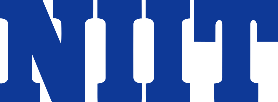


|  |
| --- |
|  |
|  |
| Amazon Product Review Analysis  2 February 2023 |



Document Control Sheet

Prepared by:

|  |  |
| --- | --- |
| Name | Contact Details |
| Saud Shaikh | Saud.shaikh0804@gmail.com |
| Sanchit Upadhyay | Sanchitupadhyays80@gmail.com |
| Shakib Hussain | shakib192@gmail.com |
| Rohit Kumar | rohit.firstblood@gmail.com |

Distribution:

|  |  |
| --- | --- |
| Name | Contact Details |
| NIIT leading online course provider | India | [www.niit.com](http://www.niit.com/) |

Reviewed by:

|  |  |
| --- | --- |
| Name | Contact Details |
| Dr. Amit Kumar | amit.1b.kumar@niit.com |

# Abstract

To develop an automated system to analyze and monitor an enormous number of reviews. By monitoring the entire review history, he will analyze tone, language, keywords, and trends over time to provide valuable insights that increase the success rate of existing and new products and marketing campaigns. The aim of this project was to evaluate and investigate the feasibility of using Machine Learning as an alternative for common problem solving strategies today. To this end, a data set was provided. From this dataset models were built in Machine Learning, to examine the machine learning algorithms available on the platform’s toolbox. After a preliminary study of the available algorithms and data review.

Thus, the study could focus on Exploratory Data Analysis, Text Preprocessing, Sentiment Analysis Machine Learning algorithms and Time Series Forecasting. The major discovery is that the machine learning approach should be suitable for these types of problems due to many aspects. not the least because of the easily grasped user interface, as well as the wide availability of algorithms within machine learning. Although it lacks features for the more advanced users and there are better alternatives available when designing models with a high-performance criterion.

Inventory Optimization and Demand Forecasting Optimize inventory management by identifying the product categories (clustering as an outcome of text processing) on the customer review data. Predict what kind of products could be in demand (Time Series Analysis).

Table of Contents

[Abstract 3](#_Toc126228796)

[Chapter 1. Introduction 7](#_Toc126228797)

[1.1 Background: 7](#_Toc126228798)

[1.2 Goal: 7](#_Toc126228799)

[1.3 Setup: 7](#_Toc126228800)

[Chapter 2. Data Preprocessing 9](#_Toc126228801)

[2.1 Importing Libraries: 9](#_Toc126228802)

[2.2 Importing Dataset: 9](#_Toc126228803)

[2.3 Merging our Two Different Dataset into one Data: XIII](#_Toc126228804)

[2.4 Checking for Null Values in Merge Dataset: 15](#_Toc126228805)

[2.5 Dropping all the unnecessary columns: 15](#_Toc126228806)

[2.6 Checking Dataset shape and reading its columns: 16](#_Toc126228807)

[2.7 Converting Datatypes to Relevant form: 17](#_Toc126228808)

[2.8 DUPLICATE VALUE TRAETMENT: 18](#_Toc126228809)

[2.9 Describing the Data: 18](#_Toc126228810)

[2.10 Dropping unnecessary Columns from Dataset: 19](#_Toc126228811)

[2.11 Changing price data type: 19](#_Toc126228812)

[2.12 Using filter in main\_Cat columns: 20](#_Toc126228813)

[2.13 Dropping and Resetting Index: 20](#_Toc126228814)

[Chapter 3. Exploratory Data Analysis (EDA) 21](#_Toc126228815)

[3.1 Univariate Analysis: 21](#_Toc126228816)

[3.2 Multivariate Analysis: 25](#_Toc126228817)

[3.3 Top 10 brands Analysis: 29](#_Toc126228818)

[Chapter 4. Text Preprocessing 30](#_Toc126228819)

[4.1Text Cleaning: 30](#_Toc126228820)

[4.2 Stopwords: 31](#_Toc126228821)

[Chapter 5. Sentiment Analysis and Machine Learning and modeling 34](#_Toc126228822)

[5.1 Defining function for sentiment Analysis: 34](#_Toc126228823)

[Chapter 5.2. Machine learning and Modeling 35](#_Toc126228824)

[5.2.1 Classification models: 36](#_Toc126228825)

[5.2.1 Multinomial Naive Bayes: 36](#_Toc126228826)

[5.2.2 Logistic Regression: 37](#_Toc126228827)

[5.2.3 Decision Tree Classification Algorithm: 37](#_Toc126228828)

[5.2.4 Random Forest Classifier: 38](#_Toc126228829)

[5.2.5 AdaBoost Classifier: 39](#_Toc126228830)

[5.2.6 Conclusion: 40](#_Toc126228831)

[5.3 Classification models (SMOTE Function): 41](#_Toc126228832)

[5.3.1 Multinomial Naive Bayes (SMOTE Function): 41](#_Toc126228833)

[5.3.2 Logistic Regression (SMOTE Function): 42](#_Toc126228834)

[5.3.3 Decision Tree Classification Algorithm (SMOTE Function): 43](#_Toc126228835)

[5.3.4 Random Forest Classifier (SMOTE Function): 44](#_Toc126228836)

[5.3.5 AdaBoost Classifier (SMOTE Function): 44](#_Toc126228837)

[5.3.6 Conclusion: 45](#_Toc126228838)

[5.3.7 Doing prediction for our future Sentiment by logistic regression: 46](#_Toc126228839)

[5.4 Clustering Models: 47](#_Toc126228840)

[5.4.1 Literature Review: 47](#_Toc126228841)

[5.4.2 Data Collection: 48](#_Toc126228842)

[5.4.3 K-Means Algorithm: 48](#_Toc126228843)

[5.4.4 Silhouette Scores: 51](#_Toc126228844)

[5.4.5 Agglomerative Clustering: 55](#_Toc126228845)

[5.4.6 Dendrogram: 56](#_Toc126228846)

[Chapter 6. Time Series Analysis 57](#_Toc126228847)

[6.1 Import Libraries: 57](#_Toc126228848)

[6.2 Sentiment analysis part according to the rating: 58](#_Toc126228849)

[6.3 Positive trend analysis: 59](#_Toc126228850)

[6.4 Decomposition chart for the Positive reviews Data: 60](#_Toc126228851)

[6.5 Stationarity in time series analysis: 61](#_Toc126228852)

[6.6 Differencing method: subtracts the current value from the previous value: 61](#_Toc126228853)

[6.7 Plot ACF PACF: 62](#_Toc126228854)

[6.8Train test split: 62](#_Toc126228855)

[6.9 ARIMA model: 64](#_Toc126228856)

[6.10 SARIMA MODEL: 66](#_Toc126228857)

[6.11 Negative trend analysis: 67](#_Toc126228858)

[6.12 Decomposition graph for Negative reviews: 68](#_Toc126228859)

[6.13 ADF test to check the stationarity of data: 68](#_Toc126228860)

[6.14 Differencing method: subtracts the current value from the previous value: 69](#_Toc126228861)

[6.15 ACF PACF Plot: 69](#_Toc126228862)

[6.16 ARMA MODEL: 70](#_Toc126228863)

[6.17 ARIMA model: 71](#_Toc126228864)

[6.18 Evaluation: 71](#_Toc126228865)

[6.19 SARIMA MODEL: 72](#_Toc126228866)

[Chapter 7. Conclusion 73](#_Toc126228867)

[7.1 Data Preprocessing Conclusion: 73](#_Toc126228868)

[7.2 EDA (Exploratory Data Analysis) Conclusion: 73](#_Toc126228869)

[7.3 Text Preprocessing Conclusion: 73](#_Toc126228870)

[7.4 Sentiment Analysis & Machine learning and Model Building Conclusion: 73](#_Toc126228871)

[7.5 Time Series Analysis Conclusion: 73](#_Toc126228872)

[7.6 Recommendation: 74](#_Toc126228873)

# Chapter 1. Introduction

In this chapter we are covering background, goals, Setup for our Amazon dataset.

### 1.1 Background:

The year was 1994 when Jeff Bezos launched Amazon from his garage. In 1995, the first product launched by Amazon was a book in 50 states and in 45 countries within 30 days. (Oberlo 2021) Within 26 years, Amazon became the world's largest online retailer and a household name. The Amazon name has become synonymous with online shopping and continues to grow by developing new products, acquisitions, and numerous service offerings to enlarge the customer base. Nowadays, almost 150.6 million people turn to the Amazon app for most everything. Several types of research have proven that customers trust Amazon. (Statista 2019) On average, the small and medium-sized businesses located in the USA sell more than 4,000 items per minute (Amazon 2019), which leads to millions of product reviews on Amazon.

Reviews tell which products and features are trending, what's in demand, what's no longer relevant, how products and competitors are doing, and much more. It's observed that a significant number of shoppers look at product reviews before they make a purchase. Survey results show that positive product reviews are a key factor for purchasing by 57 percent of Amazon buyers. (Statista, 2019) As product reviews are often the deciding factor for many customers, it's very important to have a well-automated system for monitoring them.

