Matthew Smith

Final Report:
Shopify App Store Sentiment Analysis

Problem Statement

Shopify allows developers to create and sell third party apps on their app marketplace. These apps can enhance a customers storefront allowing additional capabilities such as inventory management, marketing, store design and more. Given the wide range of app offerings on the Shopify Marketplace, are initial reviews representative of final review values and can common themes be identified amongst all apps? These themes can be provided to developers to enhance their apps or be used by the marketplace provider (Shopify) to provide developer support for features that end users desire.

Data Wrangling

The raw dataset was found on Kaggle.com and was created by web scraping the Shopify Marketplace. The portion of the dataset used is comprised of four files:

- apps - containing various categorical features of 4750 unique apps (rows)
- apps_categories - which classified each app into its respective category.
    - Apps were allowed multiple categories resulting in 7376 rows
- categories - a bridge table with the name of the categories and their respective numeric code
- reviews - containing review data and linked to the app being reviewed
    - Contained 447,317 separate reviews on 4750 apps
    - *To limit the scope of this project, the reviews dataset was subset to include the first 20,000 entries

With relevant feature data split between multiple tables, joining the data was a critical step to analyzing the data. Columns were renamed for consistency and dummy variables were created for duplicate categories before being left joined onto the apps table resulting in one dataframe containing 4750 rows while still capturing feature data for apps that had multiple app categories.

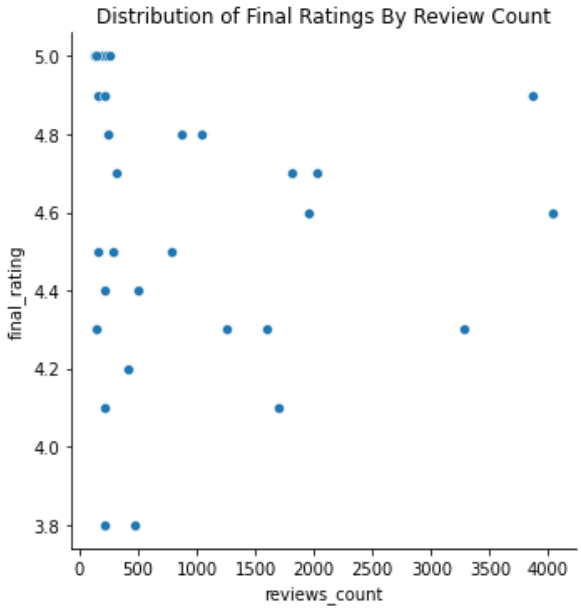
Given the large volume of reviews and potential skewing by particularly small apps, the reviews table was modified to include only the reviews of apps with over 100 ratings. An additional column was created to categorize reviews into either recommended (4+ star rating) or not recommended (<4 rating). Five columns were created to capture aggregate ratings of the first 50 reviews after each 10 reviews.

Exploratory Data Analysis

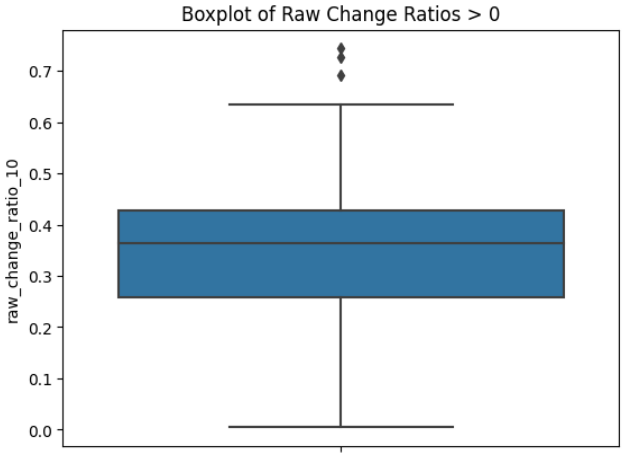
After cleaning the data, exploration of the app & review distributions was conducted. This process is summarized in two key steps. Understanding the distribution of app ratings against reviews, and understanding the sentiment of reviews as they relate to scores given.

1) Exploration of the distribution of app ratings and reviews.

