**Project Summary Report: Sentiment Analysis of Amazon Reviews**

**Introduction to the Project**

**Project Domain:** E-commerce

**Objective:** This project involves analyzing sentiments expressed in over 34,000 Amazon product reviews. The objective is to predict sentiment levels—Positive, Negative, or Neutral—based on review attributes and text. The dataset includes attributes such as brand, categories, review titles, review text, and sentiment levels.

**Objectives of the Project**

* **Understand** the sentiment expressed in consumer reviews.
* **Address** class imbalance in sentiment categories.
* **Implement** classifiers and advanced techniques for sentiment analysis.
* **Evaluate** model performance using appropriate metrics.
* **Compare** traditional machine learning algorithms with neural network approaches.
* **Explore** topic modeling techniques for clustering similar reviews.

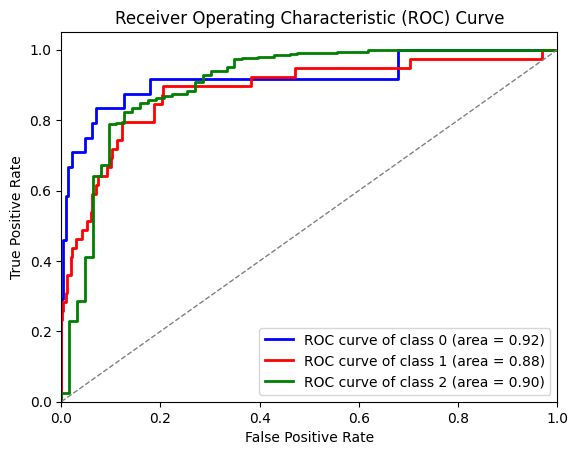
**Project Tasks**

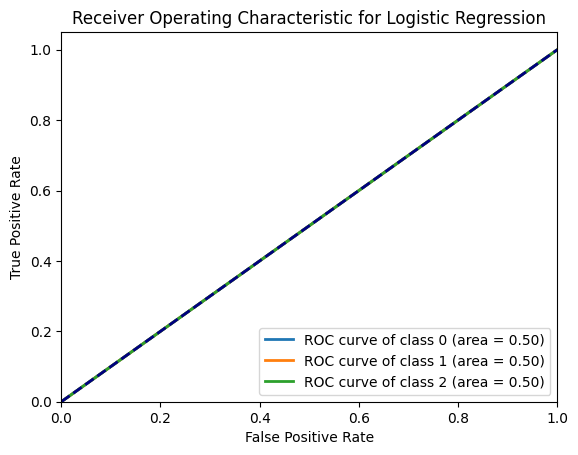
**Weeks 1 & 2: Class Imbalance Problem**

1. **Exploratory Data Analysis (EDA):**
   * **Goal:** Understand the characteristics of positive, negative, and neutral reviews.
   * **Approach:** Visualize sentiment distribution and identify patterns/trends.
   * **Class Imbalance:** Examine class counts and apply techniques to address imbalance.
2. **Feature Engineering:**
   * **Tf-Idf Transformation:** Convert reviews into numerical features suitable for machine learning.
3. **Classifier Selection:**
   * **Multinomial Naive Bayes:** Implement and train to predict sentiment.
   * Hidden Test Data Evaluation
   * Accuracy: 0.937
   * Your model's high accuracy (93.7%) is misleading due to its bias towards predicting only Class 2, resulting in zero precision, recall, and F1-scores for Classes 0 and 1. To improve, address the class imbalance by using techniques like resampling or adjusting class weights to enhance performance across all classes.
   * **Challenge:** Address potential bias towards the majority class.
4. **Tackling Class Imbalance:**
   * **Techniques:** Oversampling or under-sampling to balance class distribution.

