Table of Contents

[1. Introduction 2](#_Toc73650802)

[2. Business Understanding: Defining the Problem 3](#_Toc73650803)

[Defining the Target Variable 3](#_Toc73650804)

[3. Data Understanding 3](#_Toc73650805)

[4. Data Preparation 5](#_Toc73650806)

[Exploratory Data Analysis (EDA) 6](#_Toc73650807)

[Text preprocessing 7](#_Toc73650808)

[Visualization 8](#_Toc73650809)

[5. Modeling 12](#_Toc73650810)

[Model Evaluation 14](#_Toc73650811)

[6. Deployment 15](#_Toc73650812)

[7. Conclusion 15](#_Toc73650813)

[Next Steps 16](#_Toc73650814)

[8. References 18](#_Toc73650815)

## Introduction

Sentiment analysis, which is also known as opinion mining, studies people’s sentiments towards certain entities. [Ref 1]

In recent days, Sentiment analysis has gained much attention Sentiment analysis or opinion mining is one of the major tasks of NLP (Natural Language Processing) and Machine Learning.

It is used to determine whether the given product review is positive, negative, or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer need [ Ref 3 ]

Business Success, The Product Owner should understand the customers mindset and their opinion. Sentiment analysis is extremely important to understand the customer mindset because it helps businesses quickly understand the overall opinions of their customers.

By automatically sorting the sentiment behind reviews, social media conversations. once we understand the sentiment, Product owner can react to solve the issue or enhance the features of the product to attract the customer. [ Ref 3 ]

Without predictive model, the only way to be sure to reach all of them is to reach every single customer. Instead of that, we can build accurate predictive model and we can target much more effective way.

Python as a programming language has numerous uses such as Machine Learning, AI, mobile applications, statistical , etc. As we know, AI, machine learning, and data analysis/analytics is where it has amassed most of its popularity.

Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc. Python contains more libraries and packages is continually being employed which helped to solve complex industry issues, data-driven strategies, risk analysis etc.

## Business Understanding: Defining the Problem

The business objective remained the same: The main goal of the project is to build a sentiment analysis predictive model to predict or determine whether the product is satisfying the customer or not. I used different classification machine learning other classification models such as SVM, Logistic Regression, Multinomial Naïve Bayes, Decision Tree and k-NN. Also identify the Ensemble Classifier had the best accuracy.

### Defining the Target Variable

Defining the Target Variable : The Target variable is going to be target class. The Review prediction is going to be Positive, Negative or Neutral and Review Text and Recommended IND variable as a predictor.

## Data Understanding

The format of the dataset for this case study is CSV. The data contains below variables.

Data used in this project are online product reviews collected from Kaggle.com ( Women’s Clothing E-Commerce dataset )

https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews

**Tools and Library to be used in this Project**

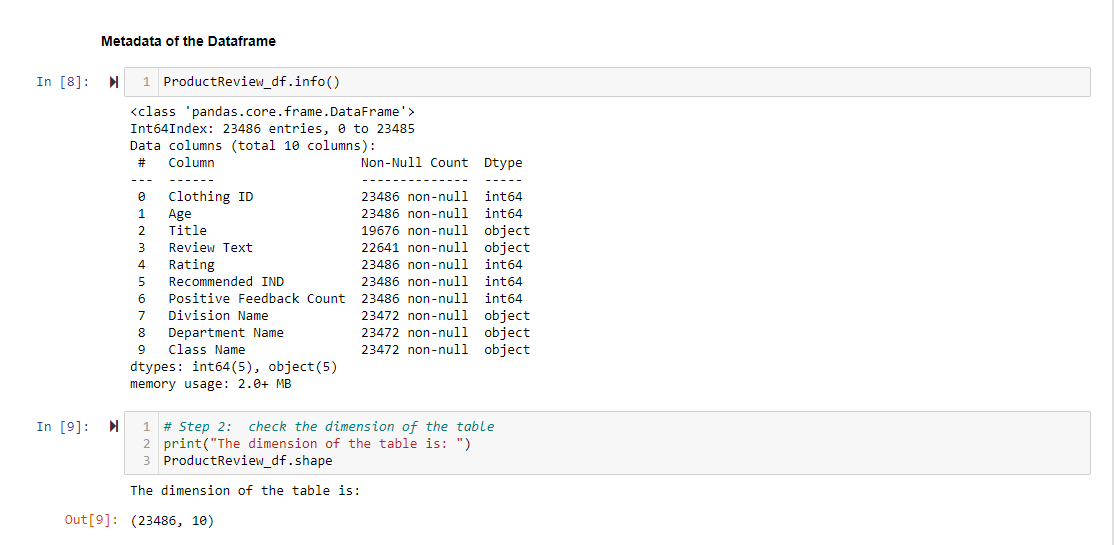
* Pandas library used to load the data from csv to Dataframe.
* Seaborn and matplotlib are used for data visualization.
* Sklearn library used for model and data preparation.



