Predicting Review Likes on Yojo.com

MACHINE LEARNING REGRESSION ASSIGNMENT

ZAKRA CHACHAR
DANIELA JAIMES
YOHANES NUWARA

4 APRIL 2024

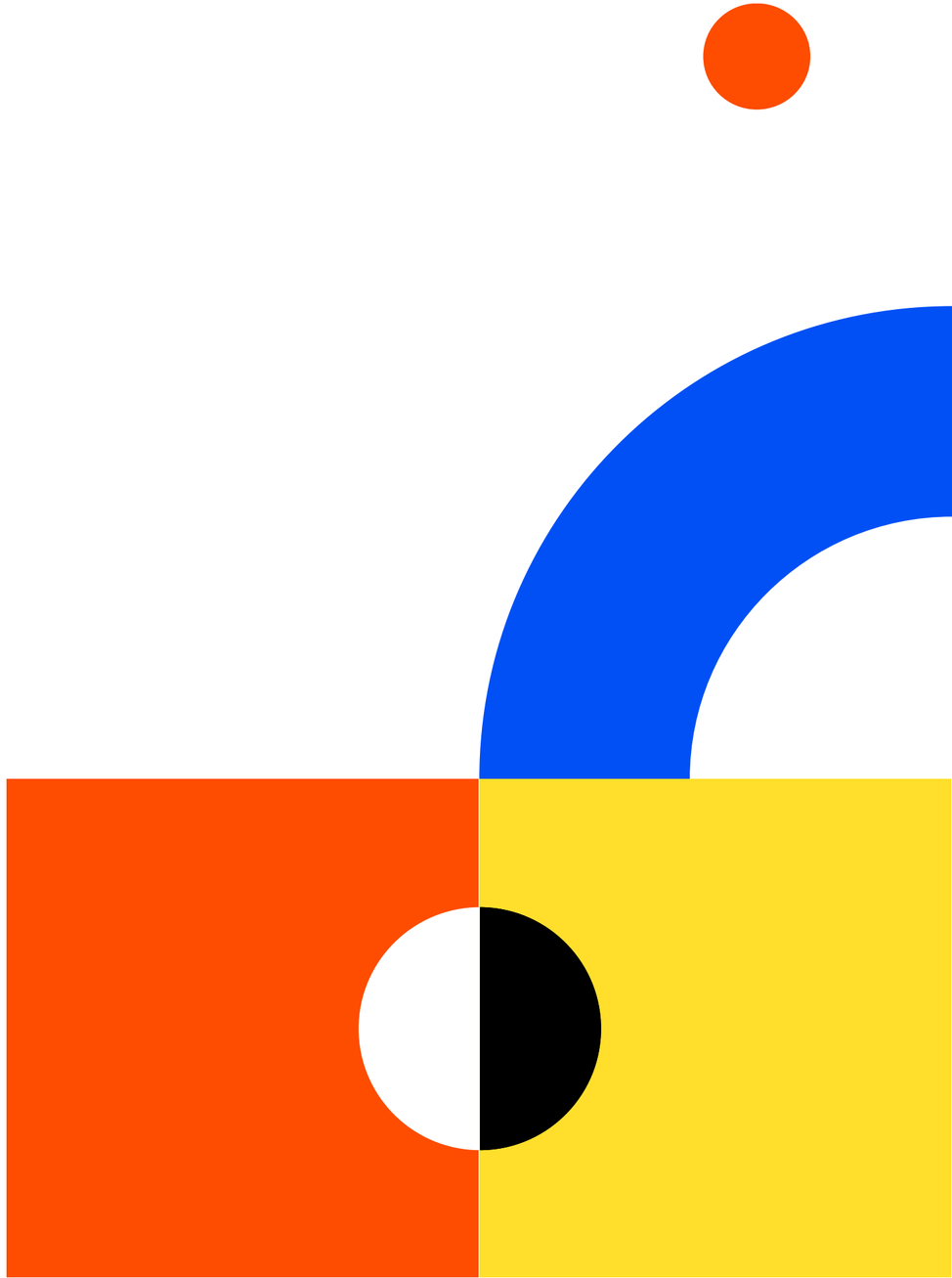
Agenda

- Introduction
- General Data Preprocessing
- Exploratory Data Analysis
  - Feature Analysis
  - Business Rules Based on Association Mining
- Model Selection and Training
  - Strategy #1
  - Strategy #2
  - Strategy #3
- Model Interpretation
- Conclusion

Introduction

This project uses machine learning to analyze textual features extracted from reviews.

- **Unveiling Customer Preferences:** Our goal is to identify key characteristics that influence how popular a review becomes.
- **Strategic Business Decisions:** These insights aim to inform strategic decisions leading to improved product visibility and customer engagement.



General Data Preprocessing

Outlier Management

Outliers were not removed for EDA and Feature Analysis

Missing Value Handling

No missing values were found in the data.

Feature Selection

Feature Selection was model-specific

Feature Engineering

- **Categorical Feature Encoding:** One-hot encoding was used to transform categorical features into numerical data.
- **Feature Scaling:** StandardScaler and PowerTransformer were used in different modeling strategies (details in next slides).

Exploratory Data Analysis (EDA)

Exploring Data Relationships:

- We used various techniques to understand the relationships between features and the target variable.

Feature Analysis:

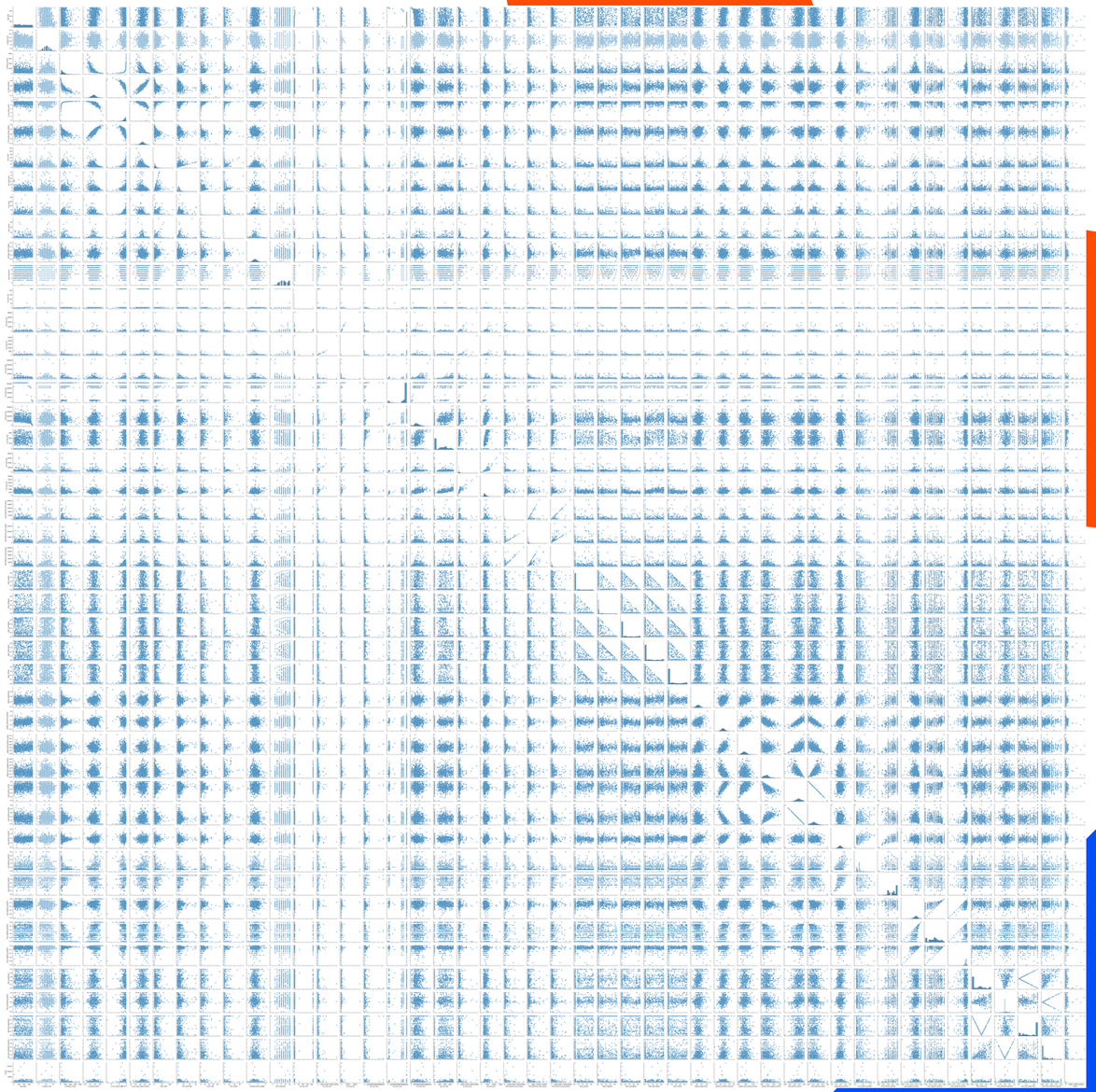
- Gained insights into the data's characteristics (47 features, 2 categorical, remaining numerical).

Feature Distribution:

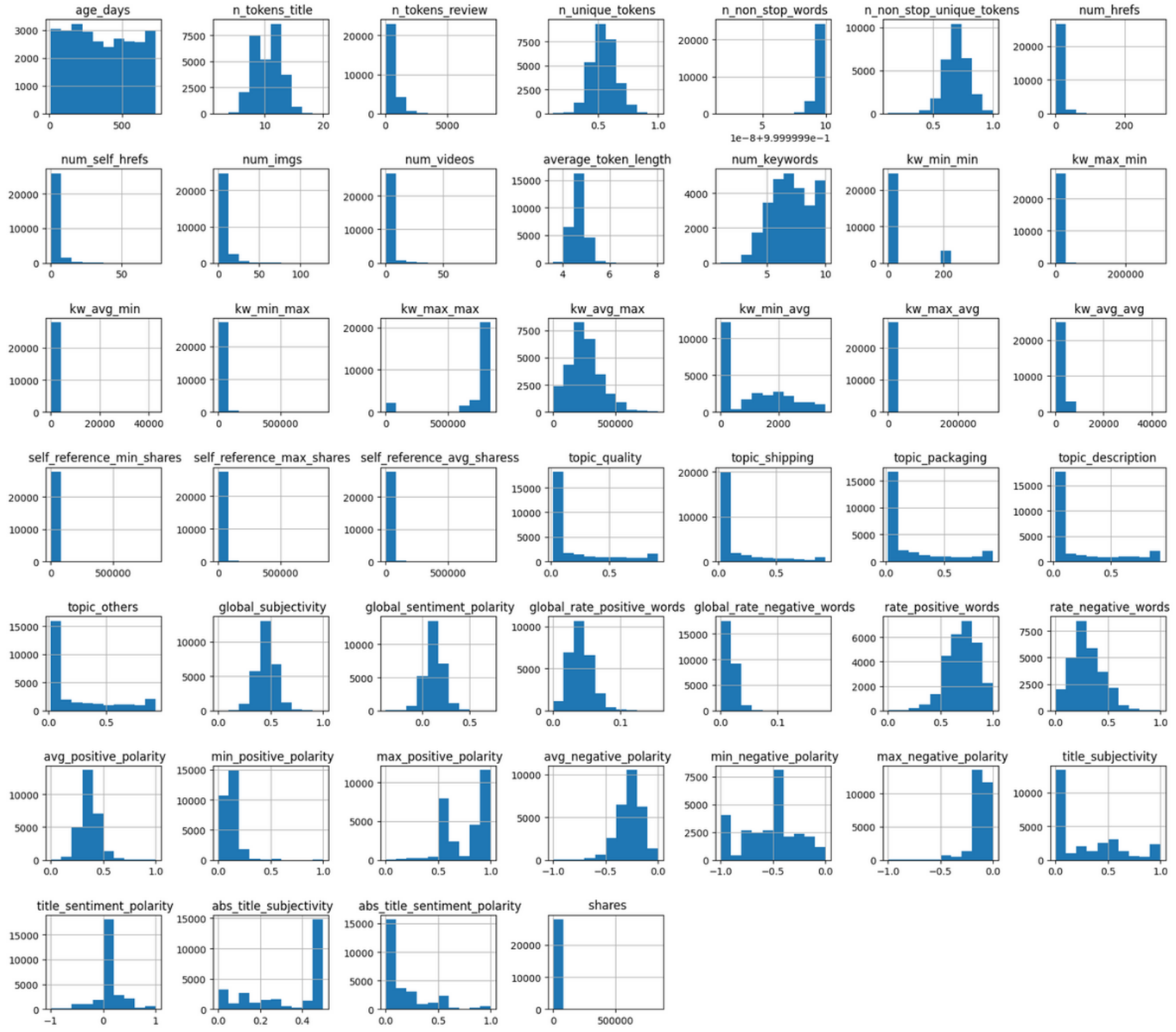
- Analyzed individual feature distributions with histograms (informed feature scaling decisions).

