Shopify App Store Capstone Analysis

By Matthew Smith

Problem Statement

Can app features such as app category, sentiment of review, or individual rating be used to accurately predict final ratings of apps?

Findings & Suggestions

Automated topic modeling can assist developers when determining which features end users value.

Further efforts should be made to gather data on predictive features such as:

- App Load Times
- Update Cadence
- Detailed App Features

Data Wrangling

Original Dataset: https://www.kaggle.com/datasets/usernam3/shopify-app-store

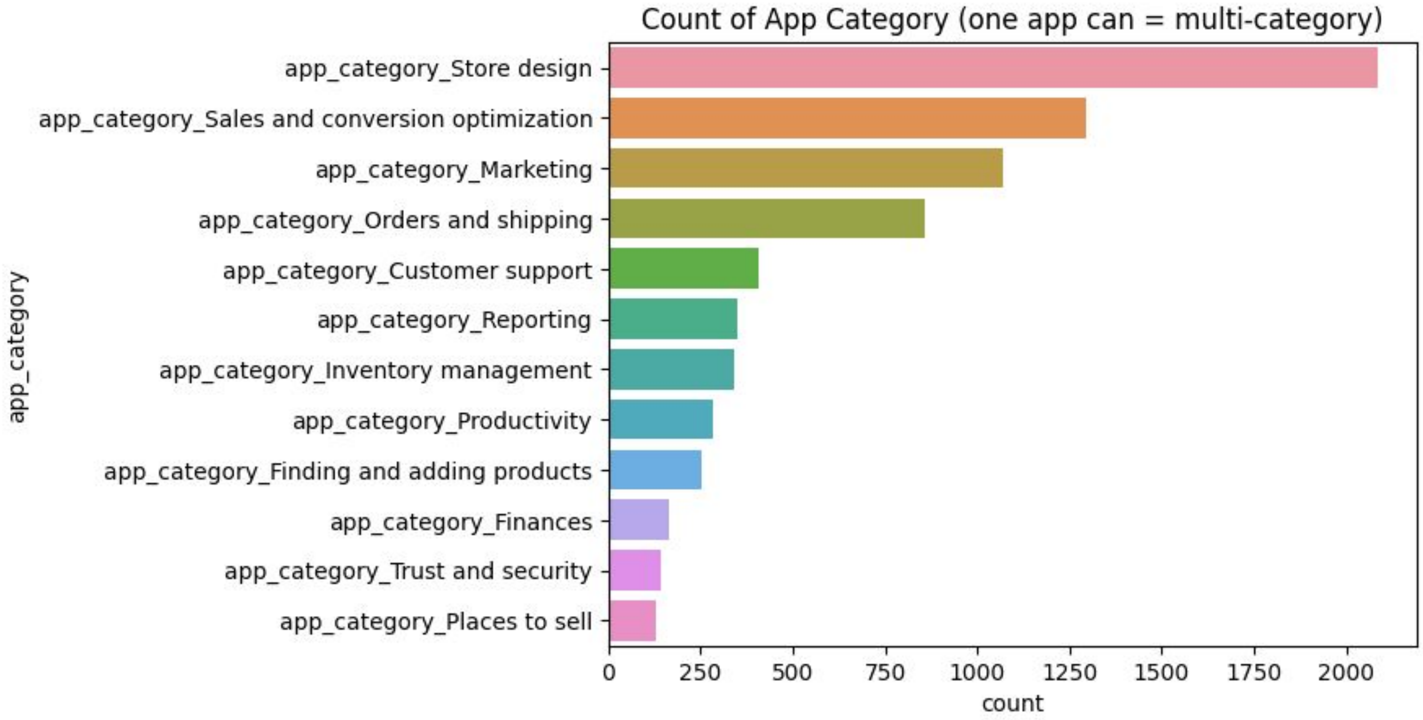
Key steps taken:

- Renamed variables for consistency between tables
- Created dummy variables to capture information from categorical variables without duplicating records