The format of the dataset for this Project is CSV. The data contains below variables.

Pandas library used to load the data from csv to Dataframe.

* Clothing ID: The specific Clothing being reviewed.
* Age: Age of the Reviewer.
* Title: Review Title.
* Review Text: detail content of the Product Review.
* Rating: Score from 1 Worst to 5 Best.
* Recommended IND: Customer recommends the product or not.
* Positive Feedback Count: Number of other customers who found this review positive.
* Division Name: Product high level division.
* Department Name: Product department name.
* Class Name: Product class name.

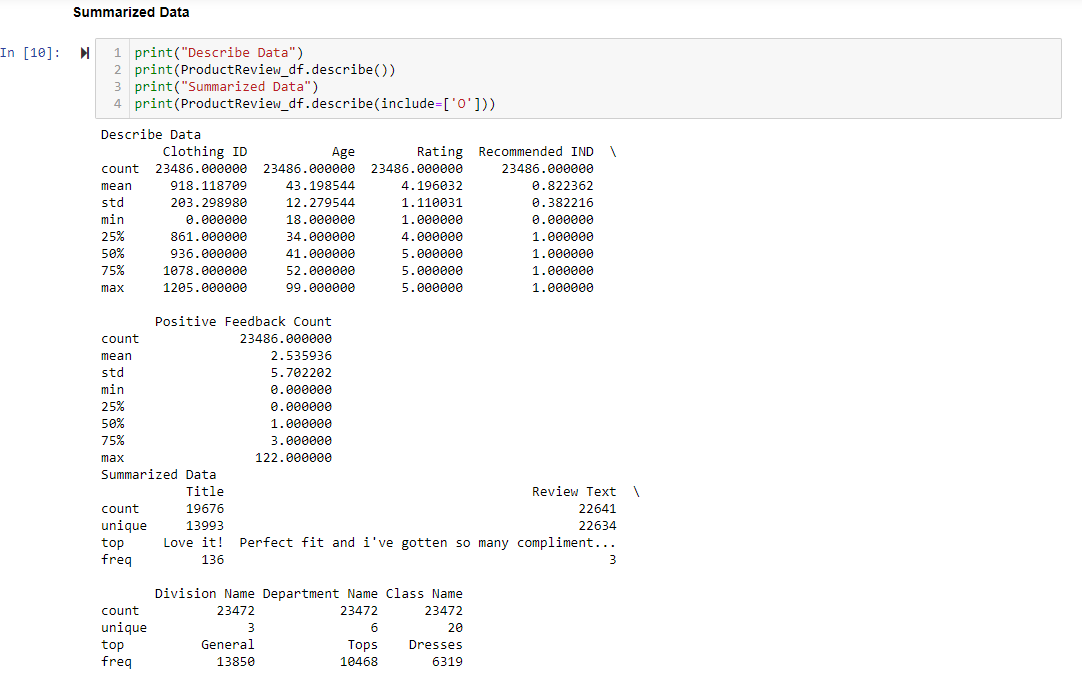


## Data Preparation

Based on the reading, “data scientists usually spend 70% of the project time here, preprocessing and exploring the data”. So, the Data Preparation is important for predictive analytics project. We need to make sure the data is suitable and well prepared for model. While preparing the Data, we need to verify the metadata of the given dataset and need to understand very clearly about data patterns, relationships etc. Review Text is going to be predictor variable and Review Text is the text collected from customers. Recommended will be used as target variable and recommended is collected from customer which defines the product 1 is recommended, 0 is not recommended.

Based on the Rating, Review and Recommended, I created one Sentiment Classifier which provides good or bad Review. As part of Text Processing, I used ‘CountVectorizer’ from the scikit-learn and CountVectorizer develops a vector of all the words in the Review Text. Then applied Logistic Regression ( binary classification ), SVM, Random Forest, Naive Bayes, and other Algorithms to find the high accurate model. Also based on the Review text, we find the most common used words for positive and most common words for Negative.

* Need to describe the dataset describe()
* Need to understand the number of columns and rows shape.
* Need to understand the summary of the data ( info )

