Amol Kaushik
kaushikamol@gmail.com

BrainStation Capstone:

Spam Review Detection

Amazon Review Dataset for Cell Phones and Accessories

December 2022

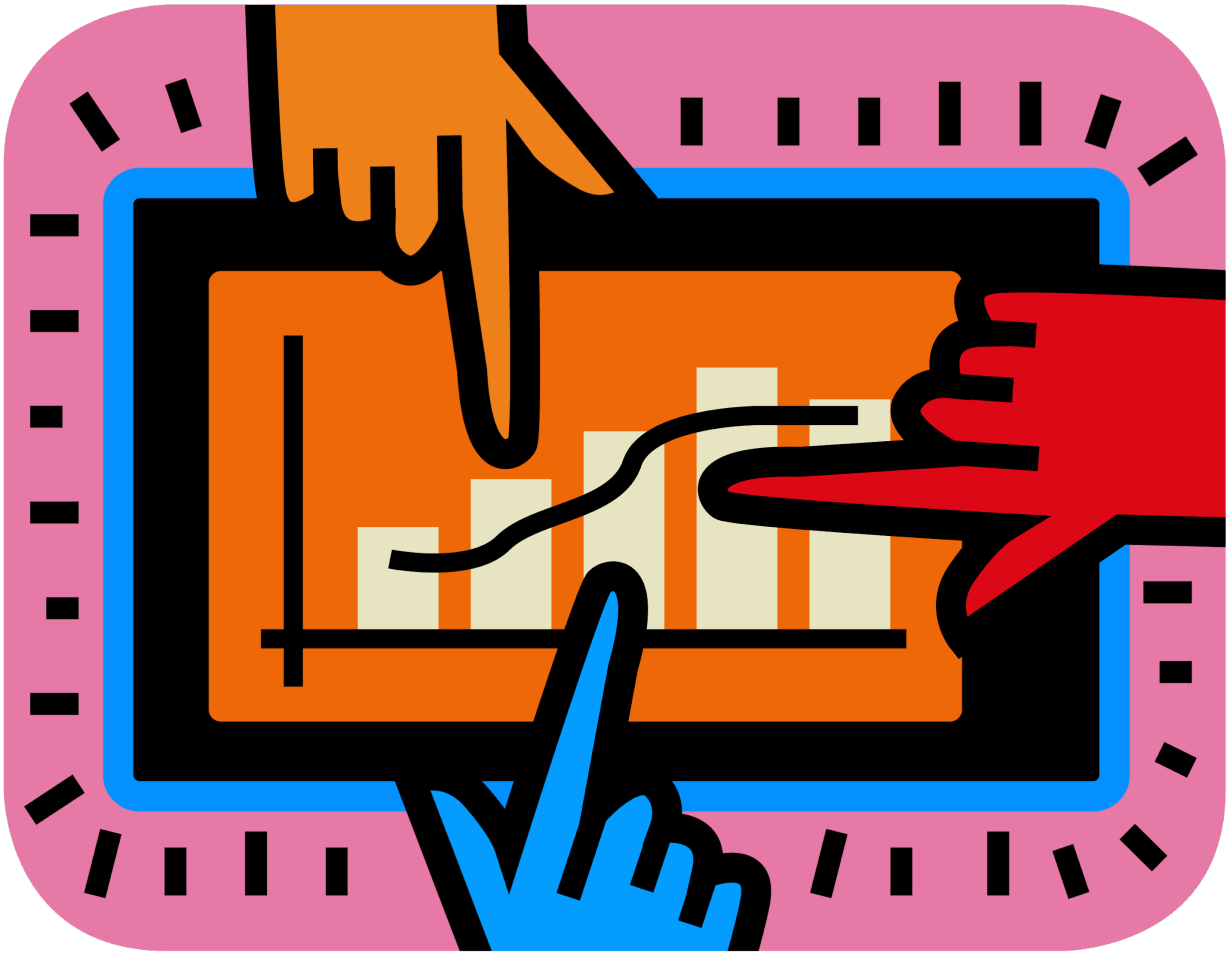

TABLE OF CONTENTS

INTRODUCTION

With the exponential rise of online shopping, especially via Amazon, there has been a steep increase in the number of fake/spam reviews posted for products. Fake reviews can be easily purchased through spam farms, whose sole purpose is to post as many spam reviews as purchased by the seller. While this is the biggest method for getting spam reviews to push your product further up the recommendation chain, another method used by sellers is by getting customers or their friends and family to post glowing reviews in exchange for a set discount on the product.