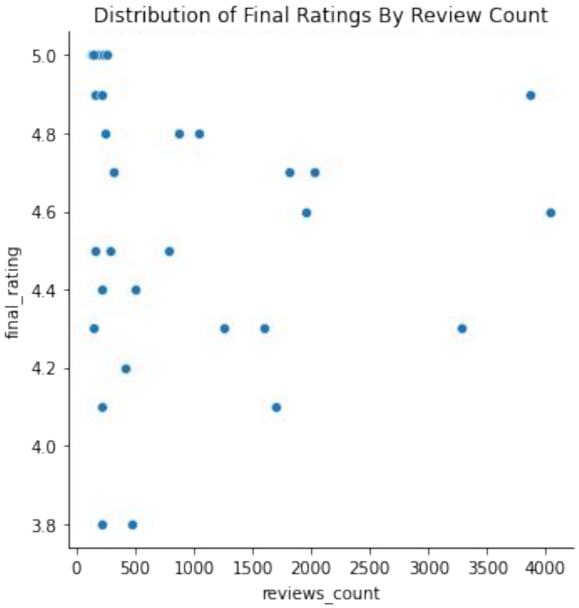
Constraints:

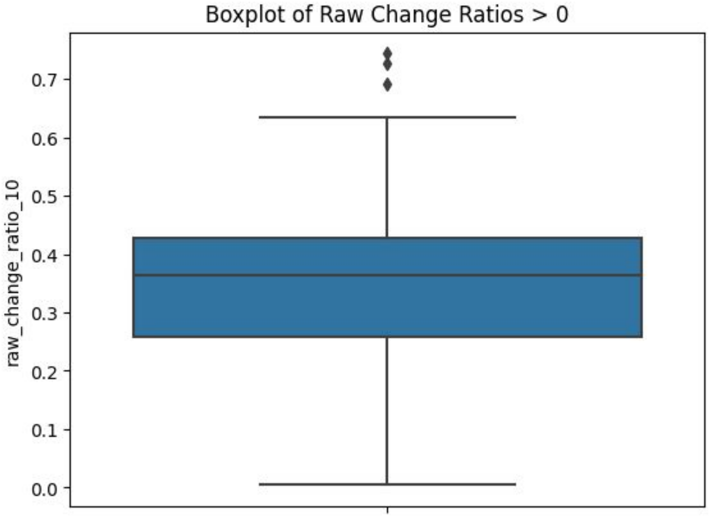
- Pricing data not used for this analysis
- First 20,000 reviews used for modeling and analysis
- Analyzed apps with 100+ reviews