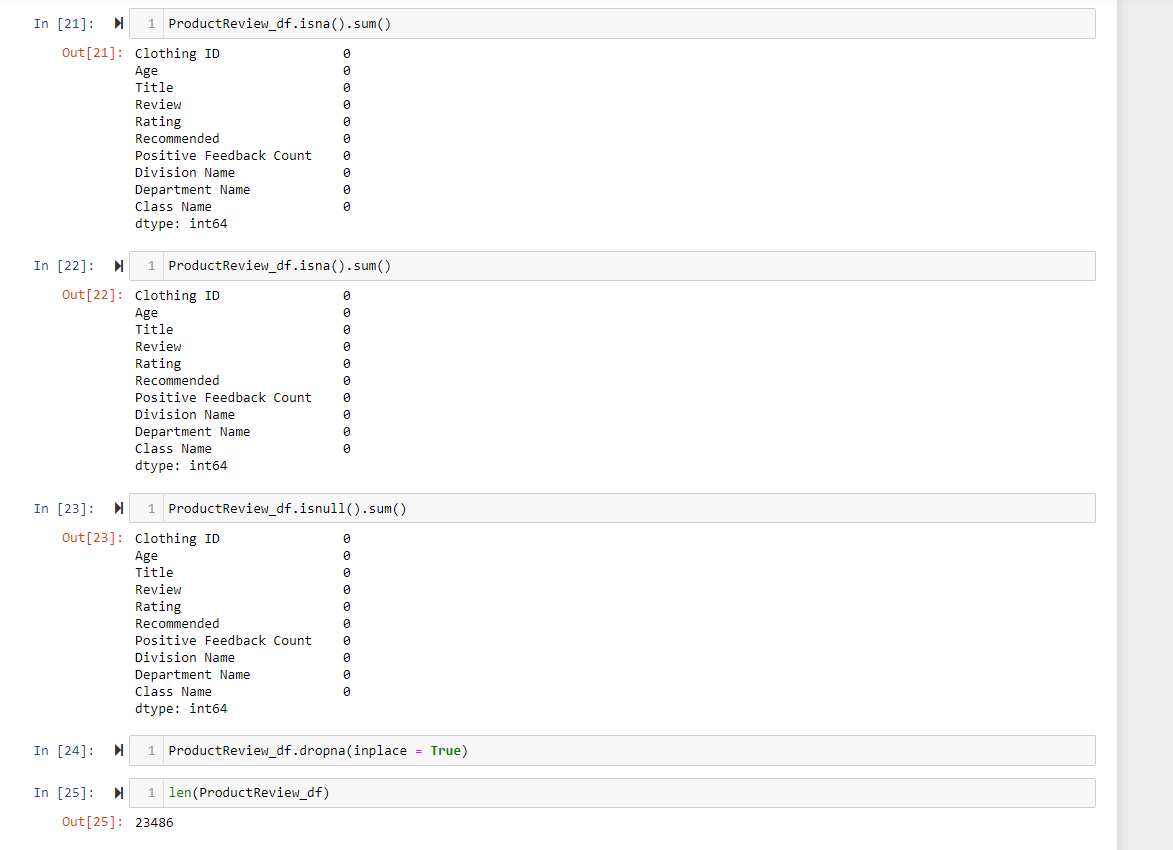


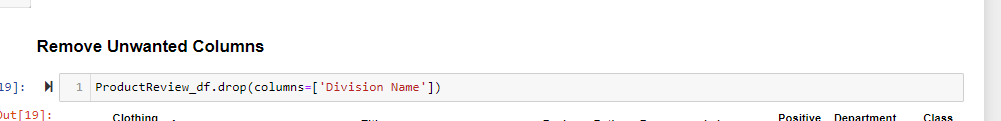
### Exploratory Data Analysis (EDA)

We need to perform EDA and that data needs to be cleaned to gain insights that may be useful in improving the performance of our model. Few of the EDA as follows

* Filling missing information (Null / NA values)
* Encoding of categorical variables
* Dropping of features (or columns)

The above Missing information handled.





### Text preprocessing

The below are steps we need to follow for the text preprocessing through Natural Language Tool Kit (NLTK)