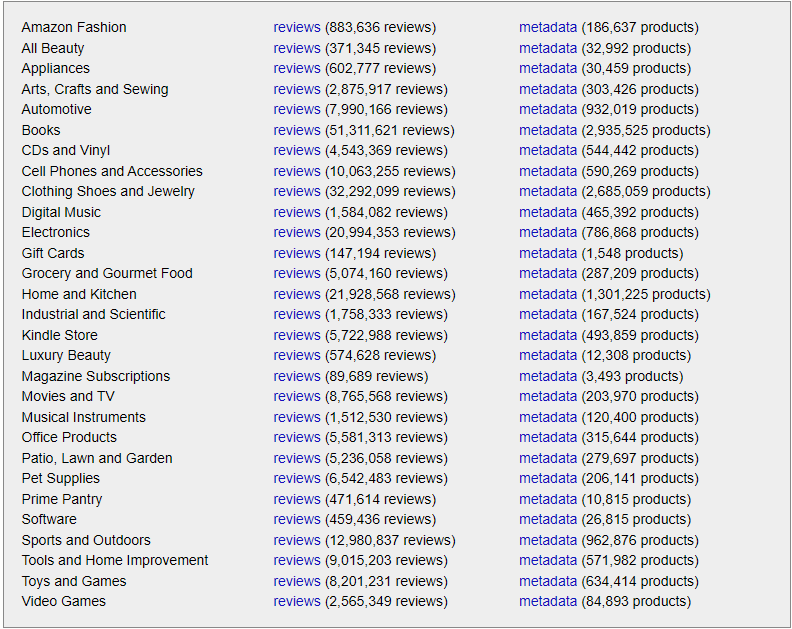
### 1.2 Goal:

The traditional manual process of Amazon product reviews is time-consuming and inefficient when millions of reviews are being posted all the time. It doesn't show trends or patterns over time, and it's tough to understand customer sentiment towards any product or its delivery. Review analysis must adjust dynamically to the changing trend.

### 1.3 Setup:

Preparing Data for Analysis Here are the data pre-process requirements for the Amazon dataset:

1. Reusability: Collecting a real dataset is very costly and time-consuming. While analyzing data, there can be a chance of losing the state of data. You should take your data backup before you start the analysis for future use.
2. Free from the anomaly: Re-sample or perform required processing to remove disintegration from data.
3. Missing value treatment: Make your data free from missing values.
4. Validity: Check against the checklist of all assumptions required to meet in data before you start the analysis.
5. Reliability: Perform required data formatting and editing to make data fit for analysis.



We have Downloaded Metadata of Musical Instruments



K-cores: (i.e., dense subsets): These data have been reduced to extract the [k-core](https://en.wikipedia.org/wiki/Degeneracy_(graph_theory)), such that each of the remaining users and items have k reviews each.

We have downloaded 5-Core data of Musical Instruments



Sentiment Analysis Using NLP:

Here is the required data analysis for the Amazon dataset:

1. Natural Language Understanding: Convert a large set of text into more formal representations, such as first-order logic structures that are easier for the computer programs to manipulate notations.
2. Information Extraction: Extract structured information from unstructured information.
3. Sentiment Analysis: Analyze the attitude or emotional state of the customers from their posted review texts. Furthermore, make sure your work should follow a minimum of three types (Positive, Negative, Neutral) of customer sentiments.

Perform product and customer segmentation using sentiment analyzed data into:

1. satisfied vs. dissatisfied customers.
2. products that may be good recommendations to a customer vs. not
3. psychographics
4. attitudes towards specific products

Sentiment Time Series Forecasting:

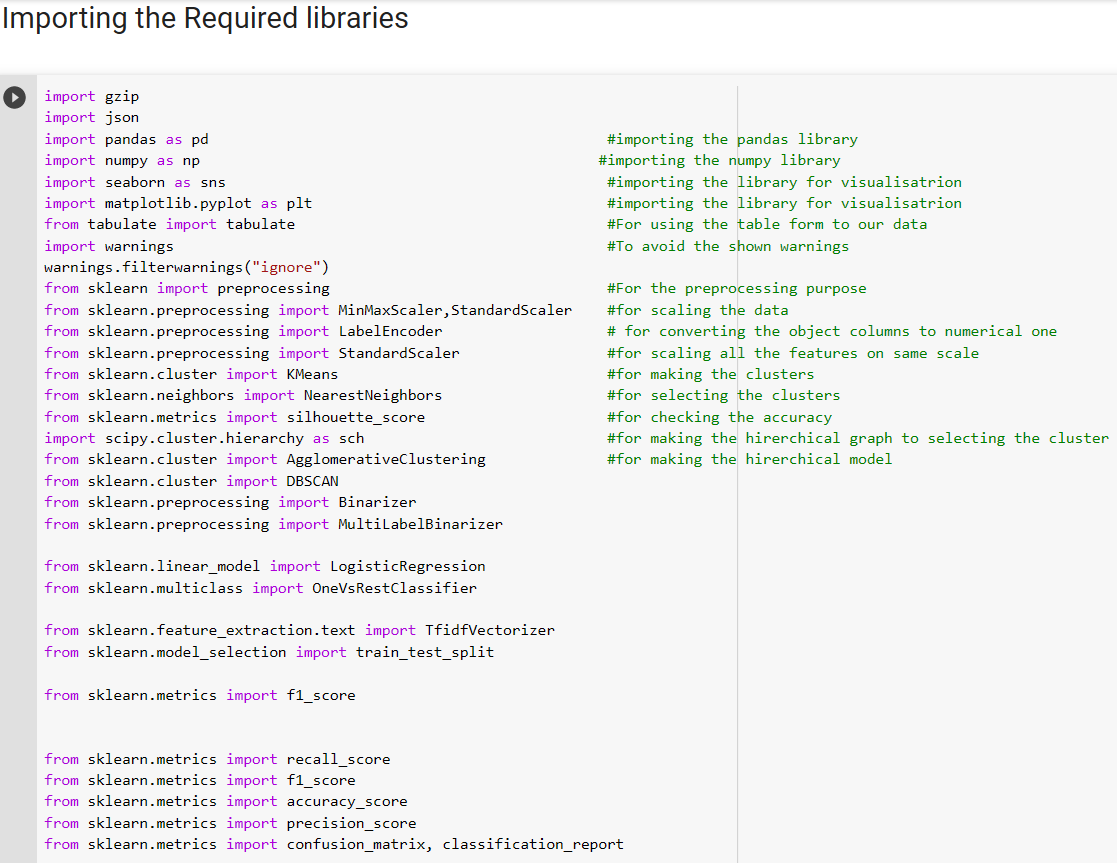
Combining time series analysis with natural language processing, we're able to show how the sentiment of unstructured text data changes over time, as well as use it to predict future data trends.

# Chapter 2. Data Preprocessing

In this chapter we are covering what libraries we have imported to do the analysis which 2 dataset we are merging checking for the missing value treatment duplicate values and dropping unnecessary columns

### 2.1 Importing Libraries:

Before doing Data preprocessing, we have to import all the required Libraries to perform further tasks for analyzing the data.



### 2.2 Importing Dataset:

5-Core Dataset:

Loading our Dataset before doing data preprocessing our data is in Jason Format and we are appending our dataset into an empty list we created and appending our data into it and making our dataset into DataFrame. First, we are importing our Musical Instrument 5-Core Dataset.

Graphical user interface, text, application

Description automatically generated

After Loading our Musical Instrument 5-Core Dataset, we are reading our dataset head which is to see our top 5 Data entries only.



Our Musical Instrument 5-Core Dataset have row 231392 and columns 12.

