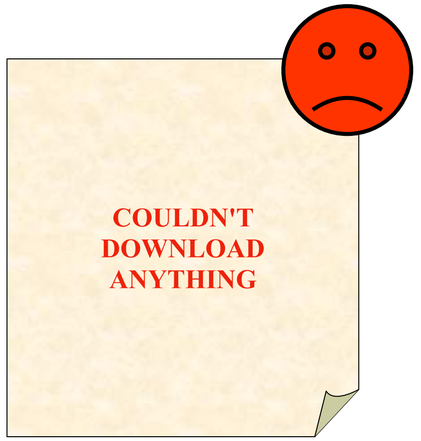
Sentiment Analysis on Amazon Product Reviews

Springboard DS Career Track Capstone 1

Pravati Swain, 15 September 2020

Introduction

- Sentiment Analysis to classify the review text as positive or negative



- Helps manufacturers to understand how consumers feel about their products and services, and hence to improve it.

Aim: Analyze and build an ML model to evaluate the positive and negative sentiment of an Amazon.com product review.

What Companies Care ?

- **E-commerce companies (Amazon.com, eBay.com) and technology companies:** to predict what people think about their product or market trend.
- **Social media companies**: to study the sentiment of social conversations.
- The machine learning algorithm could be used for any kind of business that has an online database of reviews

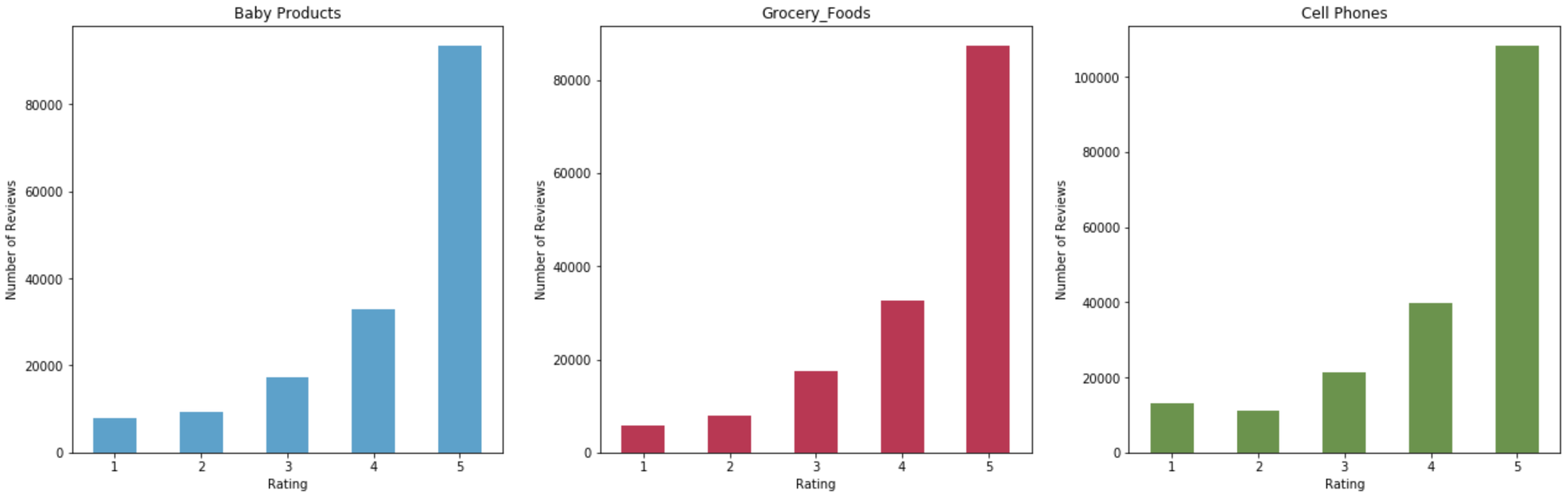
Dataset

- Baby Products, Groceries and Foods , and Cell Phones reviews datasets from Amazon.com.
- All the three datasets contain between 200,000 to 150, 000 reviews.
- Mostly used features: Review Text and Rating
- Rating is based on 5 star scaled
- Binary Classification: 5-star (1) and not 5-star (0)

