Painting Classifier

Contents

**1.** [**SAATCHI ARTS** 2](#_Toc449699929)

[**2. FEATURES OF PAINTINGS** 2](#_Toc449699930)

[**3. DATA EXTRACTION** 4](#_Toc449699931)

[**3.1 Import.io** 4](#_Toc449699932)

[**3.2 Import.io: Crawler** 4](#_Toc449699933)

[**3.2.1 Train the Crawler to extract data** 4](#_Toc449699934)

[**4. DATA BUILDING** 5](#_Toc449699935)

[**5. NEW FEATURES** 6](#_Toc449699936)

[**6. DATA DISCRETIZATION** 7](#_Toc449699937)

[**6.1 *Data Analysis:*** 9](#_Toc449699938)

[**I. Decision Tree output without new Features:** 10](#_Toc449699939)

[**Lift Graph:** 11](#_Toc449699940)

[**Lift table:** 12](#_Toc449699941)

[**II. Decision tree with New features(Open\_Edition\_Availability and Country):** 13](#_Toc449699942)

[**Lift Graph with New features (Open\_Edition\_Availability and Country):** 13](#_Toc449699943)

[**Lift Table:** 14](#_Toc449699944)

[**III. Decision Tree with new features(Year) :** 14](#_Toc449699945)

[**Lift Graph with New features(Year):** 15](#_Toc449699946)

[**Lift Table:** 15](#_Toc449699947)

[**Model used for Classification:** 15](#_Toc449699948)

[**Classification Rules:** 16](#_Toc449699949)

[**7. RESULTS INTERPRETATION:** 16](#_Toc449699950)

[**8. COMPARING THE DIFFERENT MODELS:** 16](#_Toc449699951)

[**9. SCENARIO ANALYSIS:** 16](#_Toc449699952)

[**9.1 Complete Process Flow:** 17](#_Toc449699953)

# [**SAATCHI ARTS**](http://www.saatchiart.com/paintings/fine-art)

The Saatchi Art is a leading online art gallery, offering complimentary art advisory services. Saatchi Art with its motto “Be original. Buy original” provide original art works for worldwide customers to purchase.

They offer a wide range of artworks to choose from. It has an unparalleled selection of original art works, paintings, photographs, drawings, collages, prints and sculptures in a range of prices by approximately 65,000 artists from 100 countries worldwide.

Saatchi Art founded on May 1, 2006 in Santa Monica, CA has redefined the experience of buying and selling art by making it easy, convenient and welcoming for both collectors and artists. Saatchi art has sold and shipped artwork to collectors in over 80 countries.

# **2. FEATURES OF PAINTINGS**

* **Page url**:

This feature refers to the web URL of specific paintings. Each URL directs the user to its specified painting on the Saatchi website. Every painting has a unique URL address. A customer can directly open a painting using its URL address instead of searching by its name.

* **Paint Name**:

This attribute describes the name of each painting. This feature helps in identifying the paintings according to the names assigned to them. A customer can come back to shop for a particular painting by searching using the painting name, if they do not have the URL address.

* **Artist**:

This feature describes the Artist of each artwork. This attribute helps in identifying the artist behind each painting. A customer can search paintings according to their Artist names. This feature is helpful in situations where some customers like works by a particular Artist and hence would come back to buy artworks by that Artist.

* **Price**:

Price attribute helps in determining the price or cost of each painting. This feature helps customer who visit Saatchi Art with the intention of buying paintings. The prices are listed in Dollars. The prices range from $30,000 to as low as $40. This attribute is of Float data type, so that paintings which have prices in decimal values can also be described precisely.

* **Size**:

The size feature of a painting describes its dimensions. Size of a painting is very important for a customer, as they would like to buy a particular size for their walls. Each painting on the website, displays dimension of format 20 H \* 15.9 W \* 0.7 in. We have calculated size the paintings by multiplying the numerical values of these dimensions. The size attribute is of Float datatype to store values of decimal values without losing its accuracy.