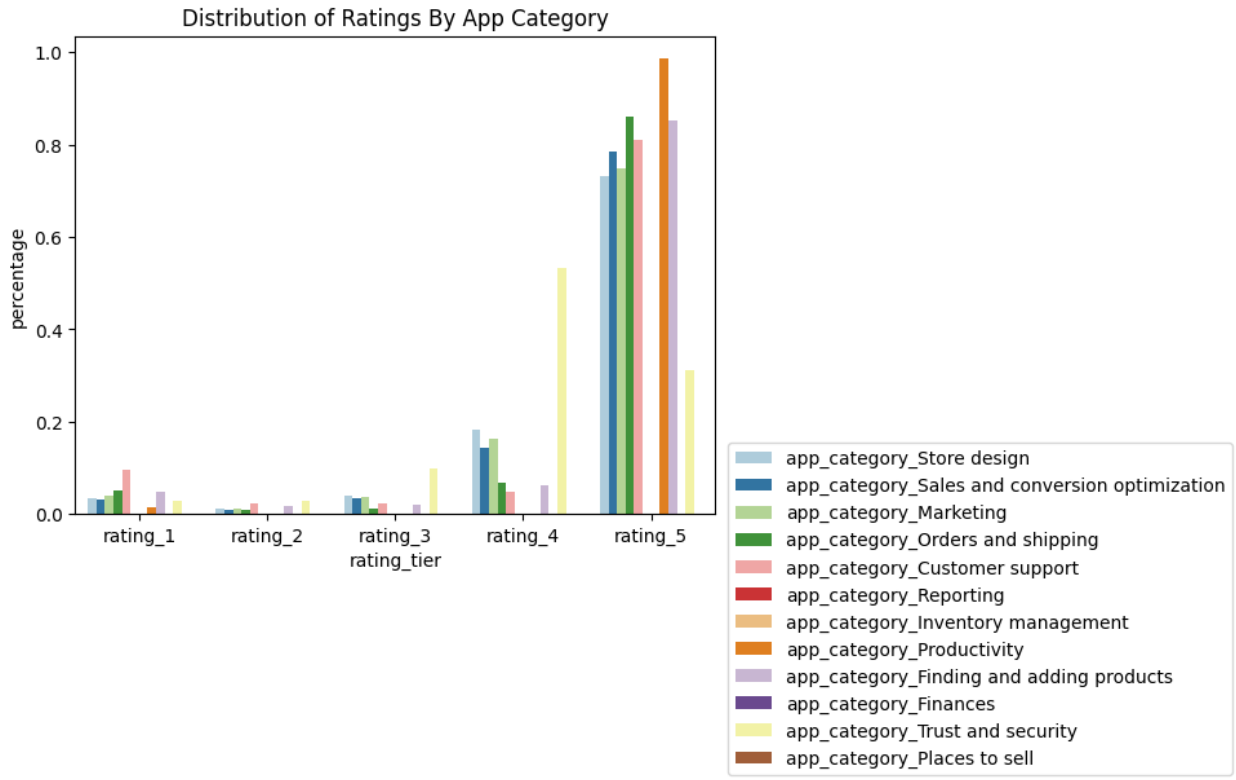
A preliminary analysis of reviews_count & final_rating was observed. Clustering around lower review counts & higher final_ratings is exhibited.



A raw change ratio was defined as the absolute value of an apps final rating less the average rating of the first ten reviews. This ratio was plotted to evaluate the magnitude of the rating change when the raw change ratio was not equal to 0. On average the change was .35 out of 5.



