Project Report:

**Sentiment Analysis and Rating Prediction Using Amazon Product Reviews** By: Nadya Malekpour

**1. Problem Definition**

**Background:** Customer reviews are essential for e-commerce platforms as they offer valuable insights into customer satisfaction and product quality. These reviews influence purchasing decisions, guide product improvements, and impact overall customer trust. Analyzing reviews manually is time-consuming and challenging, especially at scale.

**Problem Statement:** This project aims to automate the analysis of Amazon product reviews using machine learning techniques. Specifically, it seeks to classify review sentiments into positive, negative, or neutral categories and predict product ratings based on review text and other features.

**Significance:** Understanding customer sentiment and predicting ratings can help businesses enhance their products and services, leading to improved customer satisfaction. Automating this analysis can save time, provide actionable insights, and help e-commerce platforms and sellers make informed decisions.

**2. Project Objectives**

**Sentiment Analysis:** Classify review sentiments into three categories: positive, negative, or neutral. This classification helps identify customer satisfaction levels and pinpoint areas for improvement.

**Rating Prediction:** Predict star ratings based on the content of the review text and other related features. This prediction can provide a quantitative measure of customer sentiment beyond simple classification.

**Identify Key Influencing Factors:** Discover what drives customer satisfaction and dissatisfaction by analyzing which features most significantly impact sentiment and ratings. This insight can inform product improvements and marketing strategies.

**3. Analysis**

**Data Quality and Preprocessing:**

* **Missing Values and Duplicates:** Checked for missing values and duplicates in the dataset. Missing values were handled by dropping or filling with placeholders, and duplicates were removed to ensure data quality.
* **Outlier Detection:** Analyzed outliers in ratings and review lengths using the IQR method. Outliers were identified and assessed to understand their potential impact on model performance.
* **Data Validation:** Validated data types and checked for inconsistencies in reviewer and product IDs to ensure data integrity.

**Exploratory Data Analysis (EDA):**

* **Distribution of Ratings:** Visualized the distribution of star ratings to understand the overall spread of customer feedback.
* **Review Length Analysis:** Examined the distribution of review lengths, revealing insights into how much detail customers provide in their feedback.
* **Word Cloud:** Created a word cloud to visualize the most frequent words in reviews, highlighting common themes and terms associated with sentiment.
* **Trend Analysis:** Explored how ratings and review counts changed over time, identifying patterns and trends in customer sentiment.

**Feature Engineering:**

* Used TF-IDF vectorization to transform the review text into numerical features. Additionally, features such as review length were added to enhance model performance by capturing more information about the reviews.

**Data Splitting:**

* The data was split into training and testing sets for both sentiment analysis (classification) and rating prediction (regression). An 80/20 split ensured that sufficient data was available for both training the models and evaluating their performance.

**Model Selection and Hyperparameter Tuning:**

* For sentiment analysis, Logistic Regression, Naive Bayes, SVM, and Random Forest were tested. Each model was evaluated using metrics like accuracy, precision, recall, and F1-score to determine the best performer.
* For rating prediction, Linear Regression, Lasso Regression, and Random Forest were explored. GridSearchCV was used to fine-tune hyperparameters, optimizing model performance and achieving more accurate predictions.

**4. Results**

**Sentiment Analysis Results:**

* **Logistic Regression:** Achieved good performance for the positive class but struggled with class imbalance, particularly for neutral and negative sentiments.
* **Naive Bayes:** Performed well in detecting the negative class but had lower overall accuracy.
* **SVM:** Demonstrated the most balanced performance across all classes, offering a good trade-off between precision and recall.
* **Random Forest:** Showed strong performance for the majority class but failed to capture minority classes effectively, highlighting the impact of class imbalance.

**Rating Prediction Results:**

* **Lasso Regression:** Displayed higher prediction errors with an RMSE of 0.99 and low explanatory power (R² of 0.05).
* **Random Forest Regressor:** Outperformed other models with an RMSE of 0.84 and R² of 0.33, indicating it captured the data's complexity better.

**5. Discussion**

**Model Performance:**

* SVM emerged as the most balanced choice for sentiment classification, handling class imbalance better than other models, but still struggled with the neutral class.
* Random Forest was the best-performing model for rating prediction, effectively managing non-linear relationships and providing reliable results.

**Challenges:**

* Class imbalance significantly affected sentiment classification models. Despite using class weights, the models struggled with underrepresented classes, impacting overall accuracy.
* Rating prediction required careful feature engineering to capture the nuances of review text and related features.

**Key Insights:**

* Positive reviews dominated the dataset, skewing results toward the majority class and affecting model fairness.
* Textual features played a significant role in both tasks, demonstrating the importance of proper text processing and feature selection in machine learning.

**6. Evaluation and Reflection (15%)**

**Reflection on Process:**

* The project underscored the importance of data quality, preprocessing, and feature engineering. Transforming text into numerical features was crucial for model success, and hyperparameter tuning helped optimize performance.

**Ethical Considerations:**

* Bias in the dataset, such as overrepresentation of positive reviews, posed challenges. Future work should focus on addressing these biases to improve model fairness and reliability.

**Areas for Improvement:**

* Advanced techniques like SMOTE could help address class imbalance more effectively, particularly for sentiment classification.
* Exploring additional models like Gradient Boosting or XGBoost could further enhance performance and offer more robust predictions.

**Future Directions:**

* Real-time sentiment analysis could provide immediate feedback to sellers, helping them respond quickly to customer feedback.
* Implementing models to detect fake or spam reviews would enhance the platform’s integrity, ensuring that insights are based on genuine feedback.