* **Favorites**:

This feature gives information on how many customers have liked a painting. This feature along with “Views” helps in identifying a popular or trending artwork.

* **Views**:

The views feature gives the information on how many customers have viewed the painting. The more number of views would help in identifying which artwork the customers are interested.

* **Date**:

Date attribute indicates the age of a painting. Popularity of artwork can sometime depend on the age of a painting. The older a painting is, the more value it has. Many art collectors are passionate about painting which are older and come from many generations.

* **Subject**:

Subject of a painting describes the main theme of the painting. For example, some artwork depicts Architecture, Flower paintings or people, etc. Customers sometimes want to buy paintings with certain themes or subjects. And some art Collectors are passionate about particular themed artworks. An artist always show passion in painting a particular type of artworks. This feature helps in categorizing artworks in specific categories.

* **Medium**:

The medium of a painting indicates the artwork type. For example, some artworks are done in Acrylics, some are photographs, while some are fabric ink and textile, etc. The medium helps in identifying the type of paintings/material an artist used in his work. And customers can select artworks of their liking.

* **Country Code**:

This feature is used to indicate the codes of countries to which the painting belongs. This field indicates the origin of a painting.

# **3. DATA EXTRACTION**

The data needs to be extracted from the website which will be mined and further used to analyze and extract some features that the data doesn’t reflect directly. The data is about the arts and paintings in the website, the artist name, prize, size and so on. Each and every painting is associated with a unique webpage so manually extracting the information from every source is too time consuming and inefficient. To automate the data extraction and bring it in a tabular form we have used the tool **import.io.**



## **3.1 Import.io**

**import.io** is a Technology startup headquartered at London, UK and San Francisco, US. It was founded in June 2012 by David White, Matthew painter and Andrew Fogg and launched into Beta in September 2013.

import.io has won a number of startup awards including Best Startup by O'Reilly Strata Santa Clara, GigaOM and Web Summit. It is a web-based platform for extracting data from websites without writing any code. The tool allows users to create an [API](https://en.wikipedia.org/wiki/API) using its point and click interface. The free version of the tool provides with a number of features like extractor, crawler, connector and magic. The premium version come with more advanced features like storage, tests, integrations, client libraries, federation and so on. The feature used in our data extraction is Crawler.

## **3.2 Import.io: Crawler**

A **Crawler** is essentially an Extractor with a URL discovery tool. It converts an entire website's worth of pages into a structured database. The Crawler works by scanning all the links on a given page and visits all the ones that follow a particular URL template. If the links also follow a second URL template, it will also extract data from those links. A Crawler is best used when the user is looking for a lot of data spread across an entire site, for example a single product's page on an online shopping site.

### **3.2.1 Train the Crawler to extract data**

To make the crawler extract the data, first it needs to know what all attributes or columns we need from the webpage and its corresponding information and values. The web crawler systematically extracts data from multiple webpages of [www.saatchiart.com/paintings/fine-art](http://www.saatchiart.com/paintings/fine-art).

It does so by crawling, or accessing webpages and returns the data that we need for each one of them. The tool is initially trained on at least 5 webpages from where the required data needs to be extracted. Once the columns are added and rows are trained, it crawls the webpages of the site and extracts the data in a tabular format.

*Below is the screenshot of the raw data that is being crawled by the tool. The data can be exported in 4 different format of HTML, Excel, JSON & CSV:*



*Figure I* – Raw Data from import.io

# **4. DATA BUILDING**

****

We used import.io API to extract the data. We trained the webpage using the tools provided by import.io API, on certain features and achieved in extracting the exact data we desired.

By using UNIX awk commands we pulled out the data from page source. The UNIX commands were used to extract data which import.io wasn’t able to extract. From the page source, using Unix awk we extracted the exact data we needed.

Other features which were missed in import.io and Unix extraction were extracted using Python code. Using Python code we were able to extract the remaining features and build our desired data.