Outlier Detection:

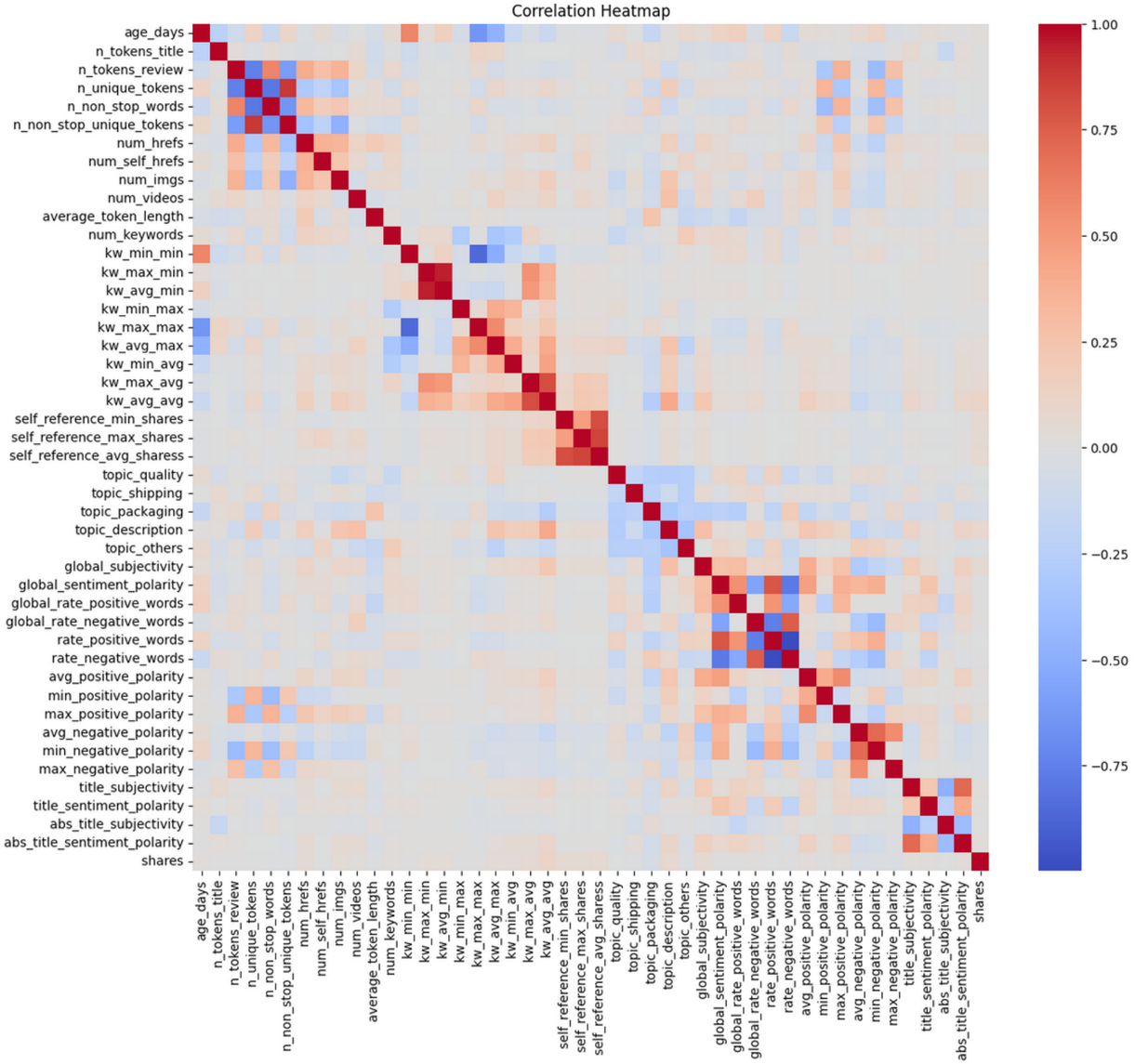
- Confirmed presence of outliers through boxplot analysis (considered for treatment).

