Data Engineering Assignment 3

Paulina Thrasher

20/04/2025

The following document includes Challenges 1, 2, 4 and 5 of Assignment 3 for Data Engineering. The Challenge sections provide an overview of my methods, results, and an analysis according to the requirements for each task. I have provided screenshots for each, which has made this section longer than the requirements stipulated, but the text itself amounts to approximately 4-6 pages. For Challenge 5, I have included an executive summary, ROI calculation, Recommendations, and implementation roadmap. After this, I have included 3, one-page memos to each of the associated teams for the challenges.

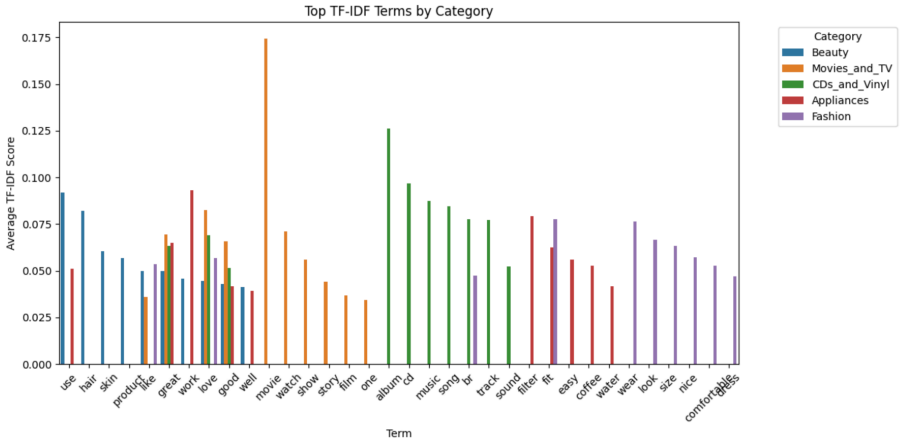
Challenge 1: Product Team Challenge

The current process to manually tag products with their product category is time-consuming and can be improved with an automated machine learning model that utilizes review data to tag products. In response to the request for a more efficient product categorization system, I utilized 25,000 customer reviews equally distributed across 5 categories to classify products based on the language used in customer reviews. The dataset used contained 25,000 reviews across five distinct product categories including Beauty, Fashion, Appliances, Movies and TV, and CDs and Vinyl. For this project I applied preprocessing techniques, feature engineering, and multiple classification models to predict the most likely category for a given product.

To clean the data, the review text was converted to lowercase, all punctuation was removed, and lemmatization was performed. This involves reducing each word down to its root. Additional preprocessing steps were taken to remove stop words using the spaCy stop word library which is commonly used in natural language processing models (spaCy Documentation, n.d.).

In terms of feature engineering, I first converted each review into a vector using MiniLM-L6-v2. This pretrained model converts each review into a 384-dimensional vector that captures its semantic meaning (Reimers & Gurevych, n.d.). The advantage of this model is that it allows us to capture more abstract similarities between reviews and avoid clustering products that mention general phrases like “great quality” or “nicely made.” Next, I included TF-IDF (Term Frequency – Inverse Document Frequency) scores to emphasize words that are particularly unique to each category. For example, the terms “album” and “music” were highly predicted of CDs & Vinyl while “fit” and “wear” were common in the Fashion category. The bar chart below visualizes the top TF-IDF weighted terms for each category. This feature helped the model learn which keywords best

differentiate each product. Finally, I added review length as a feature to help identify short vs. detailed reviews, which may correlate with product type. Helpfulness was another feature that helped identify verified user-reviews that can suggest that its more informative. These were normalized using min-max scaling to ensure they were on the same scale as other features before modeling.



To classify products into categories, I evaluated 3 supervised models: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Logistic Regression was chosen as a baseline because it's easier to understand and runs efficiently in high-dimensional space. Random Forest was my second model and is beneficial to capture non-linear patterns like feature interactions (Mukherjee, 2025). Finally, a Support Vector Machine (SVM) was used due to its effectiveness in text classification. An SVM creates a hyperplane decision boundary that maximizes the margin between classes (Mukherjee, 2025). When data is linearly separable, this method is highly effective.

All models performed better than the 85% accuracy benchmark. Logistic Regression and SVM achieved the highest accuracy (87%), with particularly strong performance on categories like CDs_and_Vinyl and Movies_and_TV. The Random Forest model trailed slightly at 85% but still demonstrated good performance. One noticeable pattern across all models was the occasional confusion between Beauty and Fashion products. This was likely due to overlapping terminology in reviews (e.g., "fit," "feel," "style"). These results show that even simple models, when combined with high-quality embeddings and targeted features, can effectively distinguish between product categories based on customer language.