Graphical user interface, text, application

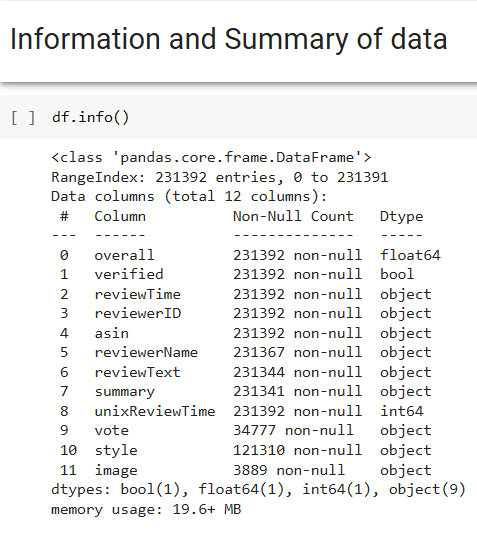
Description automatically generated

Our Musical Instrument 5-Core Dataset have following given below columns:

Graphical user interface, text, application

Description automatically generated

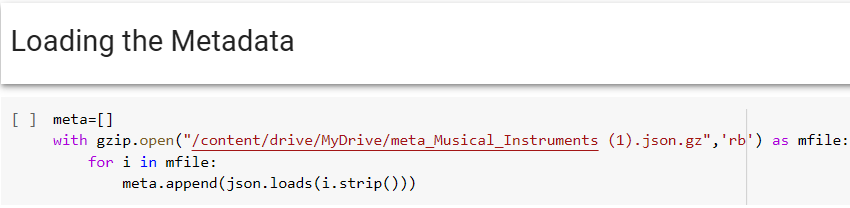
Understanding our Musical Instrument 5-Core Dataset columns data type:



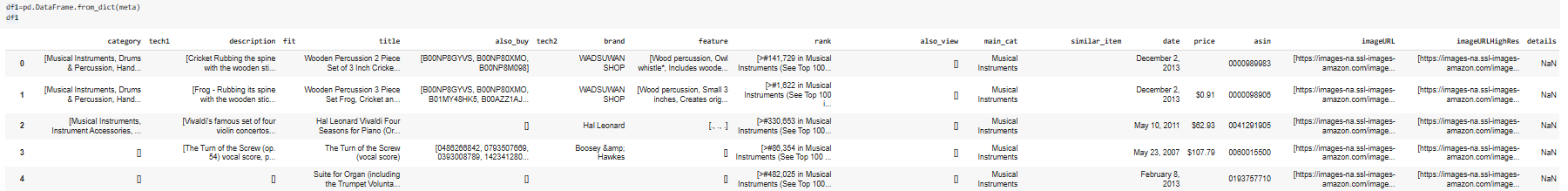
Our 5-Core Dataset have 12 columns in which overall have float data type verified have Boolean data type review Time, reviewer ID, Asin, reviewer Name, review Text, summary, vote, style, image are having Object Data type. unixReviewTime have Integer Data type.

Metadata Dataset:

Loading our Dataset before doing data preprocessing our data is in Jason Format and we are appending our dataset into an empty list we created and appending our data into it and making our dataset into DataFrame. First, we are importing our Musical Instrument Metadata.



After Loading our Musical Instrument Metadata Dataset, we are reading our dataset head which is to see our top 5 Data entries only.



Our Musical Instrument 5-Core Dataset have row 120310 and columns 19.

Graphical user interface, text, application

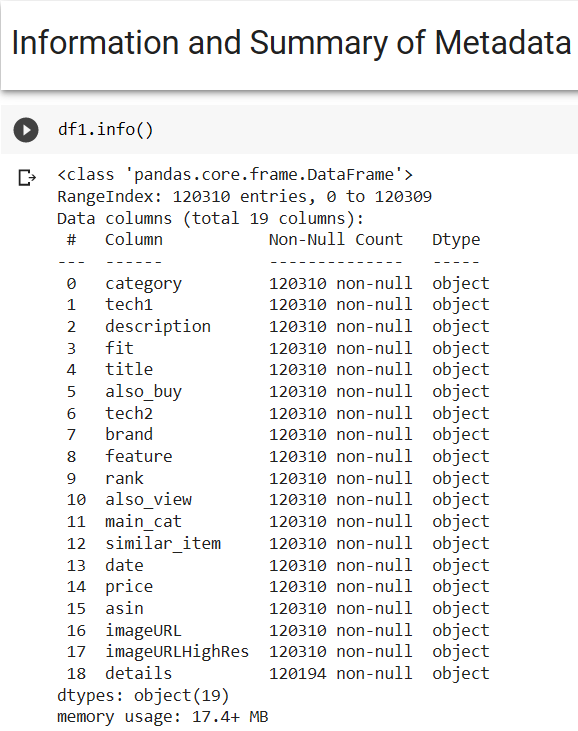
Description automatically generated

Our Musical Instrument Metadata Dataset have following given below columns:

Graphical user interface, text, application

Description automatically generated

Understanding our Musical Instrument MetaData Dataset columns data type:



Our Metadata Dataset have 19 columns in which every Columns are in Object Data types.

### 2.3 Merging our Two Different Dataset into one Data:

We are merging our 2 different Dataset into one dataset on the basis of common column Asin both

5-core and metadata have column Asin.

Graphical user interface, application

Description automatically generated

Our Musical Instrument Merge Dataset have row 286811 and columns 30.

Graphical user interface, application

Description automatically generated

Our Musical Instrument Merge Dataset have following given below columns:

Text

Description automatically generated

Understanding our Musical Instrument Merge Dataset columns data type:

Table

Description automatically generated

### 2.4 Checking for Null Values in Merge Dataset:

After merging our dataset the 1st step we do in data preprocessing is we find Null values in the columns of our merge Dataset .

Table

Description automatically generated

As we can see in the above snapshot in our merge dataset we found that we have null values in reviewer Name, review Text, summary, vote, style, image and details Columns have null values.

### 2.5 Dropping all the unnecessary columns:

Graphical user interface, text, application, chat or text message

Description automatically generated

We are dropping all the unnecessary Columns from our merge Dataset.

### 2.6 Checking Dataset shape and reading its columns:

Graphical user interface, table

Description automatically generated

After dropping unnecessary columns and doing after null value treatment we have 286650 rows and 26 columns.

Table

Description automatically generated

This are the columns left in our dataset and having no null values.

### 2.7 Converting Datatypes to Relevant form:

Review Time column in our dataset have object data type which we to convert is in DataTime data type. Category column in our dataset have values in list, we are converting list into string datatype.

Table

Description automatically generated with medium confidence

As we can see in the above snapshot we have changed our data type of review time in DataTime and category in Object.

### 2.8 DUPLICATE VALUE TRAETMENT:

Graphical user interface, application, table

Description automatically generated

Finding duplicate values in our dataset taking Asin, review text, review Name, unixreviewTime and review Id columns together we have 64802 Duplicates values and we are dropping all the Duplicate values and keeping First values.

### 2.9 Describing the Data:

Using Describe function in our dataset for the columns having float and integer data type. Where we are calculating count, mean, std(standard), min, 25%, 50%, 75% and max.

Table

Description automatically generated

### 2.10 Dropping unnecessary Columns from Dataset:

After Analyzing more about the dataset, We are further dropping the unnecessary columns from the dataset.

Graphical user interface, text, application

Description automatically generated

We are dropping All columns related to images, also\_buy, tech2, similar\_item, feature, tech1, fit, rank, date, title, review Name, description, also\_view, review Text and Summary.

We are combining Review Text and summary as Review.

After dropping all the unnecessary columns and combining columns are left with 221848 rows and 10 columns.

Graphical user interface

Description automatically generated

### 2.11 Changing price data type:

Text

Description automatically generated

Changing price data type from object to float. Defining a function to remove $ symbol from the price column values and then converting price data type from object to float.

### 2.12 Using filter in main\_Cat columns:

After analyzing the dataset main\_cat columns contents lot of unnecessary values of different categories which is not important like home theatre, electronics, toys, and games and etc. we only want musical instruments category as per our category chosen.

Graphical user interface, table

Description automatically generated

### 2.13 Dropping and Resetting Index:

Graphical user interface, text, application

Description automatically generated

Dropping all the null values and using function reset\_index () to reset our dataset index.

# Chapter 3. Exploratory Data Analysis (EDA)

In this chapter we have Performed Exploratory data analysis in python as well as TABLEAU for our dataset for Visualization we used TABLEAU. Tableau is a software for visualization and plotting graph in a professional manner.

### 3.1 Univariate Analysis:

1. We are using python in with we are defining a function for Categorical columns(info\_of\_cat) to find each columns its unique value, mode, checking is there any null values or not.
2. In another code cell we are defining a function for Continuous columns(num\_info) to find each columns its mean, median, standard deviation and checking is there any null values or not.

A picture containing calendar

Description automatically generated

1. Analyzing overall columns:

our dataset having overall columns which comes under the categorical and we are calling function which we define for categorical columns and we are finding its unique values, mode and checking its null values.

Graphical user interface, text, application

Description automatically generated

Chart

Description automatically generatedAs per the above visualization for our Overall column in our dataset 5Star Rating is having the highest percentage of 76.28% and 1Star Rating is having the lowest percentage of 0.72%.

1. Analyzing Verified column: Verification of a person (YES or NO):

our dataset having Verified columns which comes under the categorical and we are calling function which we define for categorical columns and we are finding its unique values, mode and checking its null values.

Text

Description automatically generated

Chart, pie chart

Description automatically generated

As per the above visualization for our Verified column in our dataset 90% are True and 10% are False

1. Analyzing brand column:

our dataset having brand columns which comes under the categorical.

Graphical user interface, histogram

Description automatically generated with medium confidence

The brand D’Addario have the highest count in the given visualization and the difference between 1st D’addario and 2nd brand fender have major difference and other brand don’t have major gap in count they are minor gaps.

