spam review detection

Cell Phones and Accesories - Amazon Dataset

Amol Kaushik
kaushikamol@gmail.com

Customer reviews influence purchase decisions

Stop spam reviews reaching an audience

Purchase top quality products

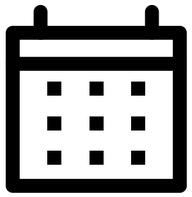
N. Hussain, H. Turab Mirza, I. Hussain, F. Iqbal and I. Memon, "Spam Review Detection Using the Linguistic and Spammer Behavioral Methods," in IEEE Access, vol. 8, pp. 53801-53816, 2020, doi: 10.1109/ACCESS.2020.2979226.

Amazon product data

Reviewer information

Review and Ratings

Date posted

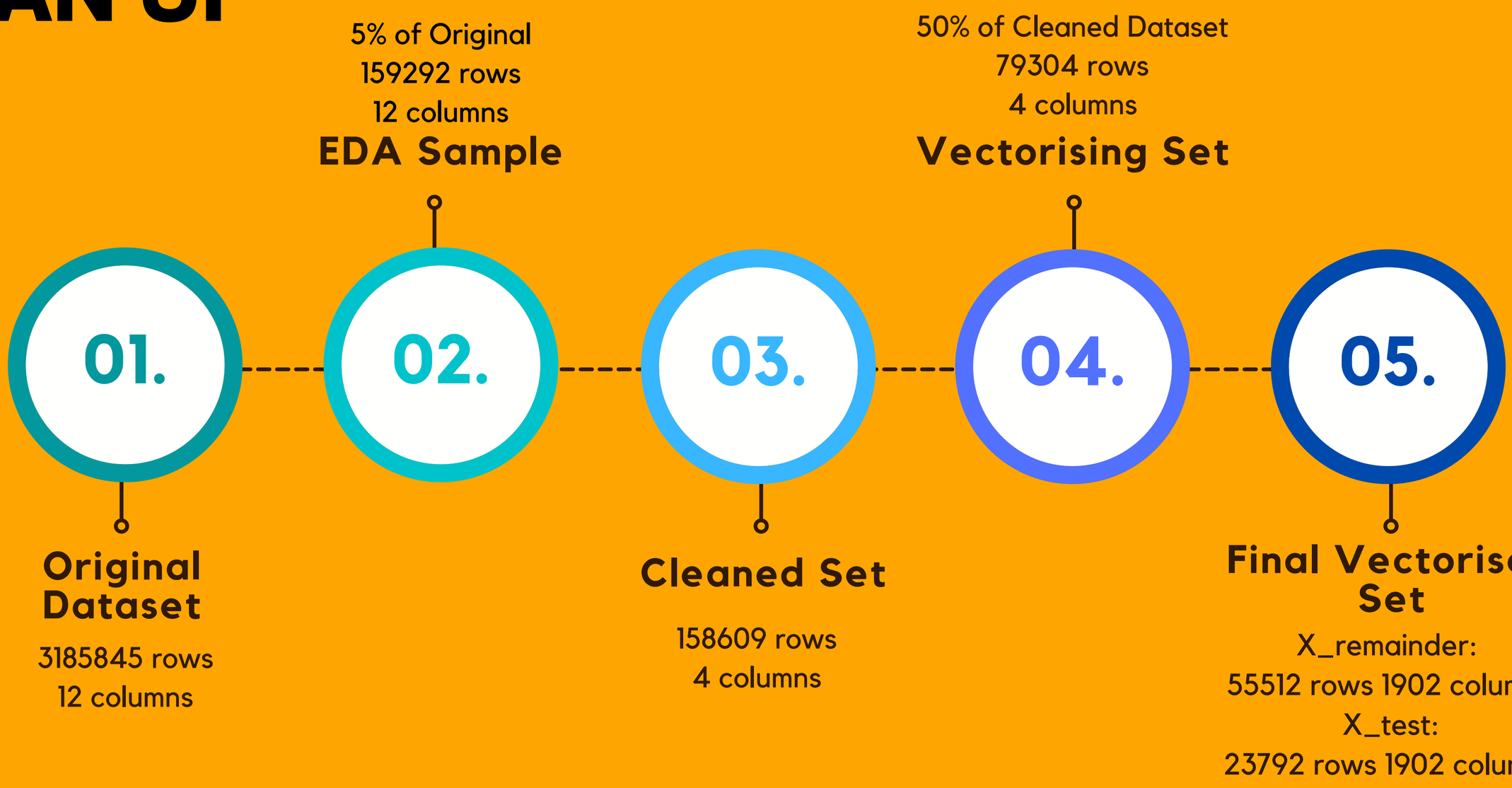


Classification

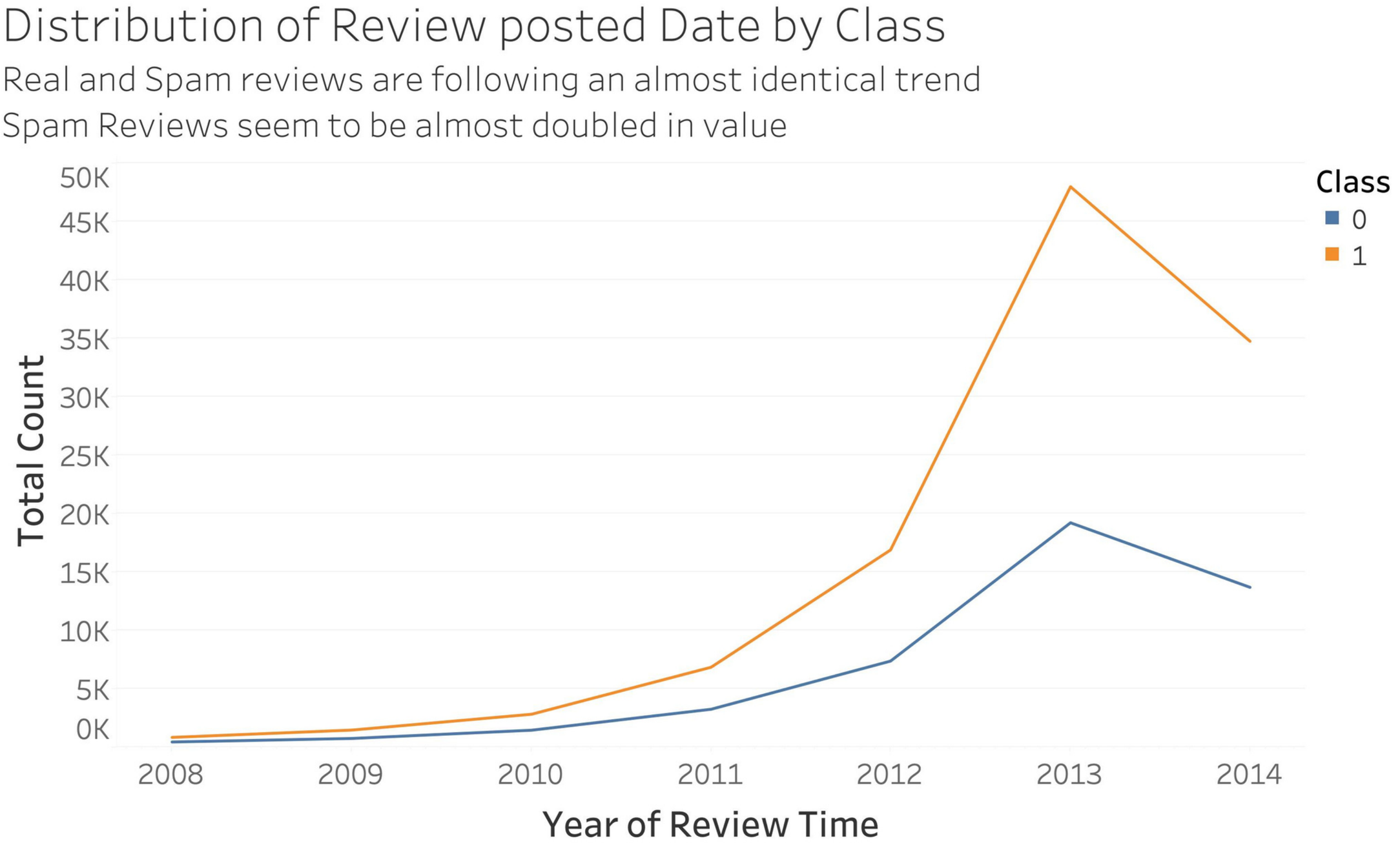
SPAM OR HAM...

- the volume control was always suspect. the sound always seemed fuzzy and unclear
- It was ok I guess - had it about a month and the front part broke on me already. Oh well
- The camo pattern had worn off the cover after 3 weeks! It is now white. I am returning it!!
- very light comes apart and breaks easily
- Cute but they don't stick
- These Otterbox case are great protectors and they look awesome as well. Neat and user friendly
- She really loved it and it is cute! Also it is really a great purchase for the price. I'm pleased.
- Perfect! Received faster than expected
- Good quality and great price.
- Awesome and very important.