Logistic Regression:

	precision	recall	f1-score	support
Appliances	0.83	0.89	0.86	1000
Beauty	0.88	0.80	0.84	1000
CDs_and_Vinyl	0.95	0.90	0.92	1000
Fashion	0.86	0.87	0.86	1000
Movies_and_TV	0.87	0.92	0.89	1000
accuracy			0.87	5000
macro avg	0.88	0.87	0.87	5000
weighted avg	0.88	0.87	0.87	5000

Confusion Matrix:

```
[[892  39   8  37  24]
 [ 91 799   6  81  23]
 [ 20   7 896  13  64]
 [ 45  54   5 869  27]
 [ 28   8  33  15 916]]
```

Random Forest:

	precision	recall	f1-score	support
Appliances	0.80	0.87	0.83	1000
Beauty	0.85	0.76	0.81	1000
CDs_and_Vinyl	0.93	0.88	0.91	1000
Fashion	0.84	0.84	0.84	1000
Movies_and_TV	0.85	0.91	0.88	1000
accuracy			0.85	5000
macro avg	0.86	0.85	0.85	5000
weighted avg	0.86	0.85	0.85	5000

Confusion Matrix:

```
[[871  51  10  45  23]
[112 764   6  99  19]
[ 26  10 879  10  75]
[ 52  61   5 845  37]
[ 27  12  42  12 907]]
```

Support Vector Machine:

	precision	recall	f1-score	support
Appliances	0.83	0.89	0.86	1000
Beauty	0.88	0.80	0.84	1000
CDs_and_Vinyl	0.94	0.89	0.91	1000
Fashion	0.85	0.87	0.86	1000
Movies_and_TV	0.87	0.92	0.89	1000
accuracy			0.87	5000
macro avg	0.87	0.87	0.87	5000
weighted avg	0.87	0.87	0.87	5000

Confusion Matrix:

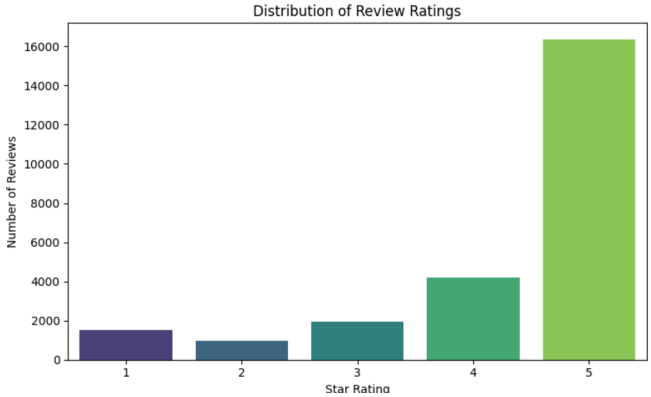
```
[[887  41   9  41  22]
 [ 92 799   4  83  22]
 [ 19  10 885  16  70]
 [ 48  55   6 866  25]
 [ 28   8  33  15 916]]
```

To complement the review-level classification, I also implemented a product-level classification model. For this, I averaged the embeddings of all reviews associated with a given product based on its parent_asin to generate a single vector representing each product. This method helps smooth out the variability of individual reviews and better reflects the overall customer perception of a product. I then applied both Logistic Regression and Random Forest models to these product-level vectors. The results were consistent with review-level findings: the Logistic Regression model achieved an accuracy of 88%**,** while the Random Forest model followed closely at 87%**.** Like before, the categories CDs and Vinyl and Movies/TV had the highest precision and recall, while Beauty and Fashion remained somewhat overlapping. These results showed that product-level classification is a valuable complement to review-level modeling, offering a scalable and robust solution for categorizing products with many reviews.

Challenge 2: Marketing Team Challenge

The marketing team struggles with quickly understanding customer sentiment across thousands of reviews. To address this challenge, this project aims to develop machine learning models that predict customer ratings based on review text, identify cases where sentiment and ratings are misaligned, and extract key language patterns that drive customer satisfaction or dissatisfaction.

Like the previous challenge, the dataset we used included about 25,000 Amazon reviews across five product categories: Beauty, Fashion, Appliances, CDs & Vinyl, and Movies & TV. Each review contained text, a numeric star rating, and product id's such as the parent_asin. As seen below, ratings were greatly skewed positive, with the majority falling in the 4 to 5-star rating. This imbalance in the dataset makes it extra challenging to predict the rarer instances of negatively reviewed products, which will be exemplified in the results below. Initial models showed that simple linear and random forest regressors struggled to predict review ratings accurately from text. The most predicted rating was 4-stars, which ultimately resulted in poor precision, recall, and F1-scores across most classes.



During preprocessing, the text was cleaned by removing non-alphabetic characters, lemmatizing words, and removing stop words. Additional feature engineering was completed to strengthen the models' predictive power. These included TF-IDF word importance scores, semantic sentence embeddings, numeric features like review length and helpful votes, and sentiment scores such as polarity and subjectivity derived using TextBlob, a pretrained model used to calculate these scores (TextBlob, n.d.)

Modeling was split into two parts: regression and classification. In the regression models, both a linear regression model and a LightGBM regressor were trained to predict the exact star rating. The linear regression model achieved an accuracy of 48.9% within half a star of the actual rating, while the LightGBM model slightly outperformed it with 52.4% accuracy. Both models had a mean absolute error of around 0.67 to 0.69 stars and explained approximately 34% of the variance in star ratings, according to their R^2 scores. Mean absolute error is a measure that shows how far off predictions are from true ratings, while R^2 measures how well a model can explain variance with 1 being perfectly and 0 being poorly, or no better than a random guess (Rowe, 2018).