1. Analyzing Category column:

our dataset having Category columns which comes under the categorical we are doing univariate analysis.

Table

Description automatically generated with medium confidence

The Category ['Musical Instruments', 'Instrument Accessories', 'Guitar & Bass Accessories', 'Picks & Pick Holders', 'Picks'] having the highest count in the above visualization.

1. Analyzing Review Time column:

In our dataset review time column is in continuous we have taken review time in years and quarter the given dataset have years from 2003 Q4 to 2018 Q3.

A picture containing line chart

Description automatically generated

As u can see line graph is in positive trend and from the year 2010 to 2018 the line is constant.

### 3.2 Multivariate Analysis:

1. Analyzing Brand VS Asin Column Year Wise:

We are doing multivariate analysis for the brand and Asin columns year wise by taking years into filter and taking Asin count. We are plotting bar graph with line graph by using dual axis ans synchronizing them and putting Asin count into color.

Graphical user interface, application

Description automatically generated with medium confidence

As per the above visualization D’Addario brand have the highest Asin count Over all the years.

1. Analyzing Brand VS Category Column Year wise:

We are doing multivariate analysis for the brand and Categorycolumns year wise by taking years into filter and taking Categorycount. We are plotting line graph for the Visualization and keeping category count in color and size.

Graphical user interface, application

Description automatically generated

As per the above visualization D’Addario brand have the highest Category count Over all the years.

1. Analyzing Brand VS Overall column Year wise:

We are doing multivariate analysis for the brand and Overallcolumns year wise by taking years into filter and taking Overallcount. We are plotting bar graph and line graph by applying dual axis and synchronizing axis for the Visualization and keeping category count in color.

Graphical user interface, application

Description automatically generated

As per the above visualization D’Addario brand have the highest Overall count Over all the years.

1. Analyzing Brand VS Price Column Year Wise:

We are doing multivariate analysis for the brand and Pricecolumns year wise by taking years into filter and taking Overallcount. We have taken price into Rows 2-time one price column have sum and another price column have Average. We are plotting bar graph for the Visualization and keeping category count in color.

A picture containing chart

Description automatically generated

As the Average price of the brand is high the number of sum price would be less as it falls under premium pricing and when the average price is low the sum of price is high as it comes under moderate pricing.

1. Analyzing Brand VS Review Column Year wise:

We are doing multivariate analysis for the brand and Reviewcolumns year wise by taking years into filter and taking Reviewcount. We are plotting bar graph and line graph by applying dual axis and synchronizing axis for the Visualization and keeping Reviewcount in color.

Graphical user interface, application

Description automatically generated

As per the above visualization D’Addario brand have the highest Review count Over all the years.

### 3.3 Top 10 brands Analysis:

Graphical user interface

Description automatically generated

The above dashboard visualizes the top 10 brand names which have highest count in review time, Asin Count, Category count, and highest overall count. As per the all the visualization the brand D’Addario is the best brand and it is recommended to the buyers because D’Addario have the highest review, Asin category and overall rating as compared to other brands and the price of D’Addario is moderate which can be easily afford by all class of buyers.

# Chapter 4. Text Preprocessing

In this chapter we are cleaning our text by defining function to clean our review column we are merging 2 columns into one column review removing StopWords removing punctuation from text and cleaned our review text from our dataset.

### 4.1Text Cleaning:

The very first step we do in text Preprocessing is that we clean our text data we remove symbols hyperlinks etc. from our Review column in our dataset we have combine our 2 columns review text and summary together into review as mentioned earlier in Data Preprocessing part.

Text, letter

Description automatically generated

Calling our define function to clean our Text data for our further analysis to make our data more small smooth and for good analysis.

We are applying the define function in review columns in our dataset we have also used lambda function for cleaning our text data.

Text

Description automatically generated

### 4.2 Stopwords:

It is the removal of unwanted text from further processing of text, for example, a, to, can, etc.

We are importing spacy library and from spacy library we are importing STOP\_WORDS

To apply Stop\_words in our review column from the dataset.

Graphical user interface, text

Description automatically generated

While removing the stop\_words we have saved necessary words in our review column we are saving necessary words because if we remove them the whole meaning of the review given by the customer may change.

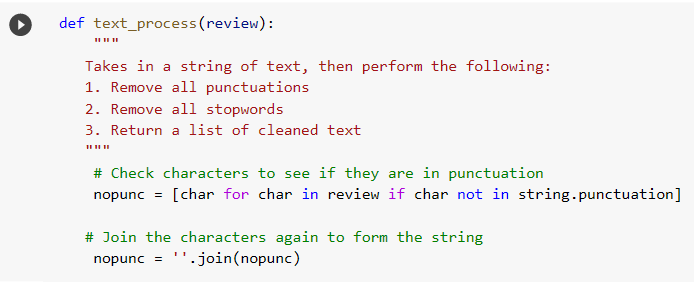
After keeping necessary words and removing all unnecessary words we are using lambda function to join back the sentences.

Graphical user interface, application, Word

Description automatically generated

4.3 Remove all punctuations:

We are defining a function to remove punctuation from our review column in our dataset.



We are making copy of our main dataset as variable name df, so that we have backup of our main dataset.

Table

Description automatically generated

# Chapter 5. Sentiment Analysis and Machine Learning and modeling

In this chapter we will implant different type of classification model to predict the future sentiment of the customer behind their review text also we will balance our data for the same.

### 5.1 Defining function for sentiment Analysis:

We are now doing Sentiment Analysis for our column overall from our dataset overall is basically our customer rating given to the product.

We are defining function for sentiment analysis where we are using if else for overall column where if overall rating is less than 3 it will fall under Negative if overall rating is equal to 3 it is Neutral and if overall rating is greater then 3 it is positive.

Graphical user interface, text

Description automatically generated

We have added a new column named as sentiment where it have values in positive, neutral and negative according to the function we have defined as above.

Graphical user interface, application

Description automatically generated

# Chapter 5.2. Machine learning and Modeling

In this step, we need develop models to predict ratings from reviews with machine learning. Here we will use Classification Algorithms. Classification algorithms are supervised ML algorithms that are used to classify, categorize, or label data points based on what it has observed in the past. There are mainly three processes classification algorithms go through:

Training

Evaluation

Smote

Before developing and evaluate models, first we split data into train and test sets, then extract all the necessary features of training and test data using the preceding feature extractors.

# Splitting Our Data into Training & Split test:

We are importing sklearn.model\_selection import train\_test\_split library.

Graphical user interface, text, application

Description automatically generated

### 5.2.1 Classification models:

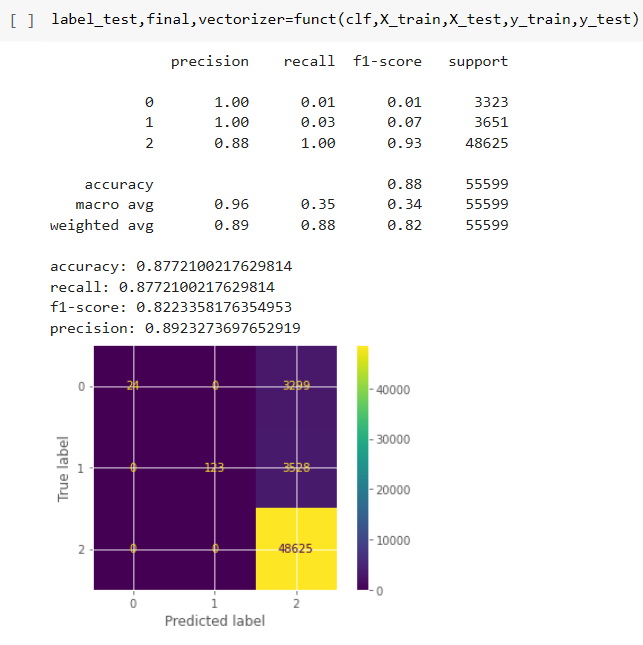
We are defining a function for our classification models where we are finding classification report, accuracy, recall, f1-score, precision and we are plotting confusion Matrix.

Text

Description automatically generated

### 5.2.1 Multinomial Naive Bayes:

*Multinomial Naive Bayes* implements the naive Bayes algorithm for multinomially distributed data and is one of the two classic naive Bayes variants used in text classification (where the data are typically represented as word vector counts). This algorithm is a special case of the popular naïve Bayes algorithm, which is used specifically for prediction and classification tasks where we have more than two classes.



### 5.2.2 Logistic Regression:

*Logistic regression*, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

Chart, treemap chart

Description automatically generated

### 5.2.3 Decision Tree Classification Algorithm:

Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules,** and **each leaf node represents the outcome.** In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed based on features of the given dataset.