EDA AND CLEAN UP



EDA/Modelling Insights



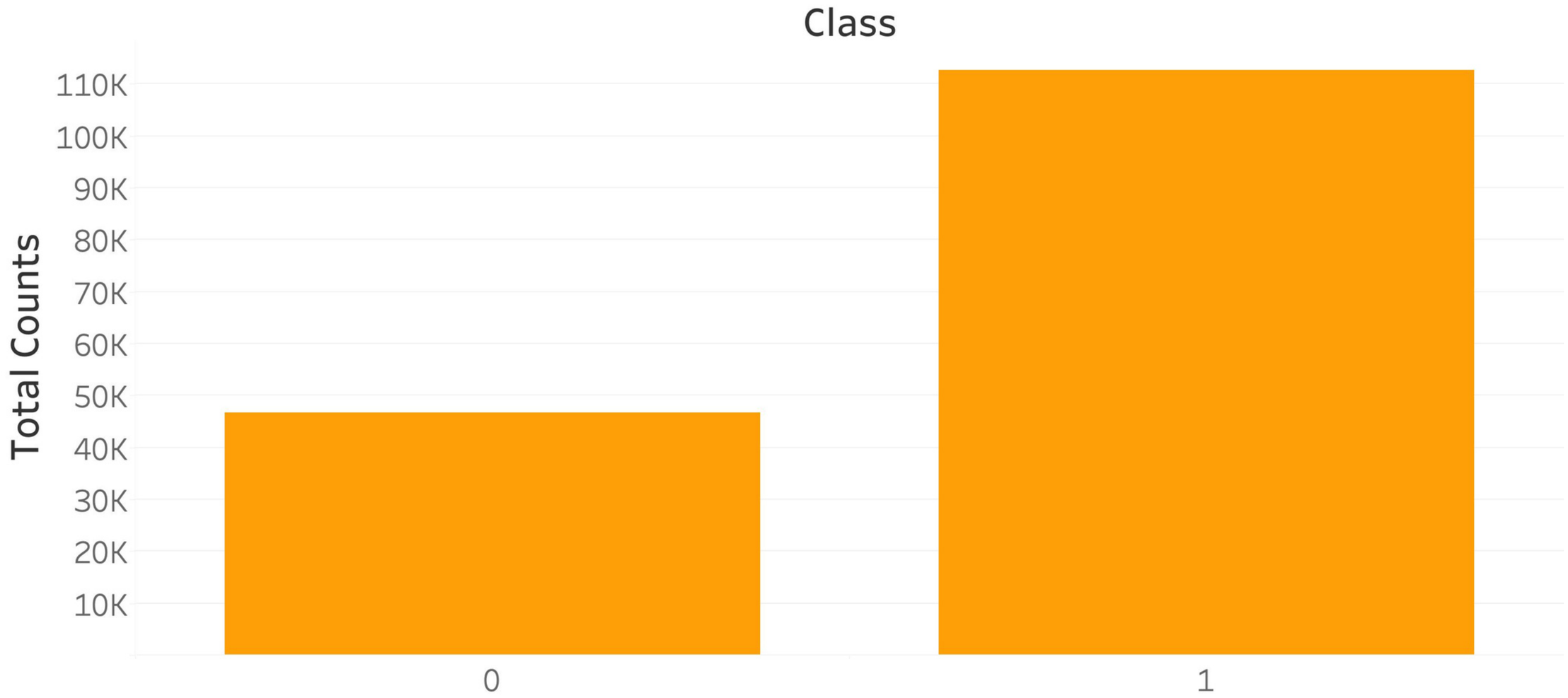
EDA/Modelling Insights

Distribution of Class

Dataset is imbalanced with a ratio of 7:3

70% of all reviews are **SPAM**

30% of all reviews are **REAL**



EDA/Modelling Insights

CountVectoriser Hyperparameters

Stop Words = "English"