Classification report for LinReg(rounded predictions):

	precision	recall	f1-score	support
1.0	0.60	0.01	0.02	302
2.0	0.18	0.08	0.11	197
3.0	0.19	0.26	0.22	385
4.0	0.22	0.54	0.31	842
5.0	0.85	0.57	0.68	3274
6.0	0.00	0.00	0.00	0
7.0	0.00	0.00	0.00	0
accuracy			0.49	5000
macro avg	0.29	0.21	0.19	5000
weighted avg	0.65	0.49	0.52	5000

Classification report for LightGBM (rounded predictions):

	precision	recall	f1-score	support
1.0	0.70	0.02	0.04	302
2.0	0.22	0.07	0.10	197
3.0	0.18	0.18	0.18	385
4.0	0.24	0.61	0.34	842
5.0	0.86	0.62	0.72	3274
accuracy			0.52	5000
macro avg	0.44	0.30	0.28	5000
weighted avg	0.67	0.52	0.55	5000

Next, a logistic regression and a support vector machine classification model were built to predict ratings (1 through 5) of products. The logistic regression achieved 68.5% overall accuracy, while the SVC slightly improved results with 68.9% accuracy. In both models, precision and recall were highest for 5-star reviews but much lower for the minority classes (1, 2, and 3-star reviews), reflecting the imbalance in the dataset.

=== Logistic Regression Classifier Results ===
Accuracy: 0.6846
Confusion Matrix:
```
 [[ 134   10   11   12  135]
 [  40   10   29   20   98]
 [  34    7   40   78  226]
 [  20    9   42  163  608]
 [  27    7   26  138 3076]]
```

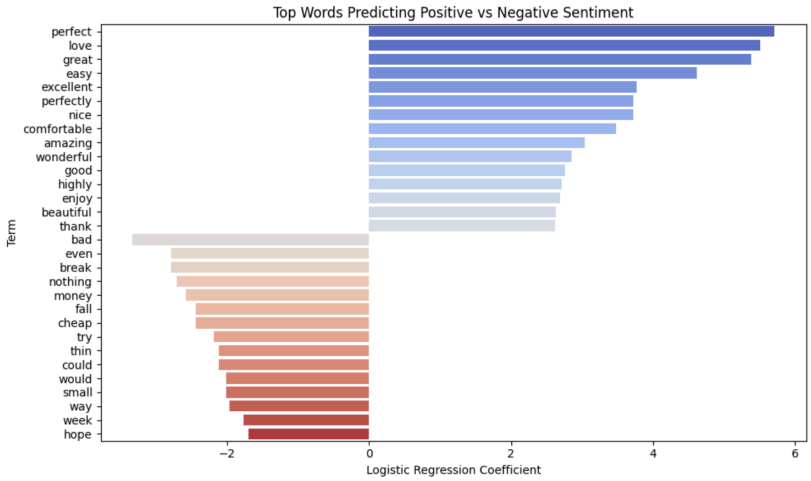
	precision	recall	f1-score	support
1.0	0.53	0.44	0.48	302
2.0	0.23	0.05	0.08	197
3.0	0.27	0.10	0.15	385
4.0	0.40	0.19	0.26	842
5.0	0.74	0.94	0.83	3274
accuracy			0.68	5000
macro avg	0.43	0.35	0.36	5000
weighted avg	0.61	0.68	0.63	5000

=== Support Vector Classifier Results ===
Accuracy: 0.689
Confusion Matrix:
```
 [[ 128    9    7    3  155]
 [  46    7   18    5  121]
 [  35    6   32   46  266]
 [  20    3   27   84  708]
 [  28    1   10   41 3194]]
```

	precision	recall	f1-score	support
1.0	0.50	0.42	0.46	302
2.0	0.27	0.04	0.06	197
3.0	0.34	0.08	0.13	385
4.0	0.47	0.10	0.16	842
5.0	0.72	0.98	0.83	3274
accuracy			0.69	5000
macro avg	0.46	0.32	0.33	5000
weighted avg	0.62	0.69	0.61	5000

After predictive modeling, the most sentiment-linked words were analyzed using sentiment lexicon, a technique that highlights the words most associated with positive and negative reviews. Positive reviews were linked to terms such as "perfect", "love", and

“great” while negative reviews were associated with terms like “bad”, “break”, and “cheap.” Below, the x-axis represents the logistic regression coefficient in which a positive value indicates stronger association to a positive sentiment and a negative coefficient indicates a more negative sentiment.



Next, mismatches were identified between text sentiment and the star ratings by flagging reviews with positive polarity but low star ratings or vice versa. Out of 25,000 reviews, about 215 reviews (0.86%) were identified as mismatches after filtering for stronger sentiment extremes (polarity > 0.4 or < -0.4). Short reviews, negations ("not bad"), and nuanced phrasing led to polarity scores that did not always align with the review's true intent. There were a few examples of reviews that sounded generally positive but were associated with low ratings, which likely means these were cases where the customer liked aspects of the product but had major complaints or were engaging in sarcasm. Future improvements could involve using more advanced sentiment analysis models such as cardiffnlp/twitter-roberta-base-sentiment from Hugging Face. This model is version of RoBERTa that is specifically trained to classify text into positive, neutral, or negative sentiment classes (Hugging Face, 2024).