Distribution of App Types on Marketplace

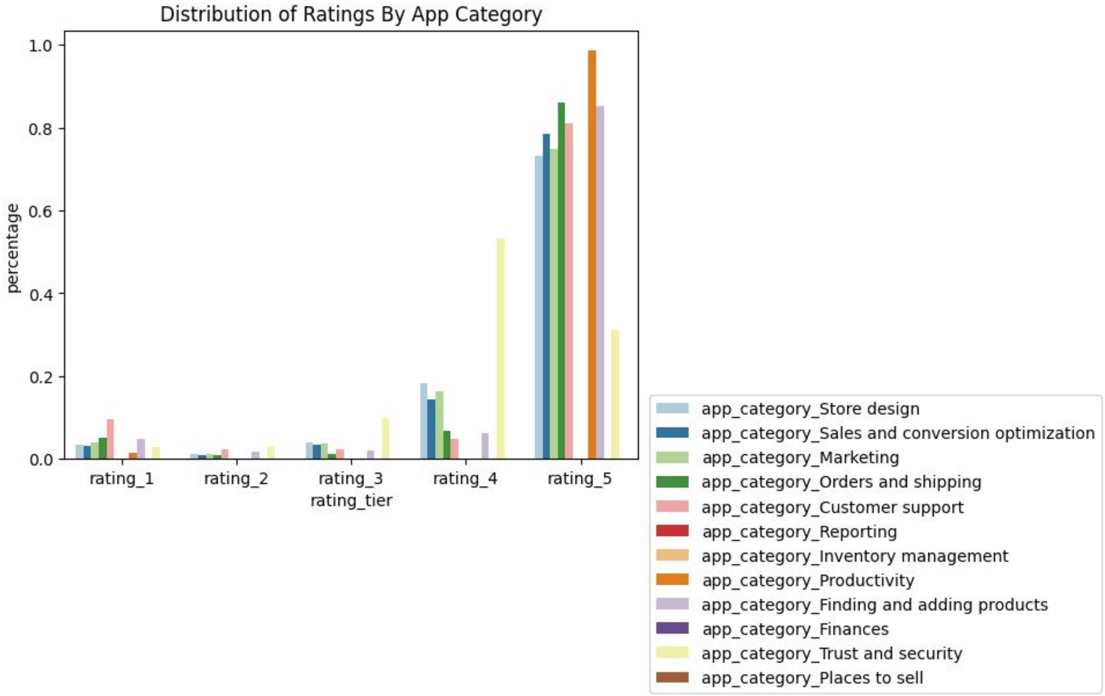


Exploratory Data Analysis

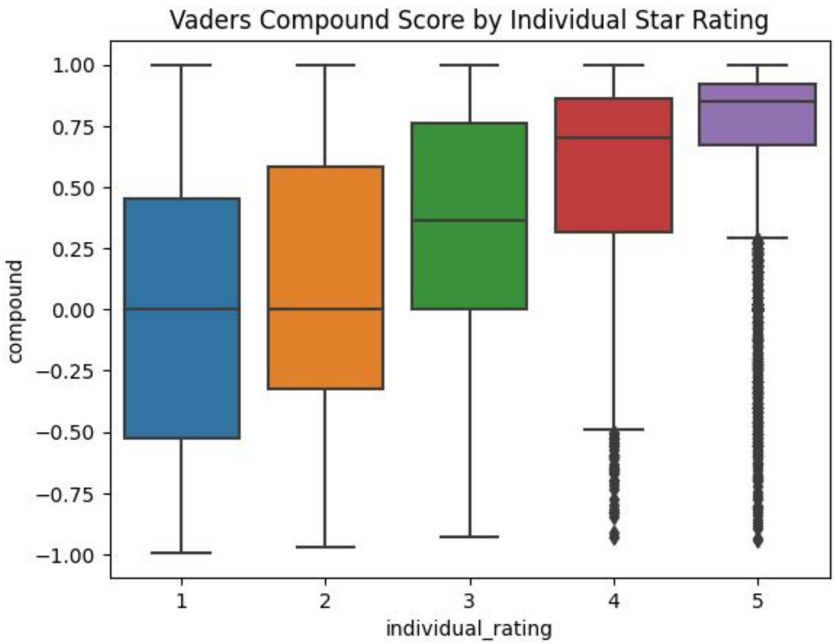




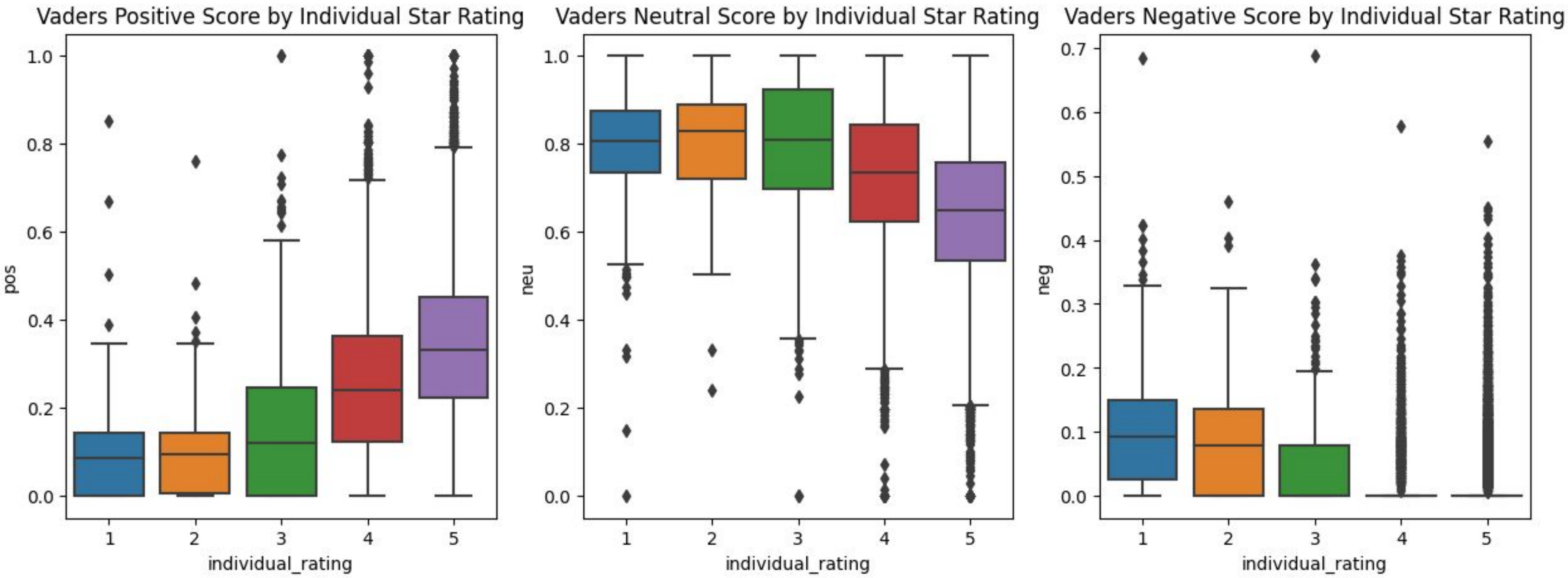
Exploratory Data Analysis (Continued)



