Machine Learning Mini Projects Report

Hadia Moosa(ID: DHC-392)

May 5th, 2025

Overview

This document summarizes three machine learning projects: Fake News Detection, Customer Segmentation, and Movie Review Sentiment Analysis. Each project addresses a different machine learning task using suitable techniques and evaluations.

1 Fake News Detection

Objective: Classify news articles as *real* or *fake* using textual data.

Approach:

- Cleaned text (lowercasing, HTML removal, punctuation, and number removal).
- Tokenization, stopword removal, stemming, and lemmatization.
- TF-IDF vectorization with 5000 features.
- Models: Multinomial Naïve Bayes and Random Forest Classifier.

Challenges:

- Class imbalance skewed model predictions.
- Naïve Bayes struggled with feature independence assumptions.
- Preprocessing was compute-intensive due to NLTK operations.

Performance:

- Naïve Bayes: Accuracy = 35%, F1-score (Fake) = 0.50
- Random Forest: Accuracy = 65%

Recommendations:

- Use transformer-based models like BERT or RoBERTa.
- Address class imbalance using SMOTE or class weights.
- Perform hyperparameter tuning and use grid search.

2 Customer Segmentation

Objective: Segment customers based on demographic and spending behavior using unsupervised learning.

Approach:

- Encoded categorical data (gender).
- Feature scaling using `StandardScaler`.
- PCA used for 2D visualization.
- K-Means clustering with Elbow Method ($k = 5$).

Challenges:

- Small dataset limited the diversity of clusters.
- Results sensitive to random initialization of cluster centers.

Cluster Insights:

- Cluster 2: Young, high spenders with moderate income.
- Cluster 1: High income but low spending behavior.

Recommendations:

- Use DBSCAN or Hierarchical Clustering for better structure discovery.
- Add behavioral features like purchase frequency or preferred categories.

3 Movie Review Sentiment Analysis

Objective: Classify IMDB movie reviews as *positive* or *negative*.

Approach:

- Text cleaning: lowercasing, HTML, punctuation, and number removal.
- Tokenization, stemming, and lemmatization using NLTK.
- TF-IDF vectorization.
- Model: Multinomial Naïve Bayes.

Performance:

- Accuracy = 86%
- F1-score = 86%

Challenges:

- Processing 50,000 reviews with NLTK was time-intensive.
- Reviews varied widely in style and complexity.

Recommendations:

- Consider deep learning (LSTM, BiLSTM, Transformers).
- Use pretrained embeddings (e.g., GloVe) or fine-tuned BERT.

Summary Table

Task	Technique Used	Accuracy	Key Challenge	Suggested Improvement
Fake News Detection	Naïve Bayes, Random Forest	35–65%	Class imbalance, text noise	Use BERT, balance classes
Customer Segmentation	K-Means Clustering	N/A	Small feature set	Try DBSCAN, enrich data
Sentiment Analysis (IMDB)	Naïve Bayes + TF-IDF	86%	Text length and noise	Use deep learning (LSTM, BERT)

Conclusion

These projects demonstrate the application of both supervised and unsupervised learning techniques to real-world datasets. While classical ML techniques like Naïve Bayes and K-Means are effective for baseline performance, advanced models and richer data can substantially improve outcomes.