Data Wrangling

- Checked for missing and duplicate values and dropped the duplicate values
- Removed short reviews of less than 3 words
- Removed non-English reviews
- Added new feature ‘label’: 5-rated reviews (1) and 1-4 stare rated reviews (0)

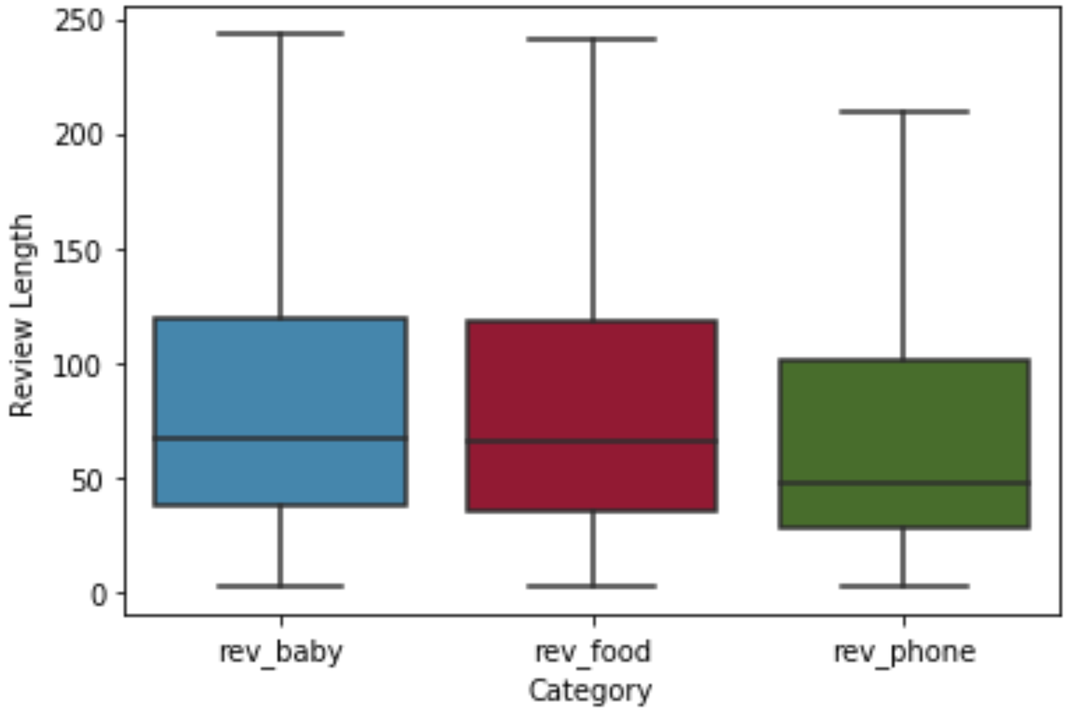
Exploratory Data Analysis



Distribution of rating by product categories

Most of the reviews are five stars in all three categories

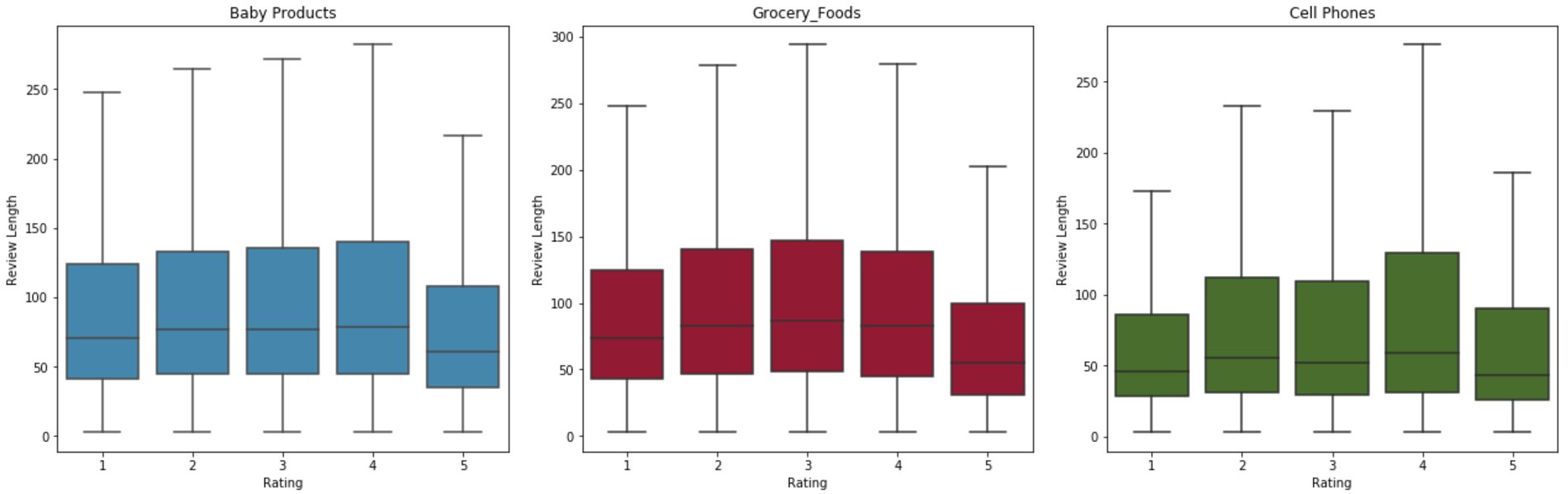
Exploratory Data Analysis



Distribution of review length by product categories

- **The cell phone reviews are shorter in length compared to the other two product categories.**
- **~ 74.8% of cell phone reviews are of length less than equal to 100.**
- **Whereas baby products and grocery food categories have about 56% and 53% of reviews of length less than or equal to 100 respectively.**

