Capstone: Sprint 0

The Problem Area: *What is your area of interest? Within it, what challenges or opportunities could your project address?*

My area of interest lies in e-commerce and retail analytics, focusing on predicting product reception and customer sentiment analysis. Accurately estimating the rating a new product will receive is crucial for brands and retailers, as it impacts inventory planning, pricing strategies, and marketing efforts. Traditional methods rely on sales data and customer feedback after a product launch, but leveraging rich metadata—such as category, price, brand, and product features—can provide early insights into consumer perception.

Major Projects:

- Predicting Product Reception (Expected Rating Prediction): Leverage structured metadata (e.g., price, main_category, features/description) to predict product popularity (avg_rating/ ratings_number), helping businesses optimize inventory and marketing strategies.
- Predicting Review Helpfulness – Amazon has a helpfulness rating system that allows users to see top-rated reviews, helping customers make informed decisions. However, newly posted reviews often appear first, regardless of quality. This project aims to build a model that predicts the helpfulness of a review before it receives votes, ensuring high-quality reviews are prioritized.
- Addressing Review Credibility Issues (Stretch Goal) – Analyze customer reviews to detect inconsistencies in ratings and potential review manipulation.
    - Predicting Review-Based Ratings – Develop a model to predict the expected rating from the review text, ensuring consistency between textual sentiment and numerical ratings. This can help identify mislabeled or inconsistent ratings, improving the reliability of product assessments.
    - Fake or Manipulated Reviews Detection – Identify fraudulent reviews that distort brand reputations and mislead consumers, affecting purchasing decisions and return rates.

This project integrates structured metadata analysis and unstructured text processing from the **Amazon 2023 dataset** to address key challenges in e-commerce analytics, including:

- extracting insights from diverse metadata
- handling review quality variability – some reviews contain more useful insights than others
- addressing data imbalance – some products have higher reviews than others

Despite these challenges, the impact is significant—better demand forecasting, improved recommendations, and greater consumer trust, making this a valuable area for innovation.

The User

- Consumers – More accurate product ratings and helpful reviews reduce misleading information, leading to better purchasing decisions and fewer returns.
- Retailers & Brands – Predicting product reception helps optimize inventory, refine marketing strategies, and protect brand reputation from review manipulation.
- E-Commerce Platforms – Enhancing review credibility ensures users see the most relevant feedback, improving trust and engagement on the platform.

The Big Idea

Machine Learning offers an automated solution to these issues by leveraging vast amounts of structured and unstructured data to uncover patterns that traditional methods might miss. This project applies ML techniques to:

- Predict Product Popularity – Estimate a product's average rating or number of reviews based on metadata like price, category, and review engagement, enabling better inventory and marketing strategies.
- Assess Review Helpfulness – Predict the helpfulness of a new review by training the mode on features like the number of helpful votes, verified purchase status, and review length to identify which reviews are most likely to assist future customers.
- Enhance Rating Consistency– Predict ratings from review text to identify discrepancies between sentiment and given star ratings.
- Detect Fake Reviews – Use anomaly detection and clustering to identify manipulated or fraudulent reviews.

ML Pipeline:

1. Data/Text Processing: Handle missing values, tokenizing text encoding categorical variables
2. Feature Engineering: Prepare numeric features (e.g., length of review, number of helpful votes) and categorical features (e.g., product category, verified purchase).
3. Model Training & Evaluation: Using Random Forest, Gradient Boosting, or Logistic Regression for regression/classification tasks, with appropriate evaluation metrics.
4. Unsupervised Anomaly Detection – Identifying review manipulation using clustering and anomaly detection techniques.

Here are a few scenarios for addressing some of the goals of the project, using the Amazon 2023 dataset:

Predicting Average Product Rating (Regression):

Target Variable: average_rating (Float between 1.0 to 5.0)

Scenario: Estimate a product's average rating using features such as price, category, and product details.

Example Features:

- price (Do expensive products receive higher ratings?)
- main_category (Do certain categories tend to get higher ratings?)
- description (Do products with detailed descriptions receive better ratings?)

Predicting Number of Ratings (Regression):

Target Variable: rating_number (Integer)

Scenario: Estimate the number of ratings a product is likely to receive based on its features.

Example Features:

- price (Do lower-priced items get more ratings due to affordability?)
- store (Are certain brands/stores driving more reviews?)
- bought_together (Are frequently bundled products getting more attention?)

Predicting Review Helpfulness(Regression)

Target Variable: helpful_vote (Integer)

Scenario: Predict the helpfulness of a new review.

Example Features:

- verified_purchase (Does verified purchase status affect helpfulness?)
- review_length (Does a longer review indicate more helpful votes?)
- rating (Does the rating given correlate with helpfulness?)

The Impact

Enhancing product popularity predictions can help retailers optimize inventory and marketing, reducing overstock and missed sales. Rating inconsistencies and fake reviews undermine trust in e-commerce, influencing purchase decisions and causing significant financial losses. Fake and misleading reviews alone contribute to billions in annual losses for consumers and businesses. By improving review reliability and detecting fraudulent activity, this project can enhance user experience, protect brands from unfair competition, and reduce consumer losses from misleading purchases. Strengthening trust in online reviews can lead to higher engagement on e-commerce platforms, ultimately benefiting the entire retail ecosystem.

The Data

The Amazon Reviews 2023 dataset provides a comprehensive source of customer reviews to train and validate machine learning models for multiple aspects of e-commerce analytics. This dataset spans various product categories, including Electronics (18.3M reviews, 1.7B metadata) and Sports & Outdoors (10.3M reviews, 1.3B metadata), making it ideal for studying product popularity, rating consistency, and review credibility.

For the stretch goals, we will explore inconsistencies in rating assignments by predicting review ratings from text, identifying potential mislabeling. Additionally, we will assess review helpfulness by predicting which reviews are most informative, ensuring that customers rely on quality feedback rather than just the most recent posts. Finally, detecting fake or manipulated reviews will help improve trust and prevent fraudulent influence on product reputation.

While the primary focus of this project is on improving rating consistency and detecting potential manipulation, it's worth noting that both categories of products exhibit seasonal demand. This can lead to variations in customer feedback and returns based on factors such as weather, sports seasons, or product launches. These seasonal patterns could be useful in understanding how review sentiment and ratings fluctuate during high-demand periods, offering additional insights into product satisfaction and potential return behavior.

An alternative dataset that can be used to address these challenges is the Yelp Open Dataset that includes business reviews.

The Alternative

Another compelling area of interest is retail sales forecasting. Accurately predicting future sales is crucial for retailers to optimize inventory, minimize losses, and improve operational efficiency. Poor forecasting can lead to overstocking, or understocking, causing missed revenue opportunities and customer dissatisfaction. By applying time series forecasting techniques, retailers can develop more reliable sales predictions. A well-optimized forecasting model is useful for inventory management and demand planning, ultimately driving higher profitability and a better customer experience.