Using Reviews to Predict Ratings of Women's Clothing Items

Marisa Mitchell

Springboard Capstone Project 2

The Problem

- E-commerce Clothing and Fashion companies often have customers rate their satisfaction with items they purchase
- Ratings influence future customers' decisions
- Ratings can help companies make decisions about:
  - Items to discontinue
  - Items to add additional choices (e.g. color, fabric)
  - Improvements to make

The Problem

- Women's E-Commerce Clothing Reviews Dataset from Kaggle (https://www.kaggle.com/nicapotato/womens-ecommerce-clothing-reviews)
- 10 variables
- 23,486 reviews

Missing Data

Title Text	3,210
Review Text	845
Division Name	14
Department Name	14
Class Name	14
Clothing ID	0
Age	0
Rating	0
Recommended Indicator	0
Positive Feedback	0

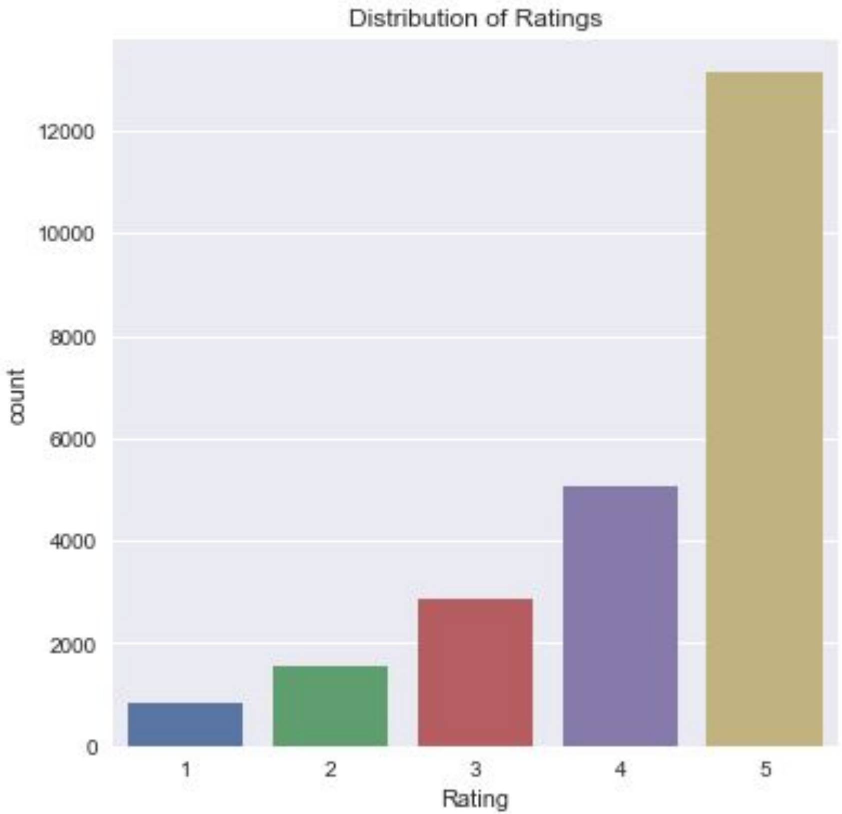
Data Cleaning Steps

- Merged Title and Review text into one feature- title_review
- Binned ratinging into three groups:
    - High- ratings of 4 and 5
    - Medium- Ratings of 3
    - Low- ratings of 1 and 2

