California State University – East Bay

Text Mining

Instructor : PENG XIE

Women's Clothing E-Commerce Reviews

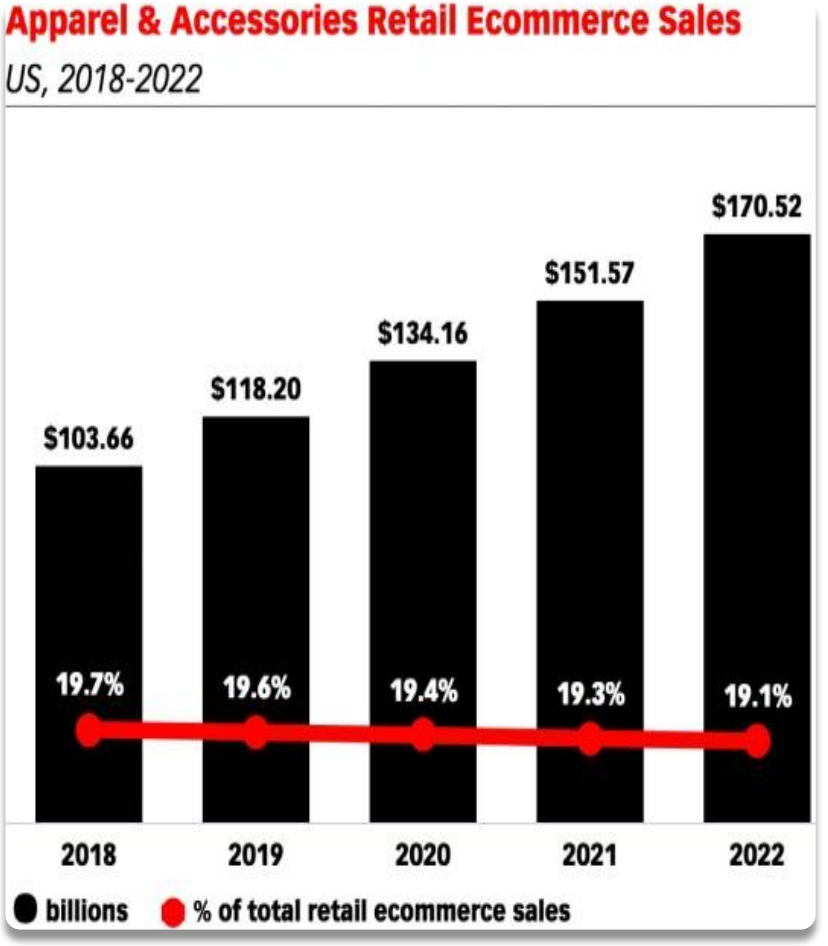
Date : July 20, 2020

Project Background

- Buying an apparel involve different factors for consideration.
- 80% of the women consumers consider reading the reviews and ratings.

Research Questions

- How can the consumers be benefitted ?
- How can the businesses make better decisions ?



Project Usefulness and Its Goal

Reviews and ratings creates the social proof
28% show the lack of trust.
42% cite the missing reviews.

Benefit the **"Sellers"** by increasing the peer competition and maximizing their business outreach

Benefit the **"Consumers"** on basis of the deep sentiment analysis done on their reviews.

Project Flow

1. Text Extraction
2. Text Pre-processing
3. Text Visualization and insights
4. Sentiment Analysis
5. Feature Extraction
6. Classification Models and Interpretation

Data Description

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

Feature	Description
Clothing ID	Integer Categorical variable that refers to the specific piece being reviewed.
Age	Positive Integer variable of the reviewers age.
Title	String variable for the title of the review.
Review Text	String variable for the review body.
Rating	Positive Ordinal Integer variable for the product score granted by the customer from 1 Worst, to 5 Best
Recommended IND	Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
Positive Feedback Count	Positive Integer documenting the number of other customers who found this review positive.
Division Name	Categorical name of the product high level division. Eg General Apetite
Department Name	Categorical name of the product department name. Eg: Jackets,Tops
Class Name	Categorical name of the product class name.Eg: Blouses,pants,Skirts

Records: 23486

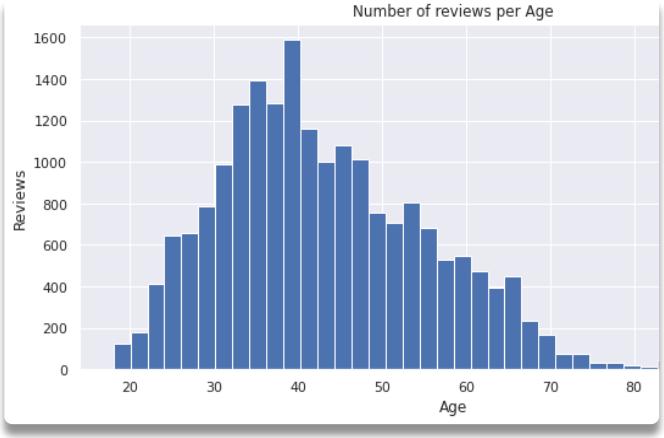
10 Features

Text Extraction and Preprocessing

- Load data from the source
- Missing Rows
- Stop words
- Lemmatization
- Punctuations
- Tokenization pos tagging

Text Visualization

- Word cloud
- Relation between different variables
- Multicollinearity



Feature Extraction

- **Count vectorizer**: A representation of the frequency of words in a review.

Good dress
Not a good dress
did not like

Good	Dress	not	did	like
1	1	0	0	0
1	1	1	0	0
0	0	1	1	1

Recommend
1
0
0

- **TF-IDF vectorizer**: A representation of TFIDF value of each word in the reviews.

Good dress
Not a good dress
did not like

Good	Dress	not	did	like
0.088	0.0880	0	0	0
0.058	0.058	0.058	0	0
0	0	0.0586	0.1590	0.1590

Feature Extraction

- **Sentiment Of Customer Reviews :**

Good dress
Not a good dress
did not like

Polarity	Sentiment
0.7	positive
-0.35	Negative
-0.10	Negative

- **Count vector + TF-IDF vector:**
- **Count vector + TF-IDF + Sentiment Of Customer Reviews**

Data Partition

- **Train Data – 70%**
- **Test Data – 30%**

Model Analysis with one feature

TF-IDF	Decision Tree 80 %	ExtraTreesClassifier (CV=5) Adaboost Classifier(CV=5) GradientBoostingClassifier	81.4% 84.9 % 85.2 %
Count Vectorizer	SVC 82 %	DecisionTreeClassifier RandomForestClassifier ExtraTreesClassifier (CV=5)	81.4 % 84.9 % 85.2 %

Model Analysis with Multiple Features

TF-IDF +Count Vectorizer	Linear SVC 87 %		
TF-IDF +Count Vectorizer + Sentiment	**Logistic Regression** 88.7 %	**K-Fold Cross Validations(100)** 88 %	**Corporate Classifier** 87.7 %

Final Model Analysis and Results

Final model features:

- TF-IDF
- Count Vectorizer of reviews
- Sentiment of Review

Final Model:

- Logistic Regression
- Linear SVC

accuracy 0.8823529411764706

	precision	recall	f1-score	support
0	0.73	0.60	0.66	1109
1	0.91	0.95	0.93	4790
accuracy			0.88	5899
macro avg	0.82	0.77	0.79	5899
weighted avg	0.88	0.88	0.88	5899

Conclusion

Business Insights

We are 90 % confident that our model will predict correctly.

How it benefits customers?

ThankYou

Q&A