



Sentiment Analysis on Amazon Fine Food Reviews

Guided By: Prof. Feinstein Zachary

Team Members:

Aakash Rami

Pratik Mulye

Shubham Heda

Tanisha Mandvikar

Agenda

- ❖ Company Overview
- ❖ Business Improvement Opportunities
- ❖ Data analytics Solution
- ❖ Data Understanding
- ❖ Data Preparation
- ❖ Sentiment understanding
- ❖ Usefulness of score among customers
- ❖ ML Modeling
- ❖ Future deployment

Company Overview

- ◆ Amazon is a multinational company which dominates in e-commerce market and with out 50% of all online sales in U.S. in 2018 by an online site.
- ◆ Amazon e-commerce also launched Amazon Fine foods services with idea of delivering food products to their customers.

Business Improvement Opportunities

- ◆ Consumer insights and food services suggestion based on preferences
- ◆ Inventory management
- ◆ Reducing food wastage

Data Analytics Solution

- ◆ Identifying customer comments and predicting helpfulness of the reviews according to score
- ◆ Classifying consumer comments into positive or negative sentiment
- ◆ Build word cloud to analyze customer reviews based on product score
- ◆ Build predictor models based on customer comments

Data Understanding

- ◆ The Dataset is acquired from SNAP-Stanford University which consists of half million for reviews for fine foods from Amazon and other products.
- ◆ Ten Columns of Dataset are:

ID	UserId	ProductId	ProfileName	HelpfulnessNumerator
Score	Time	Summary	Reviews	HelpfulnessDenominator

Data Statistics

Data Statistics	
Number of reviewers	568,454
Number of users	256,059
Number of products	74,258
Users with > 50 reviews	260
Median number of words per review	56
Time Span	Oct 1999 – Oct 2012

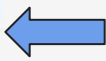
Data Preparation

	Id	ProductId	UserId	ProfileName	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time	Summary	Text
0	1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	1303862400	Good Quality Dog Food	I have bought several of the Vitality canned d...
1	2	B00813GRG4	A1D87F6ZCVE5NK	dli pa	0	0	1	1346976000	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...
2	3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia Corres"	1	1	4	1219017600	"Delight" says it all	This is a confection that has been around a fe...
3	4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	1307923200	Cough Medicine	If you are looking for the secret ingredient i...
4	5	B006K2ZZ7K	A1UQRSCLEF8GW1T	Michael D. Bigham "M. Wassir"	0	0	5	1350777600	Great taffy	Great taffy at a great price. There was a wid...

Data Preparation

◆ Pearson Correlation

	Id	HelpfulnessNumerator	HelpfulnessDenominator	Score	Time
Id	1.000000	0.001227	0.000770	0.010706	0.007912
HelpfulnessNumerator	0.001227	1.000000	0.974689	-0.032590	-0.154818
HelpfulnessDenominator	0.000770	0.974689	1.000000	-0.097986	-0.173289
Score	0.010706	-0.032590	-0.097986	1.000000	-0.062760
Time	0.007912	-0.154818	-0.173289	-0.062760	1.000000



Data Preparation

◆ Null Values and Duplicates

```
Id                0
ProductId         0
UserId           0
ProfileName      16
HelpfulnessNumerator  0
HelpfulnessDenominator  0
Score            0
Time             0
Summary          27
Text             0
dtype: int64
```

```
1 #Checking duplicate values
2 foods1_na[foods1_na['Id'].duplicated()]
```

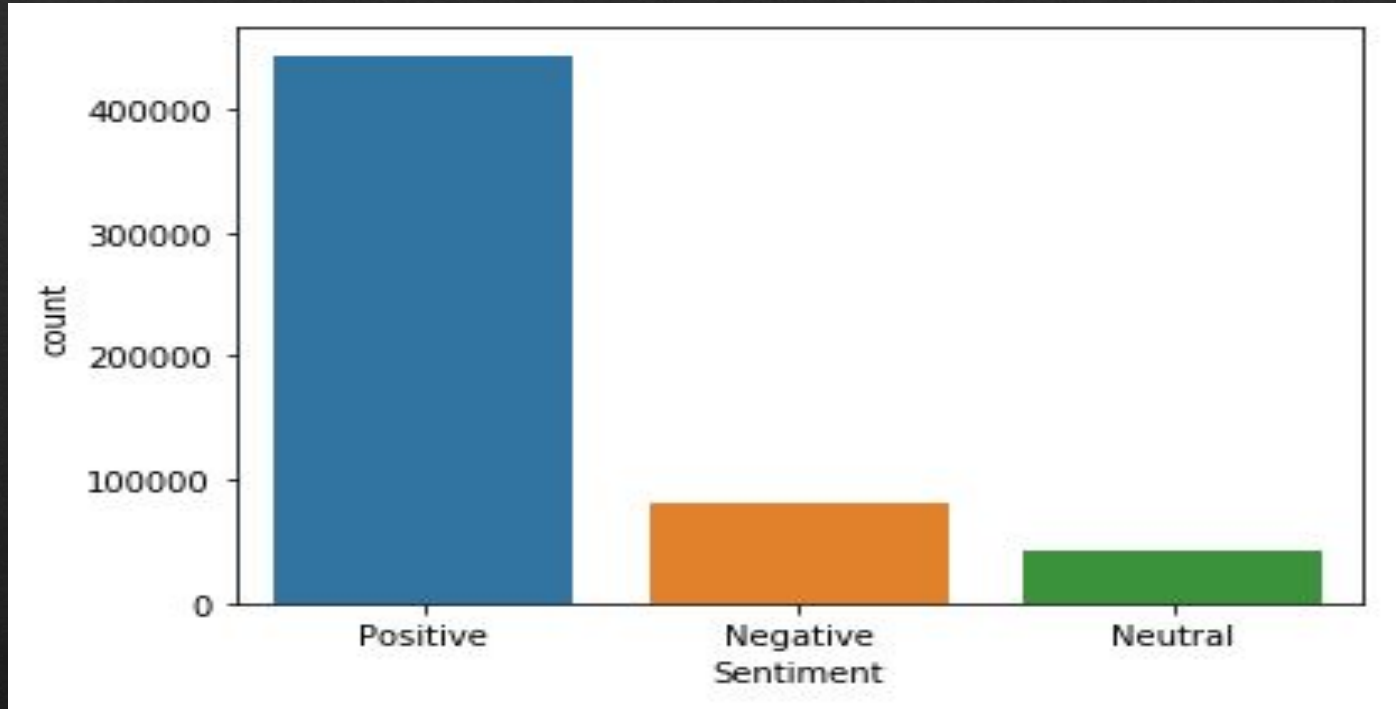
Id	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Summary	Text
----	-----------	--------	----------------------	------------------------	-------	---------	------