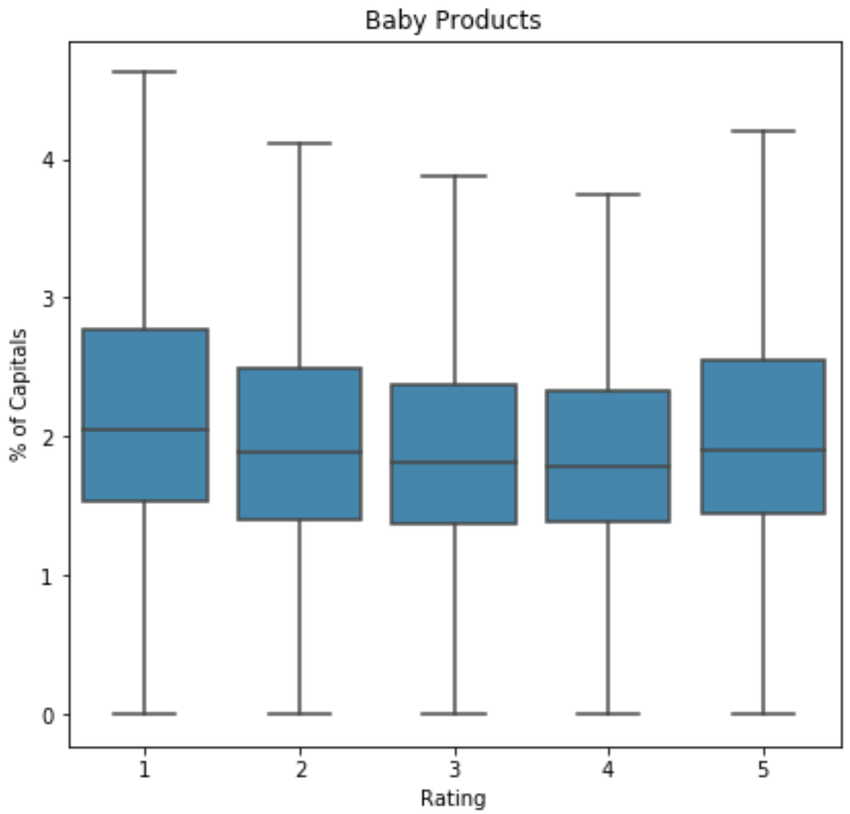
Exploratory Data Analysis



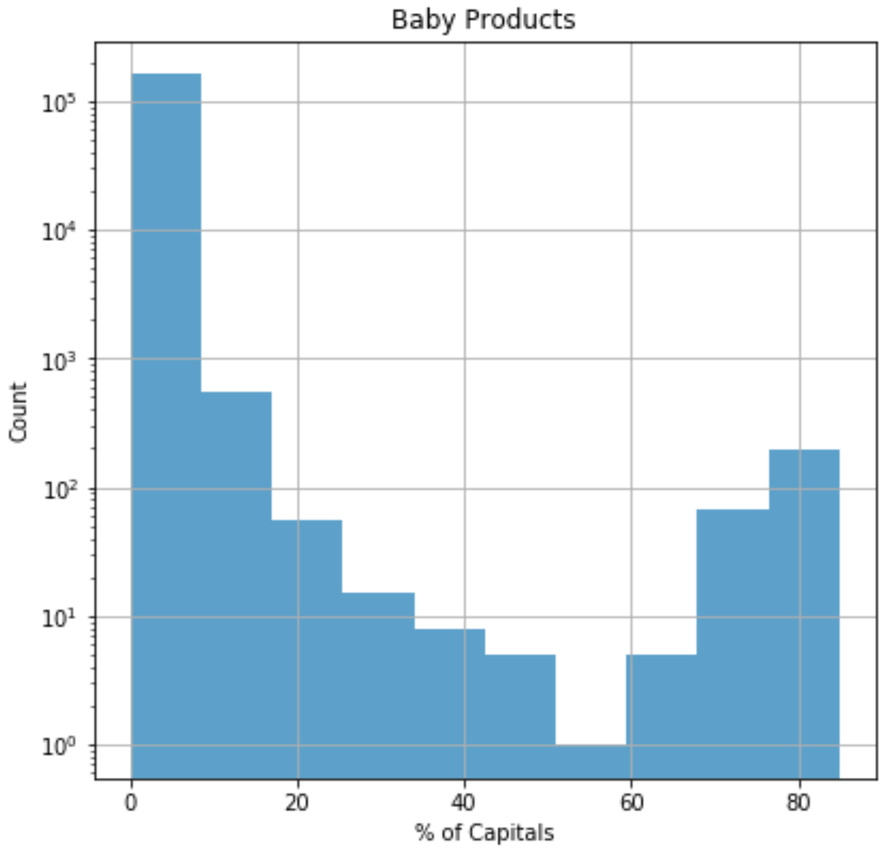
Distribution of Review Length by Rating

- **5-star rated reviews are the shortest length followed by one star rated reviews**
- **2, 3 and 4-star rated reviews are longer and show different trends in different products categories.**