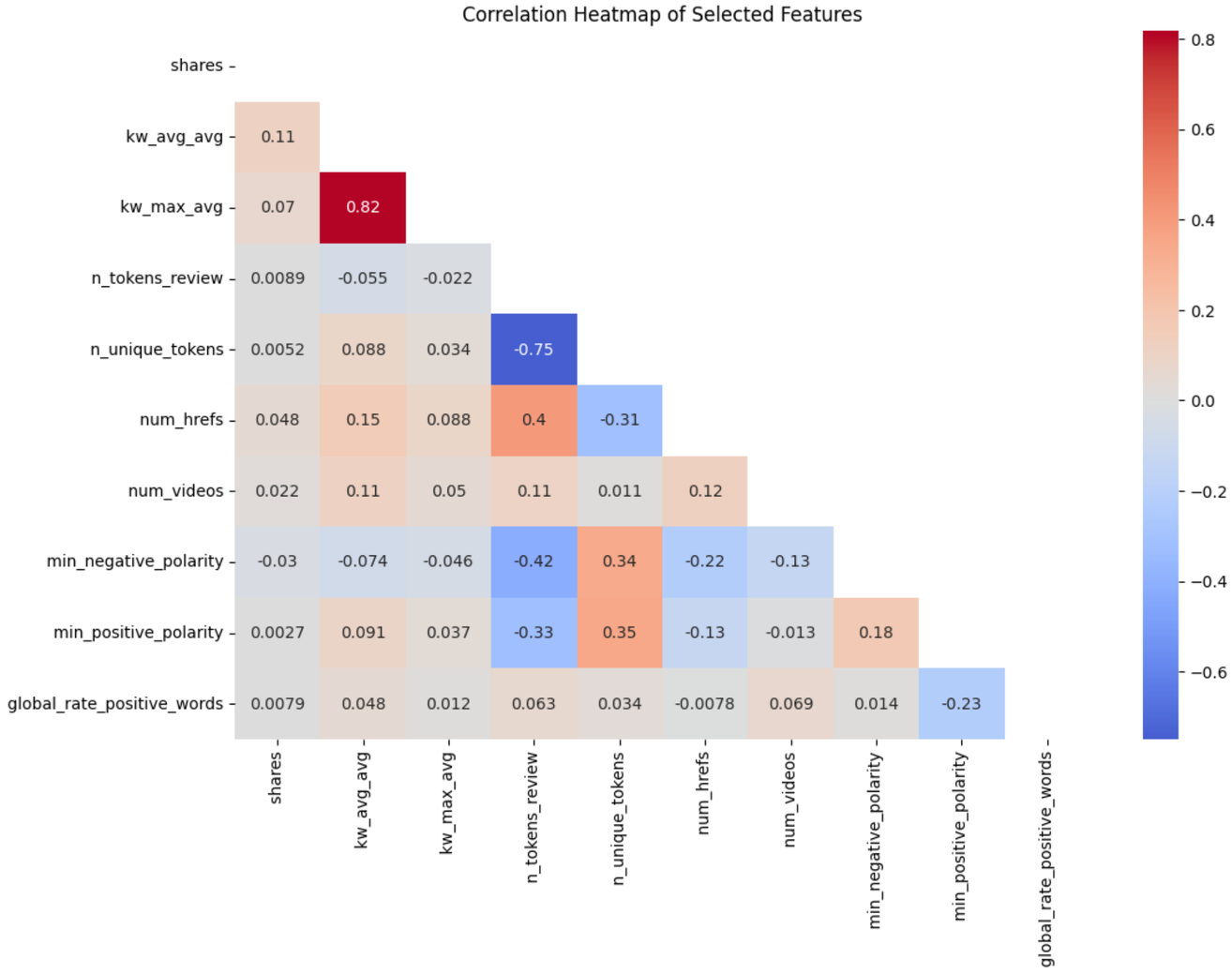


Correlation & Distribution



Correlation & Distribution





Category-Specific Insights

Business Rules Based on Association Mining

Exploring Category-Specific Trends

Employing association rule learning, we identified the most influential factors impacting shares in each category.

Actionable Business Rules

This analysis aimed to identify actionable business rules for improved performance within each category.

Identifying Relevant features

The results indicated which were the most relevant features affecting the target variable.

Category-Specific Insights

Business Rules Based on Association Mining

Category	Keyword Usage	Self-Referencing	Customer Reviews	Video Content	External Links
Entertainment	√	√ (weak)	x	x	x
Business	√	x	√ (weak)	√	√
Travel	√	√	√ (weak)	x	x
Sports	√	√ (weak)	√ (weak)	x	x
Tech	√	√ (weak)	x	x	x

Model Selection and Training

Multi-Stage Model Selection

- We explored different regression models to find the best fit for predicting review popularity.

Data Preparation

- Power Transformation: Improved data suitability for modeling.
- Correlation Analysis: Identified feature relationships.
- Outlier Removal: Removed outliers using one-class SVM.

Model Training and Evaluation:

- Category-Specific Analysis: Uncovered category-specific patterns using association rule learning.
- Feature Selection: Identified top 5 influential features.
- Model Training & Evaluation: Trained and evaluated regression models (Gradient Boosting achieved lowest MAE).

Strategy #1

Used PyCaret to evaluate different machine learning models, with Gradient Boosting showing the lowest Mean Absolute Error (MAE).

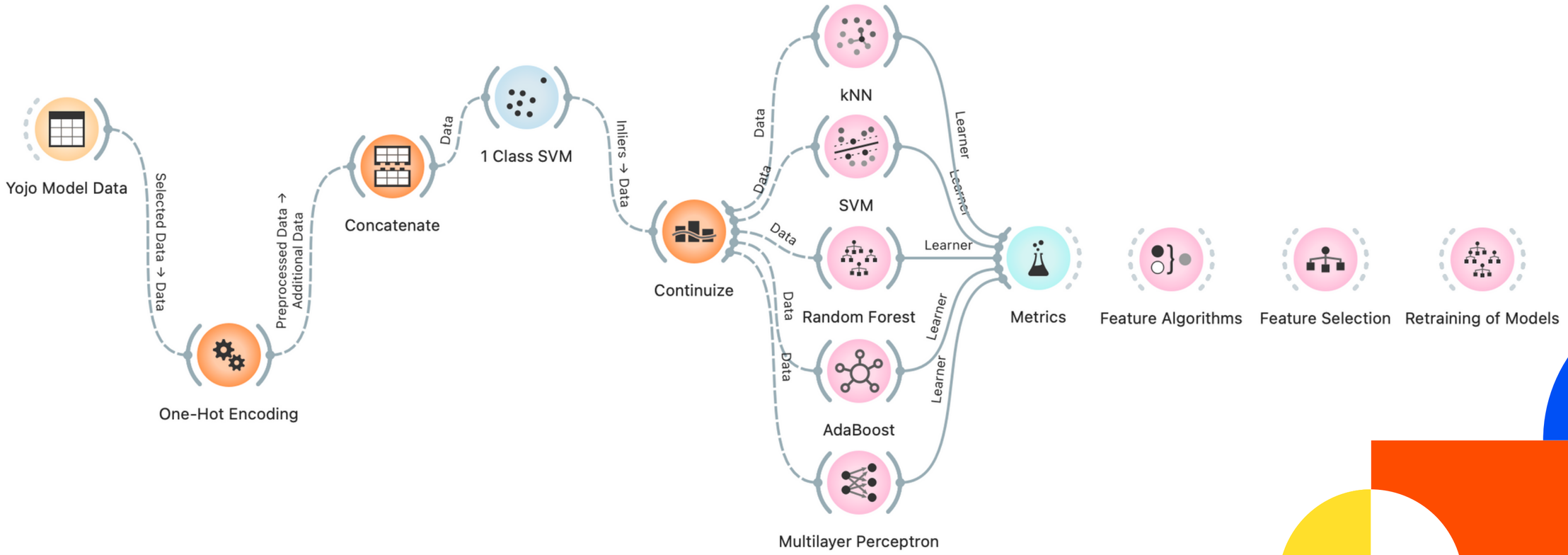
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.6354	0.6598	0.8121	0.1696	0.3990	3.8790	10.1070
lightgbm	Light Gradient Boosting Machine	0.6347	0.6636	0.8144	0.1648	0.3862	4.3800	0.9120
catboost	CatBoost Regressor	0.6344	0.6635	0.8144	0.1647	0.3875	4.4150	8.0340
br	Bayesian Ridge	0.6402	0.6691	0.8179	0.1577	0.4020	3.9136	0.0270
ridge	Ridge Regression	0.6399	0.6694	0.8180	0.1574	0.3995	4.0018	0.0260
lr	Linear Regression	0.6400	0.6696	0.8181	0.1572	0.3996	4.0040	0.5000
rf	Random Forest Regressor	0.6441	0.6719	0.8195	0.1544	0.3974	4.2126	36.2180
et	Extra Trees Regressor	0.6432	0.6731	0.8203	0.1528	0.3954	4.2384	11.5440
huber	Huber Regressor	0.6336	0.6761	0.8221	0.1489	0.3968	4.0471	0.2060
omp	Orthogonal Matching Pursuit	0.6609	0.6979	0.8352	0.1215	0.4184	3.6121	0.0220
xgboost	Extreme Gradient Boosting	0.6721	0.7365	0.8580	0.0729	0.3843	5.2828	1.4990
ada	AdaBoost Regressor	0.7179	0.7678	0.8761	0.0333	0.4014	4.4967	2.8340
lasso	Lasso Regression	0.7123	0.7958	0.8919	-0.0012	0.5681	1.0109	0.0220
en	Elastic Net	0.7123	0.7958	0.8919	-0.0012	0.5681	1.0109	0.0210
llar	Lasso Least Angle Regression	0.7123	0.7958	0.8919	-0.0012	0.5681	1.0109	0.0230
dummy	Dummy Regressor	0.7123	0.7958	0.8919	-0.0012	0.5681	1.0109	0.0300
knn	K Neighbors Regressor	0.6962	0.7981	0.8931	-0.0041	0.3894	5.4405	0.0950
par	Passive Aggressive Regressor	0.9221	1.3636	1.1626	-0.7167	0.4174	9.4358	0.0310
dt	Decision Tree Regressor	0.9205	1.4052	1.1852	-0.7692	0.4151	10.3845	0.5950
lar	Least Angle Regression	30.5742	18748.9531	44.2916	-24029.1285	0.8843	454.9368	0.0260

```
Processing:   0%|          | 0/85 [00:00<?, ?it/s]
```

```
GradientBoostingRegressor
GradientBoostingRegressor(random_state=123)
```