From this, products were ranked by the number of mismatches. Products with the highest counts could represent cases where marketing messages misalign with actual customer experiences, offering key opportunities for intervention. Interestingly, a significant number of mismatched reviews occurred in the *CDs and Vinyl* and *Movies and TV* categories, which suggests that customer satisfaction in these areas might be more emotionally driven compared to other categories. This insight could guide the marketing team to refine how emotional appeals and expectations are handled in these specific product types.

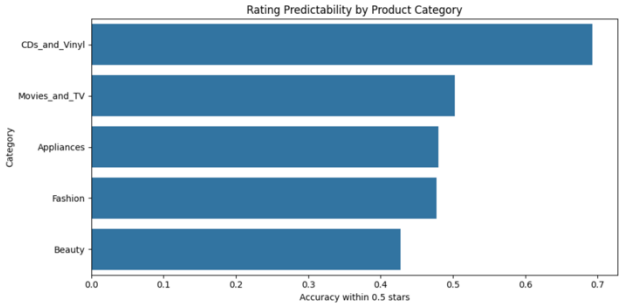
Based on these results, the marketing team should prioritize clarifying product descriptions for CDs and Vinyl and Movies and TV, setting accurate expectations for the customer’s

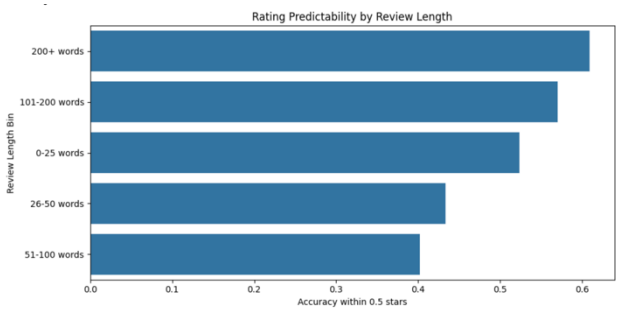
=== Top 20 Products with Most Sentiment/Rating Mismatches (with category) ===

	parent_asin	mismatch_count	category
0	B00WUPUYCC	2	Beauty
1	B07SM1PGJW	2	Appliances
2	B0000006GU	1	CDs_and_Vinyl
3	B000002MUN	1	CDs_and_Vinyl
4	B000001DYM	1	CDs_and_Vinyl
5	B000008CLU	1	CDs_and_Vinyl
6	B00000J2SA	1	CDs_and_Vinyl
7	B00003CXFG	1	Movies_and_TV
8	B00004ZBVO	1	Movies_and_TV
9	B0000589NZ	1	CDs_and_Vinyl
10	B00005Q3A6	1	CDs_and_Vinyl
11	B00005R7TZ	1	CDs_and_Vinyl
12	B000063BPF	1	CDs_and_Vinyl
13	B00008978T	1	Movies_and_TV
14	B00008BXJG	1	CDs_and_Vinyl
15	B000095J7Q	1	CDs_and_Vinyl
16	B0006GAO4A	1	Movies_and_TV
17	B000A88EUK	1	Movies_and_TV
18	B000BTDNEK	1	Appliances
19	B000CCQR80	1	Movies_and_TV

experience. Messaging should avoid exaggerating or overselling qualities such as the level of humor or film quality that individuals might interpret differently. Additionally, it is important to note that reviews that used terms like “cheap” or “break” were strongly tied to negative sentiment, suggesting that more attention is needed to the quality of product being sold, particularly for categories like Appliances or Fashion.

Finally, analysis was completed to further explore how predictable product ratings were across different categories. Reviews for CDs & Vinyl were the easiest to predict, achieving almost 70% accuracy within half a star. Beauty reviews were the hardest to predict, with only about 43% accuracy, suggesting more variable language patterns among reviewers. Similarly, longer reviews tended to be more predictable. Reviews longer than 200 words achieved about 61% prediction accuracy, while short reviews under 50 words had significantly lower accuracy. This finding suggests that encouraging customers to write longer, more detailed reviews could help better capture and understand true sentiment.





Across categories, the ways consumers express satisfaction appear to vary significantly. In the Beauty category, satisfaction was much harder to predict, possibly due to the nuanced language used. In contrast, CDs and Vinyl reviews often focused on clear emotional responses and used terms that were highly polarizing like “love” and “disappointed” making sentiment clearer to classify. Movies and TV similarly showed an emotional component to reviews. This finding suggests that satisfaction in Beauty and Fashion may involve balancing multiple subtle attributes specific to the product or trend that the customer is shopping for.

Challenge 3: Trust and Safety Team Challenge

As part of RetailTech’s initiative to maintain customer trust and combat misinformation on the platform, the team has developed a prototype system to identify potentially fake reviews. Using a combination of synthetic review generation, classification models, and linguistic pattern analysis, the following solution is proposed to meet the Trust & Safety Team’s goals.