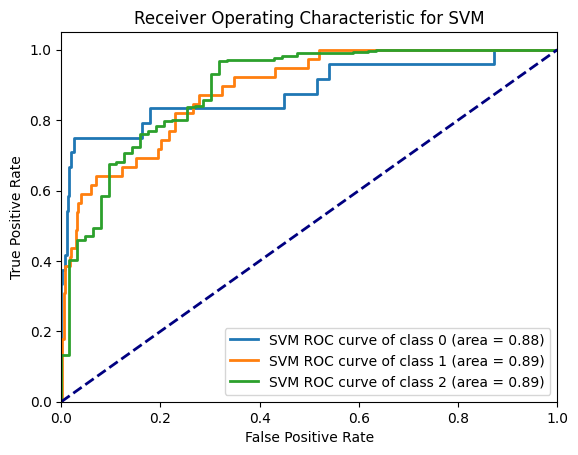
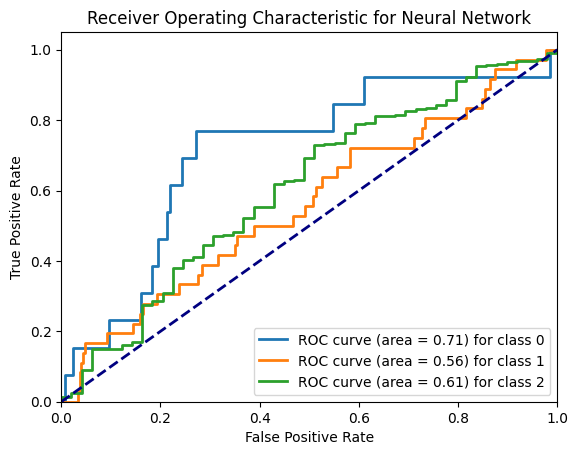
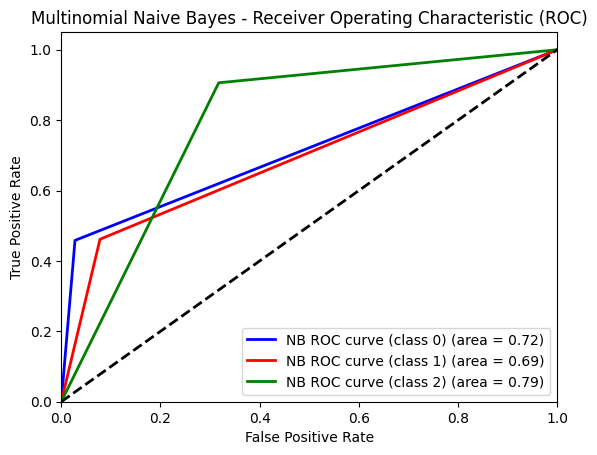
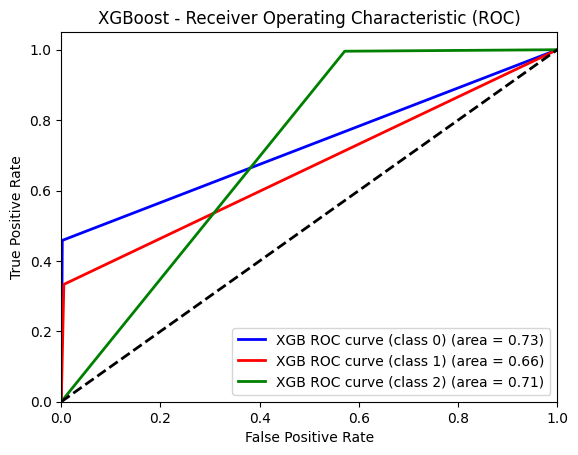