Exploratory Data Analysis



Distribution of % of Capitals by Rating

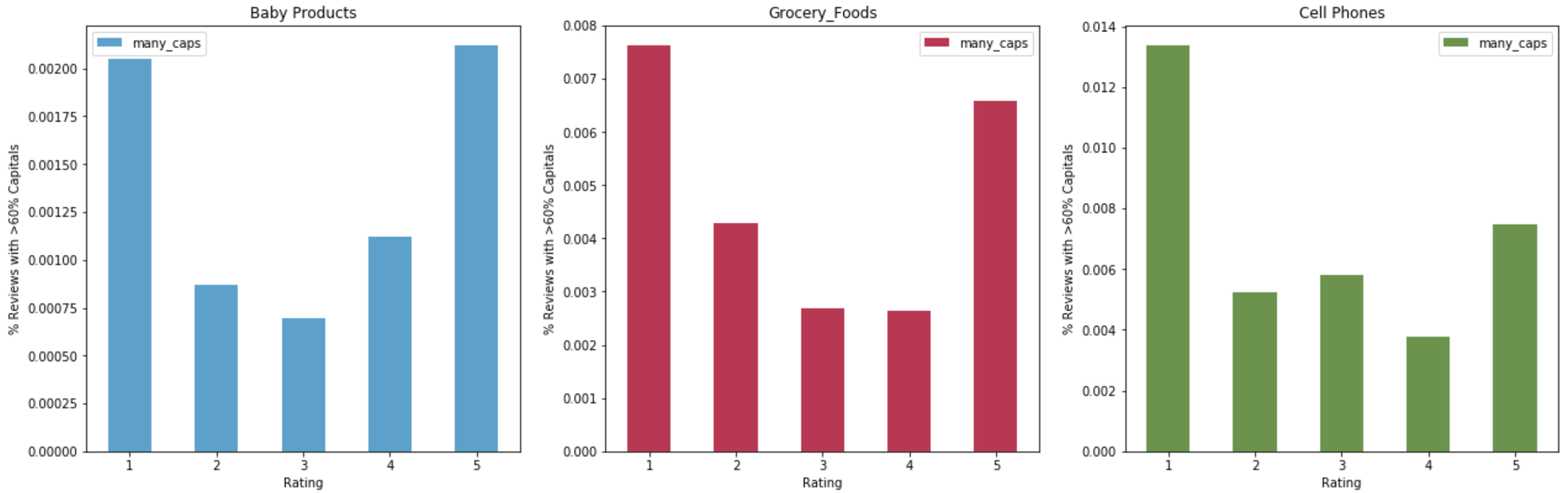


Distribution of % of Capitals

- **1-star and 5-star ratings have more uppercase letters**
- **3- star rated reviews have least upper case letters.**

The % of capitals show bimodal distribution in the three product categories

Exploratory Data Analysis



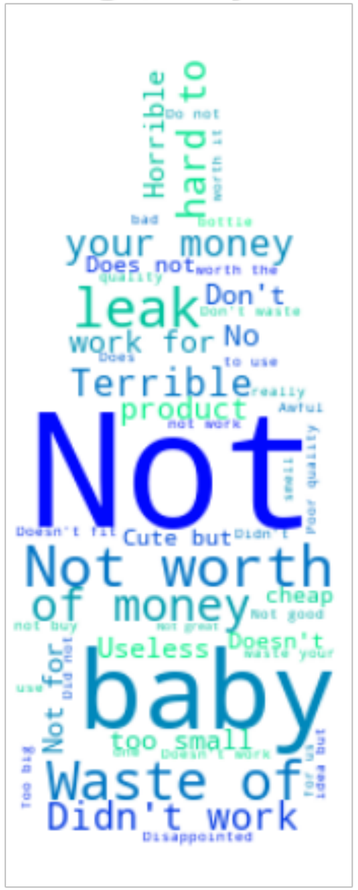
Distribution of % of capitals (> 60%) by rating

5-star reviews are more likely to contain reviews with >= 60% capital letters across product categories

Exploratory Data Analysis

High Rating (Baby Products)

Low Rating (Baby Products)



- **5-stars ratings as high and 1- 4 stars as low ratings**
- **The font size of a word indicates its frequency and importance in the review.**
- **The words used more often in the reviews are greater in font size and darker in color.**

World Clouds for high and low rated reviews

Inferential statistical data analysis

1. Comparison of review length between 5-star and not_5 star rating

Null Hypothesis: The review length for 5-star and other rated reviews are the same.

Alternate Hypothesis: The review length for 5-star and other rated reviews are different.

Baby Products: t = 41.398	**p = 0.000**
Grocery_Foods: t = 58.596	p = 0.000
Cell Phones: t = 23.899	p = 0.000