Data Preparation

- ◆ Adding column 'Sentiment' to dataframe

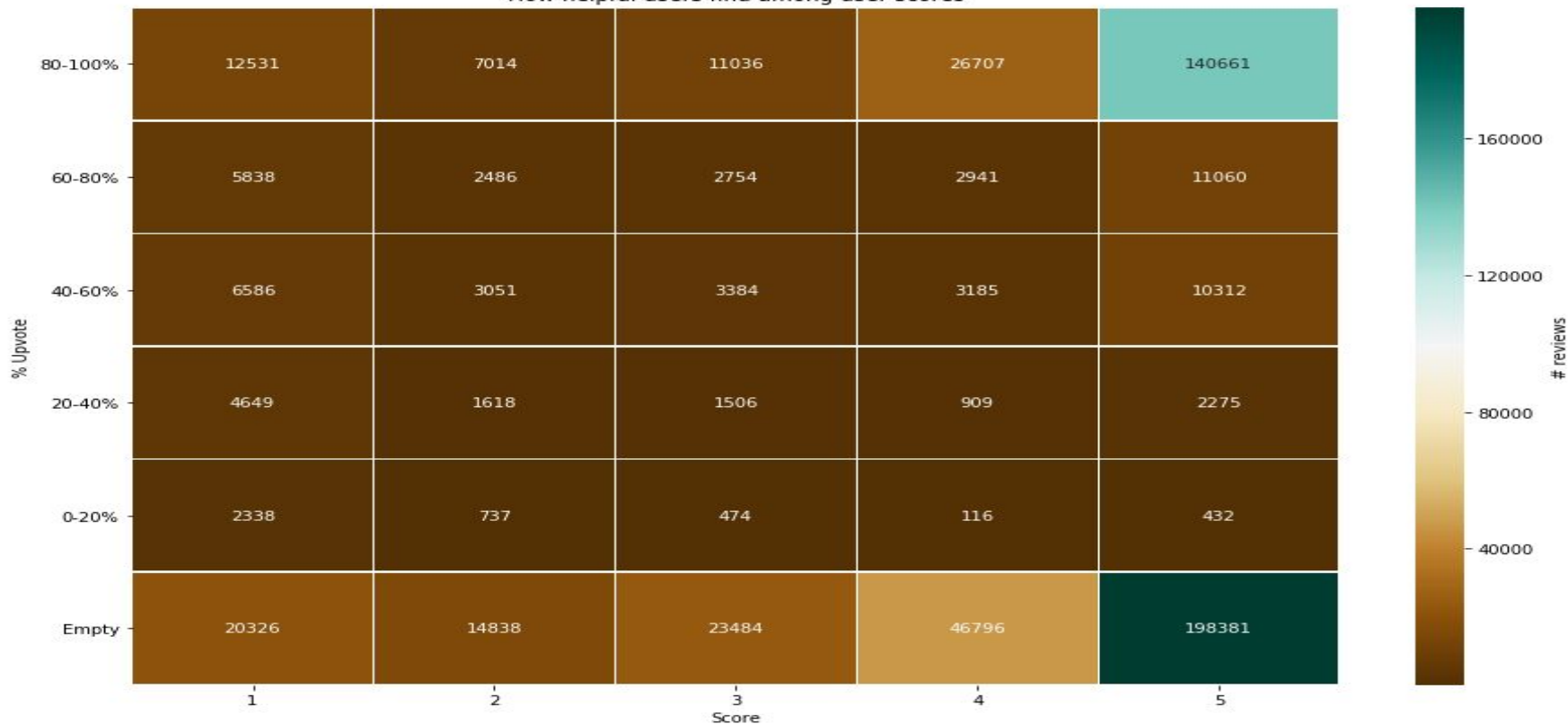
	Id	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Summary	Text	Sentiment
0	1	B001E4KFG0	A3SGXH7AUHU8GW	1	1	5	Good Quality Dog Food	I have bought several of the Vitality canned d...	Positive
1	2	B00813GRG4	A1D87F6ZCVE5NK	0	0	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	Negative
2	3	B000LQOCH0	ABXLMWJIXXAIN	1	1	4	"Delight" says it all	This is a confection that has been around a fe...	Positive
3	4	B000UA0QIQ	A395BORC6FGVXV	3	3	2	Cough Medicine	If you are looking for the secret ingredient i...	Negative
4	5	B006K2ZZ7K	A1UQRSCLF8GW1T	0	0	5	Great taffy	Great taffy at a great price. There was a wid...	Positive