Exploratory Analysis

Distribution of ratings prior to binning

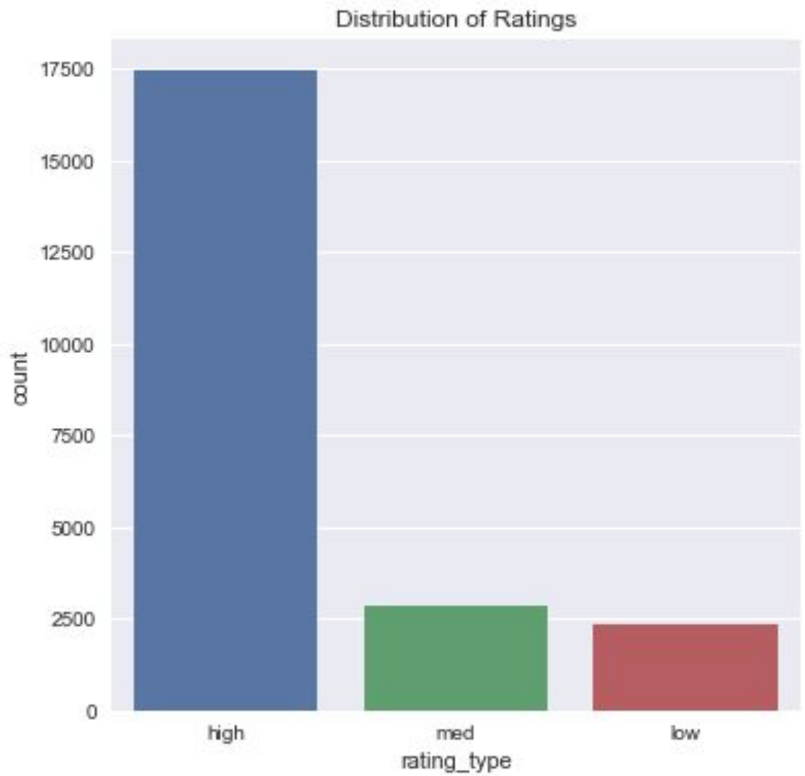
Less ratings in the lower values and increasing to the most 5 point ratings



Exploratory Analysis

Distribution of ratings after the data was binned

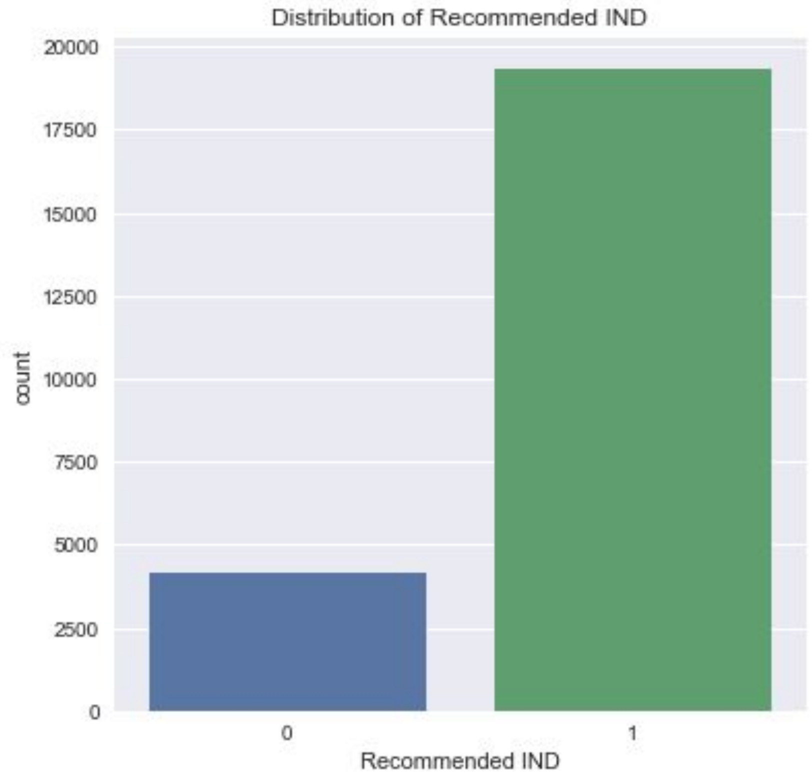
High ratings were much more frequent than the medium or low ratings