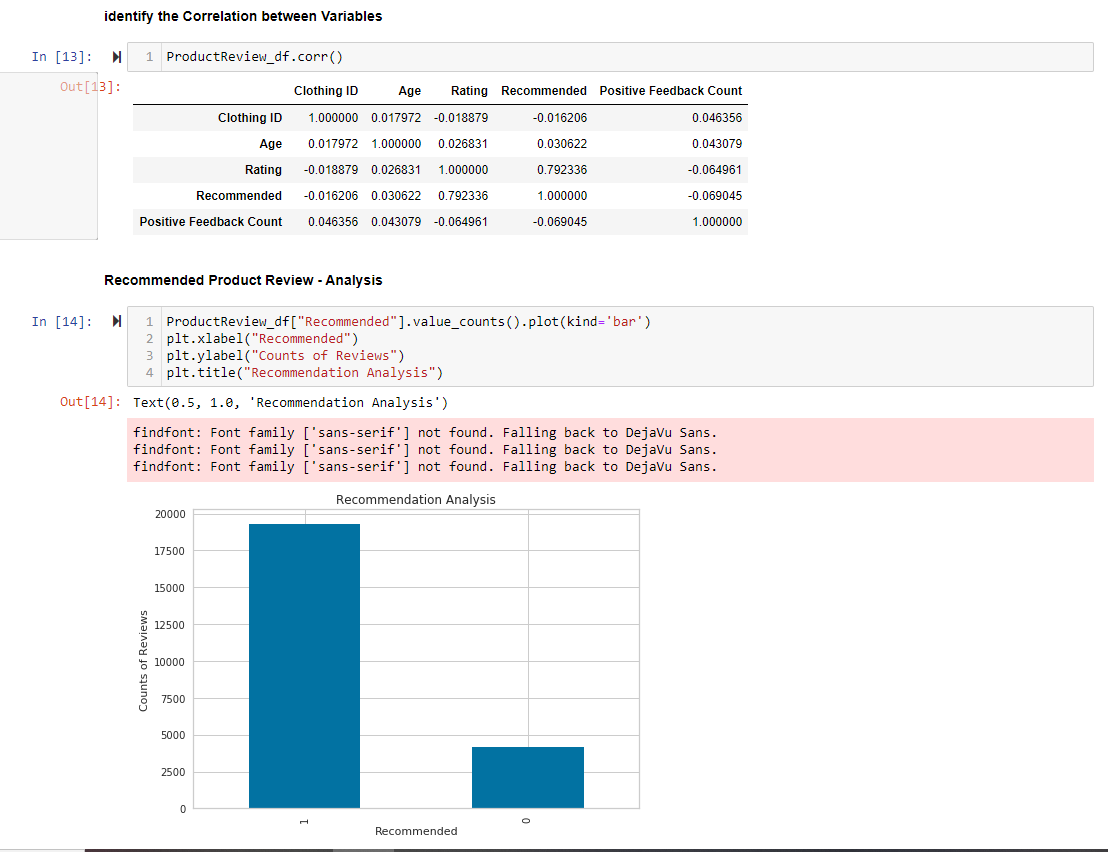
* Lowercase
* Removing Punctuation
* Tokenization
* Stopword Filtering
* Stemming



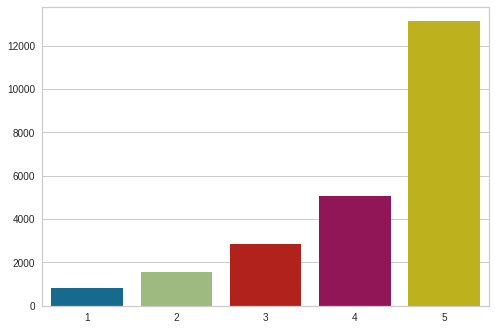
### Visualization

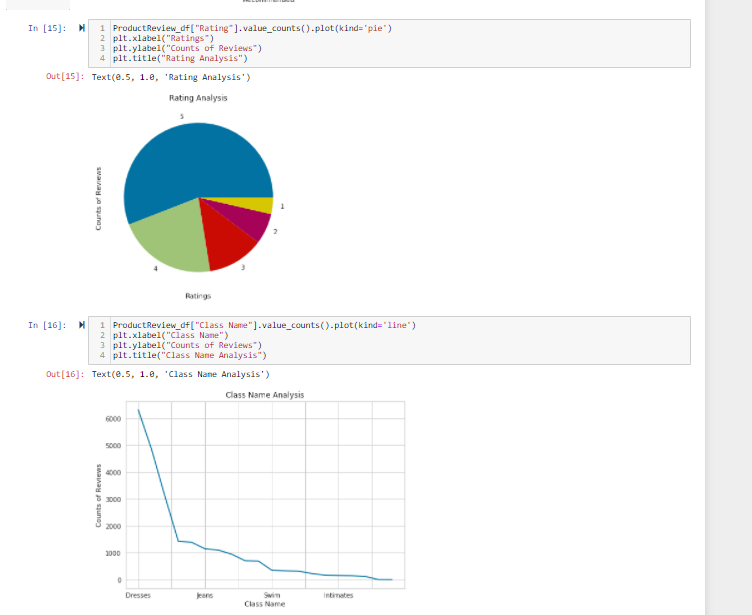
Visualization used to show behavior of data and explains insight of data.

1. Identified the Correlation between variables in the dataset and verified the count of Reviews in terms of Recommended or Not Recommended.