Sentiment Understanding



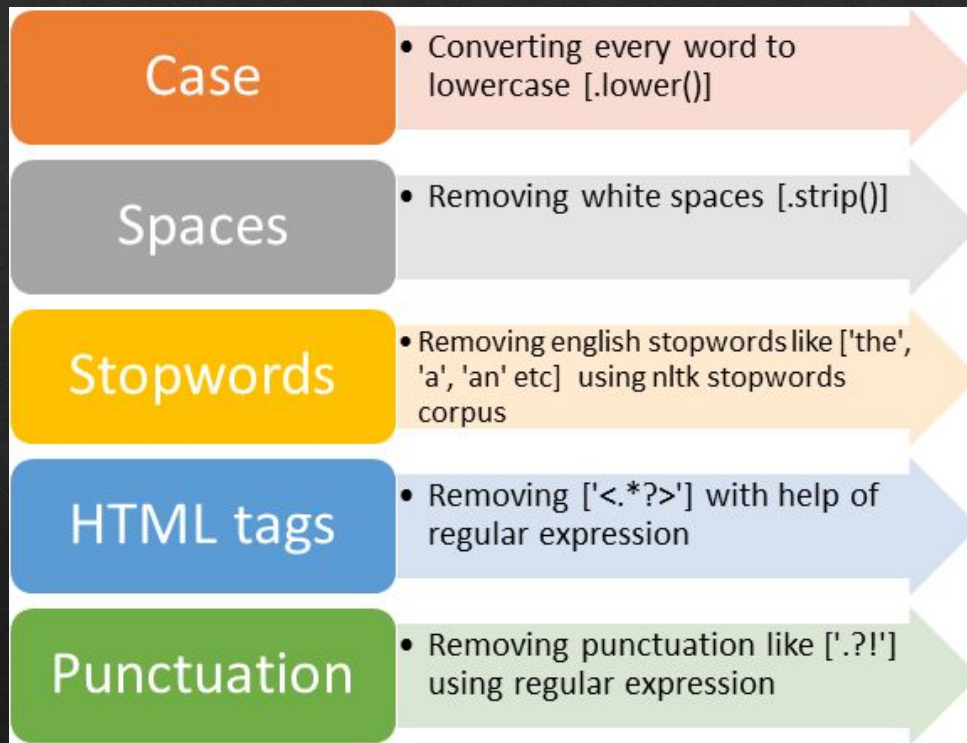
Usefulness of score among customers

How helpful users find among user scores



Data Preparation

- ❖ For text mining and word cloud:
 - Nltk library and Regular Expression



Data for ML models



Id	ProductId	UserId	HelpfulnessNumerator	HelpfulnessDenominator	Score	Summary	Text	Sentiment	Helpful %	% Upvote	Positivity
0	1	B001E4KFG0 A3SGXH7AUHU8GW	1	1	5	Good Quality Dog Food	I have bought several of the Vitality canned d...	Positive	1.0	80-100%	1
1	2	B00813GRG4 A1D87F6ZCVE5NK	0	0	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanut...	Negative	-1.0	Empty	0
2	3	B000LQOCH0 ABXLMWJIXXAIN	1	1	4	"Delight" says it all	This is a confection that has been around a fe...	Positive	1.0	80-100%	1
3	4	B000UA0QIQ A395BORC6FGVXV	3	3	2	Cough Medicine	If you are looking for the secret ingredient i...	Negative	1.0	80-100%	0
4	5	B006K2ZZ7K A1UQRSCLF8GW1T	0	0	5	Great taffy	Great taffy at a great price. There was a wid...	Positive	-1.0	Empty	1

Modeling

- ◆ Logistic Regression
- ◆ Decision Tree
- ◆ Random Forest

Logistic Regression

- ◆ Performed logistic regression to find the best hyperplane which could separate the reviews into positive and negative.
- ◆ The roc_auc_score is 93%