Exploratory Analysis

Distribution of recommended indicator

Majority of reviewers said they would recommend the item

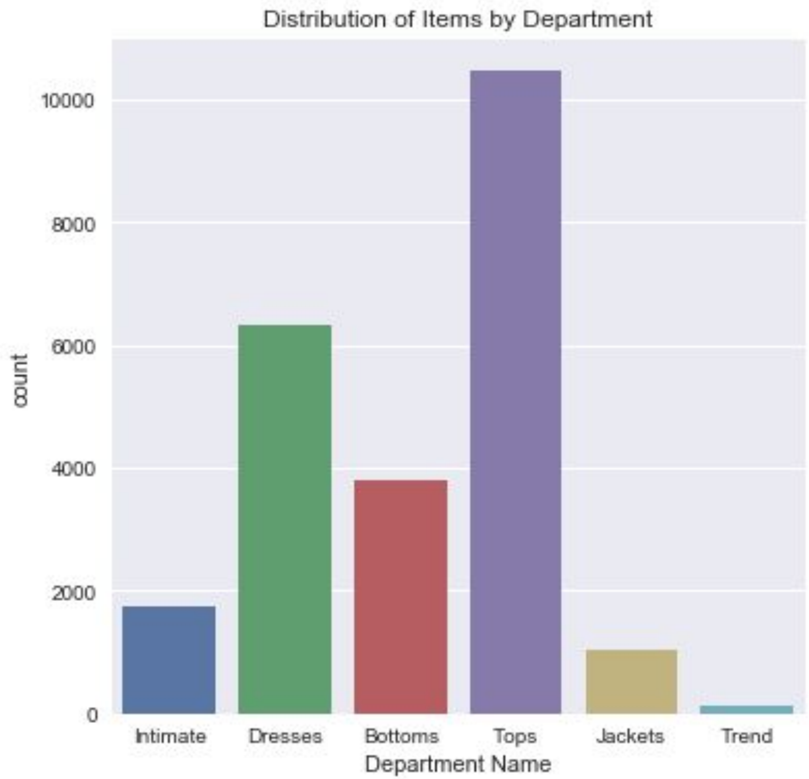


Exploratory Analysis

Distribution of items by department

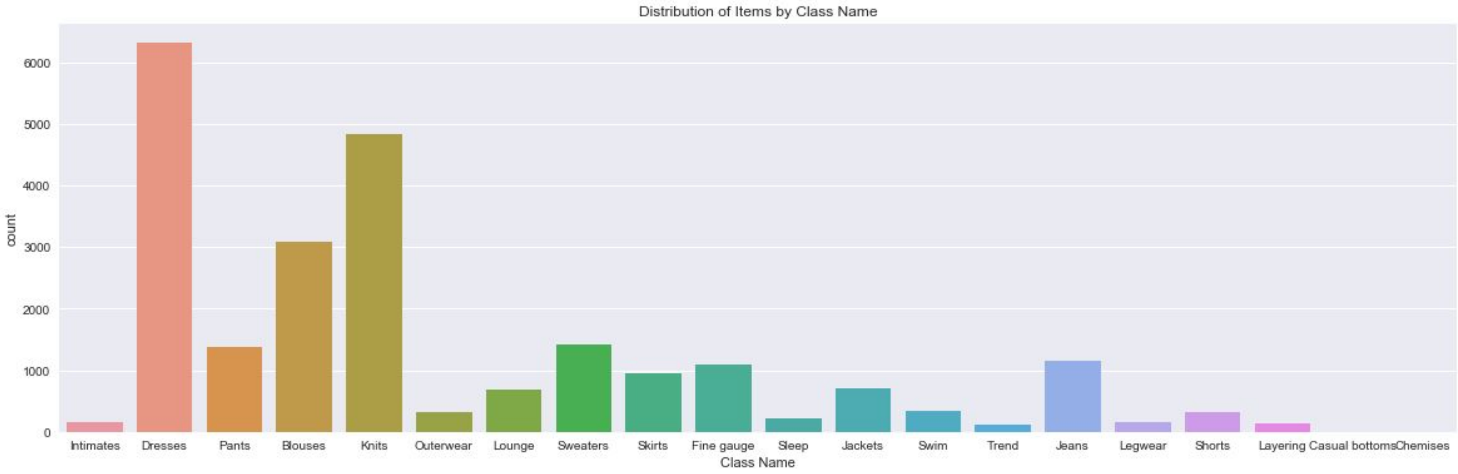
Tops were the most frequent item reviewed