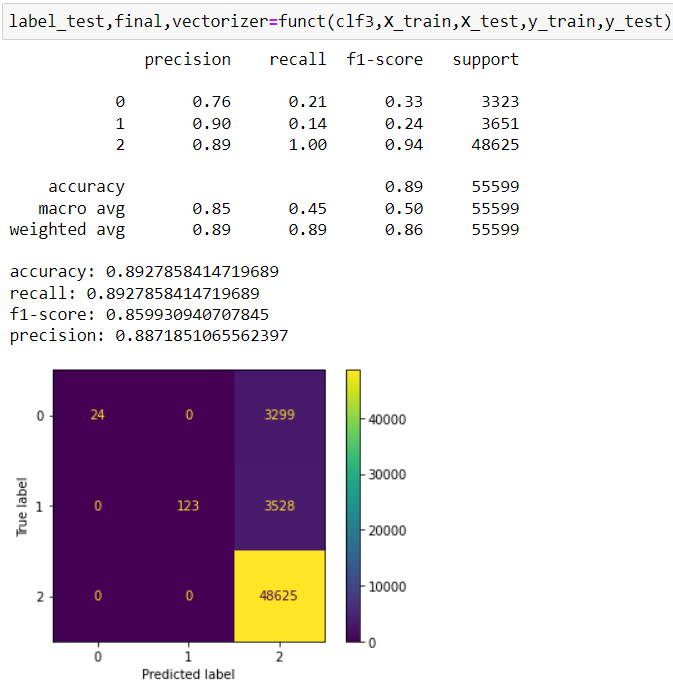
Chart, treemap chart

Description automatically generated

### 5.2.4 Random Forest Classifier:

A *random forest* is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size, but the samples are drawn with replacement if bootstrap=True (default).

In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. In addition, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. As a result of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.



### 5.2.5 AdaBoost Classifier:

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. The most common algorithm used with AdaBoost is decision trees with one level that means with Decision trees with only 1 split. These trees are also called Decision Stumps. What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points which have higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

Chart, treemap chart

Description automatically generated

### 5.2.6 Conclusion:

Navye Byes: -

accuracy: 0.8772100217629814

recall: 0.8772100217629814

f1-score: 0.8223358176354953

precision: 0.8923273697652919

Logistic: -

accuracy: 0.9069767441860465

recall: 0.9069767441860465

f1-score: 0.889477614208318

precision: 0.8893875929222128

Decision Tree: -

accuracy: 0.8569398730192989

recall: 0.8569398730192989

f1-score: 0.8532573661870194

precision: 0.849935858595231

Random Forest: -

accuracy: 0.8927858414719689

recall: 0.8927858414719689

f1-score: 0.859930940707845

precision: 0.8871851065562397

Ada Boast: -

accuracy: 0.8880555405672764

recall: 0.8880555405672764

f1-score: 0.8581457734290868

precision: 0.8588098867769088

Observation: - we get to know that the Model 2: -**Logistic Regression gives the highest values**

accuracy: 0.9069767441860465 recall: 0.9069767441860465 f1-score: 0.889477614208318 precision: 0.8893875929222128

so, Model 2 is best fit model for our Dataset.

**Now, we will try the same for our balanced datasets for which we would Balance our data with the help Of Smote Function which is describe below: -**

### 5.3 Classification models (SMOTE Function):

You connect the SMOTE component to a dataset that's imbalanced. There are many reasons why a dataset might be imbalanced. For example, the category you're targeting might be rare in the population, or the data might be difficult to collect. Typically, you use SMOTE when the class that you want to analyze is underrepresented.

Graphical user interface, text, application

Description automatically generated

### 5.3.1 Multinomial Naive Bayes (SMOTE Function):

Chart, treemap chart

Description automatically generated

### 5.3.2 Logistic Regression (SMOTE Function):

Chart, treemap chart

Description automatically generated

### 5.3.3 Decision Tree Classification Algorithm (SMOTE Function):

Chart, treemap chart

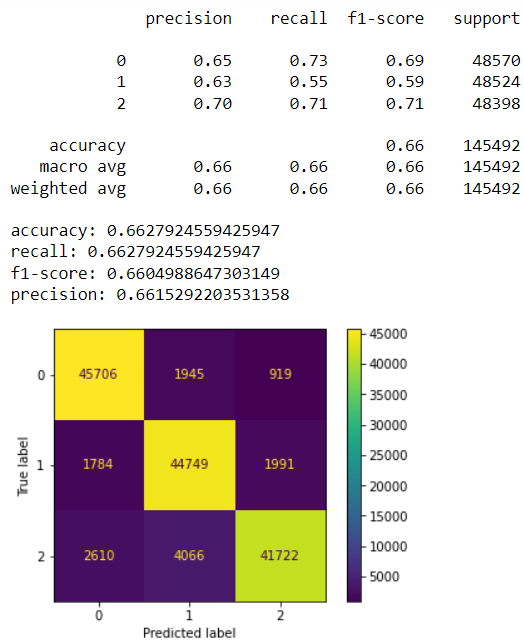
Description automatically generated

### 5.3.4 Random Forest Classifier (SMOTE Function):

Chart, treemap chart

Description automatically generated

### 5.3.5 AdaBoost Classifier (SMOTE Function):



### 5.3.6 Conclusion:

NAIVE BYES:-accuracy: 0.8992865587111318

recall: 0.8992865587111318

f1-score: 0.8988015000127976

precision: 0.9081139702275913

LOGGISTIC:-accuracy: 0.9084829406427845

recall: 0.9084829406427845

f1-score: 0.9082935176131418

precision: 0.9095375176588235

DECISION TREE:-accuracy: 0.87748467269678

recall: 0.87748467269675

f1-score: 0.87735676108388

precision: 0.8778654896630

RANDOM FOREST:-accuracy: 0.9795315206334

recall: 0.9795315206334

f1-score: 0.979449345425

precision: 0.9797251295525

ADA BOOST:-accuracy: 0.6627924559425947

,recall: 0.6627924559425947

,f1-score: 0.6604988647303149

,precision: 0.6615292203531358

# Observation:- we get to know that the Model 2 :-**Logistic Regression gives the highest values** accuracy: 0.9084829406427845 recall:0.9084829406427845 f1-score: 0.9082935176131418 precision:0.9095375176588235

So,Model 2 is best fit model for our BALANCE DataSet.