Logistic Regression

```
In [111]: #Multiple Logistic Regression
lr = LogisticRegression()
lr.fit(X_train_vectorized, y_train)
predictions = lr.predict(vect.transform(X_test))

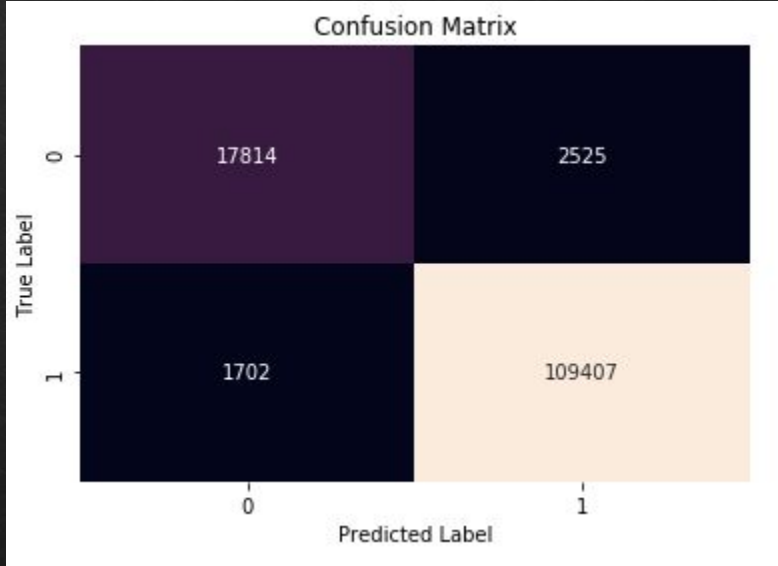
C:\Users\Heda\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:432: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
  FutureWarning)
C:\Users\Heda\Anaconda3\lib\site-packages\sklearn\svm\base.py:929: ConvergenceWarning: Liblinear failed to converge, increase the number of iterations.
  "the number of iterations.", ConvergenceWarning)

In [112]: print('AUC: ', roc_auc_score(y_test, predictions))

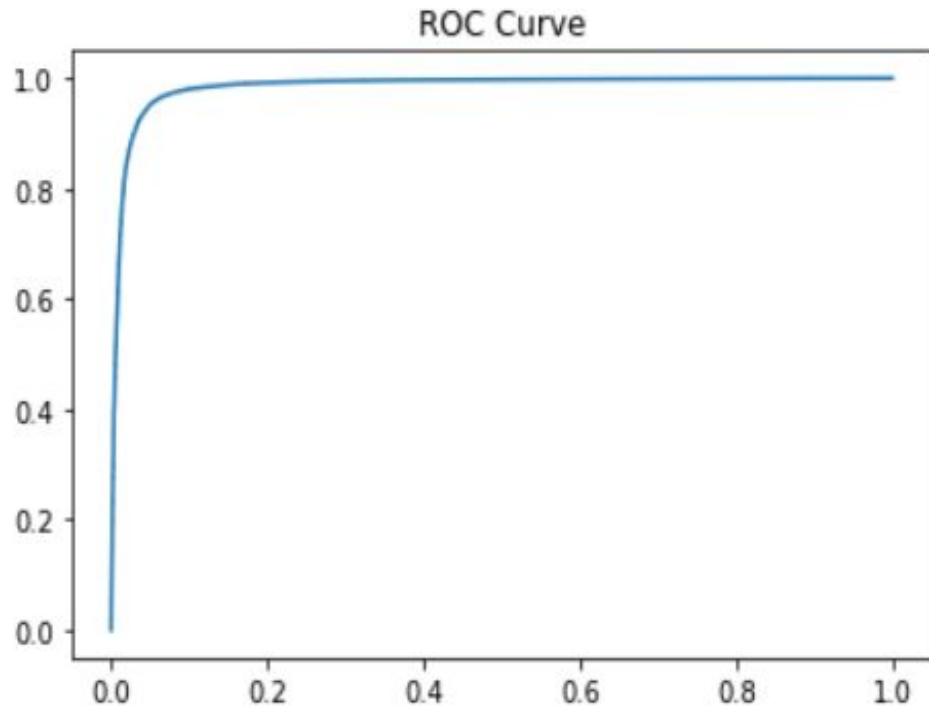
AUC:  0.9302679895369559

In [113]: conf_matr = confusion_matrix(y_test, predictions)
```

Confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.91	0.88	0.89	20339
1	0.98	0.98	0.98	111109
accuracy			0.97	131448
macro avg	0.95	0.93	0.94	131448
weighted avg	0.97	0.97	0.97	131448



ROC Curve for Logistic Regression

Decision Tree

- ◆ We have used “Reviews” as the root node for the tree.
- ◆ We got the roc_auc_score as 83.42.

Decision Tree

```
In [121]: #Decision Tree
          dt = tree.DecisionTreeClassifier()
          dt.fit(X_train_vectorized, y_train)
          predictions1 = dt.predict(vect.transform(X_test))

In [122]: print('AUC: ', roc_auc_score(y_test, predictions1))

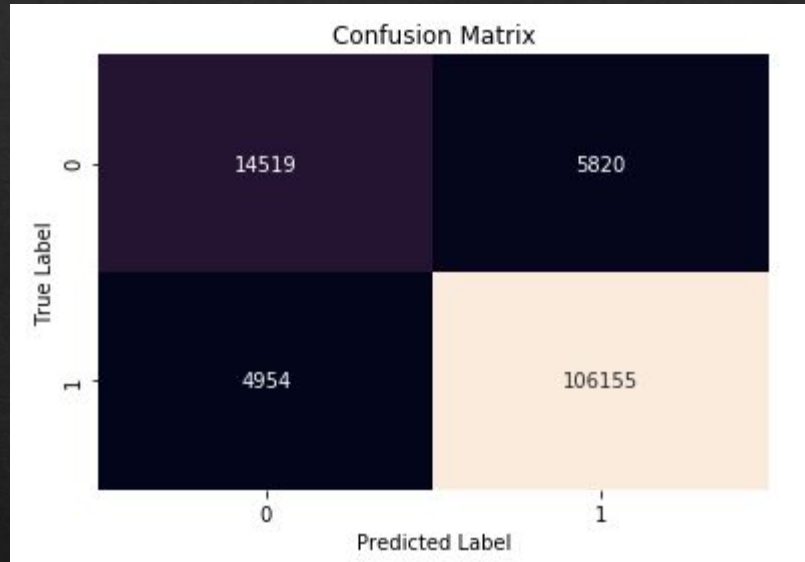
          AUC:  0.8342882749002035

In [123]: conf_matr_dt = confusion_matrix(y_test, predictions1)

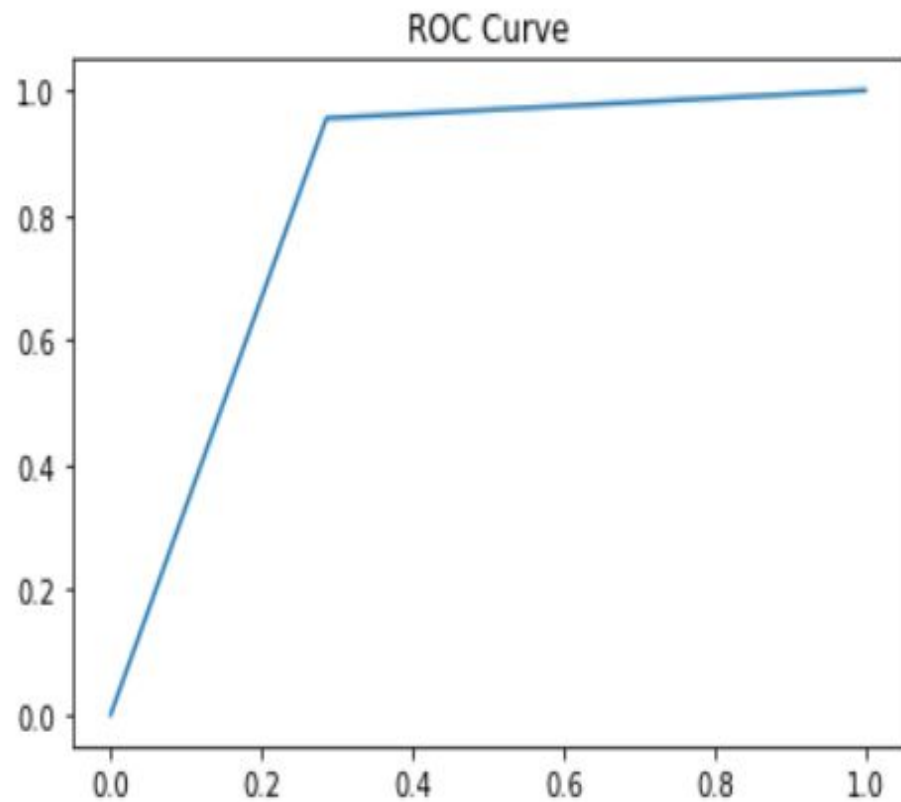
In [124]: sns.heatmap(conf_matr_dt,annot=True,cbar=False, fmt='g')
          plt.ylabel('True Label')
          plt.xlabel('Predicted Label')
          plt.title('Confusion Matrix')

Out[124]: Text(0.5, 1, 'Confusion Matrix')
```