Strategy #2

Refined the model by converting categorical features, removing outliers, and selecting key features based on Random Forest importance.



Strategy #2: Feature Selection

Feature Names	Scores
kw_avg_avg	0.0689
topic_shipping	0.0432
kw_max_avg	0.0431
topic_packaging	0.0371
global_sentiment_polarity	0.0351
topic_others	0.0335
global_subjectivity	0.0318
average_token_length	0.0306
kw_avg_max	0.0281
avg_positive_polarity	0.0279
age_days	0.0275
topic_description	0.0271

Feature Names	Scores
self_reference_avg_sharess	0.0269
topic_quality	0.0263
global_rate_positive_words	0.025
global_rate_negative_words	0.0247
n_non_stop_unique_tokens	0.0238
self_reference_min_shares	0.0232
num_hrefs	0.0225
kw_avg_min	0.0218
avg_negative_polarity	0.0216
n_unique_tokens	0.021
self_reference_max_shares	0.0206
kw_max_min	0.0204

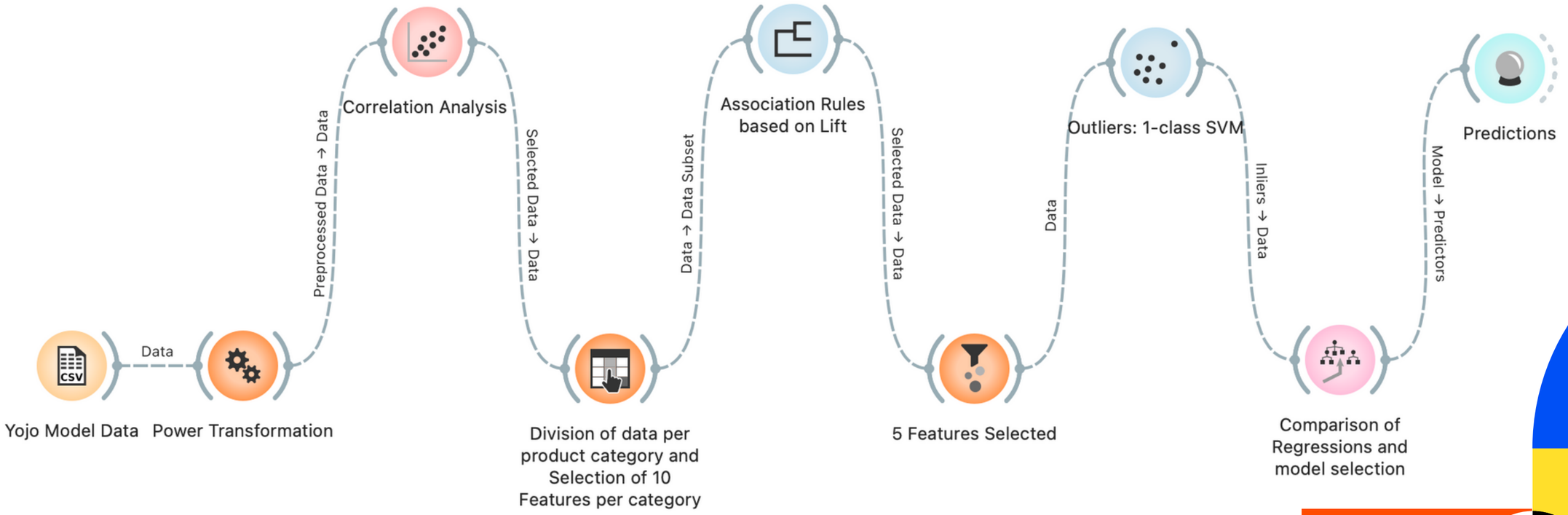
Strategy #2: Model Comparison

	Models	Parameters	Results
Before Feature Selection	KNN Regressor	Neighbours: 2-20, Parameters: 1, 2, 3	MAE train: 2365.9359440267335
			MSE train: 34097387.52236151
			RMSE train: 5839.296834582184
			r2: 0.34462507738848447
			MAE test: 2512.1529524777425
			MSE test: 45036905.84830196
			RMSE test: 6710.954168246269
			r2: 0.21218211868897974
	Random Forest Regressor	n-Estimators: 100	Mean Squared Error: 46172692.74060194
			R-squared: 0.020937617756944027
			Mean Absolute Error: 2875.0140021161224

