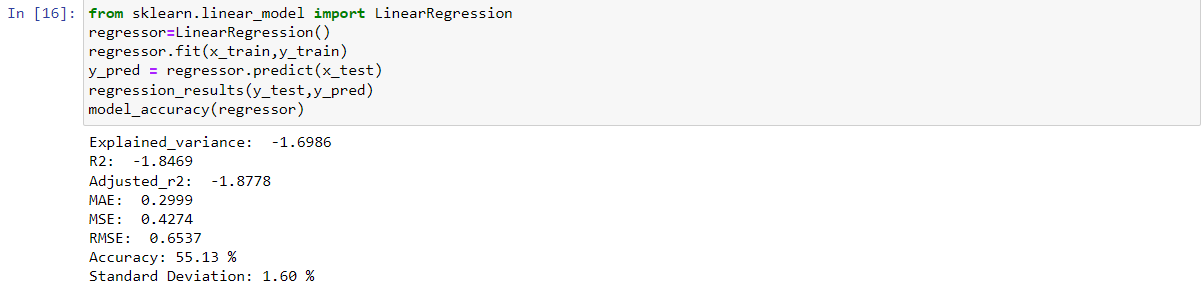
We used our train dataset to build the above models and used our test data to check the accuracy and performance of our models.

We used confusion matrix to check accuracy, , mean absolute error,mean squared error, of our models and compare and select the best model for given auto dataset of size ~ 27324 policies.

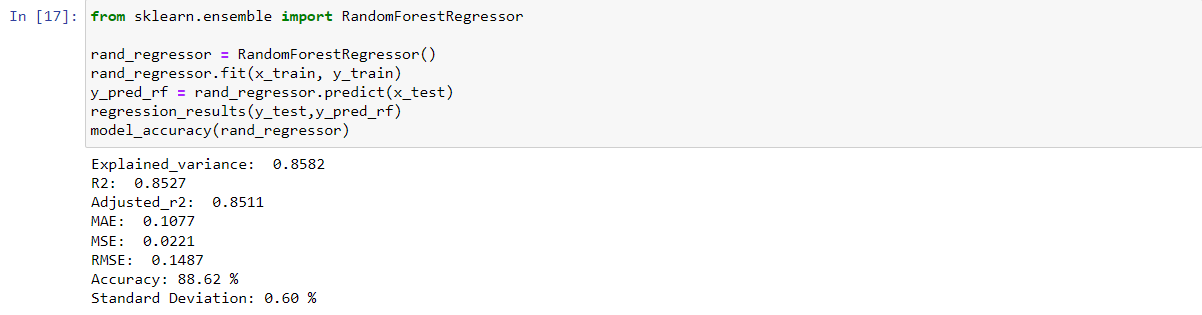
### **Different Model codes**

* The Python code for models with simple random sampling technique as follows:

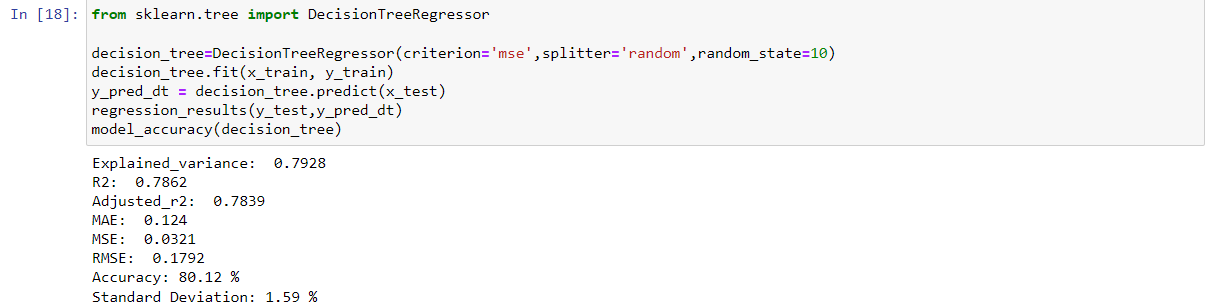
### **Linear Regression Python Code**



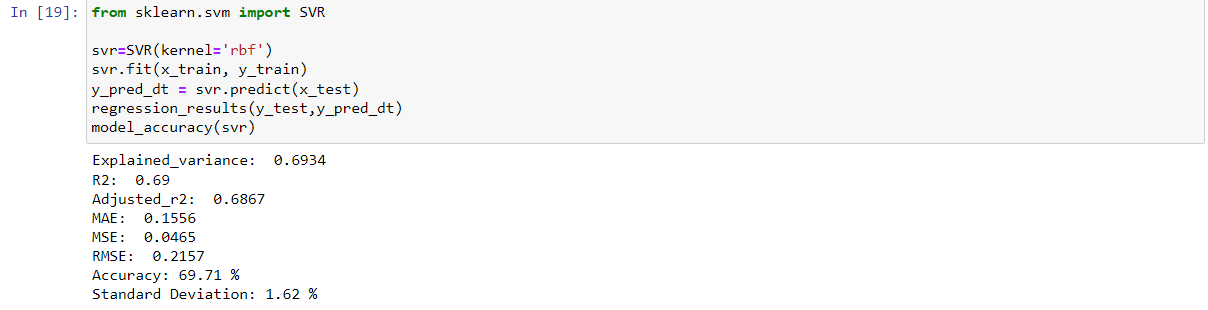
### **Random Forest Python Code**



### **Decision Tree Python Code**

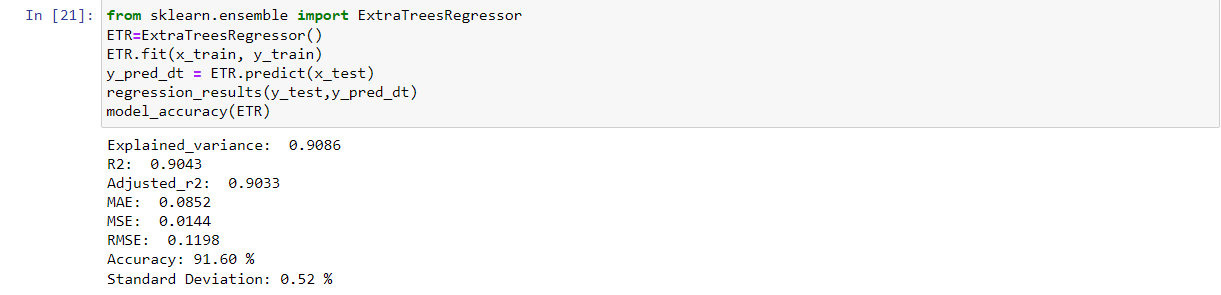


### **SVR Python Code**

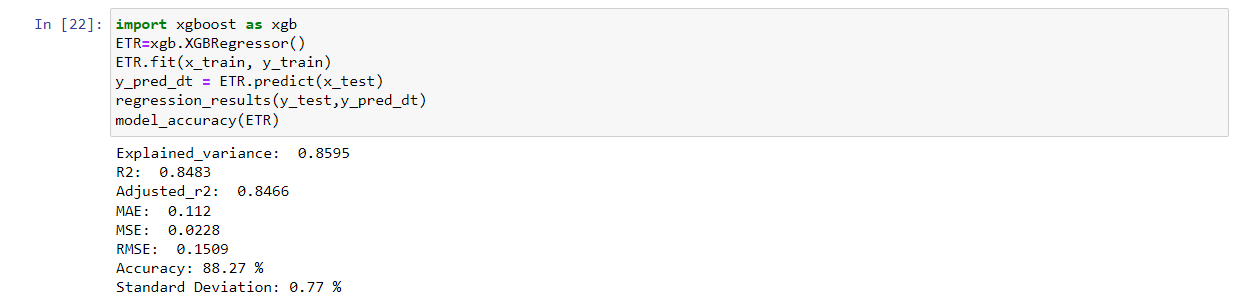


### **LGBMRegressor Python Code**

### **ExtraTreesRegressor Python Code**



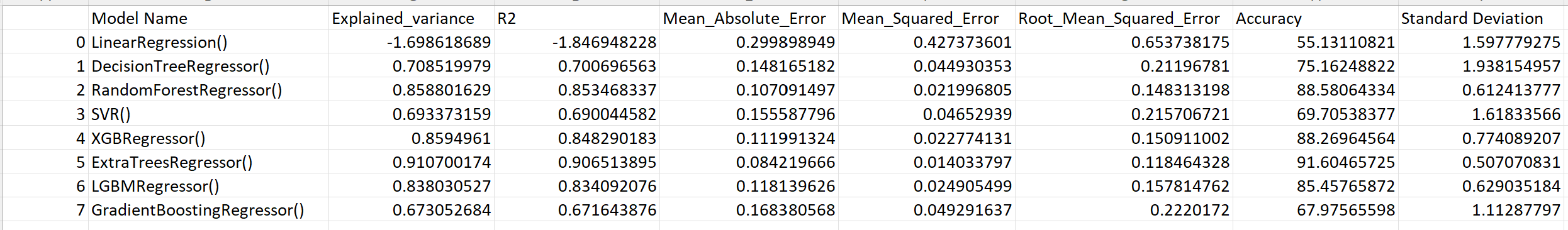
### **XGBRegressor Python Code**



### **GradientBoostingRegressor Python Code**

# **Conclusions and Future work**

The model results in the following order by considering the model accuracy, mean absolute error,mean squared error, etc.

1. **Extra Trees Regressor**
2. **Random Forest Regressor**
3. **XGBRegressor**