\*\*FINAL CONCLUSION \*\*

IMBALANCE DATASETS LOGISTIC REGRESSION MODEL VALUES: -

accuracy: 0.9069767441860465 recall: 0.9069767441860465

f1-score: 0.889477614208318 precision: 0.8893875929222128

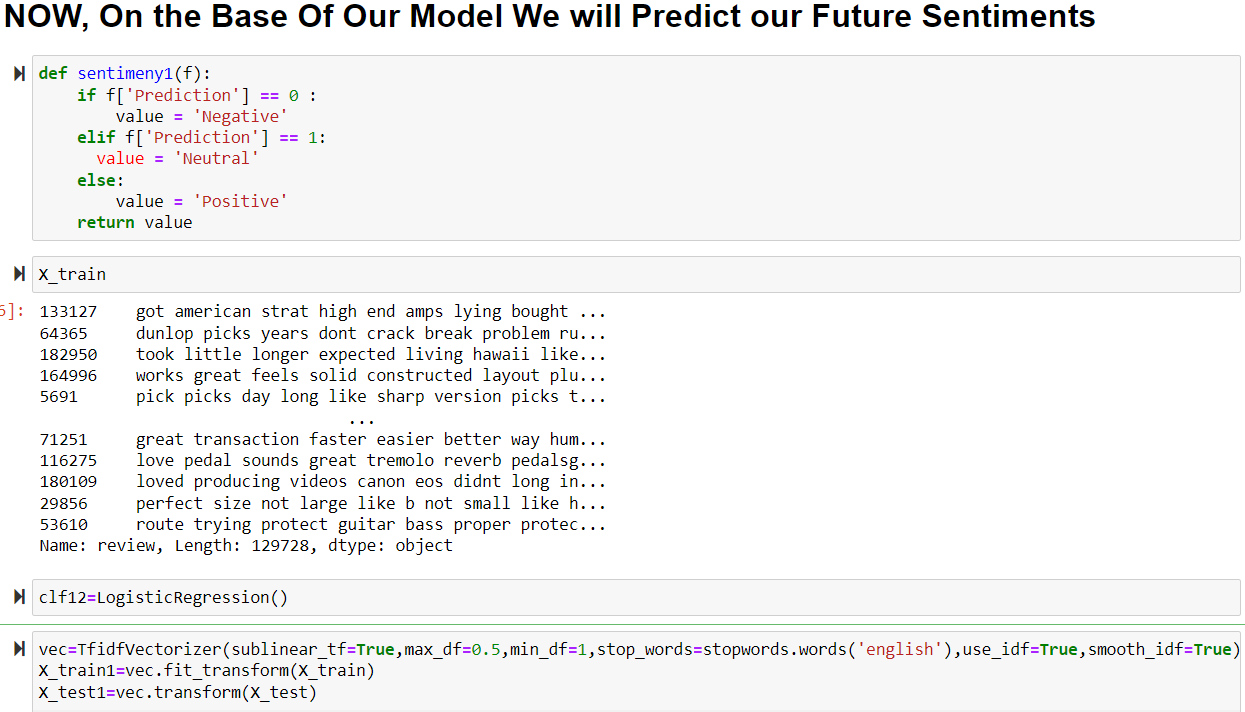
BALANCE DATASETS LOGISTIC REGRESSION MODEL VALUES: -

accuracy: 0.9084829406427845 recall: 0.9084829406427845

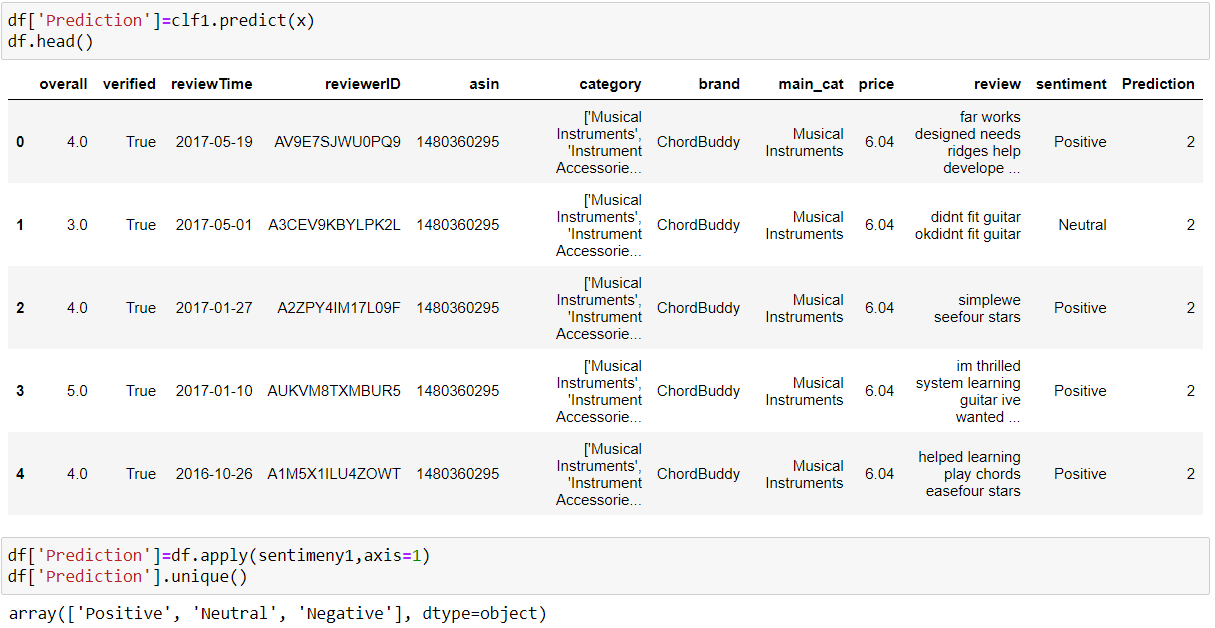
f1-score: 0.9082935176131418 precision: 0.9095375176588235

SO THERE IS NOT THAT MUCH DIFFERENCE BETWEEN THE BALANCE AND INBALANCE DATA LOGISTIC MODEL

### 5.3.7 Doing prediction for our future Sentiment by logistic regression:



We are now adding new column on the basis of sentiments as prediction.



### 5.4 Clustering Models:

Since clustering is an unsupervised classification finding the appropriate number of clusters apriori to categorize the data is a difficult problem to address. The most efficient way to learn about the number of clusters is to learn from the data itself. We address this challenge by estimating the number of clusters using methods like cross-validation and semi-supervised learning. In the project we evaluate a flat clustering algorithm on tweet and web page data sets using Mahout K-means clustering. The algorithm is initialized with random centroid points. We found, by evaluation, that the best possible number of clusters for our dataset.

### 5.4.1 Literature Review:

Hierarchical clustering builds a cluster hierarchy, or in other words, a tree of clusters. Hierarchical clustering outputs is structured and more informative than flat clustering. Hierarchical clustering algorithms are further subdivided into two types (1) agglomerative methods - a bottom-up cluster hierarchy generation by fusing objects into groups and groups into higher clusters. (2) divisive methods - a top-down cluster hierarchy generation by partitioning a single cluster encompassing all objects successively into finer clusters. Agglomerative techniques are more commonly used. Hierarchical clustering does not require knowing the pre-specified number of clusters. However this advantage came with the cost of the algorithm complexity. Hierarchical clustering algorithms have a complexity that is at least quadratic in the number of documents compared to the linear complexity of flat algorithms like k-means.

### 5.4.2 Data Collection:

Table

Description automatically generated

We are taking overall, verified, price and brand\_label columns for our clustering model and keeping brand into index.

### 5.4.3 K-Means Algorithm:

In k-mean, Each object will be represented as vector in space. Initially k points will be chosen by the algorithm randomly and treated as centers, every object closest to each center are clustered. There are several algorithms for the distance measure and the user should choose the required one

Creating Vector Files:

•Unlike Canopy algorithm, the k-means algorithm requires vector files as input, therefore you have to create vector files.

• To generate vector files from sequence file format, Mahout provides the seq2parse utility.

• After creating vectors, proceed with k-means algorithm.

K-means clustering job requires input vector directory, output clusters directory, distance measure, maximum number of iterations to be carried out, and an integer value representing the number of clusters the input data is to be divided into.

Clustering Evaluation:

Our evaluation approach is iterative and we aim to produce clustering results that can be improved over time by optimizing feature vectors and clustering algorithms. To ensure best results our approach is to: research → identify improvements → implement (or find an equivalent open source implementation) → integrate and evaluate. The end goal will be to document our experience with various clustering algorithms and techniques for identifying feature vectors using different set of data which ensures quality clustering with optimal performance.

Chart, line chart

Description automatically generated

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

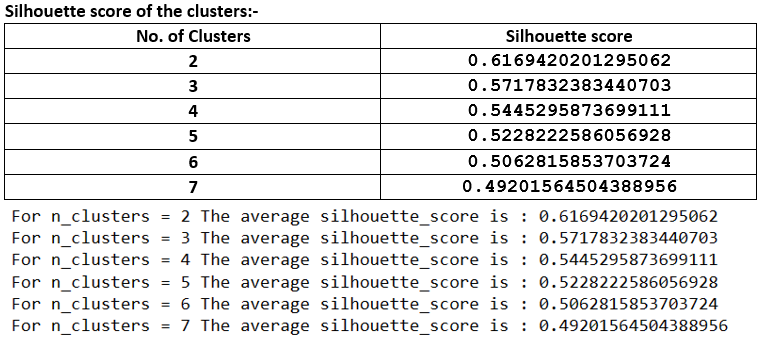
### 5.4.4 Silhouette Scores:

The goal of clustering is to ensure that the documents that are similar are clustered into their own cluster and documents that are dissimilar are clustered into separate clusters. Silhouette scoring is a well known technique to evaluate cluster results when labelled data is not available . Since clustering is an unsupervised learning we have chosen to evaluate our clustering results using Silhouette scores which does not require labelled data.

After obtaining clustering results from the work flow described in previous chapter, we categorize the the documents into their own clusters using the document ID and cluster ID mapping. Further, we compute the dissimilarity (a(i)) of each data point i with respect to all other data points within the same cluster k. In addition, we compute the lowest average dissimilarity (b(i)) from data point i to any other cluster. Silhoutte coefficient for the data point i is computed as follows:

s(i) = b(i) − a(i)/ max{a(i), b(i)}

Silhouette score for the document collection is the mean average of all coefficients computed for each of the data points. To efficiently compute the Silhouette score we use scikit learn python package. A Silhouette score of +1 represents that the documents are clustered with high quality, a score of −1 represents that the documents are clustered with poor quality. Normally, the Silhouette score for text documents will be close to zero due to the sparsity of the documents (99%). For our evaluation we assume that the Silhouette score of anything greater than zero is considered to be decent clustering result.



Chart, surface chart

Description automatically generated

Chart

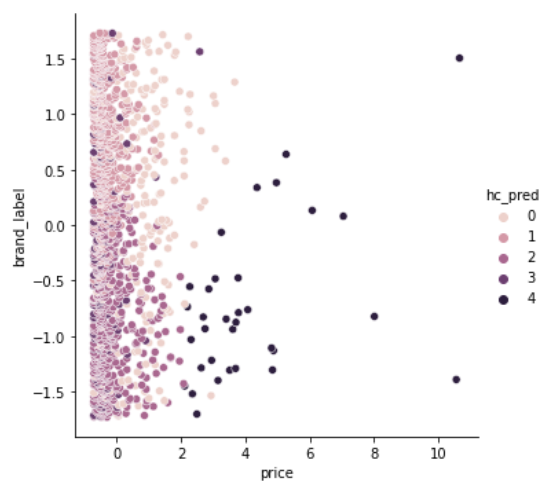
Description automatically generated

Chart, diagram

Description automatically generated

### 5.4.5 Agglomerative Clustering:

Also known as bottom-up approach or hierarchical agglomerative clustering (HAC). A structure that is more informative than the unstructured set of clusters returned by flat clustering. This clustering algorithm does not require us to prespecify the number of clusters. Bottom-up algorithms treat each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data.