	Models	Parameters	Results
After Feature Selection	KNN Regressor	Neighbours: 2-20, Parameters: 1, 2, 3	MAE train: 2436.7425886143933
			MSE train: 34468844.54969818
			RMSE train: 5871.017335155652
			r2: 0.10916628333982714
			MAE test: 2573.130621819726
			MSE test: 45493915.98639322
			RMSE test: 6744.917789446601
			r2: 0.03533064416587195
	Random Forest Regressor	n-Estimators: 100	Mean Squared Error: 46510620.388541035
			R-squared: 0.013772078379031383
			Mean Absolute Error: 2867.656944848565

Strategy #3

CHOSEN STRATEGY

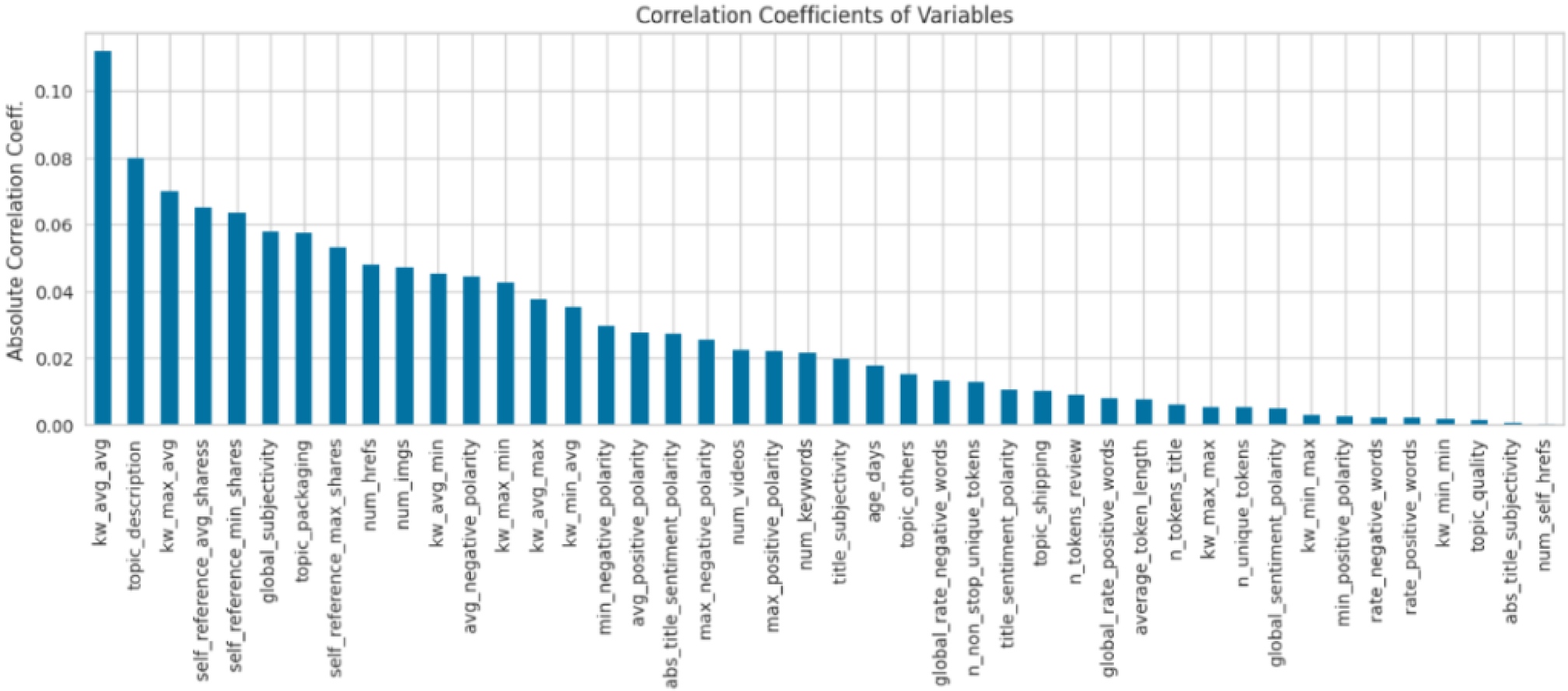


2 EXPERIMENTS IN THIS STRATEGY

- Developing one model for all category
- Developing separate model for each category

Strategy #3

CORRELATION ANALYSIS



FEATURE SELECTION

- Select 9 highest correlated features
- Calculate association rule mining and sort by Lift