4. We recommend model – **Extra Tree Regressor** with Sampling technique as a best fit for the give n BI claims dataset.

The future work to estimate the prediction of prices can be done by using these regression models.

# **5.0 References**

For Dataset: https://www.kaggle.com/datasets/alanluo418/avocado-prices-20152019

Other References:

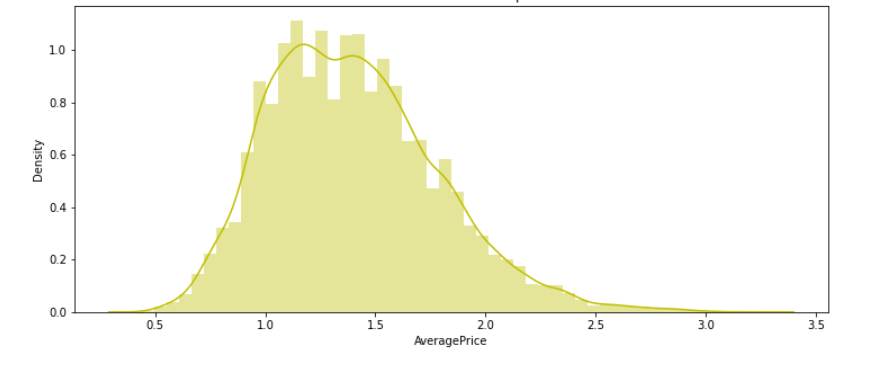
http://www.hassavocadoboard.com/retail/volume-and-price-data

# **Appendices**

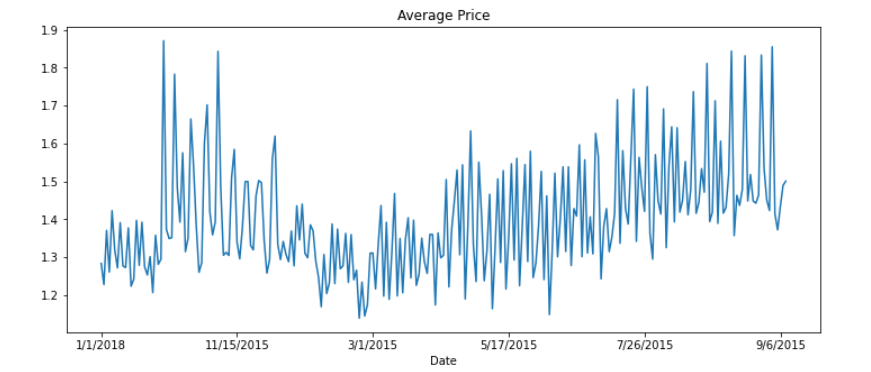
## **Python code Results**

## **List of Charts**

### **6.2.1 Chart 01: Price distribution graph**



### **6.2.2 Chart 02: Change of average price per calendar year**



### **6.2.3 Chart 03:** **Average Price in Each Region**

### **6.2.4 Chart 04: Number of conventional and organic type**