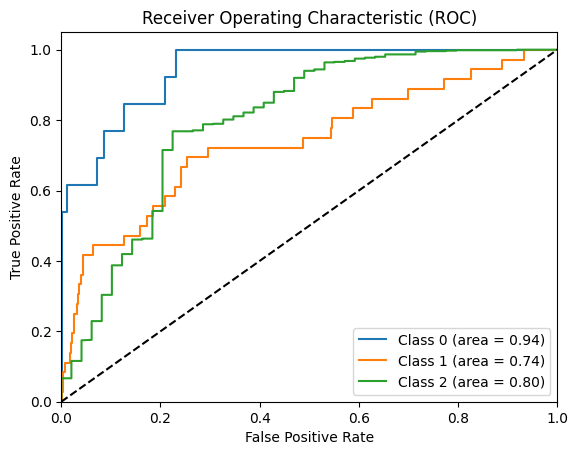
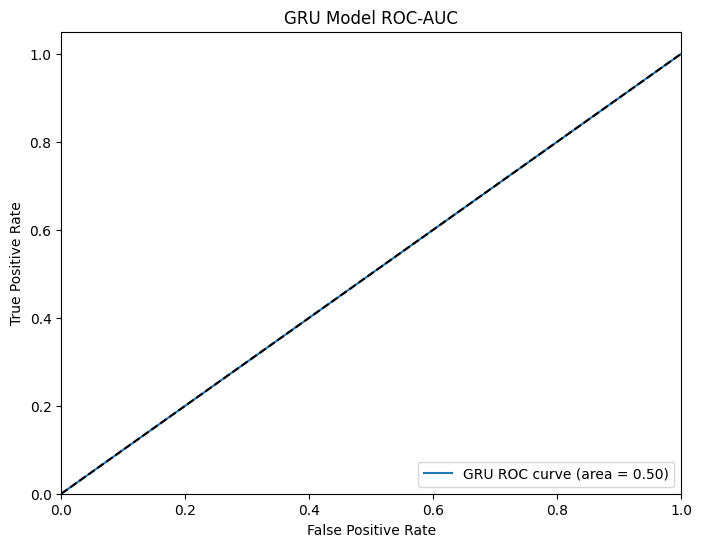
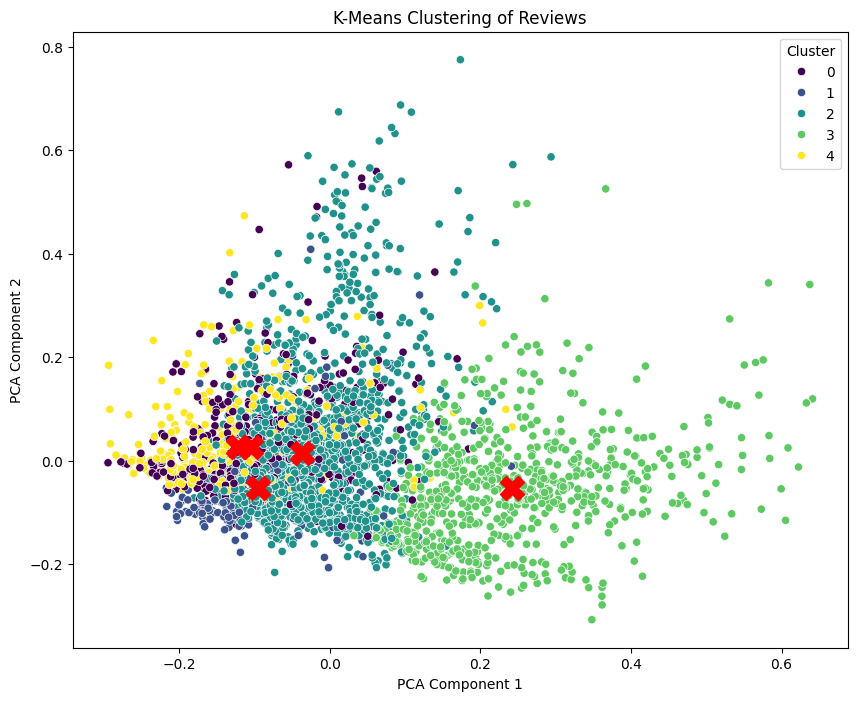
**Evaluation Metrics:**