Chart, scatter chart

Description automatically generated

### 5.4.6 Dendrogram:

A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data. A dendrogram can be a column graph or a row graph. Some dendograms are circular or have a fluid-shape, but software will usually produce a row or column graph. No matter what the shape, the basic graph comprises of the same parts.

Chart

Description automatically generated with low confidence

# Chapter 6. Time Series Analysis

In this chapter we are covering time series analysis we are applying time series models (ARMA, ARIMA and SARIMA) and we are also doing 2 year future forecasting for our sentiment(Positive, Negative and Neutral Review) for our brand.

Time-series analysis is a method of analyzing a collection of data points over a period of time. Instead of recording data points intermittently or randomly, time series analysts record data points at consistent intervals over a set period of time.

A time series model is a set of data points ordered in time, where time is the independent variable. These models are used to analyze and forecast the future.

* Features: Time series analysis can be used to track features like trend, seasonality, and variability.
* Forecasting: Time series analysis can aid in the prediction of stock prices. It is used if you would like to know if the price will rise or fall and how much it will rise or fall.
* Inferences: You can predict the value and draw inferences from data using Time series analysis.

### 6.1 Import Libraries:

Graphical user interface, text, application

Description automatically generated

### 6.2 Sentiment analysis part according to the rating:

Chart

Description automatically generated

Here we can see that majority of the review part is positive but also we having negative and neutral value

After that we have differentiated in three part

Positive

Negative

Neutral

Steps follow

Spliting data in sentiment

deco

Here taking only brand or overall rating for the forecasting the future prediction.

D’Addirio having large scale of business in amazon

Graphical user interface

Description automatically generated with medium confidence

### 6.3 Positive trend analysis:

Resampling of Data for positive reviews

Chart, bar chart

Description automatically generated

### 6.4 Decomposition chart for the Positive reviews Data:

The decomposition of time series is a statistical task that deconstructs a time series into several components, each representing one of the

A picture containing chart

Description automatically generated

A picture containing line chart

Description automatically generated

In the decomposition graph we can see that there is an increasing trend in total sales after 2014 and also the seasonality is present that means repetition of total sales in between specific time period over the years

### 6.5 Stationarity in time series analysis:

This post is meant to provide a concise but comprehensive overview of the concept of stationarity and of the different types of stationarity defined in academic literature dealing with time series analysis.

MEAN VARIANCE Co-VARIANCE

A picture containing diagram

Description automatically generated

### 6.6 Differencing method: subtracts the current value from the previous value:

DICKEY-FULLER TEST?

The Dickey-Fuller test is a statistical test used to evaluate whether a time series is stationary or not. It evaluates the null hypothesis to determine if a unit root is present. If the equation returns p>0, then the process is not stationary. If p=0, then the process is considered stationary.

Chart

Description automatically generated

### 6.7 Plot ACF PACF:

Just with a quick look over the left graph it’s easy to determine the time series is not stationary. We can see graph that after doing the smoothening of data change in the average value is not changing much and seasonal peaks over the time has been decreased.

A picture containing calendar

Description automatically generated

Here we checking AR (q) and PAC (p) value So we got p value is 0 and q values is 0.

### 6.8Train test split:

Graphical user interface, text

Description automatically generated with medium confidence

ARMA Model for Positive review

AR (4, 0, 8) model giving the lowest AIC score that’s we choose these parameters for further model training.

Plotting Actual and Predicted

Chart, line chart

Description automatically generated

Model is predicting near to the actual values but not the exact values andpredictin values following the pattern of actual values

Text

Description automatically generated

### ****6.9 ARIMA model:****

AR (4, 0, 8) model giving the lowest AIC score that’s we choose these parameters for further model training

Table

Description automatically generated

Here we can see that SARIMAX Results given the value of AR and

Prediction:

Text

Description automatically generated

Here we getting the value of MSE and RMSE

Chart, line chart

Description automatically generated

Model is predicting near to the actual values but not the exact values andpredictin values following the pattern of actual values

Checking MSE and RMSE value:

Mean squared error for ARIMA model is: 0.003300196324844027

Root Mean squared error for ARIMA model is: 0.057447335228398774

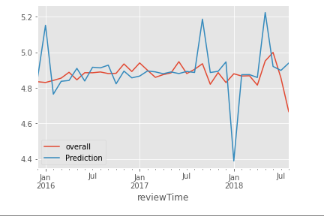
Here I got the RMSE value 0.05744 value in ARIMA Model

### 6.10 SARIMA MODEL:

The SARIMA model is an extension of the ARIMA model. The only difference now is that this model added on a seasonal component. As we saw, ARIMA is good for making a non-stationary time series stationary by adjusting the trend. However, the SARIMA model can adjust a non-stationary time series by removing trend and seasonality.

Predictions:

Here I took p value 4 and q value 8 for the prediction for 2 year

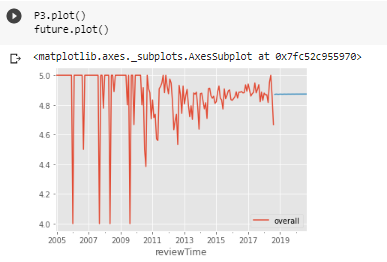


Model is predicting near to the actual values but not the exact values and predicting values following the pattern of actual values.

Table

Description automatically generated

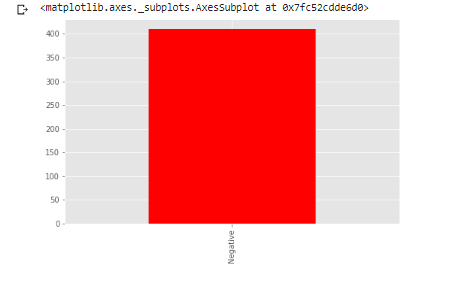
Results:



Here is our future prediction for 2 year. And it will be content for positive review there will be no effect.

### 6.11 Negative trend analysis:

Resampling for Negative reviews Data



We discussed three types of time series patterns: trend, seasonality and cycles. When we decompose a time series into components, we usually combine the trend and cycle into a single **trend-cycle** component (sometimes called the **trend** for simplicity). Thus we think of a time series as comprising three components: a trend-cycle component, a seasonal component, and a remainder component (containing anything else in the time series).

### 6.12 Decomposition graph for Negative reviews:



In the decomposition graph we can see that there is a decreasing trend in total sales after 2011 and also the seasonality is present that means repetition of total sales in between specific time period over the years

Just with a quick look over the left graph it’s easy to determine the time series is not stationary. The average value changes over time and the peaks in the seasonal periods seem to get only larger

We can see over the right graph that after doing the smoothening of data change in the average value is not changing much and seasonal peaks over the time has been decreased and trend component is no longer in data

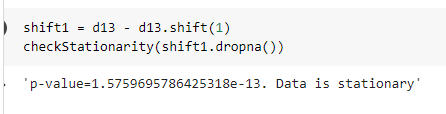
### 6.13 ADF test to check the stationarity of data:

ADF test to check the stationarity of data is not stationary

Graphical user interface, text, application

Description automatically generated

### 6.14 Differencing method: subtracts the current value from the previous value:



Here we got our data is dictionary

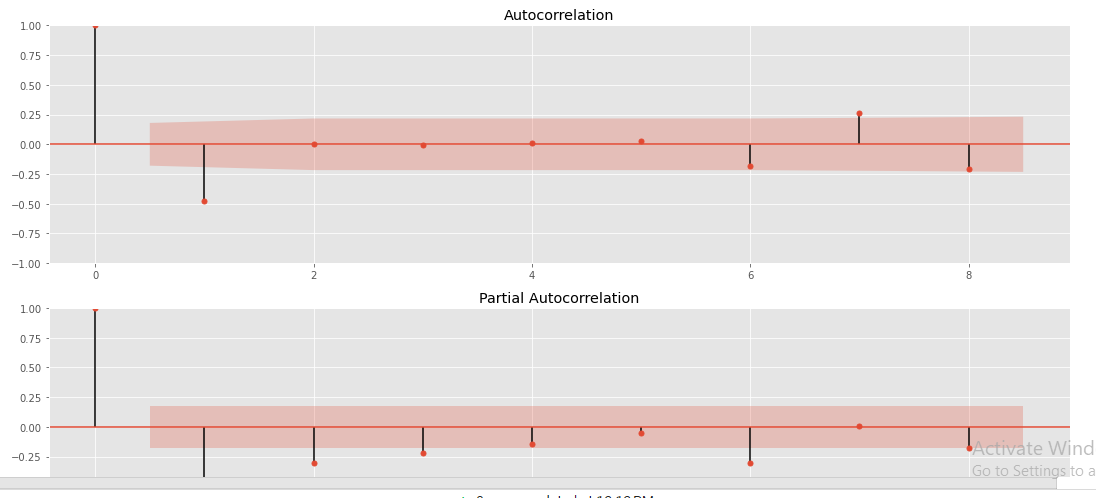
Chart, histogram

Description automatically generated

### 6.15 ACF PACF Plot:

Just with a quick look over the left graph it’s easy to determine the time series is not stationary. The average value changes over time and the peaks in the seasonal periods seem to get only larger