Trend items were least reviewed



Exploratory Analysis

Distribution of Items by Class Name



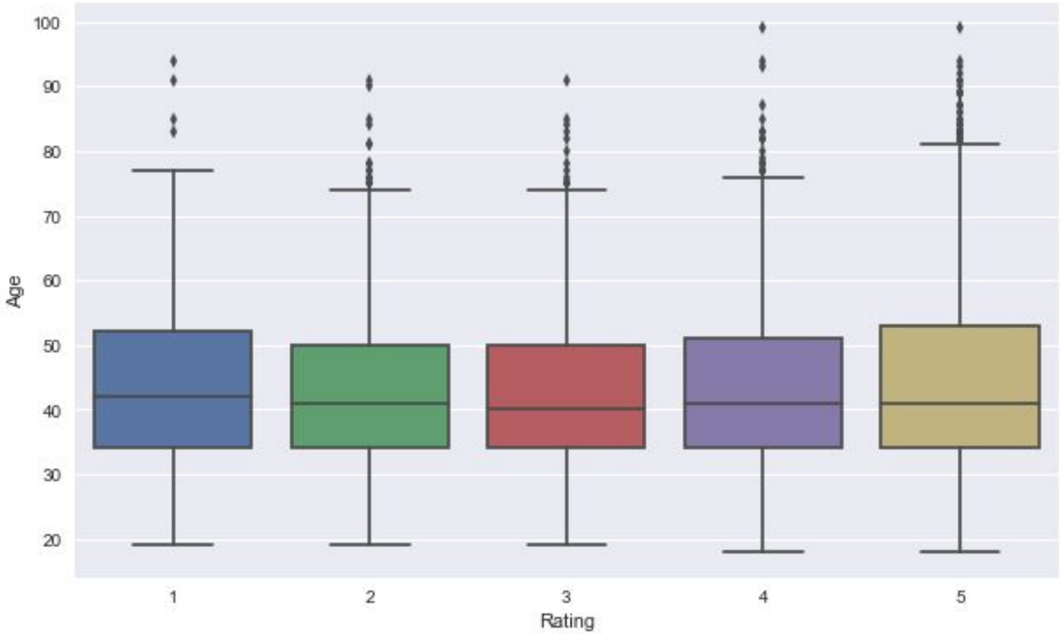
Exploratory Analysis

Distribution of iratings prior to binning for each class type

Distribution of reviews were comparable across all class types

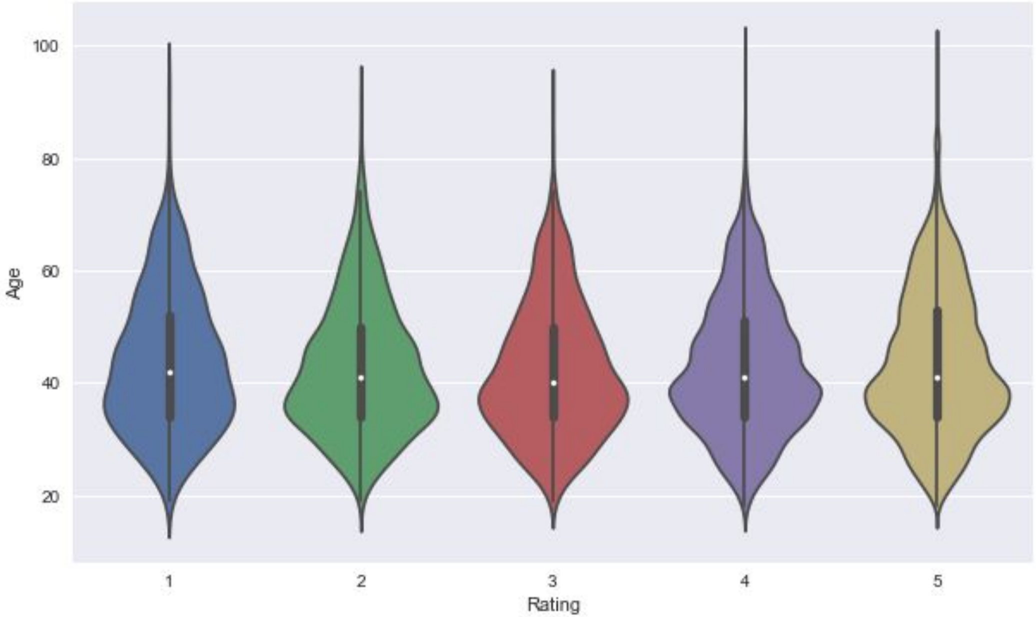
Exploratory Analysis

Boxplot showing distribution of ratings by age of reviewer



