



FROM TEXT TO TRUST: CLASSIFYING UNVERIFIED PURCHASES WITH NLP

CIS 9665 Final Report

Abstract

This study examines the classification of verified and unverified reviews in the Amazon Appliance Reviews dataset using Natural Language Processing (NLP) techniques and machine learning models to enhance decision-making and customer trust in e-commerce, creating tools with both commercial and academic applications.

Contents

Dataset, Project Motivation and Goals.....	2
Introduction to Dataset.....	2
Research into the Impact and Implications of Online Reviews	2
Research Question and Goals	3
Natural Language Processing and Model Creation.....	3
Text dataset description	3
Text Preprocessing Description	4
Natural Language Features	4
Conclusion and Discussion.....	5
Addressing Imbalanced Data	5
Models and Evaluation of Performance	5
Evidence-Based Conclusions and Practical Implications.....	5
References	7
Appendix.....	8

Dataset, Project Motivation and Goals

Introduction to Dataset

The University of California San Diego Rady School of Management hosts a wide variety of datasets ranging from Food.com paired recipes, Tradesy bartering data, EndoMondo fitness tracking data to Amazon reviews and many other topics (Himabindu Lakkaraju, 2013).

For this project, the Amazon Reviews dataset was selected. The complete dataset contains 571.54 million individual reviews which were gathered between the years 1996 and 2023. To refine the analysis and prevent potential computational limitations, the subset of Appliance reviews was selected. This subset contains around 2.1 million reviews by 1.8 million users which were made for almost 95 thousand items.

Research into the Impact and Implications of Online Reviews

According to Amazon, of the approximately 7.5 billion items sold on its platform in 2023, third-party sellers contributed “more than 60% of sales in [its] store”, totaling more than 4.5 billion items” (Amazon, 2024). The company and its many rivals have created a competitive marketplace for large and small vendors where understanding customers’ needs and complaints better can give sellers an edge over their competitors. As a result, the volume of studies on the impact of online reviews, also referred to as online word of mouth (OWOM), has rapidly expanded over the past two decades from less than 100 annually before 1990 to over 5000 annually since 2019 (see [Appendix 1](#)) (Web of Science, 2024).

Current research shows “that online brand communities [(product forums)] had a positive impact on immediate purchase intentions and sales.” Overall, negative reviews appear to have “less of an impact than expected with research showing that the impact of negative information is not as strong as the benefits of positive information.” Furthermore, “enhancing the quality of the information [is vital], since it is the entire communication quality factor (including all the four dimensions of frequency, relevance, duration, and timeliness) that impacts customer purchase behavior.” (Adjei, 2010) These findings indicate that forums and review functions lead to better sales overall, especially if these platforms are curated well to ensure quality.

Outside of communication quality and volume, reviewer valence and type have an impact on the reader. While “both customer review valence and professional critics review valence have a positive impact on the individual reviewer’s rating, the impact of customer review valence is stronger”, meaning that review readers are more likely to be influenced by other customers rather than professional critics. Another implication resulting from the research is that “OWOM could likely play an outsized role for niche brands in terms of crafting an effective marketing strategy” since “substantive differences [were] found between niche and mainstream brands [showing] that niche brands are clearly a unique and special case of brands that demand [deeper] understanding.” (Hoskins, 2021) Given that many of the third-party sellers on Amazon have a limited product catalogue and should be considered niche brands, these findings underline their need to understand review trends and sentiments.

Another strong indicator for the practical applications of sentiment analysis comes from a study of online reviews including from Amazon.com showing an “emotional review–reward effect: emotionally laden reviews lead to impulsive behaviors (in this case, spending more money on unplanned products). [Finding] a significant decrease in impulsive behaviors when reviewers share more rational content with their audience. Furthermore, lower star ratings lead to increased spending on unplanned purchases, perhaps reflecting consumers’ attempts to cope with negative product experiences”. However, the study does not

find a clear correlation between the author of a negative reviewer making a “purchase from [the same retailer], rather than from competitors.” (Motyka, 2018) This underlines the necessity for vendors to accurately understand the issues customers encounter with their products as well as their attitudes towards them.