* **Metrics:** Precision, recall, F1-score, and AUC-ROC. Emphasis on F1-score for class imbalance
* .**Undersampling:**
* The model achieves a high accuracy (94.1%) but performs poorly on Classes 0 and 1, with low precision and recall due to class imbalance, as it predominantly predicts Class 2. To improve, consider addressing the imbalance through techniques like resampling or adjusting class weights.



* **Oversampling:**
* The model's accuracy (93.7%) is high, but it only predicts Class 2, resulting in zero precision and recall for Classes 0 and 1 due to class imbalance. Address this by increasing the number of iterations for logistic regression or using techniques to handle class imbalance**.**
* ****

**Weeks 3 & 4: Model Selection and Advanced Techniques**

1. **Multi-class SVM and Neural Networks:**
   * **SVM:** Implement multi-class classification.
   * The model achieves a high accuracy (95.1%) and performs well on Class 2, but has lower precision and recall for Classes 0 and 1, indicating difficulties in distinguishing these classes. To improve, consider adjusting class weights or applying resampling techniques to better balance performance across all classes.
   * 
   * **Neural Networks:** Explore deep learning models for text classification.
   * The neural network achieves high accuracy (93.9%) but is biased towards predicting only Class 2, resulting in zero precision and recall for Classes 0 and 1. To improve, address class imbalance through resampling or class weighting techniques to ensure better performance across all classes.
   * 
2. **Ensemble Techniques:**
   * **Ensemble Methods:** Combine classifiers such as XGBoost with oversampled Naive Bayes.
   * The Multinomial Naive Bayes model achieves an accuracy of 87.8%, with high performance on Class 2 but lower precision and recall for Classes 0 and 1. This indicates a need for improvement in handling the minority classes, possibly through techniques like resampling or adjusting class weights.
   * 
   * **XBOOST**
   * The XGBoost model achieves a high accuracy of 95.7%, excelling in predicting Class 2 but with moderate performance for Classes 0 and 1. The model effectively handles Class 2, yet it shows room for improvement in the prediction of minority classes.
   * 
3. **Additional Feature Engineering:**
   * **Sentiment Score Feature:** Integrate into models for performance enhancement.
4. **LSTM Implementation:**
   * **LSTM Networks:** Apply and tune parameters for optimal performance.
   * The LSTM model achieves an accuracy of 94.4%, performing well on Class 2 but showing lower precision and recall for Classes 0 and 1. To enhance its performance across all classes, consider adjusting the model architecture or applying class weighting techniques.
   * 
   * The GRU model performs poorly with an accuracy of 1.6%, failing to learn meaningful patterns and producing only predictions for Class 2. Review the model architecture and preprocessing steps to address potential issues causing the NaN loss and ineffective predictions.
   * 
5. **Comparison:**
   * **Comparison:** Evaluate neural networks against traditional models.
6. **Optimization:**
   * **Techniques:** Grid Search, Cross-Validation, and Random Search for parameter tuning.
   * The optimized LSTM model achieves a high accuracy of 95.25% with strong performance on Class 2, though it struggles with lower precision and recall for Classes 0 and 1. It effectively identifies Class 2 but needs improvement in handling other classes.
   * The LSTM model demonstrates a high average accuracy of 93.75% across folds, excelling in predicting Class 2. However, it struggles with precision and recall for Classes 0 and 1, indicating a need for improved performance on these classes.
   * The GRU model shows very low accuracy, averaging just 2.25% across folds, and fails to predict Classes 1 and 2 effectively. It only predicts Class 0 for a few instances, indicating severe performance issues and likely requiring substantial adjustments or a different approach.
7. **Topic Modeling:**
   * **Techniques:** Apply LDA and NMF to cluster reviews based on topics.
   * The NMF model clusters reviews around tablet features, Echo and Kindle praises, product value, and suitability for gifts. The LDA model focuses on Kindle's performance, ease of use for tablets, Amazon products' pricing, Echo functionality, and Kindle as a preferred reading gift.
   * 
   * Cluster 0 features positive reviews about the ease of use and functionality of tablets and Alexa devices, highlighting their suitability for various users. Cluster 1 emphasizes satisfaction with Kindle's features, value, and performance, especially for reading and ease of use. Cluster 2 includes mixed reviews on product durability, value, and performance, with some highlighting limitations. Cluster 3 focuses on the practical value of tablets, including their price, performance, and suitability for basic needs. Cluster 4 praises Echo devices for their functionality and versatility, but includes some critiques about the screen's utility.

**Report on EDA**