Capstone Objective and Value-Add

To build a Machine Learning model to identify and differentiate between spam and real reviews. By effectively identifying spam reviews, I aim to stop or significantly reduce the sale and recommendation of sub-par quality products.
With the implementation of my model, it will have a positive impact on the lives of Amazon customers in the following ways:

- Help customers only purchase top-quality products on the basis of real customer reviews
- Reduce the headache and time/money wastage of having to return or buy new products often due to bad recommendations and spam reviews
- Improve the company ratings, and image to only be associated with the best products and recommendations

Current Spam Review Detection Methods

Currently, Amazon has their own internal ML model for spam review detection which in 2020 stopped more than 200 million suspected fake reviews. They also take action against the sellers by shutting them down (source). Amazon via AWS also offers a fraud detection service which can be implemented by various companies in identifying bad actors. This service is already in use by companies like FlightHub, Qantas, and GoDaddy (source).

BREAKDOWN OF DATA

Data Source

https://www.kaggle.com/datasets/naveedhn/amazon-product-review-spam-and-non-spam
https://ieeexplore.ieee.org/abstract/document/9027828
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.

Data Format

The Dataset can be broken down into the following categories:

- Amazon product data
- Reviewer information
- Review and Ratings
- Date posted
- Classification

Data Collection and Classification

This is a real-world dataset collected by the authors for their own Spam Review Detection project/research model. The only issue with this dataset was that it was unlabelled, and the authors manually determined the spam/real classification, an overall rating score of 4 and higher was marked as spam, while an overall rating of 3 and below was marked real. This introduces a bit of a bias within the data since the classification was determined manually instead of having been classified from the start.

DATA CLEANING AND PREPROCESSING

Starting Point

The original dataset consisted of ~3.2M rows of reviews, this is a ton of data great for modelling, but due to the limitation in computing power on my local machine, a sampled dataset consisting of only 5% of the reviews from the original dataset was created.

Cleaning

The sampled dataset was checked for any missing or duplicated data. Some rows of data were dropped due to missing information for certain features. Certain features(columns) were converted into the correct data types for the Preprocessing/EDA (Exploratory Data Analysis). Any rows with missing data were dropped.

Data Exploration

The distribution of the data was checked. It was found that there was a 7:3 ratio of spam (1) to real reviews (0) present in this dataset (Figure 1). It was also found that there was a clear distinction between spam and real reviews based on the overall rating. All spam reviews had an overall rating of 4 and higher, while all real reviews had a rating of 3 and below (Figure 2).

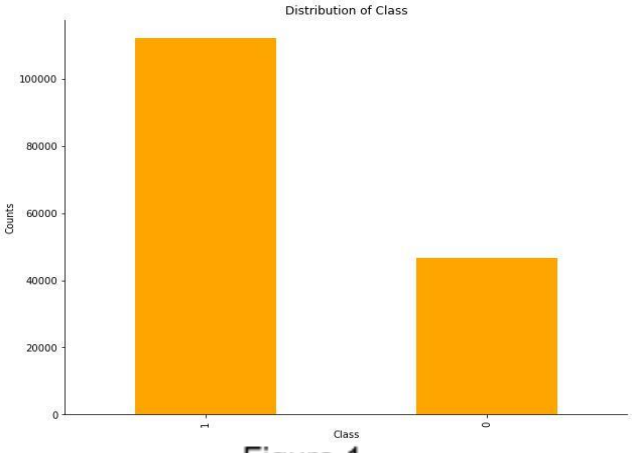

Figure 1

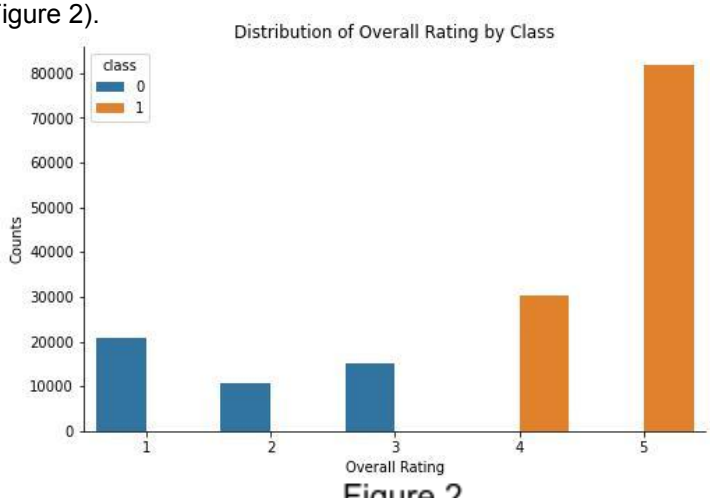

Figure 2

Correlations between the features were checked to determine the level of importance/influence certain features might have on the classification, but no such correlations were found, except with the Overall Ratings, due to this high level of influence on the classification the overall column was dropped. No seasonality or trends were found between the reviews and the dates they were posted. The time of the post was not available but this might have given some insight as to when the spam reviewers are mostly active in comparison with the real reviewers. Figure 3 shows the distribution of spam vs. real reviews by the dates they were posted, and looks to follow the same trend.