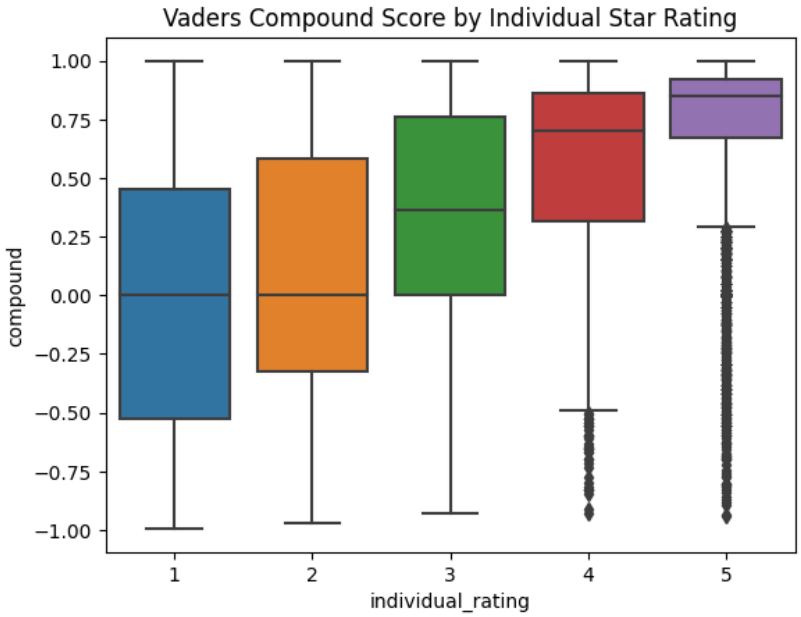
We now have an idea of the distribution of ratings, but how are reviews distributed across each category of app?



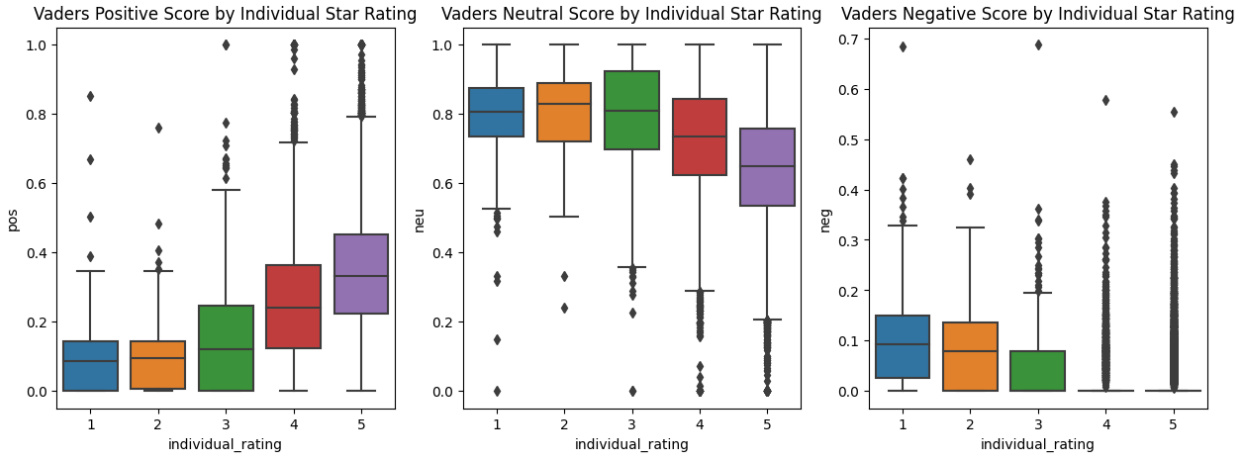
The most common types of app categories are: Store Design, Sales Conversion and Optimization, and Marketing. Reviews are still generally positive across all app categories.

2) Exploration of the sentiment of different reviews.

After establishing a baseline regarding the distribution of ratings and reviews. We can see that the data is skewed and has many more high ratings than low ratings. To extract further insight, a sentiment analysis score was created using the VADERS model and was plotted to demonstrate the overall sentiment range for each tier of rating as evidenced by the compound score.



A valuable feature of the VADERS model is that the compound score is the aggregate of the positive, neutral, and negative score for each entry.

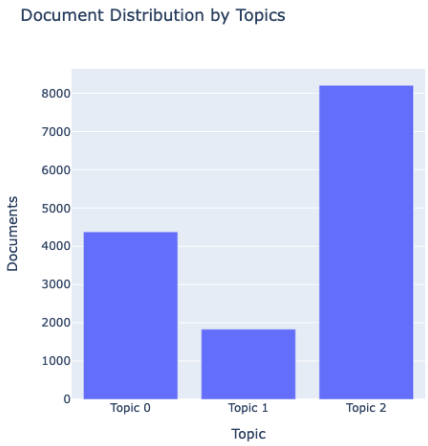


As expected, the compound score increases as the overall rating increases. Likewise, positive sentiment is more prevalent in higher reviews & negative sentiment is more prevalent in lower reviews.