The research indicates that online reviews and online word of mouth can have an outsized impact on purchasing behavior. Review valence and perceived quality of the review forum highly influence the trust customers have in the platform overall which in turn influences purchase decisions.

Research Question and Goals

Given these research findings, we pose the following research question: How can businesses leverage NLP to identify unverified purchases to enhance decision-making and customer trust in e-commerce?

To find a conclusive answer, we aim to complete three primary objectives in our project analyzing Amazon appliance reviews. By meeting these objectives through our analysis, we strive to provide valuable insights for improving customer experience and business decision making as well as offering researchers and businesses methods and tools to more easily gain insights in the future.

1. Classify Verified and Unverified Reviews using NLP Techniques:

The primary goal of this project is to classify reviews verified or unverified purchases by leveraging NLP features. These include sentiment scores, grammatical patterns, and named entity recognition. Accurate classification helps restore trust in e-commerce platforms like Amazon by detecting potentially misleading reviews thereby improving customer confidence.

2. Understand Key Text Features that Help Differentiate Verified and Unverified Reviews:

This objective focuses on analyzing important parts of the review text, for example how positive or negative the review is (sentiment), the types of words used (noun, verbs, adjectives, etc.), or whether a brand or product is mentioned. These details help uncover patterns in how verified or unverified reviews are written, making it easier to tell them apart.

3. Develop and Evaluate Machine Learning Models:

This objective focused on building and evaluating models such as Random Forests, Decision trees and Logistic Regression to accurately classify reviews. Comparing performance metrics such as accuracy, recall, precision, and PR-AUC, will help identify the most reliable model.

Natural Language Processing and Model Creation

Text dataset description

The Appliance Reviews dataset contains 2,128,605 unique reviews in a 1-5 rating system. Appendix 2 shows an overview of the dataset’s rows and columns together with their respective data types. Each review is given as an integer (star rating) by users. Averages of aggregate ratings are shown for specific items in the Amazon store but do not appear in the dataset. The dataset contains the review title and text along with metadata such as the user ID and timestamp. Additionally, information such as binary helpfulness ratings of the reviews are tracked as well (see [Appendix 2](#)).

Ratings are skewed towards the extremes, with five-star ratings making up 69.82%, and one-star ratings 11.77% of total ratings. Overall, positive ratings (more than three stars) constitute a large majority of the dataset at 79.8%. The majority of reviews for all ratings are verified (see [Appendix 3](#)).

The dataset also contains ASIN (Amazon Standard Identification Numbers) and Parent ASIN fields which are unique item identifiers that are used within Amazon's inventory system. Parent ASINs are non-purchasable seller-exclusive inventory designations for items that require child listings, whereas ASINs are specific numbers which are unique to each item and can be used to identify them in the marketplace (see [Appendix 4](#) for an overview). The majority of items receive fewer than 500 reviews while very few ASINs receive more than 8,000 and some Parent ASINs receive up to 12, 000 reviews (see [Appendix 5](#)).

Text Preprocessing Description

To more effectively analyze the text, several preprocessing methods were employed. Due to the size of the dataset and to obtain more accurate results, a random statistically significant sample of 200,000 entries was selected from the dataset. The difference of rating distributions between the complete and sample set is $\pm 0.15\%$, showing no significant changes in trends.

Text preprocessing consists of several steps starting with the expansion of word contractions into a uniform format. Words such as "don't" were transformed into their uncontracted versions like "do not". Next, every word in the dataset was converted to its lowercase equivalent to avoid words such as "Great" and "great" to be counted as two unique words, potentially skewing frequency distribution and sentiment analysis results. Individual words were then tokenized, which splits the text into smaller pieces that are easier to analyze and process as text. After tokenizing, special characters and whitespace were removed to ensure that tokens consist mostly of alphanumeric characters. The special characters "!" and "?" were kept since they give more context regarding the sentiment of the analysis. Stopwords, mainly determiners and prepositions, such as "a", "an", "the" etc. were removed since they serve a grammatical purpose but are not relevant for creating the model or understanding review sentiment. To finalize the preprocessing, a word lemmatizer was applied to reduce words such as "running" to their base form "run".