# **5. NEW FEATURES:**

The new features extracted from the Saatchi Arts are Materials, Open Edition Availability, Country, Price Distribution and Year.

* **Material**:

Material of a painting indicates the raw material used by the artist to paint on. It indicates the surface material used to paint the artwork. This feature is important from customer sense of view. A customer before purchase would like to know the canvas used in the painting and would buy according to their interests.

* **Open Edition Availability:**

Open Edition painting can have endless number of prints. They can be reproduced any number of times without exhausting. Open Edition paintings are priced lower than limited editions, because there is no limit to the number of prints as limited editions. In the dataset, the value 1 for this column indicates that the given painting is available to purchase as open edition whereas value 0 indicates its unavailability.

* **Country**:

This attributes describes the country to which a painting or artist belongs to. This field just indicates the country of the painting.

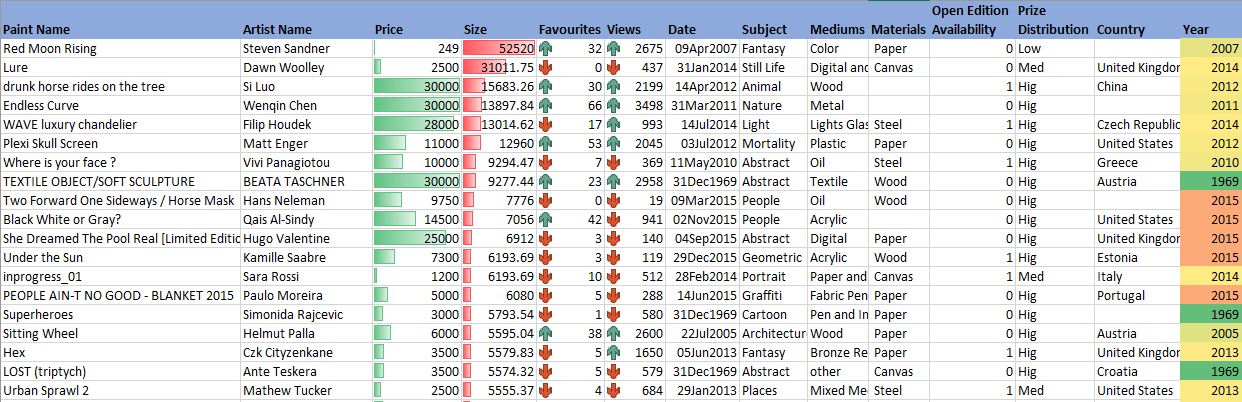
* **Price Distribution:**

Price attribute has been taken as the decision feature and split into 3 intervals: Low, Medium and High. The discretized data (using logarithmic price values) is binned into these 3 intervals around the median of the normal distributed data graph as shown in Page 9 of this report. The resulting variable has been used for our analysis.

* **Year:**

This feature indicates the year in which the artwork was created. It indicates the age of a painting. The older a painting is, the more value it has. Many art collectors are passionate about painting which are older and come from many generations.