Strategy #3

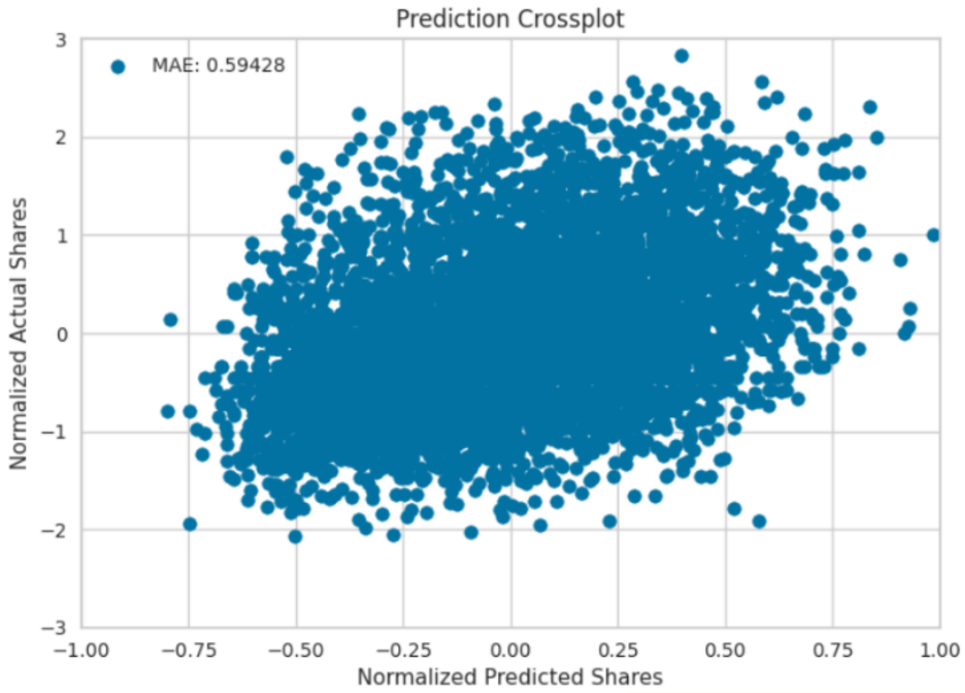
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
gbr	Gradient Boosting Regressor	0.6022	0.5760	0.7587	0.1436	0.3839	3.6169	2.9770
br	Bayesian Ridge	0.6062	0.5796	0.7611	0.1383	0.3906	3.4729	0.0170
lr	Linear Regression	0.6061	0.5797	0.7611	0.1382	0.3896	3.4968	0.5870
ridge	Ridge Regression	0.6061	0.5797	0.7611	0.1382	0.3896	3.4961	0.0160
lar	Least Angle Regression	0.6061	0.5797	0.7611	0.1382	0.3896	3.4968	0.0190
huber	Huber Regressor	0.6015	0.5842	0.7640	0.1316	0.3851	3.6380	0.0400
lightgbm	Light Gradient Boosting Machine	0.6045	0.5857	0.7651	0.1289	0.3769	4.0290	0.2620
catboost	CatBoost Regressor	0.6069	0.5888	0.7671	0.1244	0.3757	4.0891	4.7920
rf	Random Forest Regressor	0.6104	0.5918	0.7690	0.1200	0.3763	3.9475	9.9840
et	Extra Trees Regressor	0.6138	0.5989	0.7737	0.1092	0.3740	4.0985	3.5510
ada	AdaBoost Regressor	0.6348	0.6071	0.7790	0.0971	0.4014	3.2688	0.5120
omp	Orthogonal Matching Pursuit	0.6380	0.6250	0.7903	0.0708	0.4283	2.9282	0.0160
xgboost	Extreme Gradient Boosting	0.6344	0.6436	0.8020	0.0427	0.3702	4.6394	0.4660
lasso	Lasso Regression	0.6632	0.6734	0.8204	-0.0013	0.5259	1.2279	0.0160
en	Elastic Net	0.6632	0.6734	0.8204	-0.0013	0.5259	1.2279	0.0180
llar	Lasso Least Angle Regression	0.6632	0.6734	0.8204	-0.0013	0.5259	1.2279	0.0180
dummy	Dummy Regressor	0.6632	0.6734	0.8204	-0.0013	0.5259	1.2279	0.0220
knn	K Neighbors Regressor	0.6537	0.6841	0.8268	-0.0174	0.3649	5.1933	0.0740
par	Passive Aggressive Regressor	0.8627	1.2037	1.0871	-0.7863	0.4041	9.2285	0.0190
dt	Decision Tree Regressor	0.8642	1.2106	1.0999	-0.8017	0.3894	9.2551	0.1690

EXPERIMENT 3.1:
CORRELATION-DRIVEN FEATURE SELECTION

THE BEST MODEL: **GRADIENT BOOSTING**

GradientBoostingRegressor

```
GradientBoostingRegressor(learning_rate=0.04250053999178674,
                          max_features=0.5591636512007019,
                          min_impurity_decrease=1.2345158652175635e-07,
                          min_samples_leaf=4, min_samples_split=5,
                          n_estimators=245, random_state=123,
                          subsample=0.7415924079480812)
```