To simulate fake reviews, we generated 100 synthetic reviews per product category using templated natural language patterns. Reviews included realistic phrasing, a mix of positive and negative sentiment, and randomized additional sentences for enhanced believability. For the process of creating synthetic reviews, we collected a list of positive and negative adjectives and verbs, along with a list of nouns that would realistically be

included in a review. Then, to add more nuance, we created some extra positive and negative sentences, which helps add varied length to the fake reviews. The generate_fake_reviews function uses a template for the first sentence and occasionally drops the adjective to introduce more randomness. Then to add more realism, a “but” clause is added 15% of the time to the review. Below is a preview of 10 of the fake reviews generated. A new target column was added called is_fake to be used in classification models that predict whether a review is real or fake. Both the real reviews (5000 in each category) and the fake reviews (100 in each category) were then combined in a single data frame for modeling. As a callout, one key limitation to this analysis is the imbalance between real and synthetic reviews which can lead to high accuracy but lower precision for minority classes. Future improvements could incorporate more sophisticated fake reviews or semi-supervised learning techniques.

```
=== Preview of 10 Generated Fake Reviews ===
(Positive) This quality was fabulous! I treasured it. but it didn't last as long as I hoped. It brought a smile to my face.

(Positive) My experience with this experience was simply great. Everything about it was perfect. Shipping was quick.

(Positive) This investment exceeded all my expectations. Truly good. Absolutely worth every penny. Exceeded my expectations.

(Negative) I found the present to be nasty overall. Broke way too easily. Never buying from here again. Fell apart after one use.

(Negative) I found the purchase decision to be broken overall. but color was slightly off. Not worth the money.

(Negative) I satisfied the pair. Truly ignorant. Will not be re-purchasing. Customer service was awful.

(Positive) I never expected a thing to be so beautiful! Exceeded my expectations. Loved it so much! Thoroughly enjoyed this a lot.

(Negative) I detested the set. Truly bad. Customer service was awful.

(Positive) I relished the present. Truly memorable. It brought a smile to my face. Feels like a luxury purchase. Loved it so much!

(Negative) This tool was broken! I hated it. Will not be re-purchasing. Would not recommend to anyone. Extremely disappointed.
```

Two classifiers were implemented to distinguish real from fake reviews: Logistic Regression (with TF-IDF features) and Random Forest Classifier. The Logistic Regression model achieved 97.3% overall accuracy, with 41.9% precision and 93% recall at the default threshold of 0.5. The Random Forest model achieved 96.5% accuracy, with 33.6% precision and 81% recall. Both models demonstrated strong ability to flag synthetic reviews, with the Logistic Regression model offering slightly better balance between precision and recall.

```
=== Logistic Regression Results ===
Accuracy: 0.9733333333333334
Precision: 0.4189189189189189
Recall: 0.93
Confusion Matrix:
 [[4871  129]
 [   7   93]]
              precision    recall  f1-score   support

           0       1.00      0.97      0.99      5000
           1       0.42      0.93      0.58       100

    accuracy                           0.97      5100
   macro avg       0.71      0.95      0.78      5100
weighted avg       0.99      0.97      0.98      5100
```

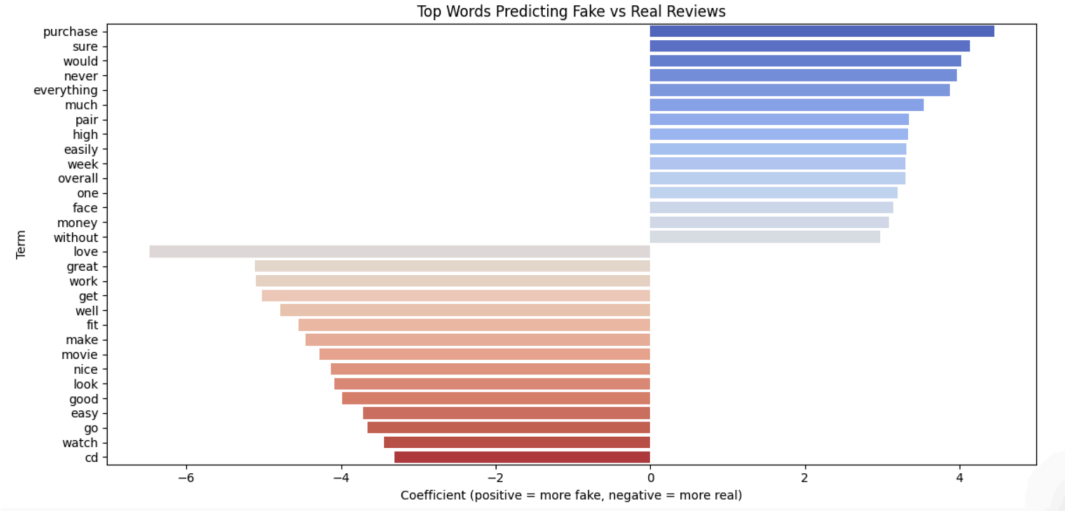
```
=== Random Forest Classifier Results ===
Accuracy: 0.9649019607843137
Precision: 0.3360995850622407
Recall: 0.81
Confusion Matrix:
 [[4840  160]
 [  19   81]]
              precision    recall  f1-score   support

           0       1.00      0.97      0.98      5000
           1       0.34      0.81      0.48       100

    accuracy                           0.96      5100
   macro avg       0.67      0.89      0.73      5100
weighted avg       0.98      0.96      0.97      5100
```