*Below is the screenshot of the cleaned dataset with new features.*

****

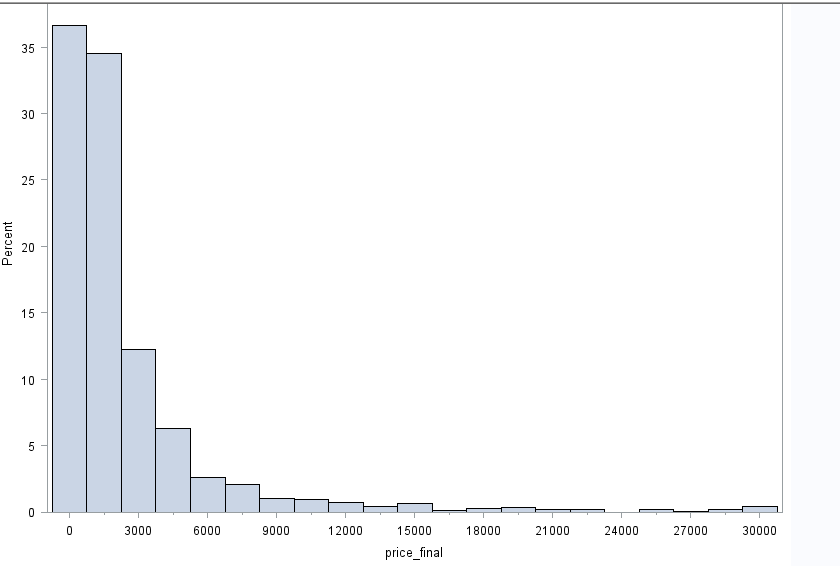
**Extracted Dataset:**

Attached below is the cleaned dataset used in the analysis.

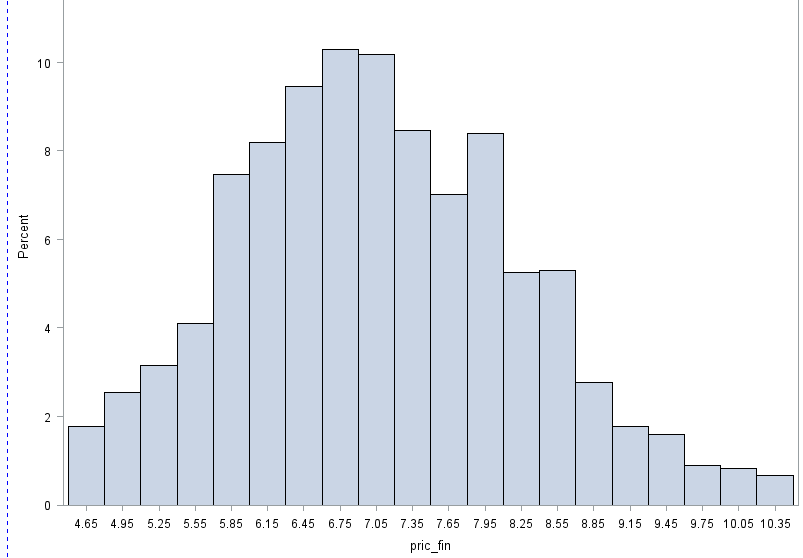
****

# **6. DATA DISCRETIZATION**

* We discretize the data around the price attribute. We implement statistical binning of data of records into Low, Medium and High categories.
* To follow up with the statistical binning process, we bin data into groups by converting continuous price values into categorical values. And then the distribution of prices is checked to ensure that they are in normalized format.
* The resulting raw data after inspection, revealed that the price data we collected is skewed as shown below.



* To bin the skewed data, we need to convert this data to have a normal distribution so that it will ease our binning process.
* To transform the data into normal distribution, we converted the price attribute values by calculating their logarithmic values; log(price).
* After changing the values, we achieve a normal distributed data resulting in a bell-shaped curve graph (indicating normal distribution of data) as shown below.



* After successfully normalizing the data, we binned the data as shown below

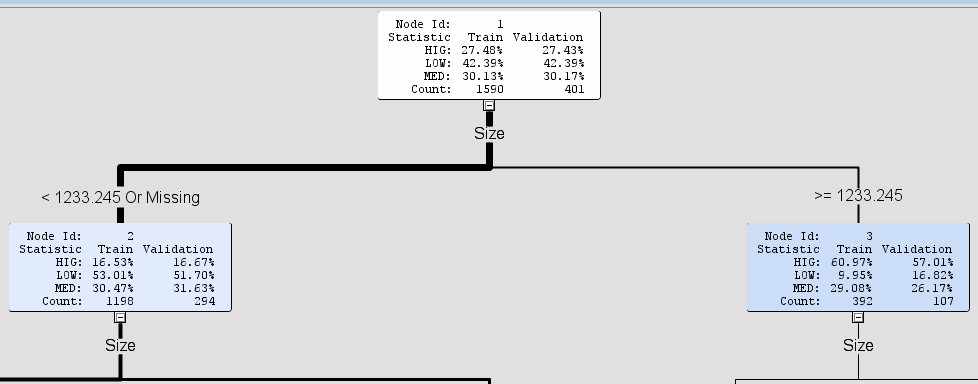
***if (pric\_fin<6.90776) then classifier='Low'; else if (pric\_fin>=6.90776 and pric\_fin<7.87778) then classifier='Medium';else if (pric\_fin>=7.87778) then classifier='High'***