Sentiment Scoring



Sentiment Scoring (Continued)

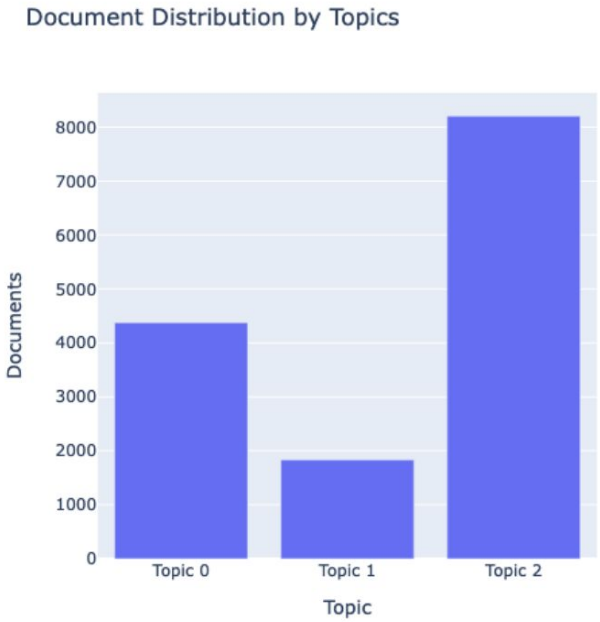


Vaders Positive Score by Individual Star Rating

Vaders Neutral Score by Individual Star Rating

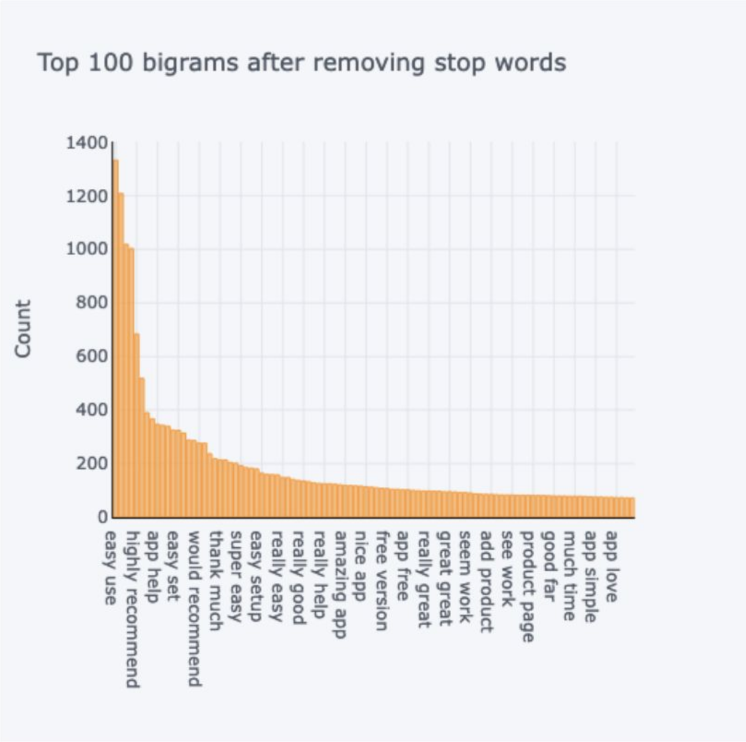
Vaders Negative Score by Individual Star Rating