In-Depth Analysis

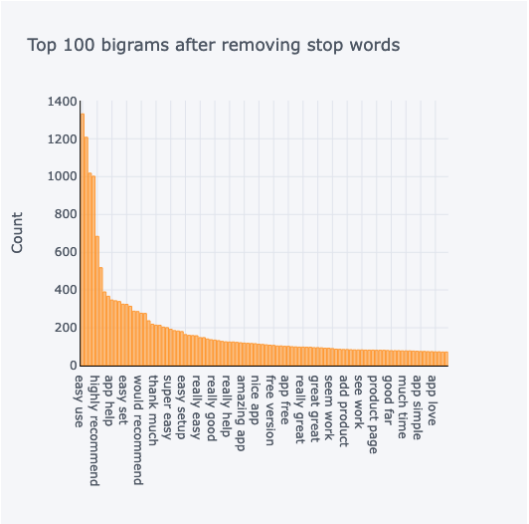
General Corpus

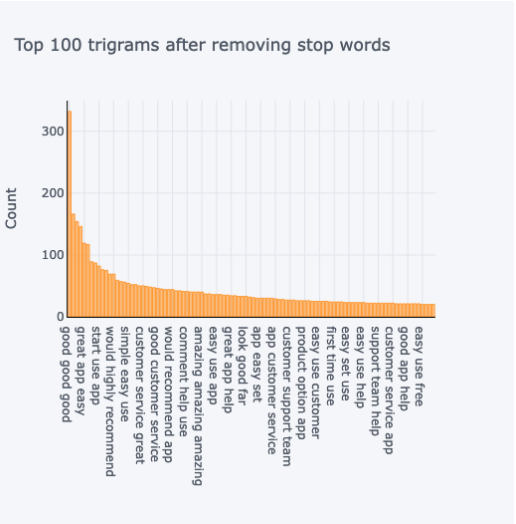
After establishing a sentiment score, topic modeling was conducted on the general corpus resulting in three topic categories.



1) Topic 0: Customer Service
    a) customer, support, app, service, work, team, issue, get, review, help
    b) 4377 items in topic 0
2) Topic 1: Product/Page Modifications
    a) product, make, order, add, option, item, price, store, find, time
    b) 1833 items in topic 1
3) Topic 2: Usability/User Interface
    a) app, use, great, easy, good, thank, recommend, really, help, work
    b) 8206 items in topic 2

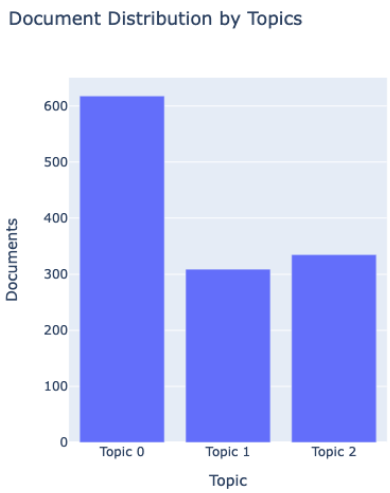
Top Bigrams & Trigrams indicate an emphasis on “simple easy use” being a reason to recommend an app.





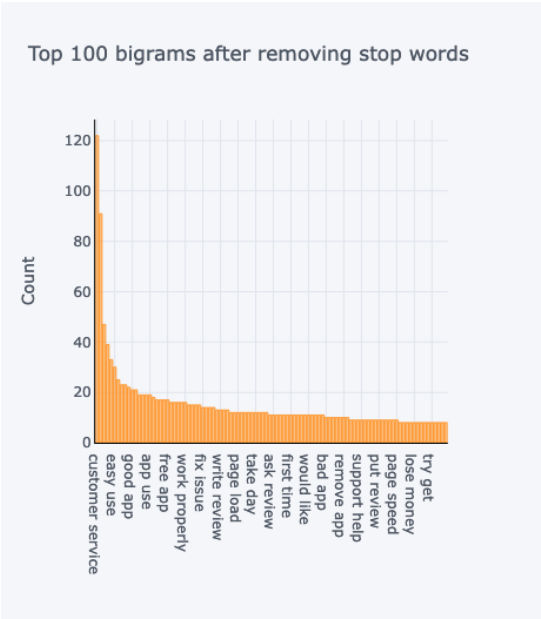
Negative Reviews (<4.0 Rating)