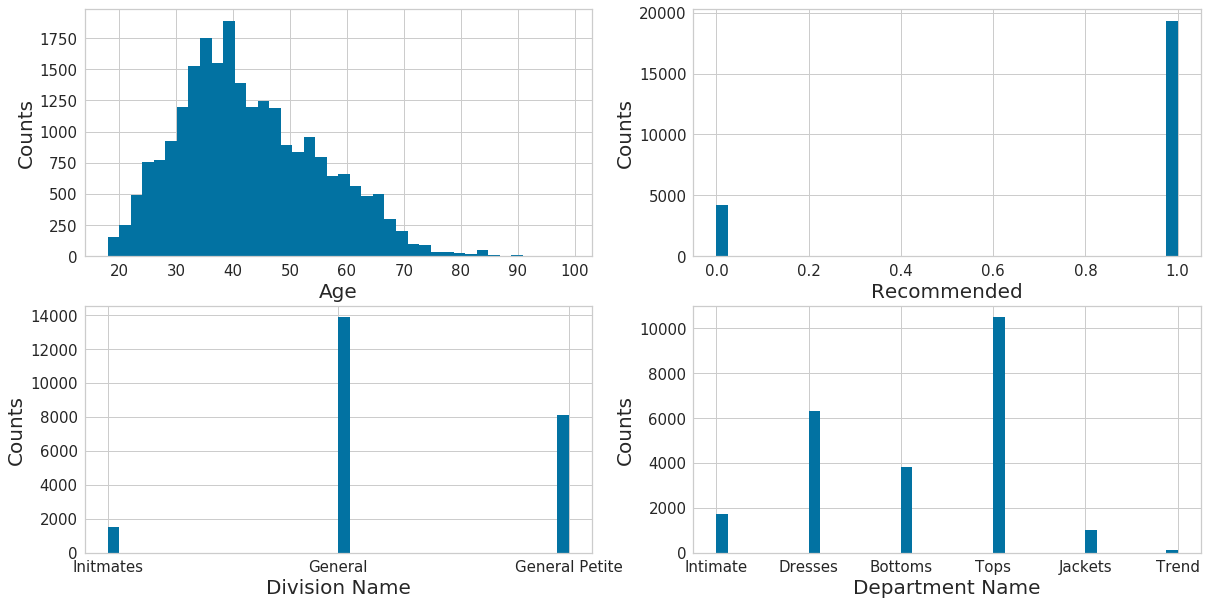


1. Analyzed the various Rating (1-4) and total number reviews in each Class ( Dress, Jeans, Swim, etc )

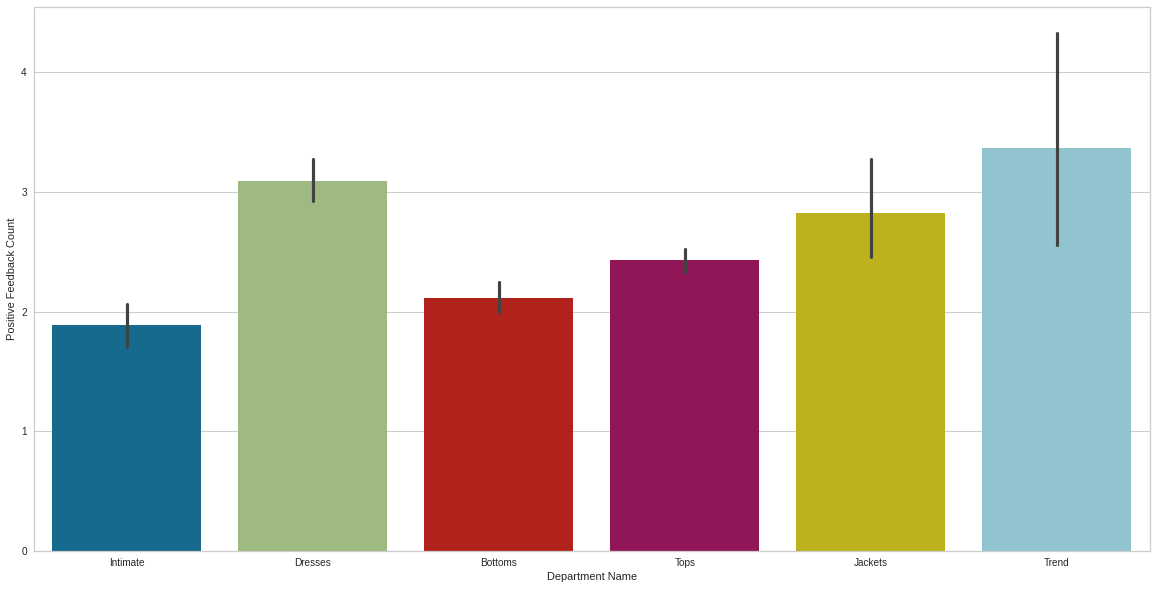




1. Performed Histogram analysis of Age, Recommended, Division and Department



1. Positive Feedback at each Department

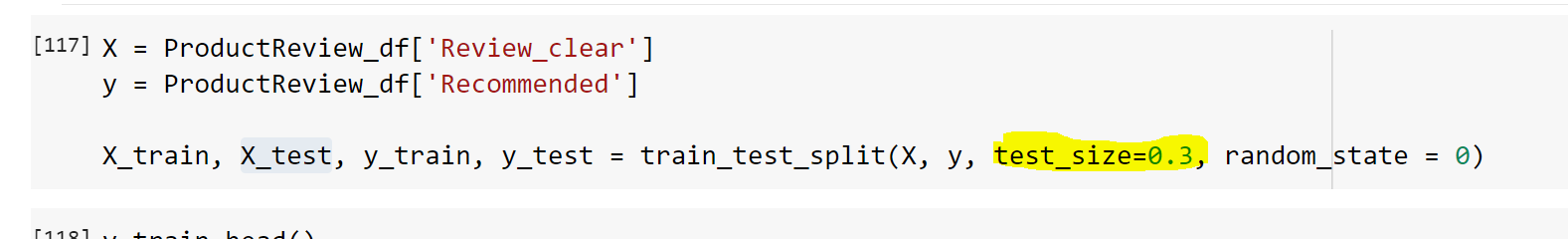