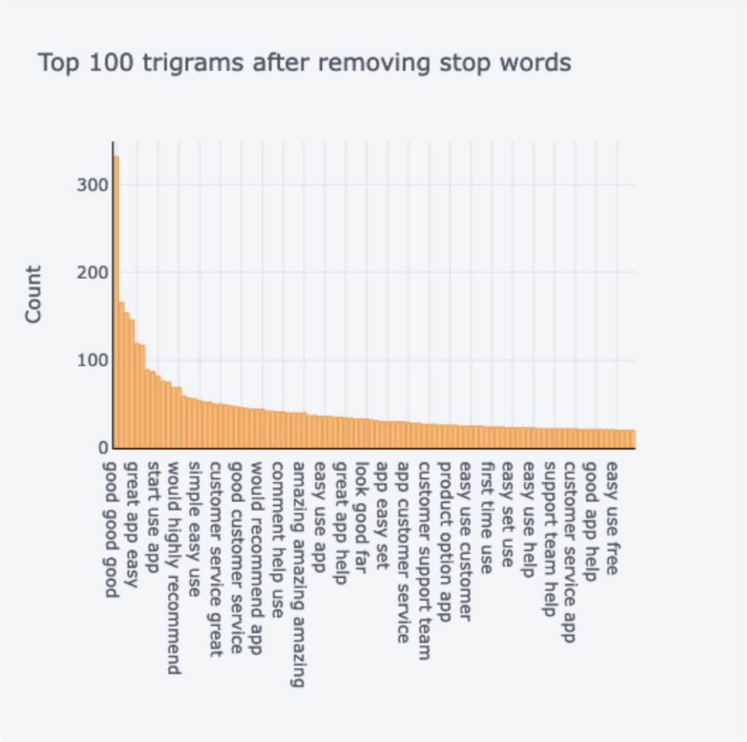
Topic Modeling



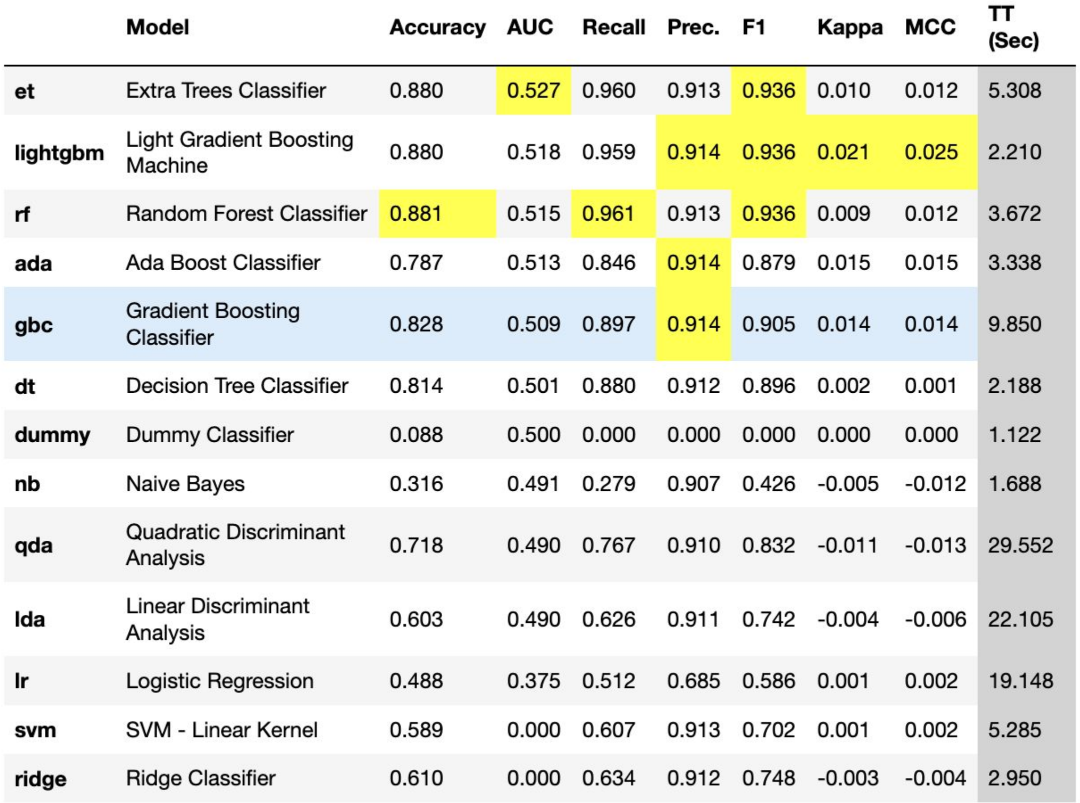
1) Topic 0: Customer Service
    a) customer, support, app, service, work, team, issue, get, review, help
    b) 4377 items in topic 0
2) Topic 1: Product/Page Modifications
    a) product, make, order, add, option, item, price, store, find, time
    b) 1833 items in topic 1
3) Topic 2: Usability/User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