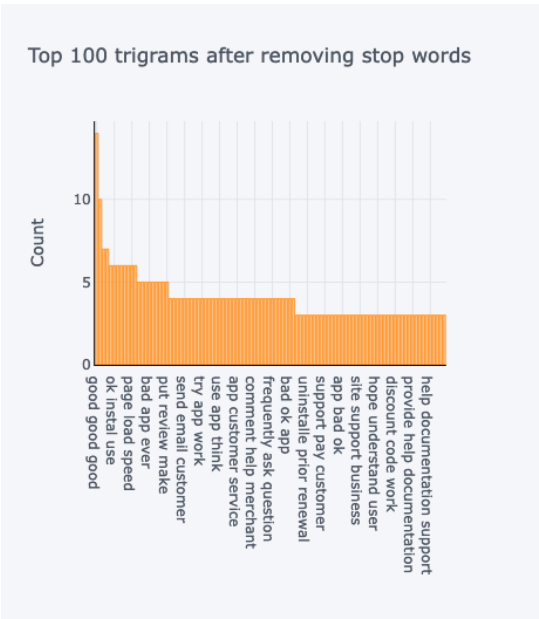
After establishing a sentiment score, topic modeling was conducted on the general corpus resulting in three topic categories.



1) Topic 0: App Functionality (Negative)
    a) app, work, support, review, use, help, time, email, customer, go
    b) 618 items in topic 0
2) Topic 1: Product/Page Modifications
    a) product, customer, order, use, app make, would, option, discount, get
    b) 309 items in topic 1
3) Topic 2: App Functionality (Positive)
    a) app, work, good, use, product, store, customer, seem, support, make
    b) 335 items in topic 2

Top Bigrams & Trigrams for negative sentiment indicate an emphasis on “customer service”, “fix issue”, and “page load speed”





Model Selection

For both the general corpus and negative reviews only, I tested 13 classification models and ordered by Area Under Curve (AUC). The target was whether an app was “recommended“ (Y/N). LightGBM was a relatively high performer in both scenarios and was tuned further.

General Corpus

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
et	Extra Trees Classifier	0.880	0.527	0.960	0.913	0.936	0.010	0.012	5.308
lightgbm	Light Gradient Boosting Machine	0.880	0.518	0.959	0.914	0.936	0.021	0.025	2.210
rf	Random Forest Classifier	0.881	0.515	0.961	0.913	0.936	0.009	0.012	3.672
ada	Ada Boost Classifier	0.787	0.513	0.846	0.914	0.879	0.015	0.015	3.338
gbc	Gradient Boosting Classifier	0.828	0.509	0.897	0.914	0.905	0.014	0.014	9.850
dt	Decision Tree Classifier	0.814	0.501	0.880	0.912	0.896	0.002	0.001	2.188
dummy	Dummy Classifier	0.088	0.500	0.000	0.000	0.000	0.000	0.000	1.122
nb	Naive Bayes	0.316	0.491	0.279	0.907	0.426	-0.005	-0.012	1.688
qda	Quadratic Discriminant Analysis	0.718	0.490	0.767	0.910	0.832	-0.011	-0.013	29.552
lda	Linear Discriminant Analysis	0.603	0.490	0.626	0.911	0.742	-0.004	-0.006	22.105
lr	Logistic Regression	0.488	0.375	0.512	0.685	0.586	0.001	0.002	19.148
svm	SVM - Linear Kernel	0.589	0.000	0.607	0.913	0.702	0.001	0.002	5.285
ridge	Ridge Classifier	0.610	0.000	0.634	0.912	0.748	-0.003	-0.004	2.950

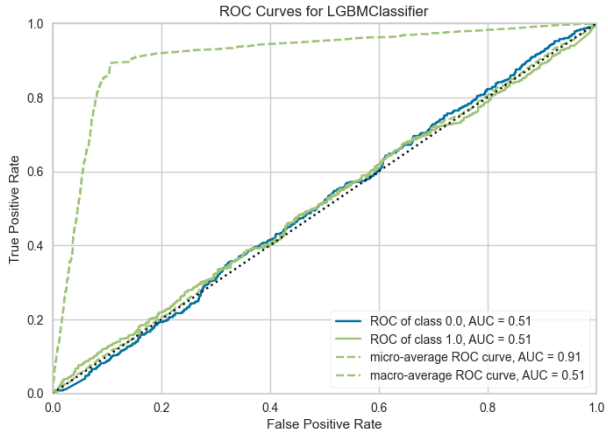
Negative Sentiment (<4.0 Rating)