Max Features = 1000

Min_df (reviews) = 40

Min_df (summary) = 15

SENTIMENT DRIVES SALES

LOVE	PERFECT	HIGHLY	BEST	PLEASED
GREAT	EXCELLENT	AWESOME	AMAZING	SATISFIED

OKAY	HORRIBLE	CHEAPLY	RETURNING	USELESS
BROKE	DISAPPOINTED	POOR	WASTE	RETURN

FINAL MODEL

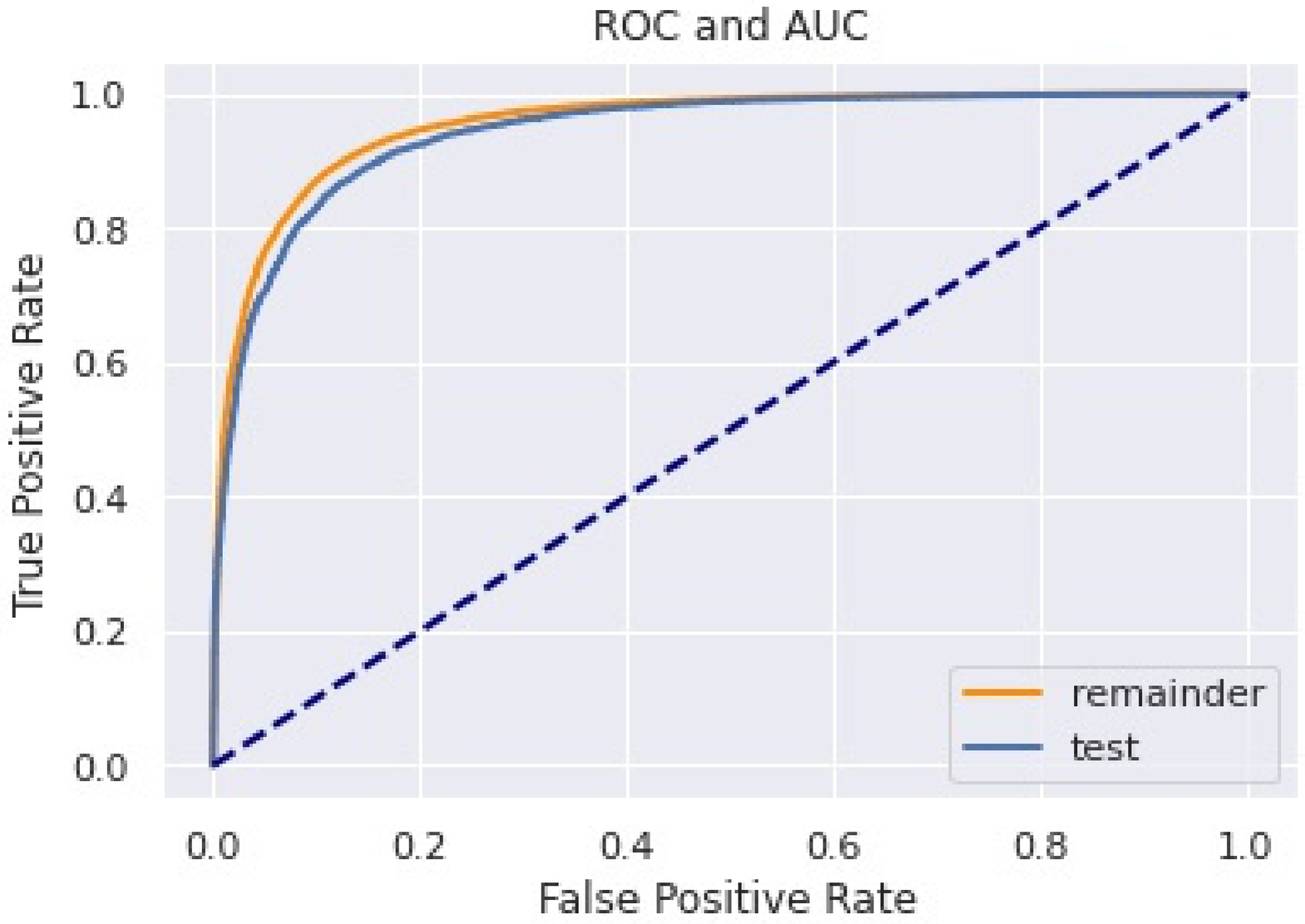
What was the best model?

Model	Default Model Parameters	Precision Score (%)	Recall Score (%)	F1 Score (%)
GS - Logistic Regression	C = 1, penalty = 'l1', random_state = 1, solver = 'saga'	90.27	93.42	91.82
GS - XGBoost	learning_rate = 0.5, max_depth = 8, n_estimators = 52, random_state = 1	88.11	93.57	90.75
Baseline - Logistic Regression	C = 1, penalty = 'l2', solver = 'lbfgs'	90.55	94.16	92.32

What was the best model?

Model	Default Hyperparameters	Remainder Accuracy	Test Accuracy
LogisticRegression	C = 1, penalty = 'l2', solver = 'lbfgs'	90.03%	88.98%

So the model works?



Remainder AUC Score = 0.95

Test AUC Score = 0.94

Next Steps

- Search for less biased data
- Engineer some features from the dataset
- Conduct a WordEmbedding analysis
- Conduct a Multi-Class analysis
- Run an extensive ML Pipeline with GridSearch (possibly via AWS due to computational limits with the local machine)
- Incorporate model into recommender system to filter products with suspicious review activity
- Create a Spam Review Detection Chrome Extention

Thank You!


kaushikamol@gmail.com


https://www.linkedin.com/in/amol--kaushik/


https://github.com/A-m-o-l-K