## Modeling

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). Hence in our Case Study we can use Logistic Regression methods to predict the outcome claim as 1 or 0 .

I have used 70%-30% split between training and Testing data.

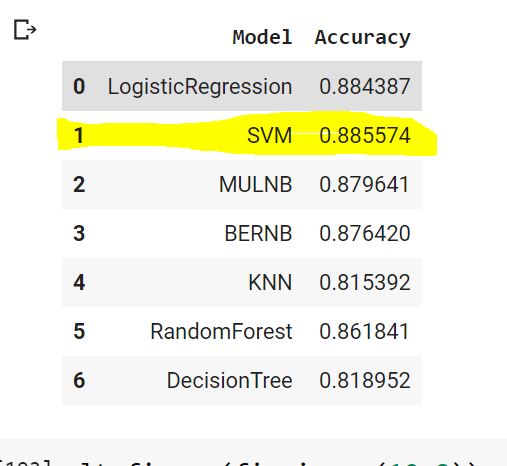
SVM, Decision Tree Regression, Random Forest, and Logistic Regression models used.



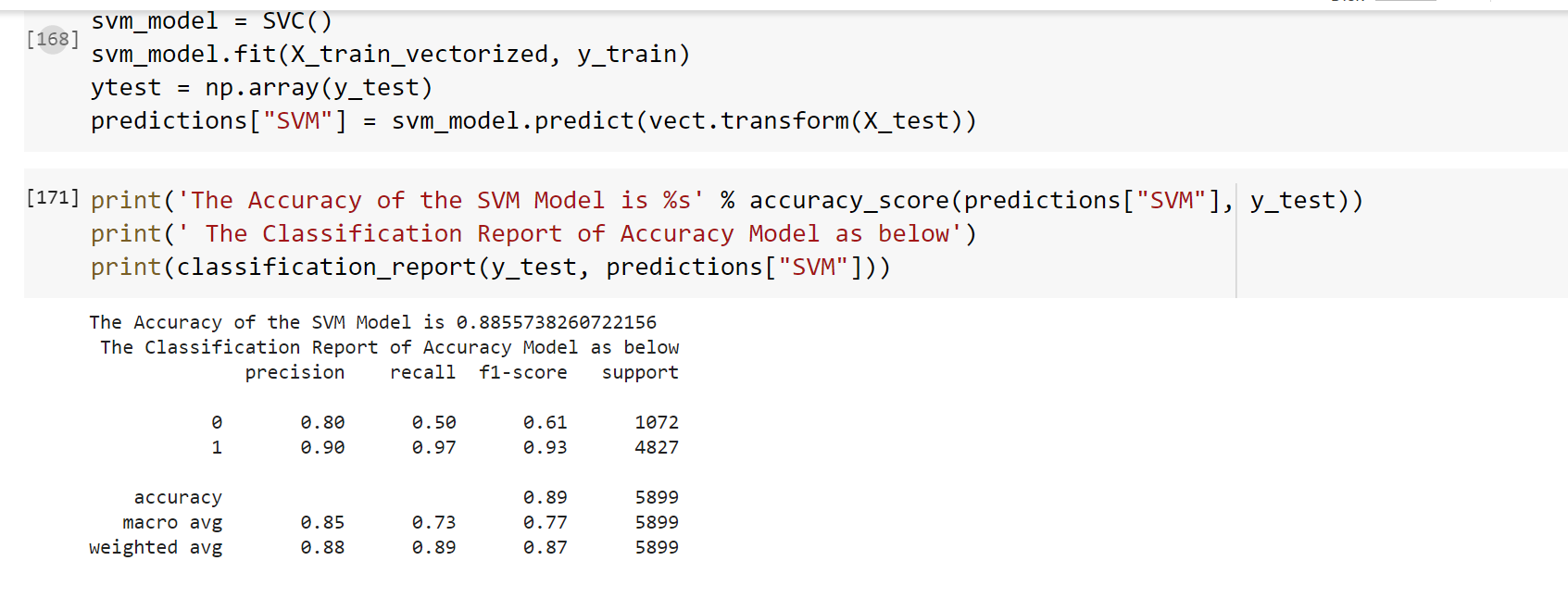
The below are models verified by training the data.