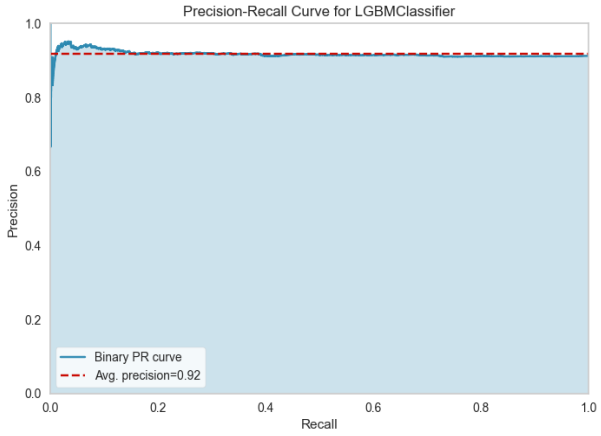
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ada	Ada Boost Classifier	0.580	0.531	0.656	0.748	0.697	0.016	0.017	0.350
gbc	Gradient Boosting Classifier	0.637	0.529	0.781	0.742	0.760	0.008	0.009	0.695
lr	Logistic Regression	0.569	0.522	0.625	0.752	0.683	0.029	0.030	0.845
dt	Decision Tree Classifier	0.619	0.521	0.731	0.749	0.740	0.029	0.029	0.595
nb	Naive Bayes	0.565	0.519	0.613	0.755	0.676	0.031	0.034	0.360
lightgbm	Light Gradient Boosting Machine	0.655	0.517	0.826	0.740	0.780	-0.009	-0.010	0.278
rf	Random Forest Classifier	0.702	0.515	0.911	0.744	0.819	0.016	0.023	0.348
et	Extra Trees Classifier	0.700	0.510	0.898	0.748	0.816	0.036	0.041	0.408
dummy	Dummy Classifier	0.259	0.500	0.000	0.000	0.000	0.000	0.000	0.155
qda	Quadratic Discriminant Analysis	0.730	0.492	0.985	0.739	0.844	-0.022	-0.062	0.570
lda	Linear Discriminant Analysis	0.506	0.481	0.515	0.739	0.606	-0.005	-0.006	0.915
svm	SVM - Linear Kernel	0.534	0.000	0.551	0.795	0.570	0.026	0.047	0.550
ridge	Ridge Classifier	0.540	0.000	0.567	0.751	0.646	0.022	0.024	0.258

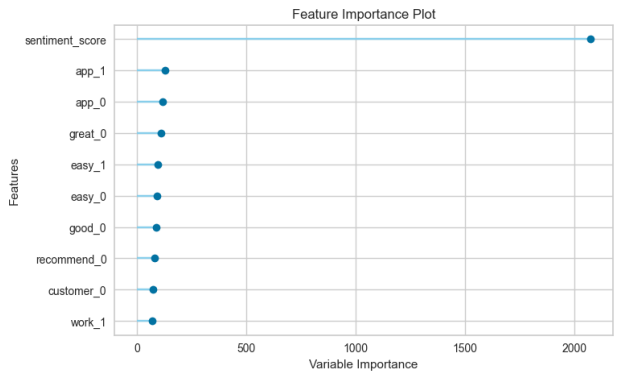
Takeaways

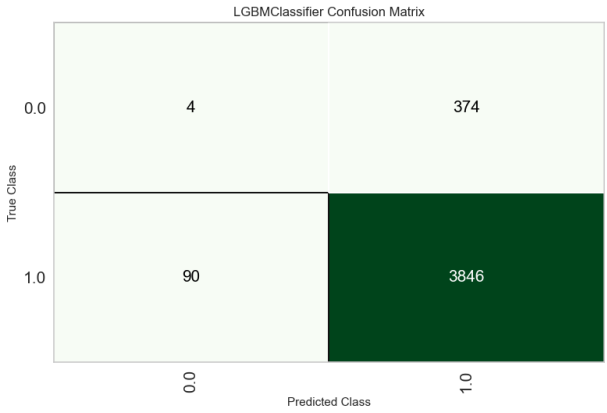
For both the general corpus and negative sentiment data, the results of each model yielded an ROC Curve of approximately 50%. This result indicates that the prediction is nearly random. The feature importance plot supports that the sentiment score was clearly the most important feature when classifying a review.