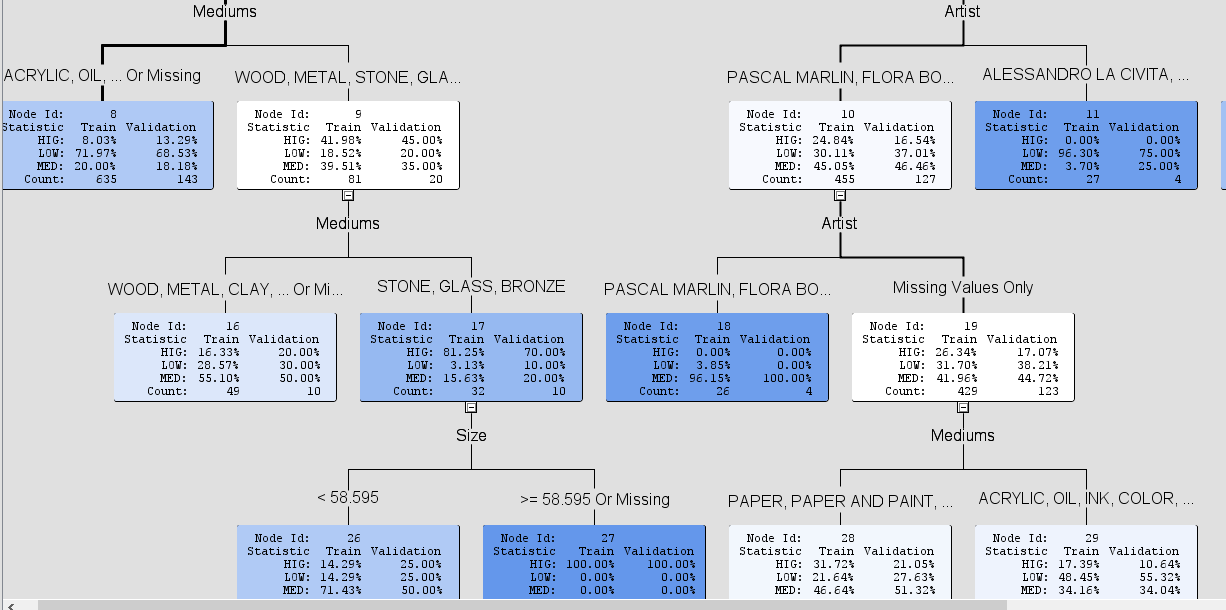
The resulting variable is used for our analysis.

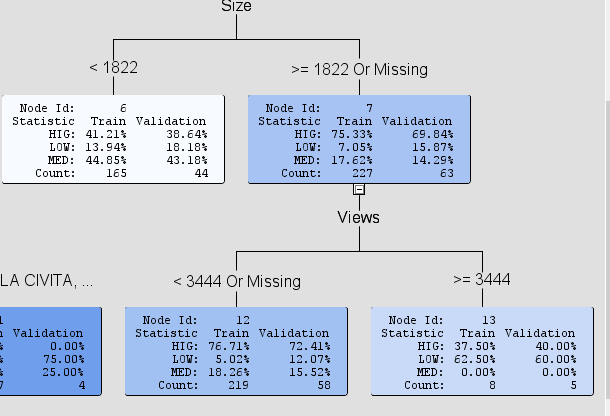
## **6.1 *Data Analysis:***

We Split the data into Train and Validation data sets. With giving 80% to training and just 20% to validation. The lift performance of the first model as seen below is more likely for both Train and Validation and we see high accuracy as well.