Confusion Matrix and Classification Report



	precision	recall	f1-score	support
0	0.75	0.71	0.73	20339
1	0.95	0.96	0.95	111109
accuracy			0.92	131448
macro avg	0.85	0.83	0.84	131448
weighted avg	0.92	0.92	0.92	131448



ROC Curve for Decision Tree

Random Forest

- ◆ When we performed Random Forest, we got roc_auc_score around 78.69.

Random Forest

```
In [127]: #Random Forest
rf = RandomForestClassifier(n_jobs=-1)
rf.fit(X_train_vectorized, y_train)
predictions2 = rf.predict(vect.transform(X_test))

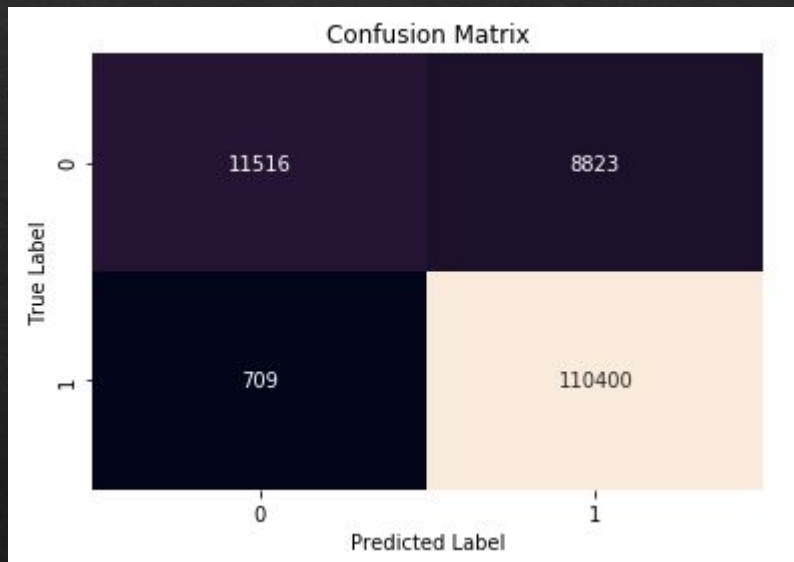
C:\Users\Heda\Anaconda3\lib\site-packages\sklearn\ensemble\forest.py:245: FutureWarning:
change from 10 in version 0.20 to 100 in 0.22.
  "10 in version 0.20 to 100 in 0.22.", FutureWarning)
```

```
In [128]: print('AUC: ', roc_auc_score(y_test, predictions2))

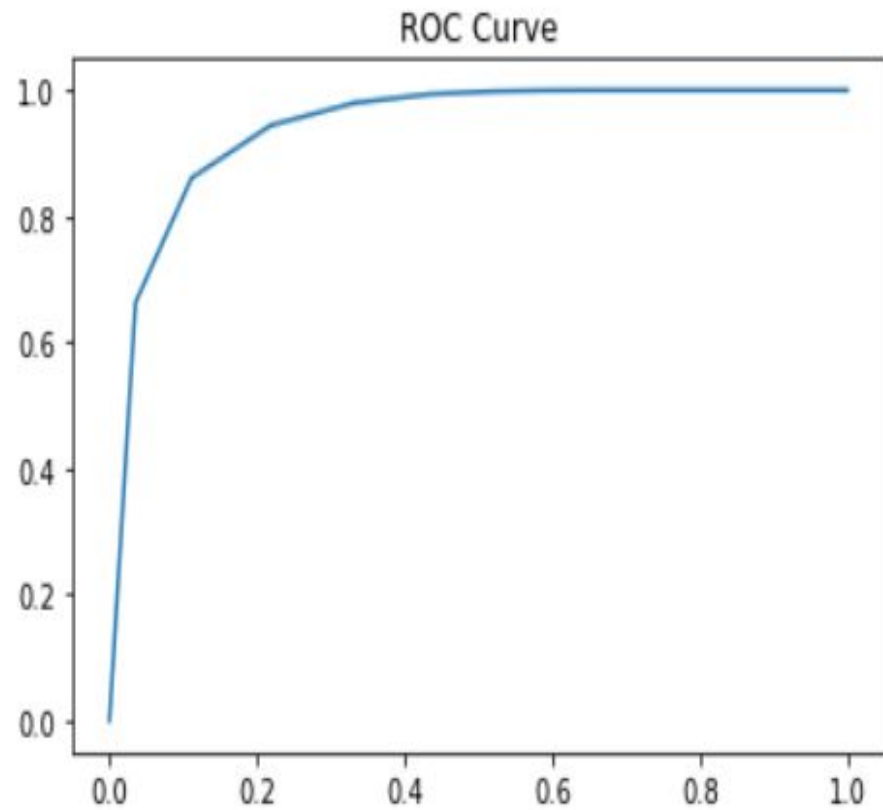
AUC:  0.7869293542832291
```

```
In [129]: conf_matr_rf = confusion_matrix(y_test, predictions2)
```