After preprocessing, the dataset is optimally prepared for analysis and creation of predictive models. The removal of non-alphanumeric characters that are not useful for sentiment analysis helps remove noise from the data. Additionally, duplicates due to deviations from base form, differences in case and contractions were removed leaving a dataset which yields accurate frequency calculations which are the basis for accurate modelling.

Natural Language Features

71 % of our features were carefully created using NLP-based methods to capture grammatical and semantic patterns that differentiate verified and unverified reviews. NLP-based features include counts of nouns, adjectives, and verbs to capture grammatical patterns, a binary indicator for named entity recognition to identify product or brand mentions, and review sentiment scores to assess the emotional tone of reviews. These features were extracted using *part-of-speech tagging*, *sentiment analysis*, and *NER techniques*. Preliminary results indicate that verified reviews tend to use more of the selected POS tags, reflecting detailed writing, while unverified reviews are less likely to mention specific brands or products

Conclusion and Discussion

Addressing Imbalanced Data

Before discussing our models, it is imperative to highlight a challenge that we faced during our research. During preliminary exploratory data analysis, we uncovered a notable imbalance in our target variable. Specifically, the majority class, verified purchases, accounted for 96% of the data, whereas the minority class, unverified purchases, only accounted for 4%. This significant imbalance in our target variable is problematic. If ignored, the models become biased, favoring the majority class, resulting in poor performance for the minority class.

As a result, we implemented a resampling strategy to offset the risk of creating a bias model. After careful consideration, we selected the Synthetic Minority Oversampling Technique (SMOTE) with a sampling strategy of 0.5, where the minority class increased to 50% of the majority class. The implementation of SMOTE increased the minority class to over 1000% while the majority class remained the same (see [Appendix 6](#)).

Models and Evaluation of Performance

We evaluated three models for our binary classification task: Random Forest, Logistic Regression, and Decision Tree. We split the resampled data into training and testing sets using an 80-20 split and applied a class weight of balanced, ensuring the models effectively handled any remaining class imbalance.

Model evaluation is done on a 0 to 1 (0% to 100%) scale with a score of 1 being the highest rating. The Random Forest was our top-performing model, leading in all metrics. It achieved the highest accuracy, being the ratio of correct classification to total classifications, of 86%. For verified purchases, it achieved a recall, the true positive rate, of 91%, meaning it correctly identified 91% of the majority class instances. It also had an unverified precision of 81%, showing that 81% of its positive predictions were accurate. Its balanced F1 scores, the balance of precision and recall, 0.89 for verified and 0.78 for unverified, demonstrated its ability to handle class imbalance effectively. In contrast, logistic regression was the weakest model, with the lowest metrics overall, and it had an unverified precision of just 55%. The Decision Tree performed reasonably well but did not surpass Random Forest in any metric, particularly in identifying unverified purchases (see [Appendix 7](#)).

While we did consider accuracy when selecting the best model, the metric can be misleading Especially in our case due to the imbalance target variable. Consequently, we decided to gauge the performance of our models using a complimentary metric, the Precision-Recall Curve, to identify the minority class (unverified purchases). As with the other metrics, a score close to 1 indicates strong performance, while a score below 0.5 suggests poor performance in the classification task. In other words, a score below 0.5 is worse than a random guess which has a 50% chance of correctness on average. The PR-Curve scores: 0.86 for the Random Forest, 0.78 for Decision Tree and 0.66 for the Logistic Regression model (see [Appendix 8](#)).

Evidence-Based Conclusions and Practical Implications

To draw conclusions, we must first understand why the Random Forest model outperformed the other models. The model combines predictions from single trees to reach a singular result. The ensemble nature of Random Forest has key technical advantages that contributed to its top performance. The first advantage is the aggregation of multiple trees, leading to more accurate predictions and improved generalization across different data subsets. Furthermore, it handles noisy data more effectively as

indicated by its F1-score for both classes. Lastly, the majority voting mechanism balances class performance, capturing the characteristics of each class despite imbalanced data. By achieving the best precision and recall scores for unverified purchases along with other metrics, the Random Forest model directly supports our research question about improving decision-making and customer trust in e-commerce.