We can see over the right graph that after doing the smoothening of data change in the average value is not changing much and seasonal peaks over the time has been decreased and trend component is no longer in data

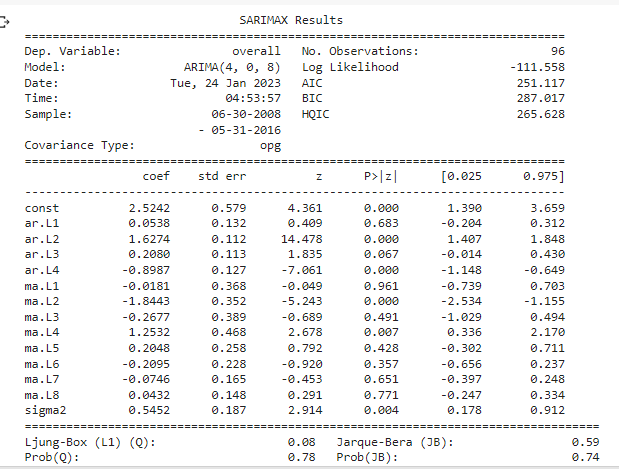


Here we getting the chart for AR or PAC for apply models

### 6.16 ARMA MODEL:

Checking the AIC score with arm model

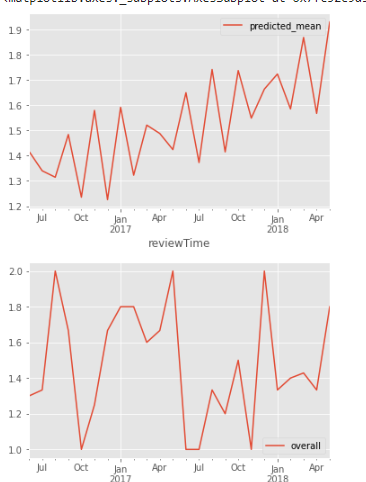
AR (4, 0, 8) model giving the lowest AIC score that’s we choose these parameters for further model training



Mean squared error for ARMA model is: 0.1339696055912096

Root Mean squared error for ARMA model is: 0.366018586401305

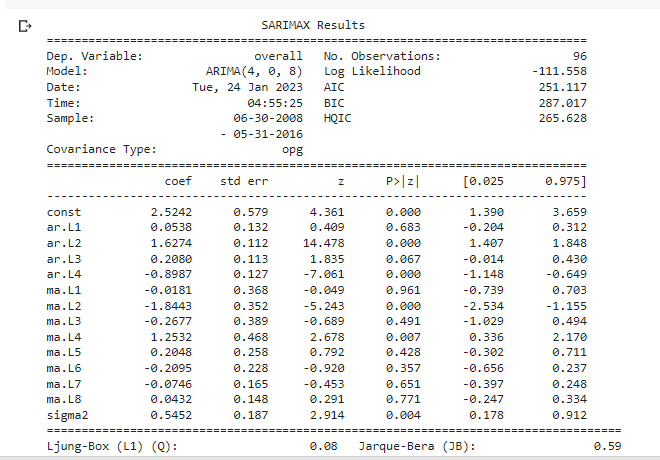
PREDICTION:



Model is predicting near to the actual values but not the exact values andpredictin values following the pattern of actual values

### 6.17 ARIMA model:

AR (4, 0, 8) model giving the lowest AIC score that’s we choose these parameters for further model training

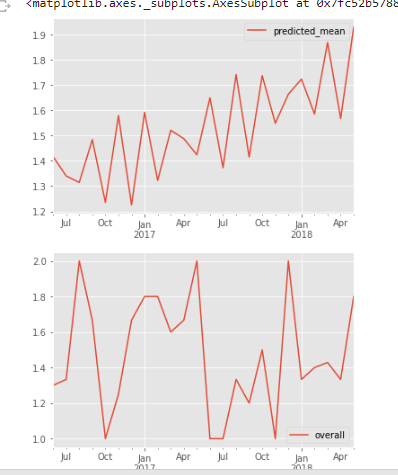


### 6.18 Evaluation:

Mean squared error for ARIMA model is: 0.1339696055912096

Root Mean squared error for ARIMA model is: 0.366018586401305

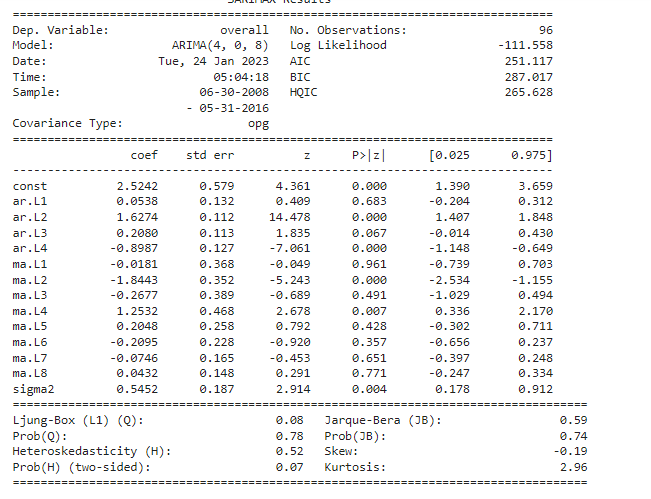
PREDICTION:



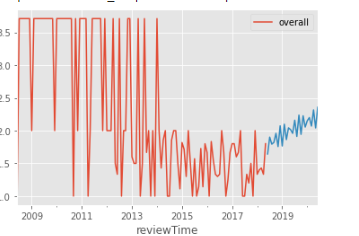
Model is predicting near to the actual values but not the exact values andpredictin values following the pattern of actual values

### 6.19 SARIMA MODEL:

Applying surimi model in negative trend we used P value is 4 and q value 8



PREDICTION:



For the future prediction in negative feedback will be increase amazon will get losses in music instrument product

# Chapter 7. Conclusion

### 7.1 Data Preprocessing Conclusion:

* 1. Merging our 2 dataset (Metadata and 5-core) into one dataset on common columns .
  2. Checking missing values in the dataset and giving missing value treatment.
  3. Checking for Duplicate values and deleting all the duplicate values from our dataset.
  4. After merging our dataset we have 30plus columns we deleted all unnecessary columns from the dataset and in our final dataset we are left with 221848 rows and 10 columns.

### 7.2 EDA (Exploratory Data Analysis) Conclusion:

According to the EDA which we did it analyize:

1. Brand D’Addario is performing good and Brand RedSkyTrader is not performing good 2.
2. Majority of the customer given Overall rating 5
3. The number of positive review is higher then negative and neutral reviews.
4. From the year 2004 till 2010 the number of review increases and from 2010 till 2018 review count was constant.

### 7.3 Text Preprocessing Conclusion:

1. We merged 2 columns (ReviewText and Summary) into one column Review
2. We defined a function to clean text by removing hyperlinks signs symbols emojis and etc.
3. We have define a function to remove all punctuation from the text

### 7.4 Sentiment Analysis & Machine learning and Model Building Conclusion:

According to sentiment Analysis we get to know that

1. Overall Positive Sentiment behind the customer reviews are maximum in count

2. Most loyal customer is the one who has A2NYK9KWFMJV4Y reviewer id

3. Behringer is the Brand who has maximum false in it as it gets the maximum no negative review

### 7.5 Time Series Analysis Conclusion:

1. For time series analysis we are considering reviewtime and Overall columns
2. We are doing time series anlysis on sentiments (Positive, Negative and Neutral).
3. Checking for data stationary applying 3 time series models (ARMA, ARIMA and SARIMA).
4. As per the future forecast our Positive review will be constant.
5. Our Negative review will increase over the period.
6. Our Neutral review will also increase over the period.

### 7.6 Recommendation:

1. To avoid the decrease in sale we can improve the quality of a product by initializing new innovative or custom themes.
2. we can apply some discounts and offers on the product
3. Fast action we would suggest manufacture company to keep hold on the production of this products for a while