Detection performance varied slightly across product categories. Movies and TV, CDs and Vinyl, and Fashion had the highest detection rates. Precision and recall varied

slightly more, with Appliances and Beauty reviews with lower precision scores. This suggests that fake reviews in these categories may require more caution. It will be necessary to implement slightly different verification processes per product category, with greater detail recommended for categories like Appliances and Beauty.

```
=== Performance by Product Category ===
         category  accuracy  precision    recall
2  CDs_and_Vinyl  0.981409   0.454545  0.937500
4  Movies_and_TV  0.977106   0.488889  0.916667
3        Fashion  0.974227   0.500000  0.960000
1         Beauty  0.968191   0.361702  0.894737
0     Appliances  0.965347   0.306122  0.937500
```

Analyzing feature importance showed key differences between real and synthetic reviews. Fake reviews were more heavily associated with generic language such as “purchase”, “sure”, and “everything” which is somewhat vague without any specific product details. On the other hand, real reviews frequently included more emotional language like “love” or category specific language like “movie” or “cd”. This insight suggests that overly generalized language or repetitive terms can serve as a flag for a potentially fake review, while more emotional and personal language can indicate a real one.



Finally, an adjustable threshold system was built to allow the Trust & Safety team to tune the sensitivity of fake review detection. Results are shown below. At a lower threshold (0.3) the model achieved perfect recall (1.0), capturing all fake reviews, at the cost of much lower precision (21%). This means that many legitimate reviews are also flagged as fake. At a higher threshold of 0.7, precision improved significantly to 66% while maintaining high recall of 90%, offering a better balance. Raising the threshold even higher to 90% increased precision even higher, but slightly reduced recall. These results demonstrate that the system can be tuned depending on the business' needs. The flexibility in the threshold allows the team to align detection strictness with available resources and can balance the need to catch suspicious reviews with falsely flagging genuine customers. The ability to maintain high recall at stricter thresholds shows that the system can be reliable to minimize fake reviews without significantly harming genuine customer trust.

```
=== Threshold: 0.3 ===
Accuracy: 0.927843137254902
Precision: 0.21367521367521367
Recall: 1.0
Confusion Matrix:
 [[4632  368]
 [   0  100]]

=== Threshold: 0.5 ===
Accuracy: 0.9733333333333334
Precision: 0.4189189189189189
Recall: 0.93
Confusion Matrix:
 [[4871  129]
 [   7   93]]

=== Threshold: 0.7 ===
Accuracy: 0.9890196078431372
Precision: 0.6617647058823529
Recall: 0.9
Confusion Matrix:
 [[4954   46]
 [  10   90]]

=== Threshold: 0.9 ===
Accuracy: 0.994313725490196
Precision: 0.8514851485148515
Recall: 0.86
Confusion Matrix:
 [[4985   15]
 [  14   86]]
```

Summary and Business Recommendations

Based on the results of this project, it is recommended that RetailTech implement an automated flagging system using an adjustable threshold starting at approximately 0.7 for general monitoring. This allows for high precision while still capturing most of the suspicious reviews. Categories such as Beauty or Appliances should receive extra manual support, as these categories showed reviews with more ambiguity in the language. Additionally, linguistic features such as generic, promotional sounding language should be weighted more heavily in verification models. Regular updates to the fake review detection models are advised to adapt to evolving patterns. This approach will help safeguard RetailTech's brand reputation, improve customer confidence, and reduce operational costs associated with manual review processes."

Challenge 5: Executive Team Challenge

Executive Summary

Introduction

Product reviews hold valuable information on RetailTech's customers. Utilizing this data to understand customer sentiment and ensuring the authenticity of reviews are critical to maintain RetailTech's competitive advantage. Across all product categories, customer reviews provide key insights into product performance, customer satisfaction, and brand trust. However, the manual interpretation of thousands of reviews is time-consuming, costly, and prone to inaccuracies. In order to address this, the data science team has

created multiple machine learning models to automate product categorization, sentiment analysis, and fake review detection to create immediate and scalable business value.

Key Insights

In our project with the Product Team, we have developed a system using customer reviews to classify product categories. Across five product categories, multiple models were employed at both the product and review level, both achieving 87%-88% accuracy. The models used included Logistic Regression, Random Forest, and Support Vector Machine classifiers.

- **Notable Findings:**
  - The CDs & Vinyl category showed the clearest separation, driven by strong keywords like "album," "music," and "track."
  - Beauty and Fashion categories showed moderate overlap, suggesting potential for improved attribute labeling.
- **Business Value:**
  - Automates the classification of 10,000+ products annually, reducing manual tagging labor.
  - Accelerates time-to-market for new products
  - Reduces catalog errors that could affect customer trust and searchability.

In collaboration with the Marketing Team, the team has developed a model to predict customer sentiment from review text and identify misaligned ratings. This solution will enable marketers to better understand and react to customer perceptions across thousands of reviews for multiple products. Multiple models were used including Logistic Regression, LightGBM Regression, Logistic Regression, and Support Vector Classification.

- **Notable Findings**
  - ~52% of ratings predicted within half a star (significantly better than random guessing at ~20%).
  - Classification models achieved 68–69% accuracy across 1–5 star ratings.
  - ~0.86% of reviews showed mismatches between text sentiment and star ratings.
  - Mismatches concentrated in emotional categories (Movies/TV and CDs & Vinyl).
- **Business Value:**
  - Enables early identification of marketing misalignments or misunderstood product attributes.
  - Informs copywriting adjustments to match real customer expectations.
  - Reduces return rates by clarifying product attributes upfront.