Confusion Matrix and Classification Report

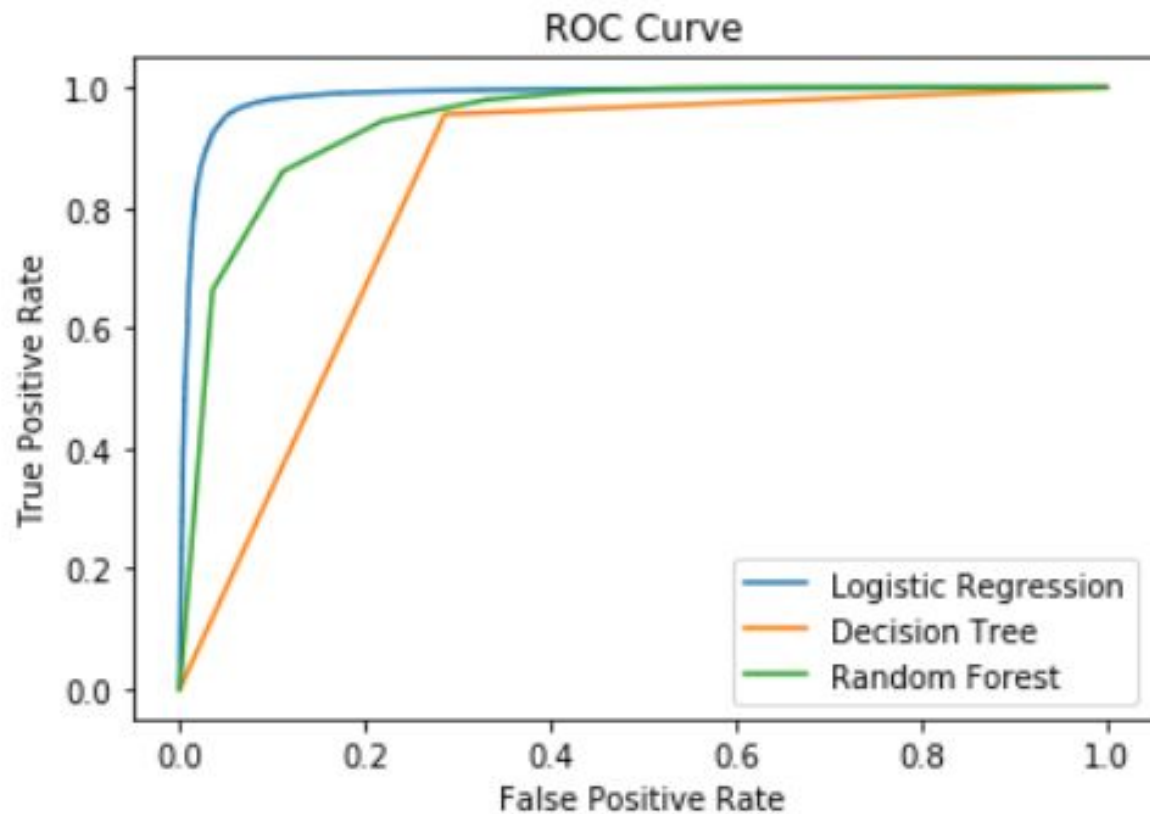


	precision	recall	f1-score	support
0	0.94	0.57	0.71	20339
1	0.93	0.99	0.96	111109
accuracy			0.93	131448
macro avg	0.93	0.78	0.83	131448
weighted avg	0.93	0.93	0.92	131448



ROC Curve for Random Forest

- ◆ Logistic regression is better since it gives more accuracy. A high threshold value results in a point at bottom left and low threshold value results in a point at top right.
- ◆ Apart from this, it does not require too many computational resources, it is highly interpretable and it does not require input features to be scaled.
- ◆ On the other hand Random forest requires more computational resources, owing to the large number of decision trees joined together. Due to their complexity, they require more time to train than other algorithms.
- ◆ Also in case of decision tree algorithm, a slight change in the data can cause a large change in the structure of the decision tree which causes instability. Decision tree often requires more time to train the model.