Exploratory Analysis

Violin plot showing distribution of ratings by age of reviewer

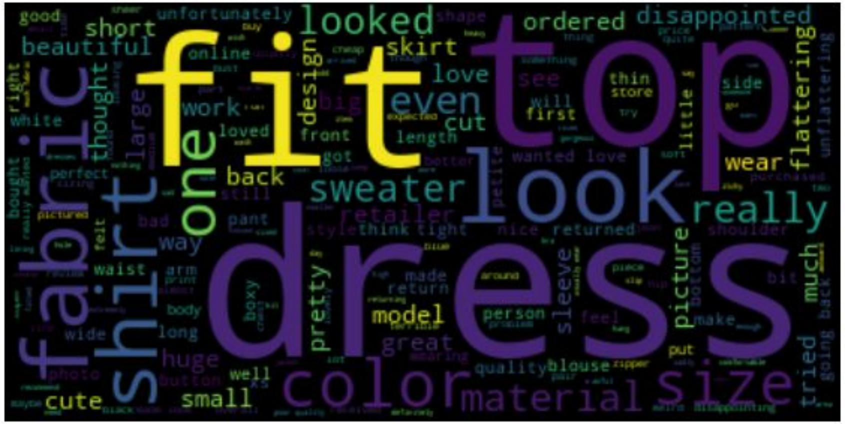


Exploratory Analysis

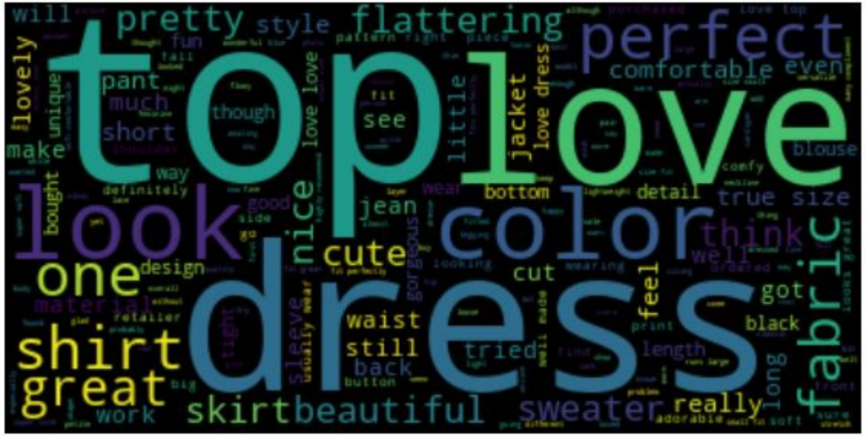
Word cloud of all words in the title_review variable

Exploratory Analysis

Word cloud of all words in the title_review variable for reviews that were low or high