The following categories of data were dropped:

- Amazon product data
- Reviewer information
- Date posted

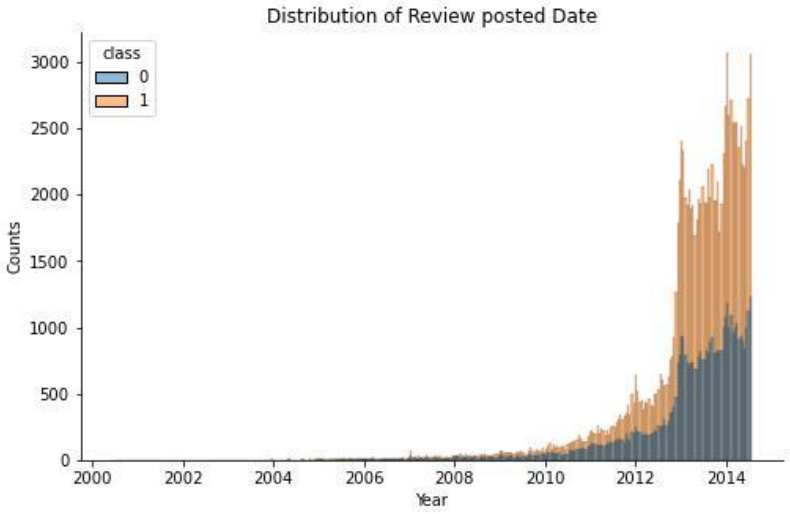

Figure 3

MODELLING AND INSIGHTS

Modelling procedure:

- Data Transformation (splitting up columns)
- Vectorization (converting the unique words from the summary and reviews into individual features)
- Scale Data
- Run Model
- Check Model
- Manual Model Hyperparameter Optimization
- Save the final vectorized dataset (used in advanced modelling)
- Advanced Modelling (ML Pipeline and GridSearch)
- Best Model Evaluation

Models Used:

- Logistic Regression
- SVM (Support Vector Machines)
- Random Forest
- XGBoost

The main reason for selecting these specific models is that they work very well with big datasets that have high dimensionality (lots of features), and they can be easily fine-tuned/optimized to get the best result possible.

Models Evaluation Procedure:

- Score the model for accuracy on the dataset
- Check the confusion matrix to identify the recall (how many spam reviews out of all spam reviews were identified correctly) and precision (how many reviews were correctly predicted to their actual class out of all the predictions for that class)
- Check the variance of the precision and recall based on the adjusted classification threshold (the min coefficient value that helps determine if the review is spam or real)
- Plot Receiver Operating Characteristic and Area Under the Curve (this helps identify the probability that the model will correctly classify any spam review chosen at random)

Insights from Modelling

It was interesting to see that after splitting up the reviews into unique words, the words with the highest indication of a spam review were very positive in nature (Figure 4), while the exact opposite can be said about the real reviews. The words most indicative of real reviews were mostly negative or neutral in nature (Figure 5).