ROC Curve for all models

Evaluation

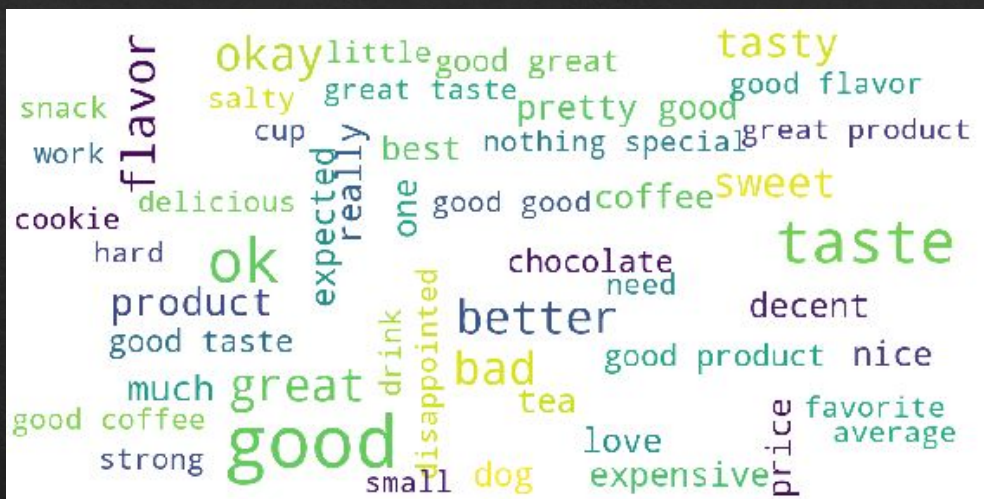
- ◆ To analyse the word pattern used by customers, we developed a word cloud on the summary of reviews provided by customers.
- ◆ This helped us in understanding the significant words that make the product positive or negative.



Review Score One



Review Score Two



Review Score Three



Review Score Four



Deployment

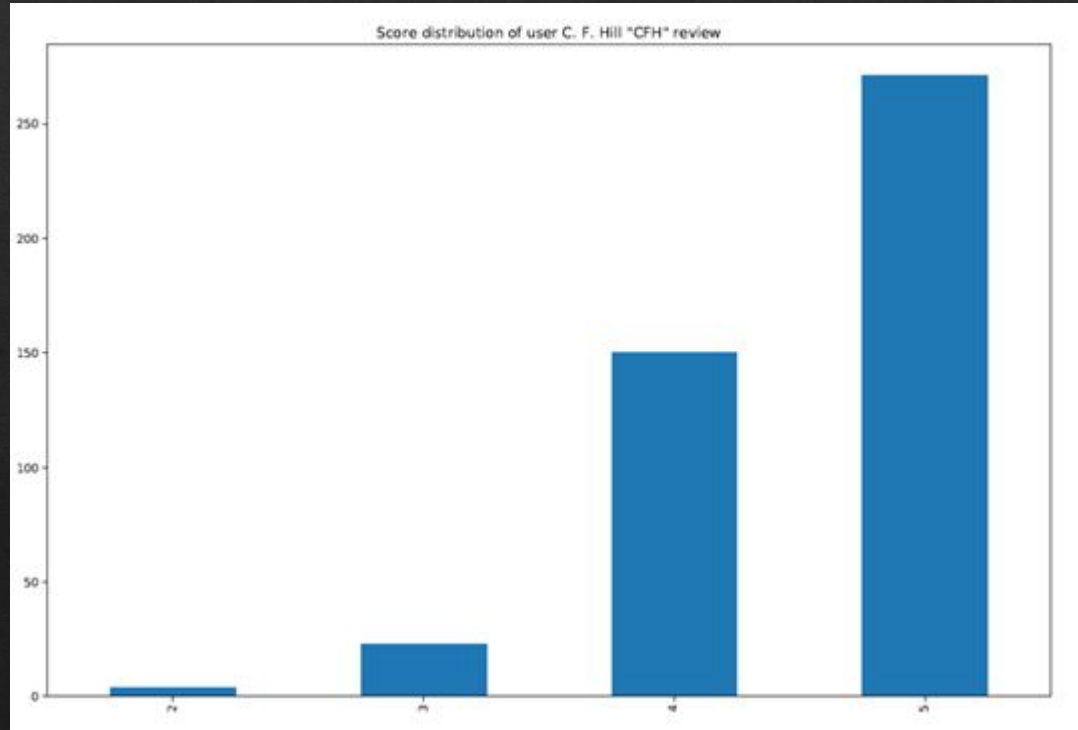
- ◆ Analysis is done on users, determine their liking for food based on reviews and their ratings
- ◆ Helps Amazon to do customer oriented marketing
- ◆ It will help Amazon to grow their customers

Below stats are the customers with who has given the maximum number of ratings and reviews to Amazon food products

UserId	ProfileName	Score count	Score mean
A3OXHLG6DIBRW8	C. F. Hill "CFH"	448	4.535714
A1YUL9PCJR3JTY	O. Brown "Ms. O. Khannah-Brown"	421	4.494062
AY12DBB0U420B	Gary Peterson	389	4.647815
A281NPSIMI1C2R	Rebecca of Amazon "The Rebecca Review"	365	4.841096
A1Z54EM24Y40LL	c2	256	4.453125
A1TMAVN4CEM8U8	Gunner	204	4.833333
A2MUGFV2TDQ47K	Lynrie "Oh HELL no"	201	3.751244
A3TVZM3ZIXG8YW	christopher hayes	199	1.000000
A3PJZ8TU8FDQ1K	Jared Castle	178	4.601124
AQQLWCMRNDFGI	Steven A. Peterson	176	3.954545

Below bar graph represents analysis of single user.