* Logistic Regression
* SVM
* MultinomialNB
* BernoulliNB
* KNN
* RandomForest
* DecisionTree

I have seen SVM provides the accuracy as 89% which is more when compared to other models.

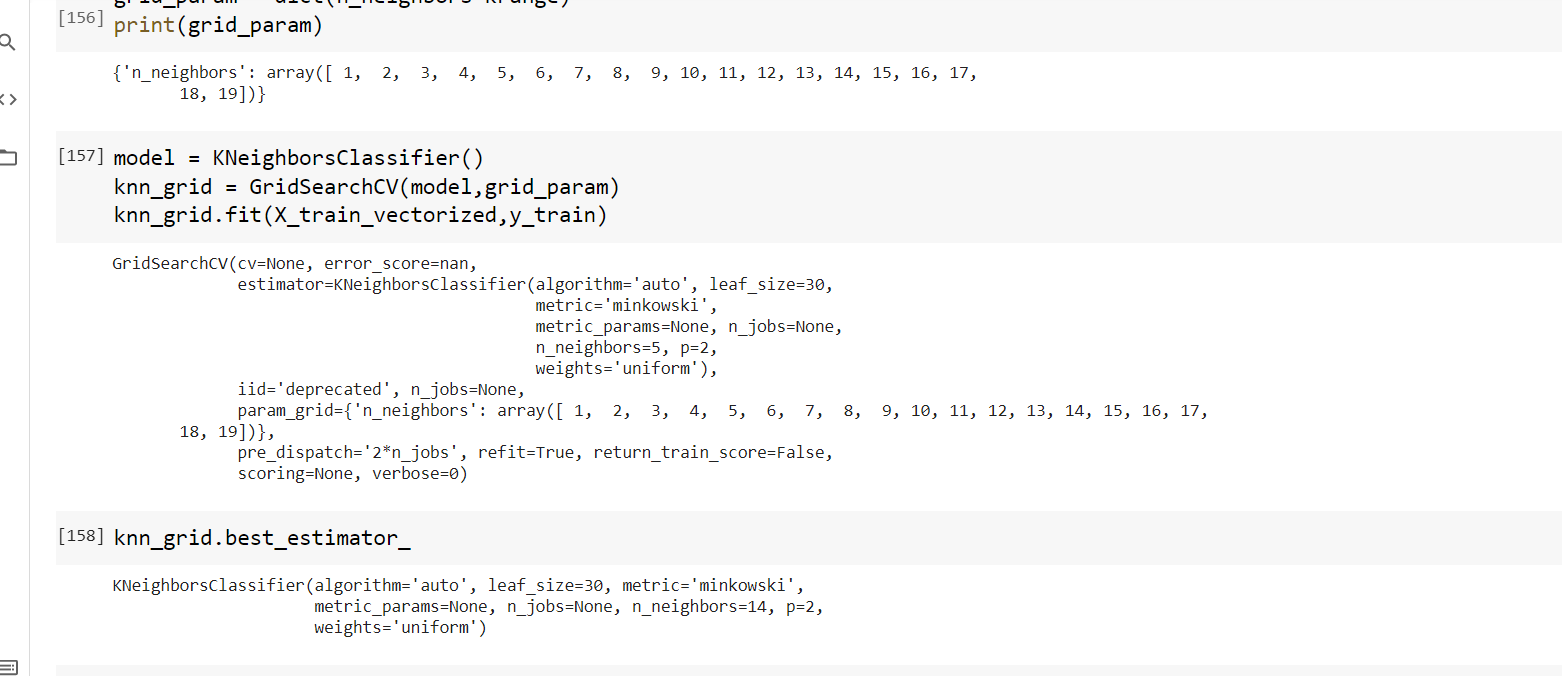


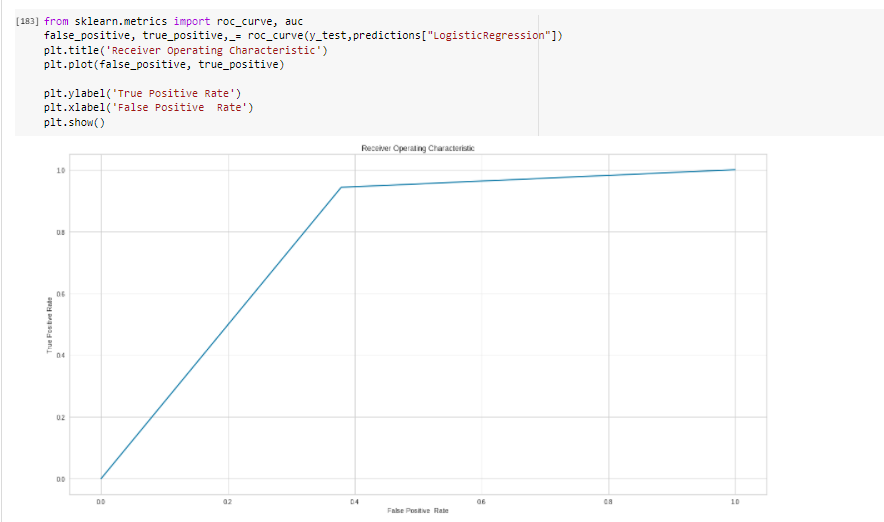
As part for the model development, I verified the Accuracy of the model and generated the Classification Report to view the Precision, recall, F1-Score and Support details.



### Model Evaluation

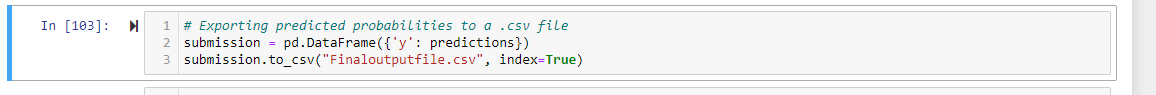
* Performed **Grid Searches** to identify the best hyperparameters for Random Forest Classifier



* Plotted **ROC curve** to understand the AUC value.
* Based on the ROC curve we can see that SVM performs better than Linear, Decision Tree, Logistic Regression to predict the Recommended Text.
  + - * Test Accuracy : 0. .885574

## Deployment

The individual branches decided to what degree the models were used to influence what changes were made within their branches. The Deployment of Machine Learning or Predictive Analytics Solution, so the deployment will be dependent on how it will be utilized.



## Summary and Conclusion

* Loaded the data from .csv files and performed the cleaning and verification.
* Performed Exploratory Data Analysis on dataset to come up with insights such as trends over the target variables, distribution of counts, Histogram comparison. Removed the missing values, removed the unwanted columns.
* 80% reviews were highly rated (5, 4 )
* Dress Department received more reviews and intimates received very less.
* Age between 30 to 50 provided more reviews.
* Positive feedback in Trends received more.
* Handled the NA values and NULL Values.
* Preprocessed the Text data and identified the Positive and Negative comments
* Lowercase
* Removing Punctuation
* Tokenization
* Stopword Filtering
* Stemming
* Split the data set into Training – Testing in the ration of 70% : 30%
* Build the machine learning algorithms – SVM, Random Forest,  Multinomial Naïve Bayes, Bernoulli NB,K-NN, Logistic and DecisionTree by using training datasets.
* Performed the accuracy of every 6 ML Algorithms and identified that SVM is the best fit since the Accuracy of the ML is 89%
* Performed Grid Searches to identify the best hyperparameters for Random Forest Classifier and Plotted ROC curve to understand the AUC value.