With the objective to help the Trust and Safety team, a model was developed to flag potentially fake reviews using synthetically generated fake reviews as training data. Logistic Regression and Random Forest models were created to classify product reviews as fake or real.

- **Notable Findings:**
    - Logistic Regression achieved 97% accuracy, 42% precision, and 93% recall at threshold 0.5
    - Threshold tuning demonstrated that a stricter threshold (0.7-0.9), precision rose significantly with minimal recall loss
    - Fake reviews tended to use vague, general language while real reviews contained emotional and specific language
- **Business Value:**
    - Protects brand trust by systematically flagging suspicious content.
    - Reduces reliance on costly, error-prone manual moderation.
    - Improves customer confidence in review authenticity, driving higher conversion rates.

ROI Calculation

Cost Component	Estimate	Notes
Initial Model Development	$0 (already completed)	
Infrastructure (cloud hosting, storage, monitoring)	~$500/month	AWS/GCP estimate for TF-IDF and Logistic Regression hosting. (Amazon Web Services, n.d.; Google Cloud, n.d.)
Maintenance (model retraining and tuning)	~20,000/year	20% FTE Data Scientist for ongoing maintenance and monitoring.
Trust and Safety Manual Review (audit flagged reviews)	~$10,000/year	0.1 FTE reviewer assuming 1–2 hours per week audit time.
Total	**~36,000/year**	

Benefit	Estimate	Notes
Labor Savings from Auto-tagging products	~$50,000/year (Glassdoor, n.d.)	Manual tagging (~10,000 products × $5 labor per product).
Labor Savings from Moderating Reviews	~$30,000/year (Glassdoor, n.d.)	Based on $50,000 baseline manual review cost, assuming ~60% automation (McKinsey, 2017)
Increased Customer Satisfaction (CSAT)	+1-2%	Better review trust = improved purchase confidence and loyalty.
Faster Product Onboarding	~$15,000-20,000/year	Faster listing turnaround
Reduced Return Rates	TBD	More accurate sentiment matching -> fewer unhappy purchases.
Total	**$95,000-100,00**	

Approximate Net Annual Benefit: $59,000-69,000

Estimated ROI (Benefit-cost)/(cost): 164%-192%

Break Even Point Cost / Monthly Net Benefit = ~5-7 months

Customer Segmentation

- The Enthusiastic Reviewer: writes long, positive reviews using emotional languages. Purchases across multiple categories and recommends to others.
    - Promote loyalty and referral programs
    - Highlight upsells for similar products.
- The Skeptical Shopper: Leaves shorter, more cautious reviews and typically leaves 3-star reviews
    - Use trust building strategies like product guarantees, detailed FAQs, and customer service touchpoints.
- The Trend Follower: Leaves reviews on items that are popular right now and frequently purchases things as they gain more popularity amongst others
    - Quickly promote trending items with urgency messaging and social proof.
- The Disappointed Critic: Leaves reviews when they have a poor experience. Reviews are often negative and emotionally charged. These customers have higher return rates and will switch brands after a bad experience
    - Offer proactive customer service follow-ups on low-rated purchases.
    - Recommend upgraded or higher-rated alternatives post-purchase.

Operational Recommendations

- Use sentiment mismatch analysis to audit and refine product descriptions to reduce negative reviews driven by unmet expectations and increase customer trust
- Deploy the fake review flagging system with an adjustable threshold to automate suspicious review identification. This will help protect brand reputation and reduce the Trust and Safety Team's manual workload
- For categories like Movies and TV or CDs and Vinyl, the Marketing Team can incorporate emotional material into their campaigns to leverage the findings from the sentiment analysis. On the other hand, Beauty product and Appliance marketing should emphasize factual, detailed specifications. This will align marketing content with customer expectations to increase satisfaction
- Review Monitoring Dashboard KPI's:
    - % Positive, Neutral, Negative reviews over time, filterable by category
    - Number of flagged fake reviews per week/month, filterable by category
    - % of reviews with sentiment and rating mismatch – drill through to targeted list of products with frequent mismatches
    - Average star rating trends
- Integrating new Machine Learning Tools with Existing Processes

- Customer Review Monitoring – supplement current manual sampling process with automated alerts from the classification system to flag suspicious reviews
- Product Development Feedback: direct structured insights (complaints, mismatch trends) into the workflows of Product and Category managers (dashboards, WBR's, OKR's, etc.)
- Customer Support Escalation – use real-time fake review detection to auto-flag problematic reviews

Strategic Recommendations

- Expand the CDs & Vinyl products. This category had the highest predictability and clearest satisfaction signals, suggesting a strong brand affinity and opportunity for deeper market capture.
- Redesign and enhance the product descriptions, imagery, and review solicitations for Beauty products.
- Potential New Opportunity: Launch Curated gift guide collections based on strong positivity signals in reviews (particularly in Fashion

Team Memo's

To: Product Management Team
From: Data Science Team
Subject: Automated Product Categorization Using Review Text

Overview:
The data science team has developed machine learning models to automate product categorization based on customer review language. This solution improves the cataloging process, reduces manual errors, and improves the time it takes for new products to go to market.