General Corpus

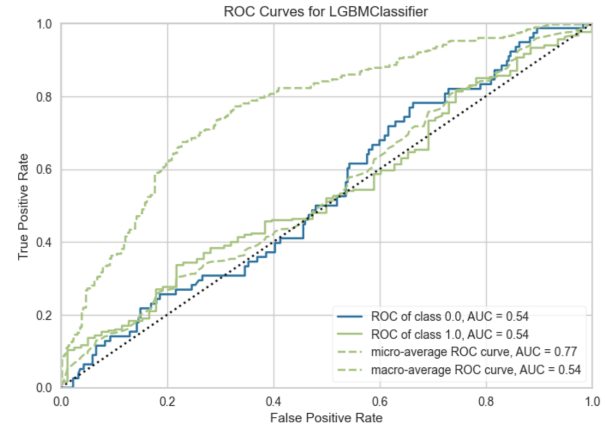


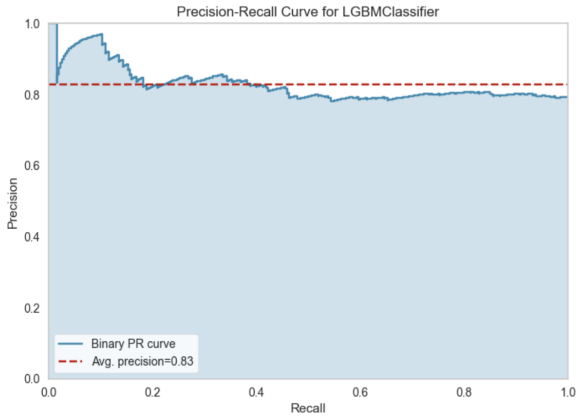


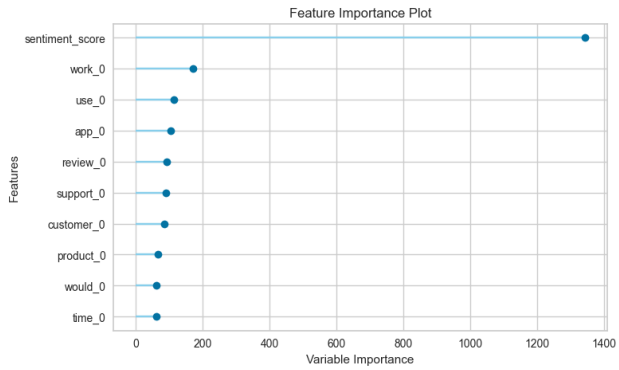


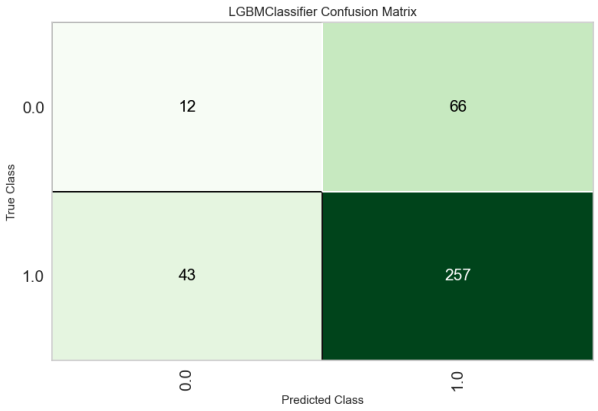


Negative Sentiment









Future Research

As the model currently exists, its predictive power is minimal. However, this can be improved upon in future iterations by expanding on the feature store of the dataset. Some potential features that are not currently available would be: app load times, update cadence, additional information on the apps features.

Although the model's predictive power is limited, the topic modeling can be used to automate extraction of insights which could provide developers with immediate feedback on their apps. This could help developers improve their app, which could result in more positive reviews in future updates. This feedback loop may improve customer satisfaction with Shopify's app marketplace.