- **p_value is less than 0.05 we reject the null hypothesis**
- **We conclude that the word counts for 5-star rated reviews are significantly shorter than all 1–4-star (not5) rated reviews.**

Inferential statistical data analysis

2. Comparison of %caps between 5-star and not_5 star rating

***Null Hypothesis*: The % of capitals for 5-star rated reviews are the same as other rated reviews.**
***Alternate Hypothesis*: The % of capitals for 5-star rated reviews are different from the other rated reviews**

Baby Products: t = 12.134	**p = 0.000**
Grocery_Foods: t = 13.586	p = 0.000
Cell Phones: t = 12.302	p = 0.000

- **p_value is less than 0.05 we reject the null hypothesis**
- **We conclude that the % of capitals for 5-star rated reviews are higher than the other rated reviews.**

Machine Learning Highlights

Preprocessing → **Vectorization** → **Model Tuning**

- Removing URL
- Keeping only alphabets
- Lowercase all text
- Tokenization
- Removing stopwords
- Lemmatization

- Vectorizer selection
- Compared CountVectorizer and TfidfVectorizer with a Multinomial Naive Bayes Model
- Select the vectorizer with highest ROC-AUC score

- Fitted and tuned 3 classifiers: Logistic Regression, Multinomial Naive Bayes and Random Forest Trees.
- Tune with GridSearchCV
- Compare ROC-AUC scores