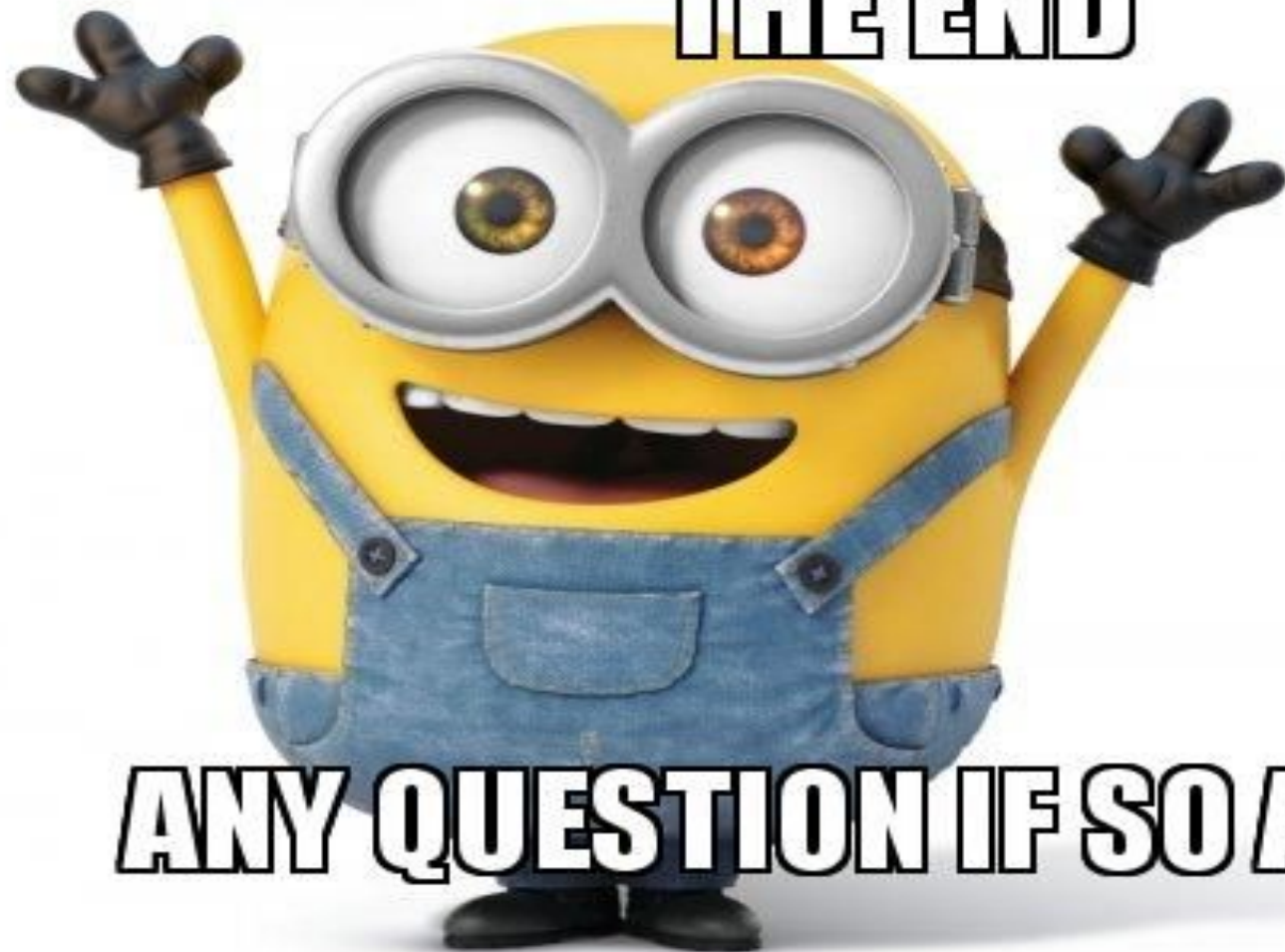
Bar graph represents ratings given by the user for food products.



References

1. SNAP: Web data: Amazon Fine Foods reviews, SNAP: Web data: Amazon Fine Foods reviews. (2019). Snap.stanford.edu.
2. Logistic Regression In Python, Logistic Regression In Python. (2019). Medium.,
<https://towardsdatascience.com/logistic-regression-python-7c451928efee>
3. Decision Tree In Python, Decision Tree In Python. (2019). Medium.,
<https://towardsdatascience.com/decision-tree-in-python-b433ae57fb93>
4. Decision Trees in Python with Scikit-Learn, Decision Trees in Python with Scikit-Learn. (2018). Stack Abuse.
<https://stackabuse.com/decision-trees-in-python-with-scikit-learn/>
5. An Implementation and Explanation of the Random Forest in Python, An Implementation and Explanation of the Random Forest in Python. (2018). Medium.
<https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76>
6. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.21.3 documentation, 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.21.3 documentation. (2019). Scikit-learn.org.,
<https://scikit-learn.org/stable/modules/generated/sklearn>
7. Understanding Data Science Classification Metrics in Scikit-Learn in Python, Understanding Data Science Classification Metrics in Scikit-Learn in Python. (2019). Medium.,
<https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3bc336865019>
8. (Tutorial) Generate Word Clouds in Python, (Tutorial) Generate Word Clouds in Python. (2019). DataCamp Community.
<https://www.datacamp.com/community/tutorials/wordcloud-python>
9. Text Analytics for Beginners using NLTK, Text Analytics for Beginners using NLTK. (2019). DataCamp Community.
<https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk>

THE END



ANY QUESTION IF SO ASK ME