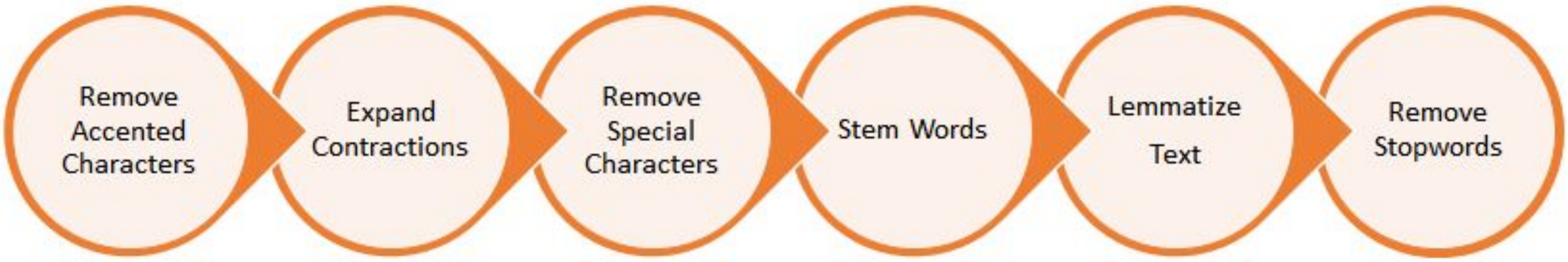


Low Ratings

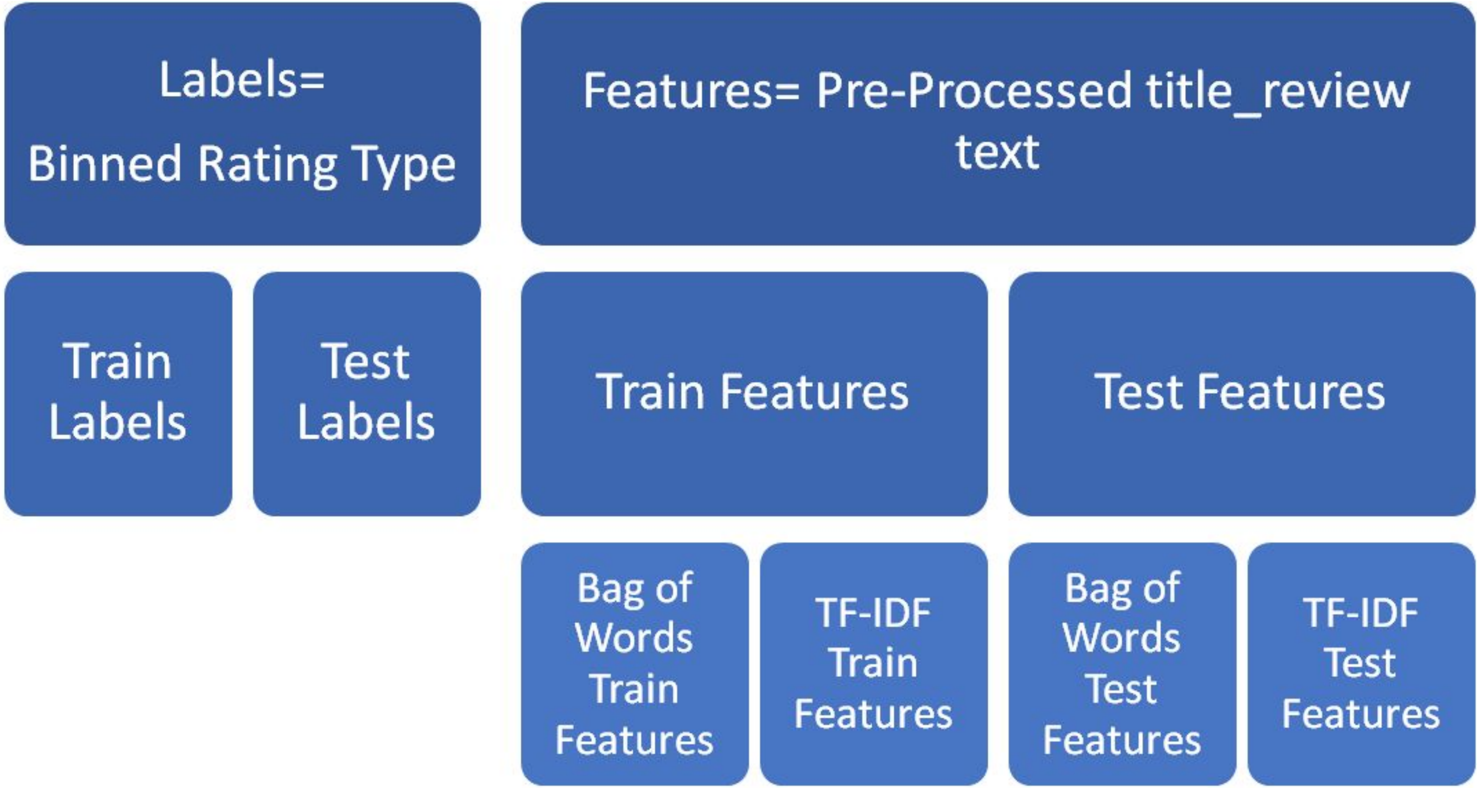


High Ratings

Text Preprocessing



Feature Engineering



Machine Learning Algorithms Used for Review Classification

Model Evaluation Performance Metrics

Bag of Words

Model	Accuracy Before Tuning	Accuracy After Tuning
KNN	0.77	
Logistic Regression	0.83	0.83
Linear SVM	0.80	
Decision Tree	0.74	
Random Forest	0.79	
Gradient Boosting	0.80	