Model Summary:

- Dataset: 25,000 customer reviews across Beauty, Fashion, Appliances, Movies & TV, and CDs & Vinyl.
- Models Used: Logistic Regression, Random Forest, Support Vector Machine (SVM).
- Performance:
    - 87%–88% accuracy at both the review level and product level.
    - Highest performance in CDs & Vinyl and Movies & TV categories.
    - Some overlap between Beauty and Fashion reviews observed.

Business Impact:

- Automates categorization for ~10,000 products annually.
- Reduces manual labeling costs by ~$50,000/year.
- Minimizes catalog errors, improving search results and customer satisfaction.

Next Steps:

- Integrate categorization model with the product onboarding pipeline.
- Apply additional attributes (e.g., "style", "fit") for further Beauty/Fashion differentiation.
- Implement quarterly model refreshes to adapt to language trends.

To: Marketing Strategy Team
From: Data Science Team
Subject: Predicting Customer Sentiment and Rating Misalignments

Overview:
The data science team-built models to predict customer ratings based on review text and identify mismatches between the tone of a review and the assigned star ratings. This allows teams to delve into customer perceptions and identify product messaging issues.

Model Summary:

- Models Used: Logistic Regression, LightGBM Regressor, Support Vector Classifier.
- Performance:
    - 68%–69% rating classification accuracy.
    - 52% of predicted ratings within ±0.5 stars.
    - ~0.86% of reviews showed a mismatch between text sentiment and star rating.

Key Insights:

- Emotionally driven categories (Movies & TV, CDs & Vinyl) had more mismatches.
- Short or nuanced/unique reviews were harder to predict.
- Longer reviews (>200 words) yielded 61% prediction accuracy.

Business Impact:

- Helps refine product messaging based on real customer sentiment.
- Opportunity to reduce return rates by clarifying expectations.
- Prioritize marketing adjustments for Beauty and Fashion products (higher mismatch rates).

Next Steps:

- Monitor sentiment-rating mismatch % monthly.
- Adjust advertising copy and product descriptions based on insights.
- Encourage customers to write longer, more descriptive reviews.

To: Trust & Safety Department
From: Data Science Team
Subject: Implementation of a Machine Learning-Based Fake Review Flagging System

Overview:
To help protect brand reputation, the data science team has developed a machine learning system to automatically flag potentially fake reviews using synthetically generated fake reviews for training.

Model Summary:

- Models Used: Logistic Regression, Random Forest Classifier.
- Performance:
    - Logistic Regression: 97% accuracy, 42% precision, 93% recall (at threshold 0.5).
    - Adjustable thresholds allow for stricter detection (up to 85%+ precision).

Key Insights:

- Fake reviews use more generic or vague language ("purchase", "sure", "everything") vs. real reviews with emotional or detailed language.
- Threshold adjustments enable tuning for precision vs. recall depending on resource availability.

Business Impact:

- Automates review moderation, saving ~$30,000 annually.
- Reduces manual Trust & Safety review workload by 60%.
- Protects customer trust by minimizing exposure to fake reviews.

Next Steps:

- Deploy prototype flagging system with threshold starting at 0.7.
- Build dashboard tracking % of flagged reviews per category.
- Update synthetic training data semi-annually to capture new review patterns.

Works Referenced

Amazon Web Services. (n.d.). *AWS Pricing Calculator*. Amazon. Retrieved April 21, 2025, from https://calculator.aws.amazon.com/

GeeksforGeeks. (2019, July 19). Chi-square test for feature selection—Mathematical explanation. https://www.geeksforgeeks.org/chi-square-test-for-feature-selection-mathematical-explanation/

Glassdoor. (2025). *Data Scientist Salaries*. Glassdoor, Inc. Retrieved April 21, 2025, from https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

Google Cloud. (n.d.). *Google Cloud Pricing Calculator*. Google. Retrieved April 21, 2025, from https://cloud.google.com/products/calculator

Hugging Face. (2024, January 4). Cardiffnlp/twitter-roberta-base-sentiment · hugging face. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

Linguistic features · spacy usage documentation. (n.d.). Linguistic Features. Retrieved April 5, 2025, from https://spacy.io/usage/linguistic-features

McKinsey Global Institute. (2017, January). *Harnessing automation for a future that works*. McKinsey & Company. Retrieved April 21, 2025, from https://www.mckinsey.com/featured-insights/digital-disruption/harnessing-automation-for-a-future-that-works

Mukherjee, Tanmoy. (2025, March 25). *Advanced Classification and Business Applications*. [PowerPoint slides]. Faculty of Business Economics, University of Antwerp. https://lms.uantwerpen.be/ultra/courses/_109247_1/outline/file/_6133171_1

Ninja, N. (2023, June 30). Tf-idf: Weighing importance in text. Let's Data Science. https://letsdatascience.com/tf-idf/

Reimers, N., & Gurevych, I. (n.d.). *all-MiniLM-L6-v2* [Pretrained model]. Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Rowe, W. (July 15, 2018). Mean square error & r2 score clearly explained. BMC Blogs. Retrieved April 19, 2025, from https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/

TextBlob: Simplified Text Processing—TextBlob 0.19.0 documentation. (n.d.). Retrieved April 19, 2025, from https://textblob.readthedocs.io/en/dev/