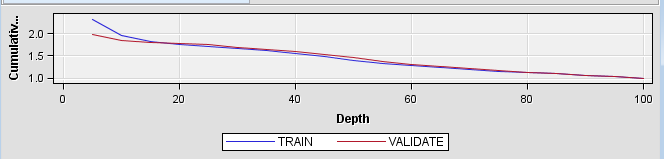
### **I. Decision Tree output without new Features:**



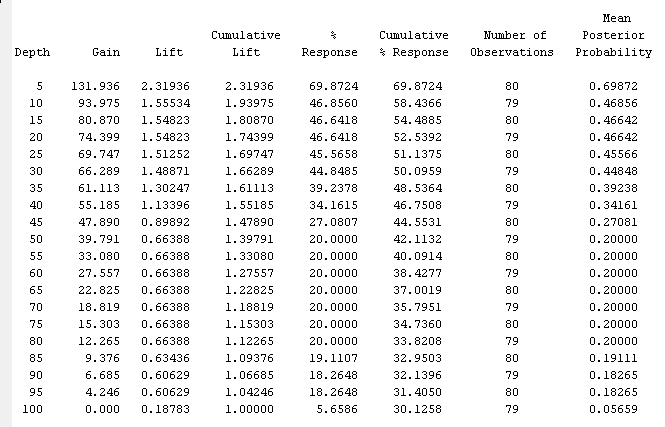




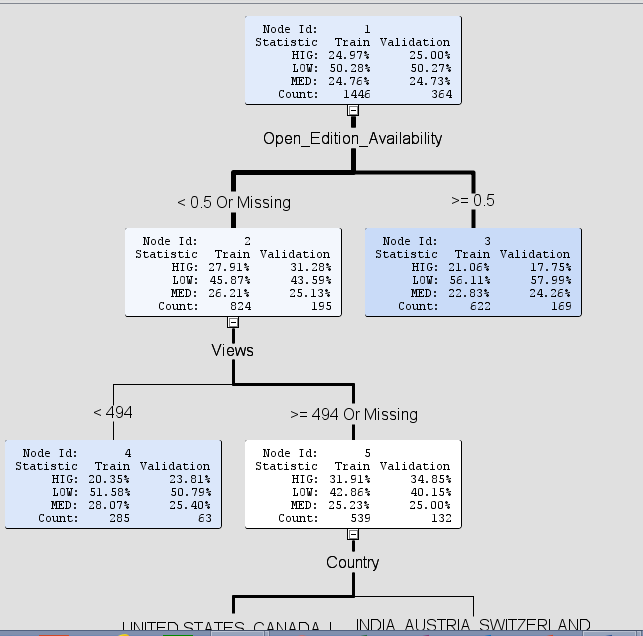
### **Lift Graph:**



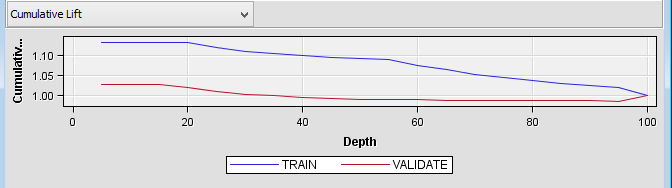
### **Lift table:**



### **II. Decision tree with New features(Open\_Edition\_Availability and Country):**

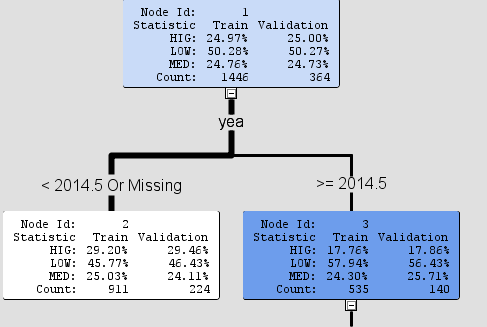


### **Lift Graph with New features (Open\_Edition\_Availability and Country):**

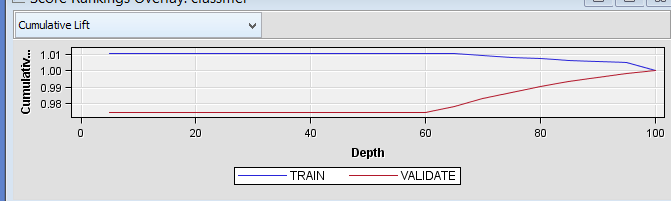


### **Lift Table:**

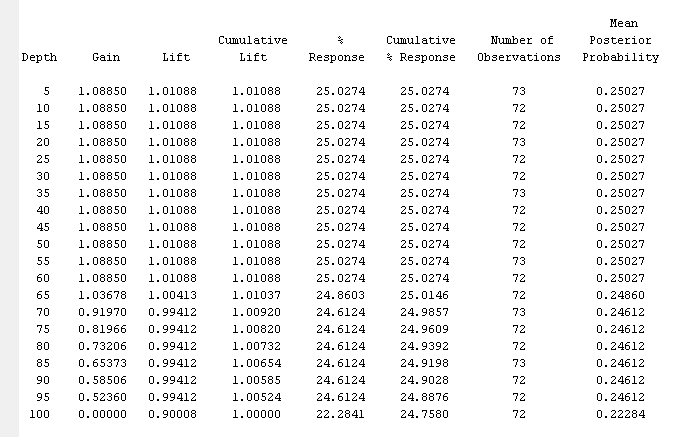
### **III. Decision Tree with new features(Year) :**



### **Lift Graph with New features(Year):**

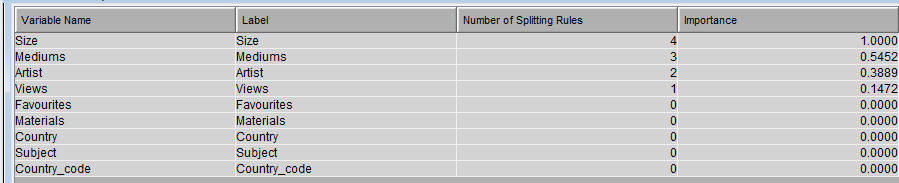


### **Lift Table:**



### **Model used for Classification:**

We use decision tree as our classification algorithm. Decision tree classifies based on Entropy and Information Gain. The top 4 important variables based on Information is gain is as shown below.



### **Classification Rules:**



## **7. RESULTS INTERPRETATION:**

The below result is completely based on the Decision Tree-1 because both the lift performance was better than the other two models.

The Most predominant node in the tree is for Size<1233.245 and Medium in (Acrylic, Oil). The other variables though are part of the rules are not significant considering the number of painting that fall under each rule.

On the other hand whenever the Medium is Stone or Glass or Bronze there is high probability that the Painting will fall into high price category. And the price is mostly low when the Artist name is Alessandro La Civita.

## **8. COMPARING THE DIFFERENT MODELS:**

Our selected Model as shown above has an accuracy of 69%. Though the other two model has an accuracy more that the selected model with 75% each, the lift performance for the depth of 100% of the data is very low compared to our selected model. As you can see from the lift table of all the three results, the probability of selecting a painting of any category is 30% using the first model in a depth of 100% of data, but at the same time for the rest of the two models the probability is just 24%. And also our selected model behaves best at the depth of 30% with a probability of getting the necessary data is 44%. But even at the depth of 30% the probability of other two model increases only by 1% with 25%.

## **9. SCENARIO ANALYSIS:**

Let us say a buyer wants to select a painting from a randomly sampled set and wants to end up getting a painting with low price. The using our model if you randomly sample and select just the top 30% of the data and apply our Model, then there is 44% chance that he/she will get a painting with price low. But on the other hand if he doesn’t apply our model and randomly select from 30% of data, the chance of getting a low price painting is just 30%.

### **9.1 Complete Process Flow:**