TF-IDF

Model	Accuracy Before Tuning	Accuracy After Tuning
KNN	0.79	
Logistic Regression	0.83	0.83
Linear SVM	0.83	0.83
Decision Tree	0.74	
Random Forest	0.79	
Gradient Boosting	0.81	

Hyperparameter Tuning

- Tuned the top three models using GridsearchCV
- Logistic regression models were tuned on
    - Penalty- l1, l2
    - C- .001, .01, 1, 10, 100
- Linear SVM model was tuned on
    - Class_weight- balanced, none
    - C- .001, .01, 1, 10, 100, 1000

Confusion Matrix for Top 3 Models

Logistic Regression
Bag of Words

	high	low	med
high	5019	78	168
low	188	361	147
med	400	189	243

Logistic Regression
TF-IDF

	high	low	med
high	5102	57	106
low	250	325	121
med	475	175	182

Linear SVM TF-IDF

	high	low	med
high	5028	72	165
low	174	371	151
med	410	191	231

Classification Reports for Top 3 Models

Logistic Regression
Bag of Words

	precision	recall	f1-score	support
high	0.90	0.95	0.92	5265
low	0.57	0.52	0.55	696
med	0.44	0.29	0.35	832
micro avg	0.83	0.83	0.83	6793
macro avg	0.64	0.59	0.61	6793
weighted avg	0.81	0.83	0.81	6793

Linear SVM TF-IDF

	precision	recall	f1-score	support
high	0.90	0.95	0.92	5265
low	0.59	0.53	0.56	696
med	0.42	0.28	0.34	832
micro avg	0.83	0.83	0.83	6793
macro avg	0.63	0.59	0.61	6793
weighted avg	0.81	0.83	0.81	6793

Logistic Regression
TF-IDF

	precision	recall	f1-score	support
high	0.88	0.97	0.92	5265
low	0.58	0.47	0.52	696
med	0.44	0.22	0.29	832
micro avg	0.83	0.83	0.83	6793
macro avg	0.63	0.55	0.58	6793
weighted avg	0.79	0.83	0.80	6793

Feature Weights of Logistic Regression TF-IDF

y=high top features		y=low top features		y=med top features	
Weight	**Feature**	**Weight**	**Feature**	**Weight**	**Feature**
+9.568	perfect	+7.922	horrible	+4.831	however
+8.000	compliment	+5.922	disappointed	+4.255	meh
+6.783	comfortable	+5.737	poor	+3.992	ok
+6.771	great	+5.463	awful	+3.732	oversize
+6.307	happy	+4.970	disappointment	+3.196	seem
+6.022	love	+4.897	disappointing	+3.161	not
+5.178	glad	+4.501	ill	+3.007	excited
+5.161	perfectly	+4.337	unflattering	*… 192 more positive …*	
… 275 more positive …		*… 179 more positive …*		*… 174 more negative …*	
… 204 more negative …		*… 186 more negative …*		-3.199	love
-5.047	not	-4.396	nice	-3.226	comfy
-5.162	disappointment	-4.436	gorgeous	-3.258	glad
-5.233	return	-4.530	lovely	-3.268	happy
-5.727	horrible	-4.713	beautiful	-3.402	versatile
-5.791	cheap	-5.330	soft	-3.506	flattering
-5.798	bad	-5.600	happy	-3.775	boot
-6.392	meh	-5.754	compliment	-3.898	classic
-6.490	unflattering	-6.133	love	-4.031	perfectly
-6.626	disappointing	-6.288	great	-4.097	great
-7.153	awful	-6.302	perfect	-4.430	comfortable
-8.851	poor	-6.718	comfortable	-5.887	compliment
-9.224	disappointed	-7.899	little	-7.891	perfect

Conclusions

- Logistic Regression using TF-IDF features is recommended
- High ratings are associated with being comfortable, getting compliments, and customers being glad and happy while not being associated with returns, being cheap, unflattering, or poor
- Medium ratings are associated with being oversized and ok while not being associated with words like love comfy, glad, happy, and flattering
- Low ratings are associated with disappointment, unflattering, poor, and horrible while not being associated with being nice, gorgeous, beautiful, soft, complement, and comfortable

Next Steps

- Examine feature weights for each clothing category
- Examine feature weights individual clothing ID’s
- Apply oversampling or undersampling techniques to deal with the unbalanced ratings
- Include other dataset features in addition to the text features