Topic Modeling (Continued)


Top 100 bigrams after removing stop words


Top 100 trigrams after removing stop words

Model Selection



	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

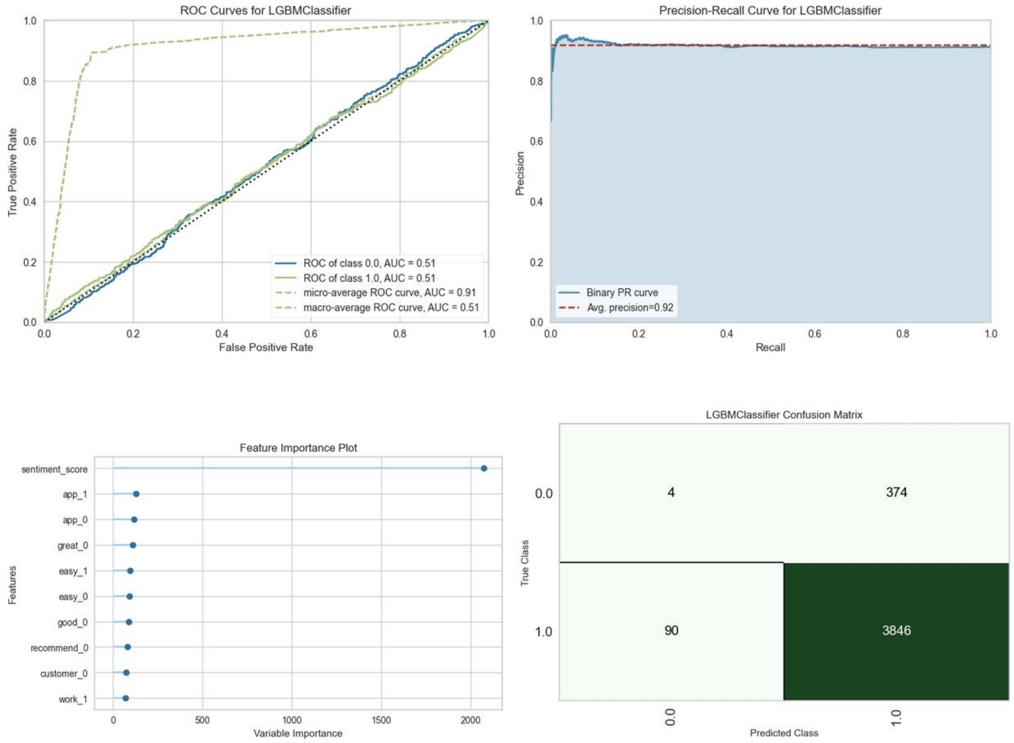
Takeaways

- ROC Curve approximately 50%

- Sentiment Score has highest variable importance

- Additional predictive features needed to boost model performance

General Corpus



Future Research

Dataset needs more predictive features to enable better classification. This data is likely already available but needs to be gathered and tagged appropriately.

Some features that would be helpful are:

- App load times
- Update cadence
- Additional information on the apps features

Questions & Answers