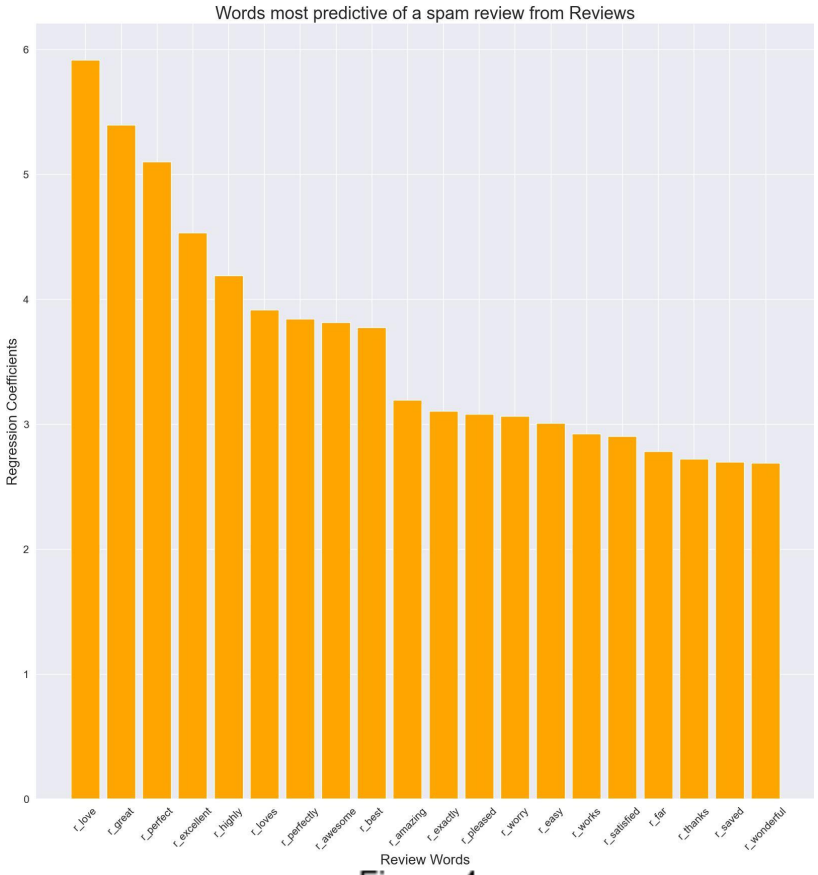

Figure 4

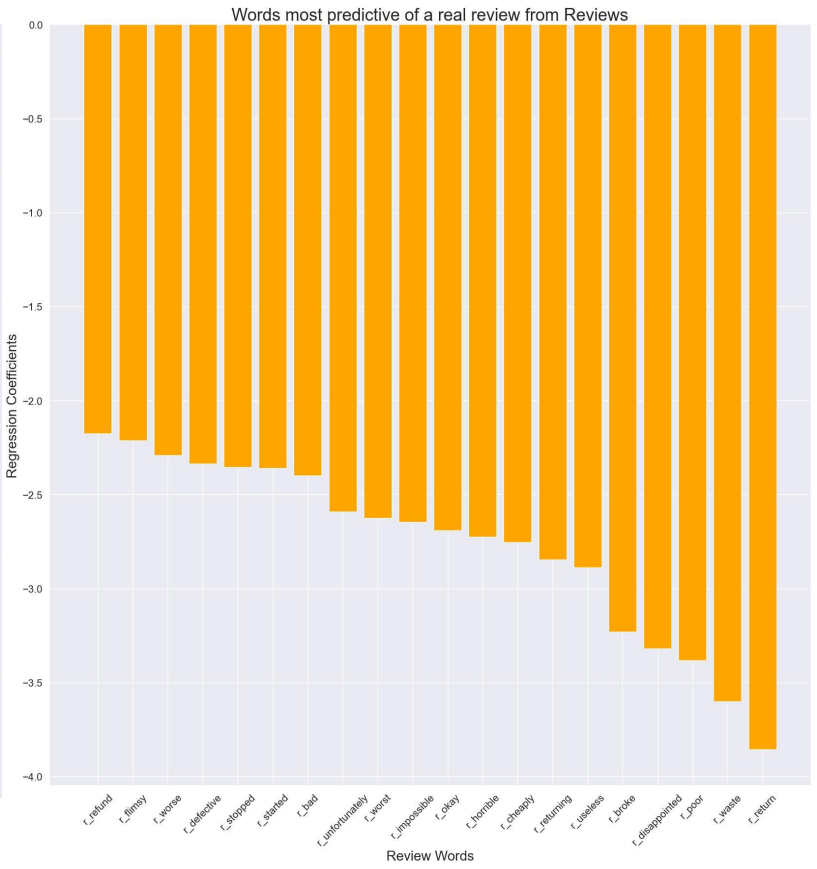

Figure 5

Final Model and Results

After running multiple models with a GridSearch for the best model with optimal hyperparameters, and evaluating the best results. The Baseline Logistic Regression Model with the default hyperparameters turned out to be the best-performing model. The Model Results were:

- Remainder Set Accuracy = 90.03%
- Test Set Accuracy = 88.98%

CONCLUSION

Based on all of the modelling conducted, the Baseline Logistic Regression model with the default hyperparameters was the most accurate and successful at identifying spam reviews in the dataset.

Moving forward, these are some of the improvements I would make to this project:

- Search for a less biased/manually classified dataset
- Conduct more feature engineering
- Conduct a Word Embedding analysis
- Run a GridSearch with more models and hyperparameters (possibly on AWS)