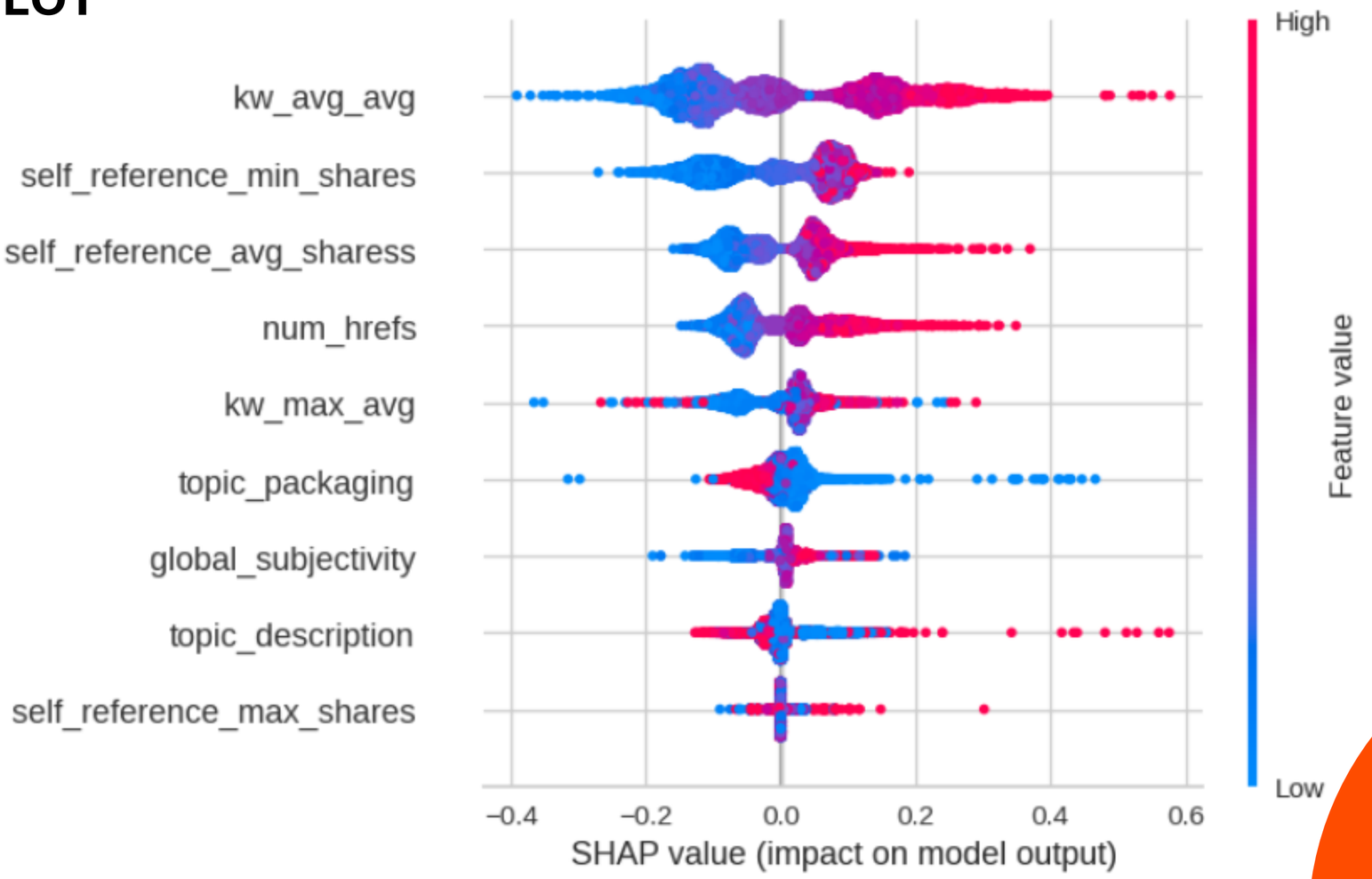
Strategy #3

EXPERIMENT 3.2:
PRODUCT CATEGORY-SPECIFIC MODEL

Category	Model	MAE (10-fold normalized)	Hyperparameters after tuning
Sport	Gradient boosting	0.6188	GradientBoostingRegressor(learning_rate=0.06038336630691028, max_depth=1, max_features=0.8354811479247907, min_impurity_decrease=1.5020779639698281e-07, min_samples_leaf=5, n_estimators=130, random_state=123, subsample=0.47953466882116846)
Travel	Gradient boosting	0.5461	GradientBoostingRegressor(learning_rate=0.015608951881726757, max_features=0.48102517310295906, min_impurity_decrease=9.182830657629517e-09, min_samples_leaf=4, min_samples_split=10, n_estimators=293, random_state=123, subsample=0.3669410765576156)
Tech	Gradient boosting	0.6390	GradientBoostingRegressor(learning_rate=0.04810016648012334, max_depth=1, max_features=0.7796767673507085, min_impurity_decrease=3.216567759382981e-08, min_samples_leaf=5, min_samples_split=6, n_estimators=217, random_state=123, subsample=0.2043783400695087)
Business	Gradient Boost	0.6141	GradientBoostingRegressor(learning_rate=0.032764035013193415, max_depth=2, max_features=0.5150643002168847, min_impurity_decrease=0.0013620836290466568, min_samples_leaf=3, min_samples_split=8, n_estimators=196, random_state=123, subsample=0.33620443949321777)
Entertainment	Gradient Boost	0.7310	GradientBoostingRegressor(learning_rate=0.02345748438403326, max_depth=1, max_features=0.7618622680626514, min_impurity_decrease=7.4662124572421e-09, min_samples_leaf=5, min_samples_split=3, n_estimators=235, random_state=123, subsample=0.20035816613002366)
Cleaning	Bayesian ridge regression	0.701	BayesianRidge(alpha_1=0.0003301199549309903, alpha_2=1.029851770894848e-09, compute_score=True, fit_intercept=False, lambda_1=0.9975973597393759, lambda_2=1.679686975465842e-06)
Other	Gradient boosting	0.8640	GradientBoostingRegressor(learning_rate=0.1048779636865015, max_depth=1, max_features=0.42105782577735423, min_impurity_decrease=2.1815170651677707e-08, min_samples_leaf=3, min_samples_split=6, n_estimators=156, random_state=123, subsample=0.37195141927520237)

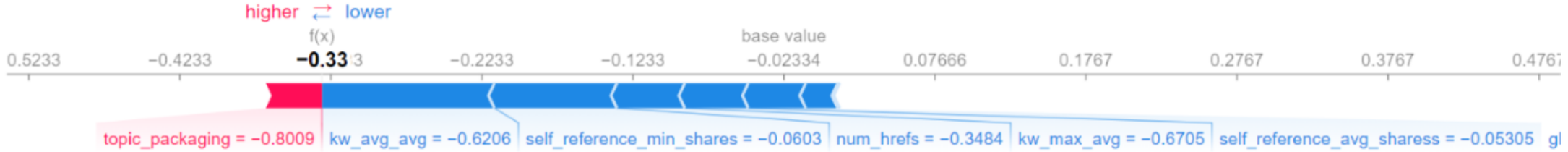
Model Interpretation (SHAP)

BEESWARM PLOT

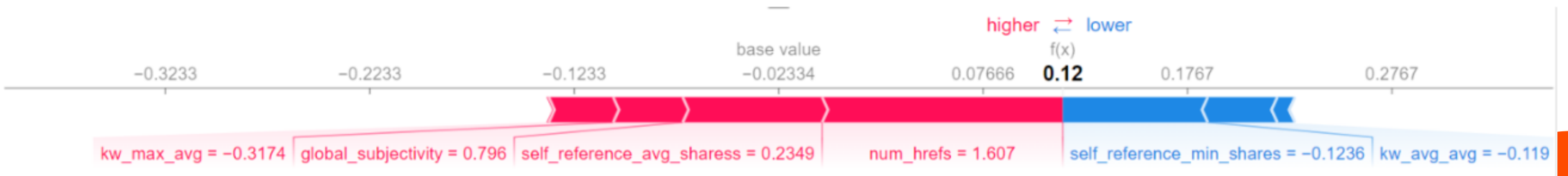


Model Interpretation (SHAP)

FORCE PLOT OF OBSERVATION #275 (LOW SHARES)



FORCE PLOT OF OBSERVATION #42 (HIGH SHARES)



Conclusions

BUSINESS RECOMMENDATIONS FOR IMPROVED PERFORMANCE

Keyword Research & Optimization

Conduct thorough research and use high-performing keywords in titles, descriptions, and metadata.

Content Quality & Self-Referencing

Develop high-quality content (text, images, videos) and explore self-referencing strategies where applicable.

Customer Reviews & Sentiment Analysis

Encourage reviews, respond to feedback, and utilize sentiment analysis to understand customer preferences.

Thank you!