INNOMATICS
RESEARCH LABS

Machine Learning - Project Report Document

Student Name	Sada Vijay, Dawa Phuti Lepcha
Batch	AI Elite 18
Project Name	Sentiment Analysis
Project Domain	
Type of Machine Learning	Supervised ML
Type of Problem	Classification
Project Methodology	CRISP-DM
Stages Involved	• Data Collection and Understanding • Data Preparation • Model Building • Model Training

	• Model Evaluation

Business Understanding:

Sentiment Analysis of Amazon reviews is aimed at deciphering the sentiments expressed by customers in their textual feedback and predicting corresponding scores. As one of the largest online marketplaces, Amazon accumulates a vast repository of customer reviews spanning a wide array of products.

Understanding the sentiments conveyed in these reviews is essential for Amazon and its sellers to gauge customer satisfaction levels, identify product strengths and weaknesses, and make data-driven decisions to enhance overall customer experience.

By leveraging these insights, businesses can drive continuous improvement, enhance product quality, and foster stronger customer relationships, ultimately bolstering their competitiveness in the dynamic landscape of e-commerce.

Problem Statement: The objective is to develop a sentiment analysis model for Amazon reviews, predicting the sentiment scores from textual reviews.

Stage 1: Data Collection and Understanding

a) **Data Collection:** The data was provided to us by the client.

b) **Data Understanding:**

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

Data includes:

- Reviews from Oct 1999 - Oct 2012 - 568,454 reviews
- 256,059 Users and 74,258 products
- 260 users with > 50 reviews

Here are the features and their descriptions:

1. idlist: Unique row number.
2. productid: The id of the product which is being reviewed.
3. Userid: The id of the customer who reviewed the product.
4. ProfileName: The name of the customer in the profile.
5. HelpfulnessNumerator: The number of customers who found that review helpful.
6. HelpfulnessDenominator: The number of customers who indicated whether they found the review helpful or not.
7. Helpfulness: HelpfulnessNumerator/ HelpfulnessDenominator
8. Score: The rating of the product from 1 to 5.
9. Time: Timestamp for the review.
10. ReviewSummary: Summary of the Reviews of the product given by the customers.
11. ReviewText: The review of the product given by the customers.

S No	**Feature Name**	**Data Type**
1	idlist	Int64
2	productid	Object
3	Userid	Object
4	ProfileName	Object
5	HelpfulnessNumerator	Int64
6	HelpfulnessDenominator	Int64
7	Helpfulness	Float64
8	Score	Int64
9	Time	Int64
10	ReviewSummary	Object
11	ReviewText	Object

Stage 2: Data Preparation

a) Exploratory Data Analysis:

S No	Type	Feature Names	Observation
1	Missing Values	ProfileName ReviewSummary ReviewText	There were null values.
2	Duplicates	ReviewText ReviewSummary	There were a lot of duplicates.
3	Outliers	Helpfulness	We found 2 outliers.

b) Data Cleaning/wrangling:

S no	Type of Cleaning	Technique	Feature Name	Reason
1	Missing value	Drop	ReviewText, ReviewSummary, ProfileName	They contributed to only 0.0249%.
2	Duplicates	Drop	ReviewText, ReviewSummary	They are unnecessary.
3	Unmeaningful words and Stopwords	Drop, Replaced with “ “	ReviewText, ReviewSummary	They do not provide any useful information.
4	Encoding	WordNetLemmatizer	ReviewText, ReviewSummary	Used Lemmatization for pre-processing on texts to retain the true meaning
5	Scaling	TF-IDF, BoW	ReviewText, ReviewSummary	Used both TF-IDF and BoW for representing the textual data as numerical vectors.

Stage 3: Model Building:

S No	Type of Problem	Algorithm Name
1	Classification	KNNeighbors Classifier
2	Classification	Logistic Regression
3	Classification	SVC
4	Classification	Random Forest Classifier
5	Classification	Decision Tree Classifier
6	Classification	XG Boost Classifier
7	Classification	Naïve Bayes Classifier

1. **Logistic Regression:** Logistic regression is a statistical method used to predict the probability of an event happening, such as whether an email is spam or not. Unlike linear regression, it works well for situations where the outcome is binary (yes/no) instead of continuous.

2. **SVC:** A support vector classifier (SVM) excels at finding the best separation line between categories in your data. It prioritizes a wide margin between the classes, making it effective even for complex datasets.

3. **KNeigbors Classifier:** The K-Nearest Neighbors (KNN) classifier predicts a data point's class by analyzing the labels of its closest neighbors in the training data, making it simple to understand and effective for various classification tasks.

4. **Decision Tree Classifier**: A decision tree classifier is a machine learning method that uses a tree-like structure to classify data. It asks a series of questions about the data's features, branching out based on the answers, until it reaches a final leaf node that predicts the class.

5. **Random Forest Classifier:** Random Forest Classifier is a machine learning algorithm that combines multiple decision trees for stronger predictions. By training a "forest" of trees on random subsets of data, it reduces the risk of overfitting and improves overall accuracy.

6. **XGBoost Classifier**: XGBoost, short for eXtreme Gradient Boosting, is a popular machine learning algorithm known for its efficiency and performance in supervised learning tasks, particularly in structured/tabular data and gradient boosting frameworks.

7. **Naive Bayes Classifier**: The Naïve Bayes classifier is a probabilistic machine learning algorithm based on Bayes' Theorem with an assumption of independence between features. Despite its simplicity and "naïve" assumption, it's surprisingly effective in many-real world applications, especially in text classification and sentiment analysis.

Stage 4: Model Training:

S No	Algorithm Name	Metric used for Evaluation
1	KNN Classifier	Accuracy
2	Logistic Regression	Accuracy
3	Support Vector Classifier	Accuracy
4	Random Forest Classifier	Accuracy
5	Decision Tree Classifier	Accuracy
6	XG Boost Classifier	Accuracy
7	Naïve Bayes Classifier	Accuracy

Stage 5: Model Evaluation:

ReviewText:

S No	Algorithm Name	Metric Score(TF-IDF)	Metric Score(BoW)
1	KNN Classifier	0.607267	0.557800
2	Logistic Regression	0.712800	0.690933
3	Support Vector Classifier	0.711600	0.700733
4	Random Forest Classifier	0.661867	0.664733
5	Decision Tree Classifier	0.578733	0.589733
6	XG Boost Classifier	0.697467	0.696133
7	Naïve Bayes Classifier	0.644333	0.687267

ReviewSummary:

S No	Algorithm Name	Metric Score(TF-IDF)
1	KNN Classifier	0.601333
2	Logistic Regression	0.655333
3	Support Vector Classifier	0.658000
4	Random Forest Classifier	0.637000
5	Decision Tree Classifier	0.570133

Challenges Faced:

While identifying the missing values inside the reviewtext and reviewsummary columns, we've also observed that there were many broken html tags which didn't carry any meaning. We have also faced issues while training the models since some models took more time.

Conclusion:

From the above Accuracy results we can observe that Logistic Regressor, Support Vector Classifier and XGBoost Classifier has the highest acccuracy Score when compared to all the other models i.e. they have more than 0.69 accuracy score.

And TF-IDF and BoW gives nearly the same accuracy score for all the models except for Naïve Bayes, for which BoW has higher accuracy and KNN Classsifier, for which TF-IDF has better accuracy score.

And ReviewText gave a better result compared to ReviewSummary.