To answer the research question, the following practical implications can be drawn from the project. Combining the classification model with the parent ASINs or ASINs which exist in the dataset, could allow Amazon to monitor products for irregularly high instances of unverified reviews. Since we know that across two million reviews 96% are verified, any individual product or product tree that shows a strong deviation from this ratio could indicate a seller having purchased positive reviews or being the target of artificial negative reviews. This could allow the company to intervene before any damage to consumer trust occurs due to misleading reviews. The same principle can be applied to other platforms which have an established ratio of verified to unverified reviewers.

The sentiment analysis which was created as part of the project could be a useful tool for sellers on the platform to understand their customers' opinions towards the product. Especially knowing factors of differentiation and aspects that need to be improved can be vital for sellers to improve their sales.

Furthermore, unexpected changes in overall sentiment could indicate quality control issues or changing expectations in the market. Given the competitive environment of online retail spaces, it can give one vendor an edge over the competition, especially with an easily implementable model.

Outside of commercial applications, researchers could use the model to further discern the research into the impact of online reviews and online word of mouth. Understanding the influence of verification status on consumer sentiment could add another layer to understanding how the reading of reviews and ratings influences purchasing behavior. Given the prolific nature of the field for both business and academia, the tools created in this project should find many useful applications.

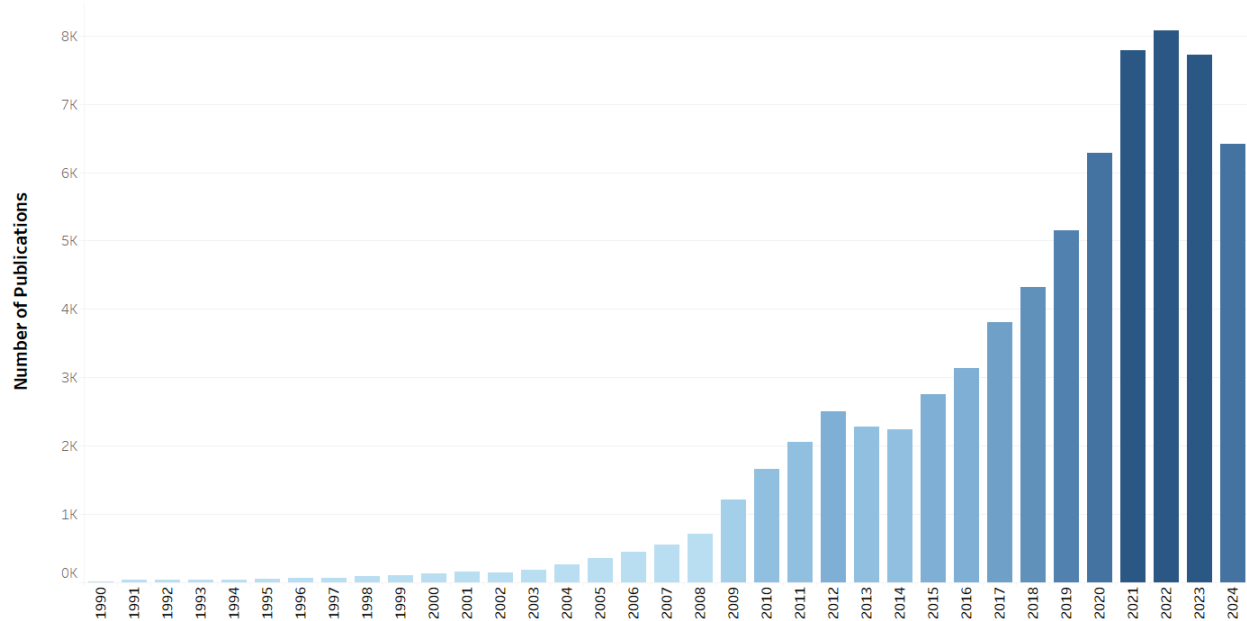
References

- Adjei, M. T. (2010, October). The influence of C2C communications in online brand communities on customer purchase behavior. *Journal of the Academy of Marketing Science*, 38(5), 634-653. doi:<https://doi.org/10.1007/s11747-009-0178-5>
- Amazon. (2024, May 9). *10 things to know about Amazon's 2023 small business empowerment report*. Retrieved from Amazon.com, Inc.: <https://www.aboutamazon.com/news/small-business/amazon-2023-small-business-empowerment-report>
- Himabindu Lakkaraju, J. M. (2013). *Understanding the interplay between titles, content, and communities in social media*. Retrieved from https://cseweb.ucsd.edu/~jmcauley/datasets.html#amazon_reviews
- Hoskins, J. G. (2021, November). The influence of the online community, professional critics, and location similarity on review ratings for niche and mainstream brands. *Journal of the Academy of Marketing Science*, 49(6), 1065-1087. doi:<https://doi.org/10.1007/s11747-021-00780-4>
- Motyka, S. G. (2018, November). The emotional review–reward effect: how do reviews increase impulsivity? *Journal of the Academy of Marketing Science*, 46, 1032-1051. doi:<https://doi.org/10.1007/s11747-018-0585-6>
- Web of Science. (2024, November 05). *Web of Science*. Retrieved from Clarivate: <https://www.webofscience.com/wos/woscc/analyze-results/651bb6c7-9900-46f2-b0f9-d5aa9da86850-012a8f3a46>

Appendix

Appendix 1:

Number of Publications with Topic "Online Reviews" by Year



Source: <https://www.webofscience.com/wos/woscc/analyze-results/651bb6c7-9900-46f2-b0f9-d5aa9da86850-012a8f3a46>

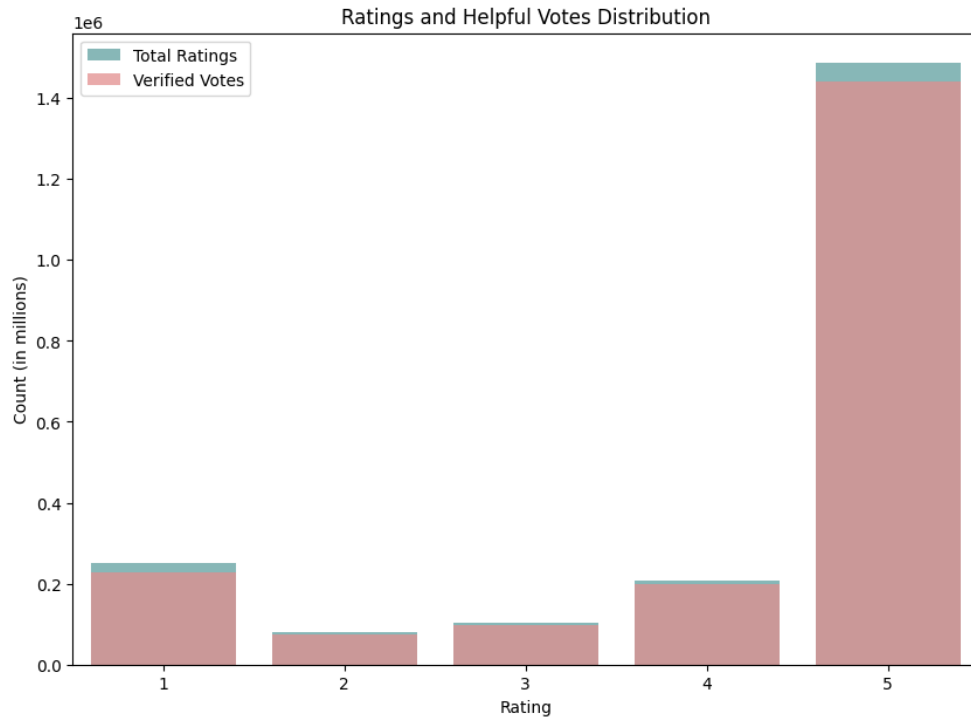
Appendix 2:

The dataset contains: (2128605, 10) rows and columns.