* Exported predicted probabilities into the .csv file. The Y value is the prediction of Rating based on the Customer Review.

### Next Steps

What are your recommendations?

This predictive system is built for product review, and we can expand this approach to deal with Product like Amazon, Flipchart by using Product Reviews API data or many social media sites application programming interfaces (API) data ( example twitter / Facebook )

Why should someone in the business care about this solution?

With the increasing market competition, sentiment analysis has become the need of the time. Even established brands like Amazon are actively using this technique to improve the consumer experience.

**Improve Business**

Positive feedback can be used for product promotion and negative feedback can be used to improve product. ( We can fix the issue if any )

**Customer Engagement**

For Customer long term relationship, we need to win the trust of the customers. Hence receiving the feedback and approaching the solution will improve the Customer Engagement.

What are some of the potential challenges or additional opportunities that need to be explored?

1. Faced some Challenges to prepare the data for Modeling input but used Vector Transform to handle these challenges

2. We can extend this approach to other online and real time Rest API sources to make it more useful.

Sentiment analysis is an continues process to understand what customer think about products or brand. Maximum possible product reviews are collected from different sources such as review sites, social media platforms, app stores, and eCommerce stores to gather user sentiment data. [ Ref 7]

## References

1. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die by Eric Siegel

Page number – 166 and 167

1. Applied Predictive Analytics by Dean Abbott ( Principles and Techniques for the Professional Data Analyst )
2. Sentiment analysis using product review data, Journal of Big Data

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0015-2

1. Sentiment Analysis, Wikipedia

https://en.wikipedia.org/wiki/Sentiment\_analysis.

1. Sentiment Analysis: A Definitive Guide, MonkeyLearn

https://monkeylearn.com/sentiment-analysis/

1. https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.html
2. https://www.revuze.it/blog/sentiment-analysis-using-product-review-data