Vectorization

Baby_products Dataset		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.742	min_df = 1, alpha=1
TfidfVectorizer	0.832	min_df = 1, alpha =1
CountVec w/ GridSearch	0.828	min_dfÂ = 0.001, alpha =5
TfidfVec w/ GridSearch	0.842	min_df =0.001, alpha =1
Grocery_foods Dataset		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.758	min_df = 1, alpha=1
TfidfVectorizer	0.839	min_df = 1, alpha=1
CountVec w/ GridSearch	0.819	min_df = 50, alpha=0.1
TfidfVec w/ GridSearch	0.834	min_df = 50, alpha=5
Cell_phones Dataset		
Vectorizer	**ROC-AUC**	**Best Parameters**
CountVectorizer	0.747	min_df = 1, alpha=1
TfidfVectorizer	0.833	min_df = 1, alpha=1
CountVec w/ GridSearch	0.802	min_df = 50, alpha=0.001
TfidfVec w/ GridSearch	0.824	min_df = 50, alpha=10

Comparison of vectorizers with a Multinomial Naive Bayes Model

TfidfVectorizer worked best and used for all the classifier

Model Comparison

Comparison of three machine learning models fitted with TfidfVectorizer for three datasets

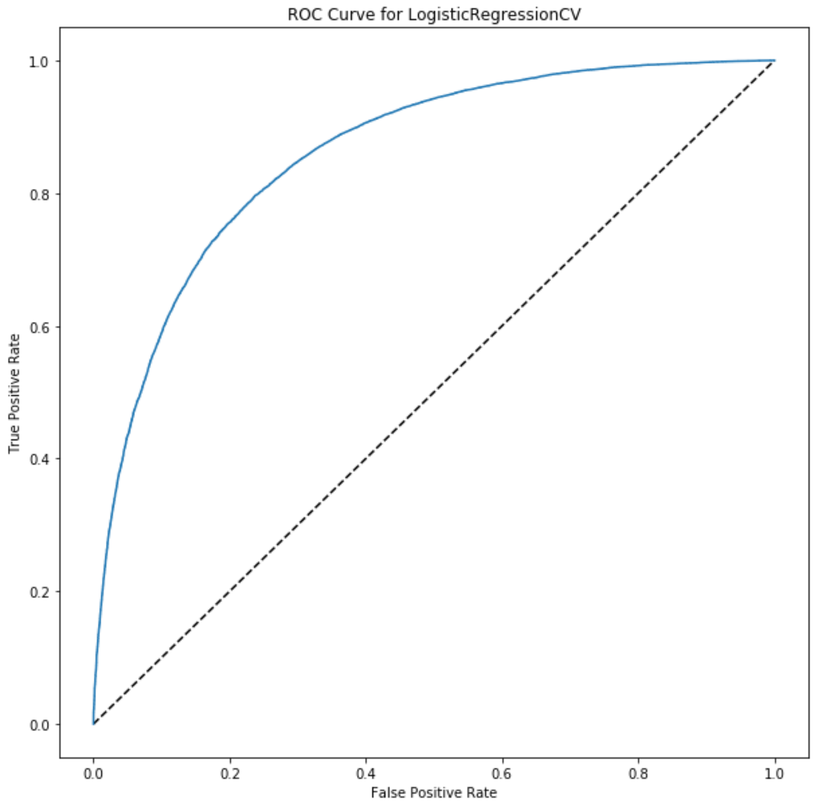
Baby_products Dataset		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.842	alpha =1, fit_prior = True
LogisticRegressionCV	0.859	C= 0.85, max_iter=1000
RandomForestClassifier	0.835	max_depth= 100, max_feature= sqr, n_estimators= 500
Grocery_foods Dataset		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.834	alpha=5, fit_prior=True
LogisticRegressionCV	0.856	C= 0.85, l1_ratio= 0
RandomForestClassifier	0.825	max_depth= None, max_features= sqrt, n_estimators=100
Cell_phones Dataset		
Classifier	**ROC-AUC**	**Best Parameters**
MultinomialNB	0.824	alpha =10, fit_prior = True
LogisticRegressionCV	0.846	C= 0.85, l1_ratio= 0
RandomForestClassifier	0.819	max_depth= None, max_features= auto, n_estimator= 100

Best Classifier:

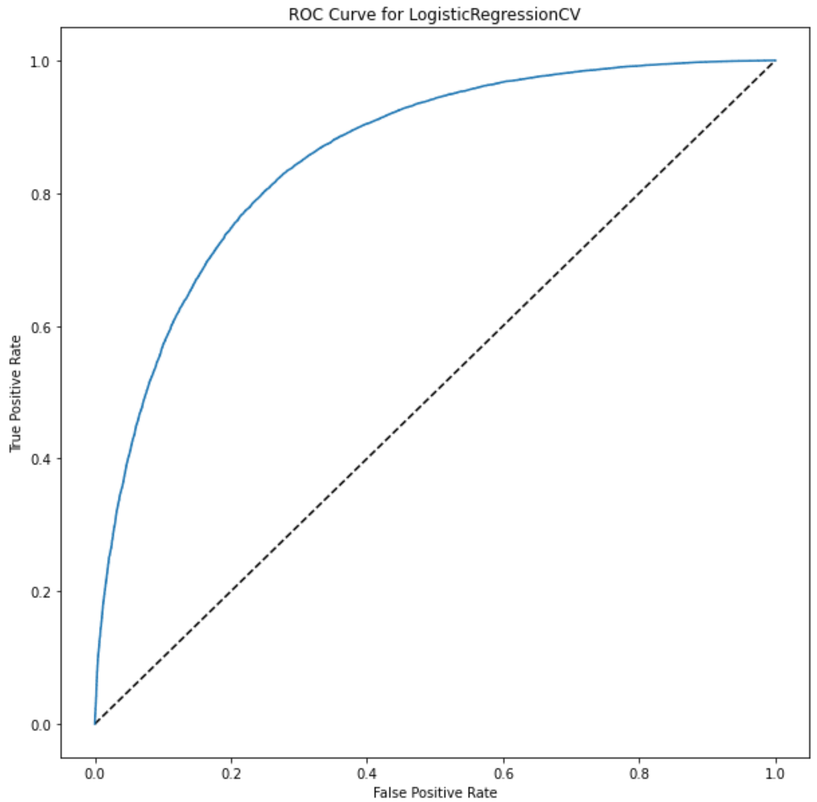
Logistic Regression for all the 3 product categories

Best Classifier: Logistic Regression
ROC Curve

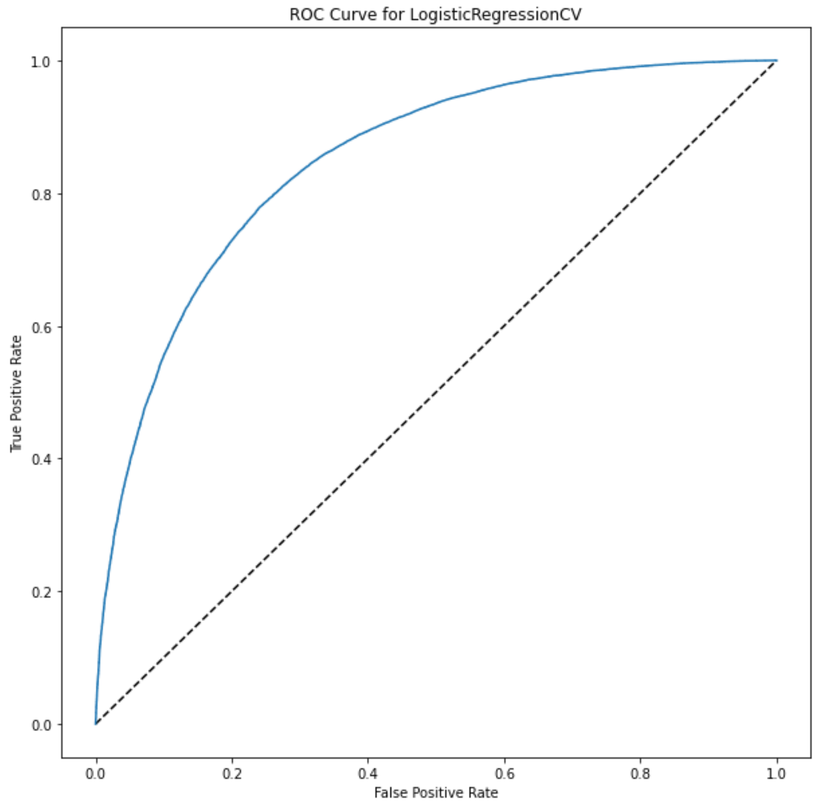
Baby_products Dataset



Grocery_foods Dataset



Cell_phones Dataset



Improve Classification: Thresholding

Goal: Maximize Precision

- High Rated Reviews are common
- Minimize number of false positive
- Classify more ‘low’ reviews correctly
- Sellers and customers both benefit from accurate information in review classification

Improve Classification: Thresholding

Business Case 1: Product Market Comparisons

Baby_products Dataset

Threshold	Confusion matrix	balanced_accuray_score
0.5 (Default Threshold)	[[14110Â 6052] Â [4269 23768]] ---------- [['TN' 'FP'] Â ['FN' 'TP']]	0.774
0.58 (Best Threshold)	[[15683Â 4479] Â [6172 21865]]	0.779

Grocery_foods Dataset

Threshold	Confusion matrix	balanced_accuray_score
0.5 (Default Threshold)	[[13452Â 5776] Â [4012 22077]]	0.773
0.56 (Best Threshold)	[[14488Â 4740] Â [5231 20858]]	0.776

Cell_phones Dataset

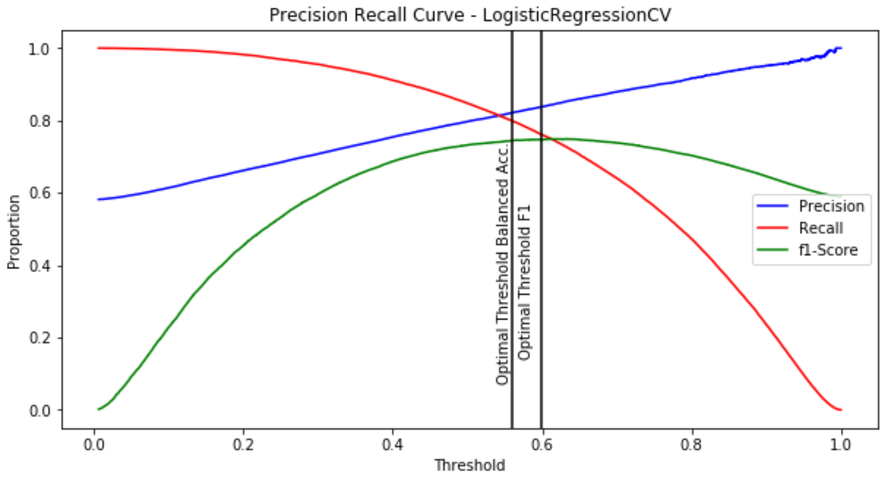
Threshold	Confusion matrix	balanced_accuray_score
0.5 (Default Threshold)	[[18222Â 7318] Â [5863 26760]]	0.767
0.54 (Best Threshold)	[[19207Â 6333] Â [7010 25613]]	0.769

Improve Classification: Thresholding

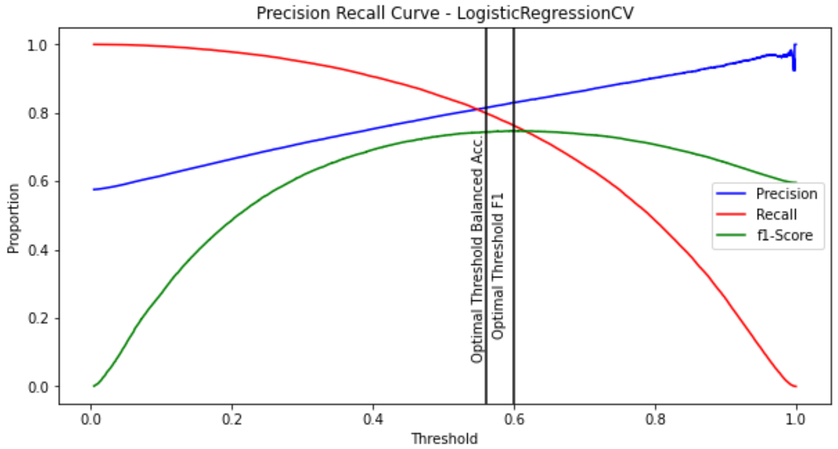
Business Case 2: Customer Satisfaction

(Target the negative or low rated reviews)

Baby_products Dataset



Grocery_foods Dataset



Cell_phones Dataset