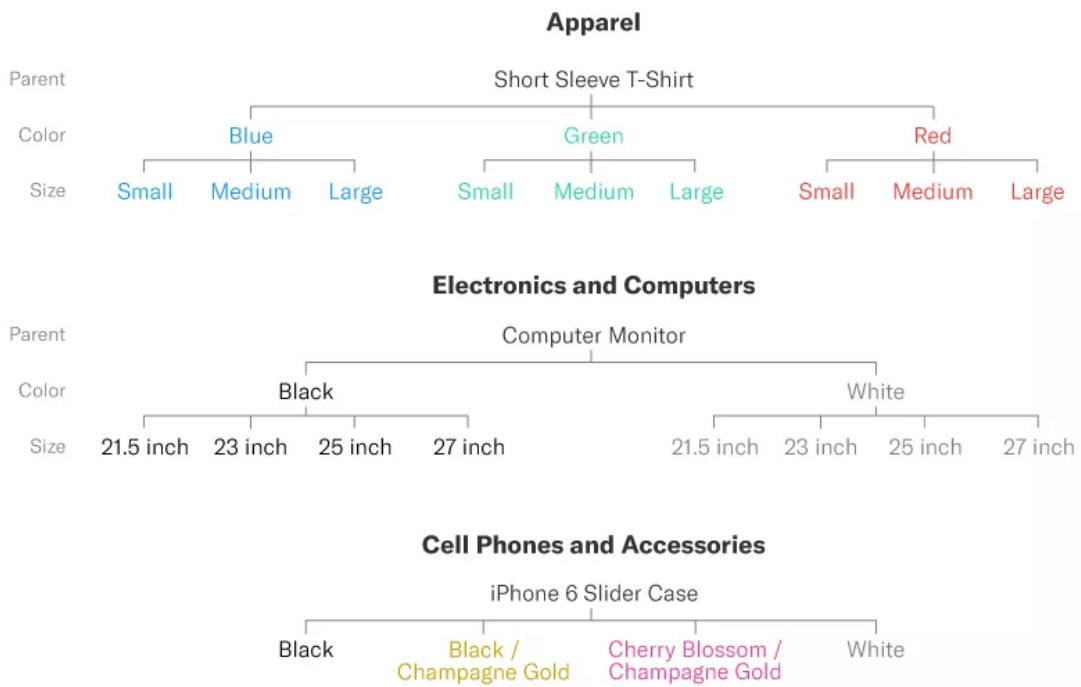
The contained collums have the following datatypes:

```
rating          float64
title           object
text            object
images          object
asin            object
parent_asin     object
user_id         object
timestamp       int64
helpful_vote    int64
verified_purchase bool
dtype: object
```

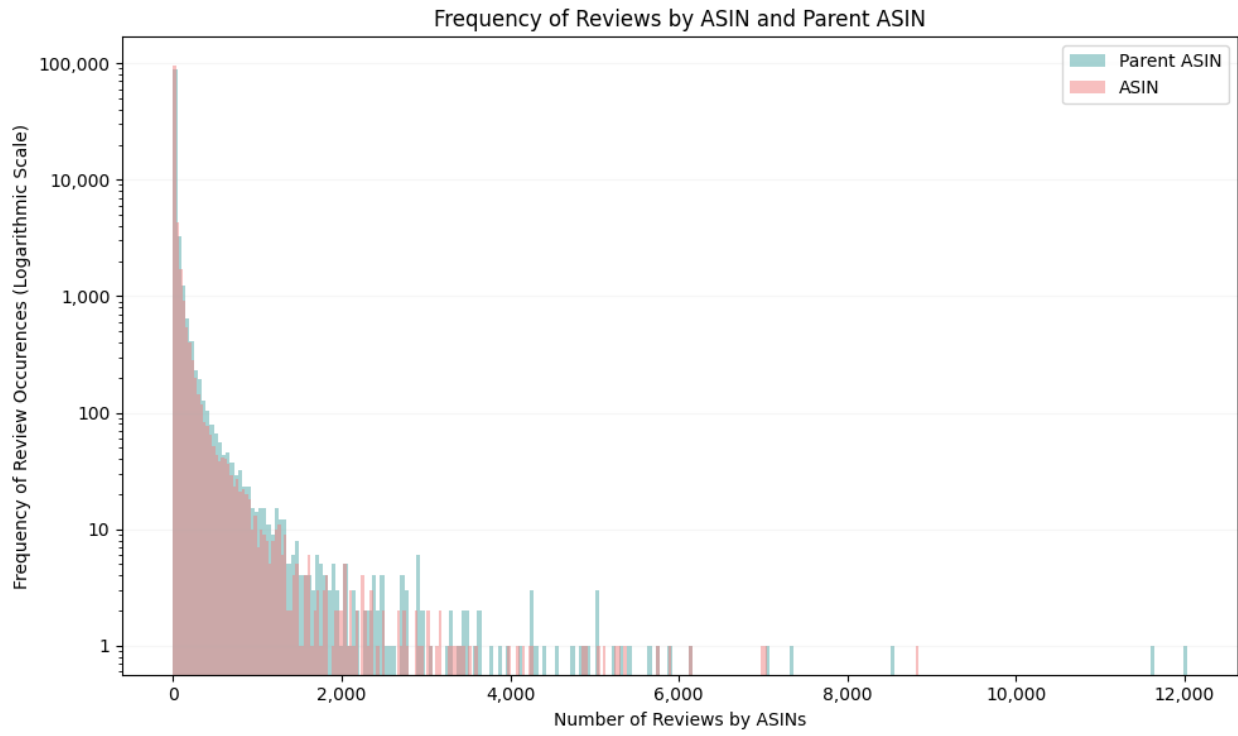
Appendix 3:



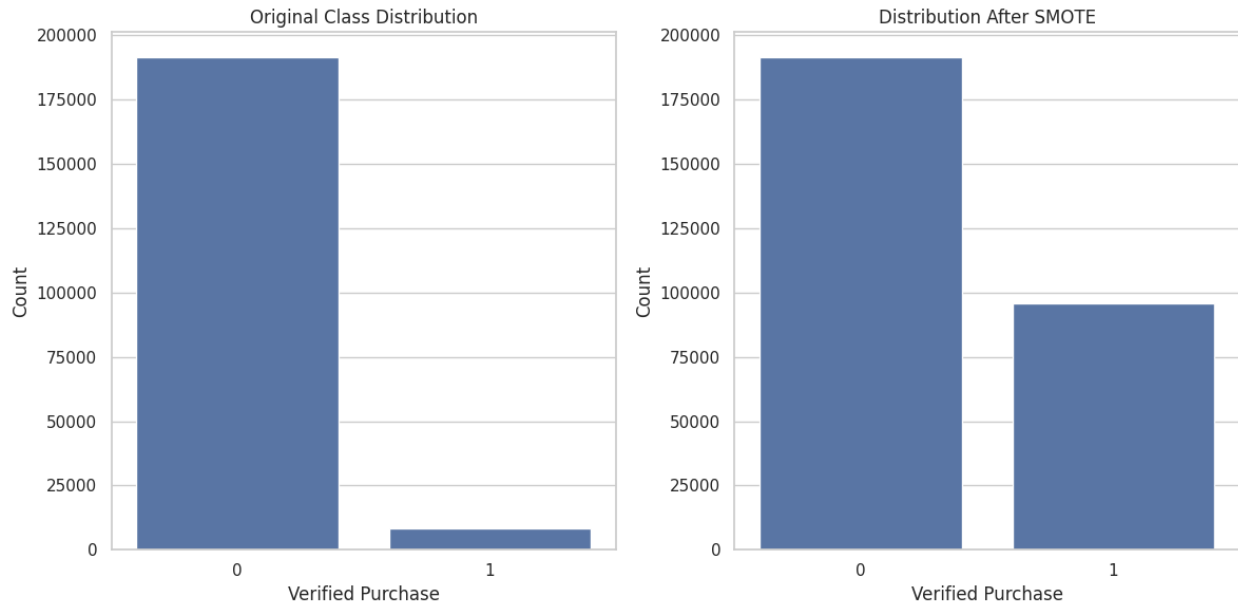
Appendix 4:



Appendix 5:



Appendix 6:



Appendix 7:

Model Performance Comparison

Model	Verified Purchases			Unverified Purchases			Overall
	Precision	Recall	F1	Precision	Recall	F1	Accuracy
Random Forest	88%	91%	89%	81%	75%	78%	86%
Decision Tree	87%	88%	88%	76%	73%	75%	83%
Logistic Regression	79%	76%	77%	55%	59%	57%	70%

Appendix 8:

