GITAM
(Deemed to be university)
School of Technology, Hyderabad
Department of Computer Science and Engineering

Customer profiling, segmentation, and sales prediction

Project Batch No: P17-07

Name of Student(s) and Full Reg No(s):
B. Mahitha – HU21CSEN0600088
R. Sahasra – HU21CSEN0600184
Y. Jaswanth – HU21CSEN0600026
B. Rohan – HU21CSEN0600181

Guide Name: Dr. T. Sasi Vardhan

GITAM
DEEMED TO BE UNIVERSITY

CONTENTS

- Introduction
- Abstract
- Review of literature Survey
- Problem statement
- Objectives
- Requirement analysis/Dataset
- Identification of Tools/Technologies
- Design/Flowchart of complete project
- Implementation
- Algorithm
- References

INTRODUCTION

- To develop an efficient system for customer profiling and segmentation through clustering algorithms such as K-Means and DBSCAN.

- To conduct sentiment analysis of customer reviews with VADER to analyze customer satisfaction and preferences.

- To forecast future sales patterns using regression models (Linear Regression, Decision Tree Regression, Random Forest Regression) and time-series forecasting methodologies (ARIMA, LSTM).

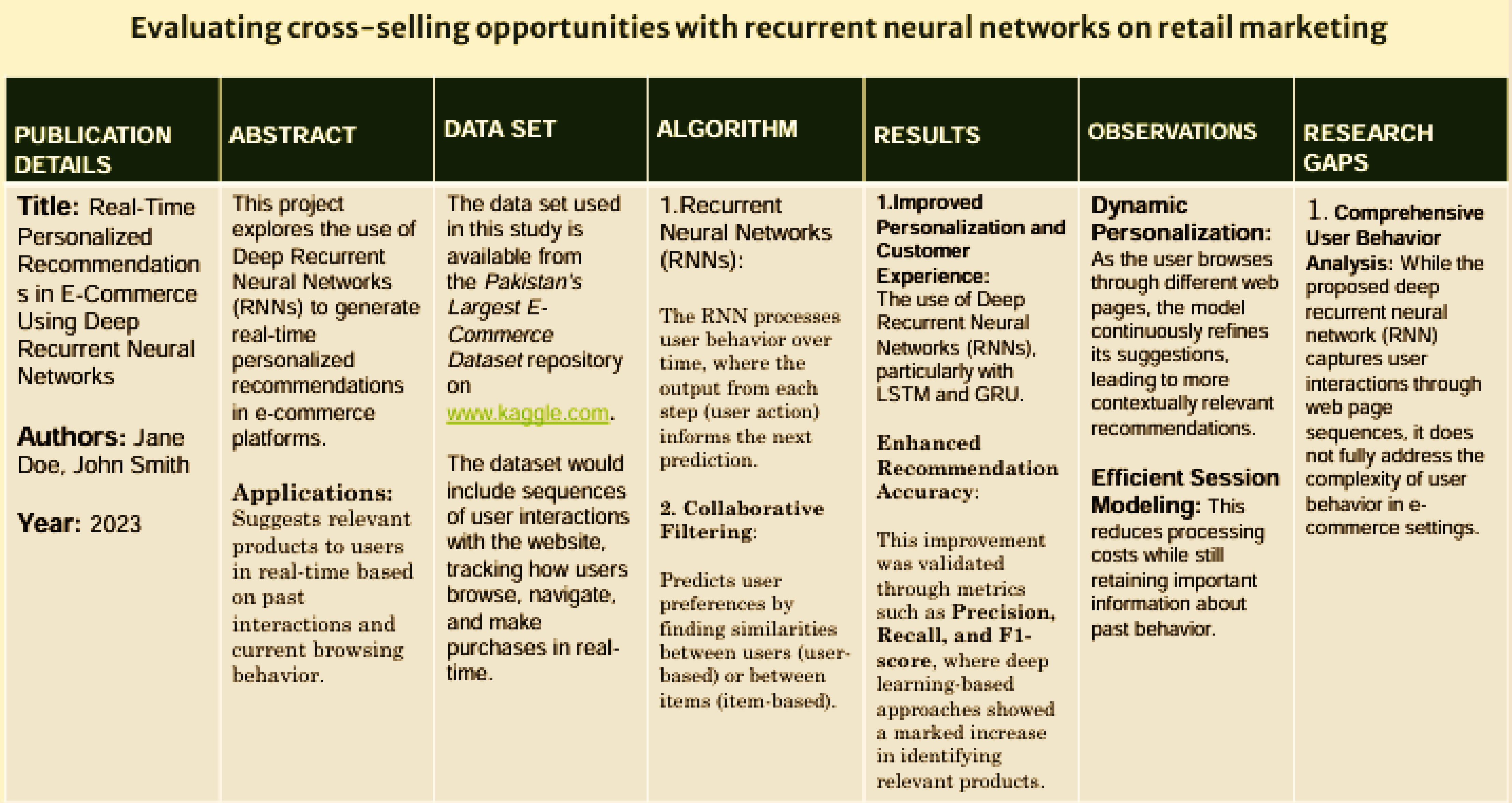
- To combine customer segmentation and sentiment analysis with sales forecasting models for better forecasting and targeted marketing campaigns.

ABSTRACT

- Customer behavior knowledge and forecasting sales trends accurately are the biggest challenges for organizations in today's competitive age. The system proposed not only identifies distinct customer segments but also predicts future sales with greater accuracy, allowing key insights for effective targeted marketing strategies and inventory management. Experimental results validate the effectiveness of using customer sentiment and profiling data in prediction models.

- The project revolves around the use of clustering algorithms like K-Means and DBSCAN for effective customer segmentation, VADER-based sentiment analysis for customer review comprehension, and multiple regression models like Linear Regression, Decision Tree Regression, and Random Forest Regression for effective sales forecasting. Time-series forecasting models ARIMA and LSTM are also employed to predict sales trends over time.

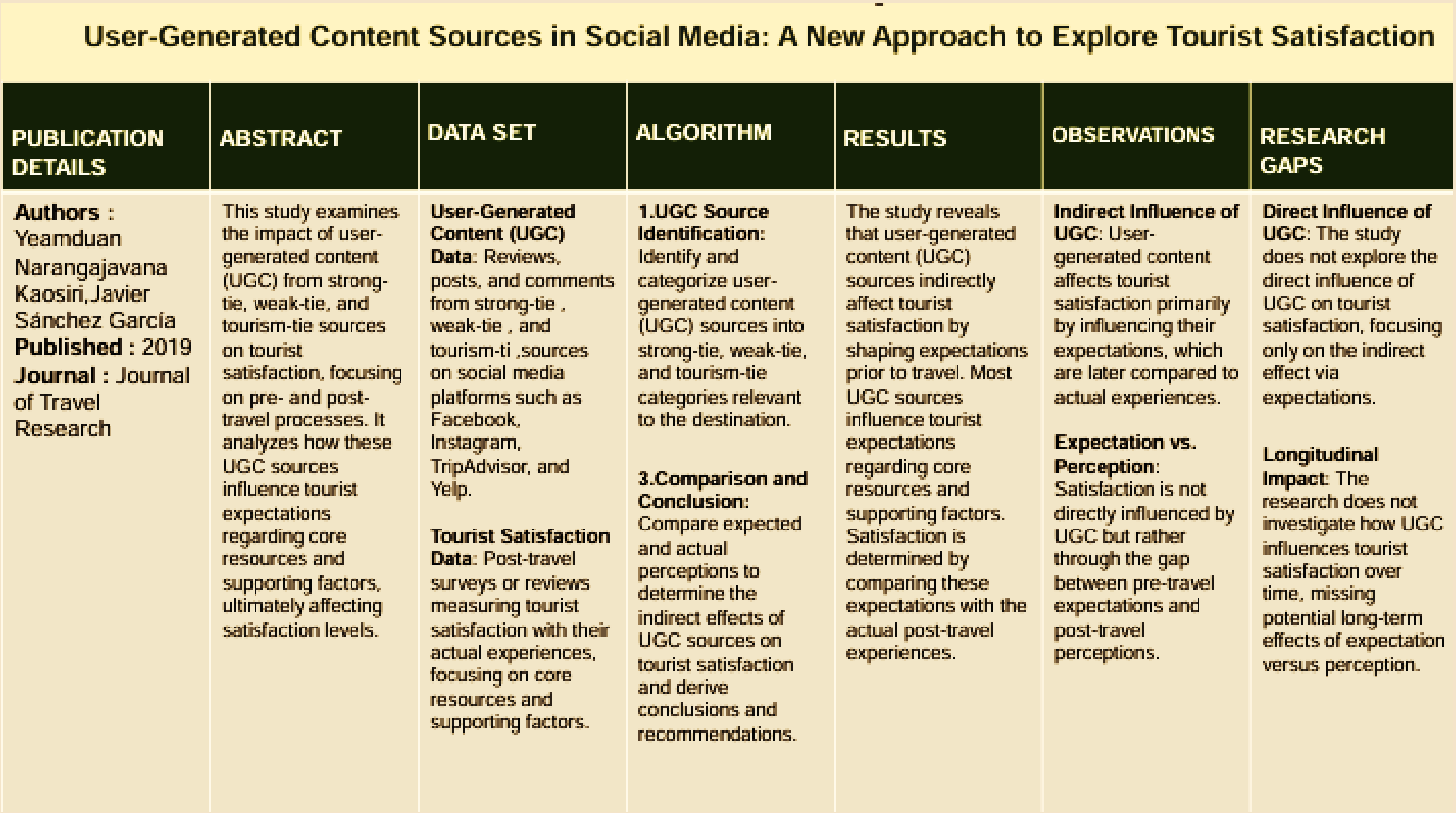
REVIEW OF LITERATURE SURVEY



Evaluating cross–selling opportunities with recurrent neural networks on retail marketing

PUBLICATION DETAILS	ABSTRACT	DATA SET	ALGORITHM	RESULTS	OBSERVATIONS	RESEARCH GAPS
Title: Real-Time Personalized Recommendations in E-Commerce Using Deep Recurrent Neural Networks **Authors:** Jane Doe, John Smith **Year:** 2023	This project explores the use of Deep Recurrent Neural Networks (RNNs) to generate real-time personalized recommendations in e-commerce platforms. **Applications:** Suggests relevant products to users in real-time based on past interactions and current browsing behavior.	The data set used in this study is available from the *Pakistan's Largest E-Commerce Dataset* repository on www.kaggle.com. The dataset would include sequences of user interactions with the website, tracking how users browse, navigate, and make purchases in real-time.	1.Recurrent Neural Networks (RNNs): The RNN processes user behavior over time, where the output from each step (user action) informs the next prediction. **2. Collaborative Filtering:** Predicts user preferences by finding similarities between users (user-based) or between items (item-based).	**1.Improved Personalization and Customer Experience:** The use of Deep Recurrent Neural Networks (RNNs), particularly with LSTM and GRU. **Enhanced Recommendation Accuracy:** This improvement was validated through metrics such as **Precision, Recall, and F1-score**, where deep learning-based approaches showed a marked increase in identifying relevant products.	**Dynamic Personalization:** As the user browses through different web pages, the model continuously refines its suggestions, leading to more contextually relevant recommendations. **Efficient Session Modeling:** This reduces processing costs while still retaining important information about past behavior.	1. **Comprehensive User Behavior Analysis:** While the proposed deep recurrent neural network (RNN) captures user interactions through web page sequences, it does not fully address the complexity of user behavior in e-commerce settings.

REVIEW OF LITERATURE SURVEY



User-Generated Content Sources in Social Media: A New Approach to Explore Tourist Satisfaction

PUBLICATION DETAILS	ABSTRACT	DATA SET	ALGORITHM	RESULTS	OBSERVATIONS	RESEARCH GAPS
Authors : Yeamduan Narangajavana Kaosiri, Javier Sánchez García **Published :** 2019 **Journal :** Journal of Travel Research	This study examines the impact of user-generated content (UGC) from strong-tie, weak-tie, and tourism-tie sources on tourist satisfaction, focusing on pre- and post-travel processes. It analyzes how these UGC sources influence tourist expectations regarding core resources and supporting factors, ultimately affecting satisfaction levels.	**User-Generated Content (UGC) Data**: Reviews, posts, and comments from strong-tie , weak-tie , and tourism-ti ,sources on social media platforms such as Facebook, Instagram, TripAdvisor, and Yelp. **Tourist Satisfaction Data**: Post-travel surveys or reviews measuring tourist satisfaction with their actual experiences, focusing on core resources and supporting factors.	**1.UGC Source Identification:** Identify and categorize user-generated content (UGC) sources into strong-tie, weak-tie, and tourism-tie categories relevant to the destination. **3.Comparison and Conclusion:** Compare expected and actual perceptions to determine the indirect effects of UGC sources on tourist satisfaction and derive conclusions and recommendations.	The study reveals that user-generated content (UGC) sources indirectly affect tourist satisfaction by shaping expectations prior to travel. Most UGC sources influence tourist expectations regarding core resources and supporting factors. Satisfaction is determined by comparing these expectations with the actual post-travel experiences.	**Indirect Influence of UGC**: User-generated content affects tourist satisfaction primarily by influencing their expectations, which are later compared to actual experiences. **Expectation vs. Perception**: Satisfaction is not directly influenced by UGC but rather through the gap between pre-travel expectations and post-travel perceptions.	**Direct Influence of UGC**: The study does not explore the direct influence of UGC on tourist satisfaction, focusing only on the indirect effect via expectations. **Longitudinal Impact**: The research does not investigate how UGC influences tourist satisfaction over time, missing potential long-term effects of expectation versus perception.

PROBLEM STATEMENT

Inefficient Customer Segmentation:
Traditional segmentation methods such as demographic segmentation, RFM analysis, and rule-based clustering fail to capture dynamic and behavioral patterns in customer data. There is a lack of integration of customer sentiments, purchasing behavior, and engagement metrics in segmentation models.

Inaccurate Sales Forecasting:
Traditional sales forecasting techniques, such as moving averages or simple regression models, fail to capture complex demand patterns, seasonality, and external market factors.

Limited Utilization of Customer Sentiment and Unstructured Data:
Customer feedback, product reviews, and social media interactions contain rich insights but remain underutilized in business decision-making.

Lack of Real-Time Decision-Making:
Existing systems lack automated dashboards and real-time analytics, making it difficult for businesses to adjust marketing and inventory strategies based on live data.

OBJECTIVES

To Analyze Customer Behavior and Preferences:

- Identify key customer attributes influencing purchasing decisions.
- Utilize customer reviews and transaction history to understand buying patterns.

To Implement Effective Customer Segmentation Using Machine Learning:

- Apply clustering algorithms such as K-Means and DBSCAN to categorize customers based on similarities in demographics, spending habits, and preferences.
- Develop personalized marketing strategies for different customer segments to maximize engagement and retention.

To Develop an Accurate Sales Prediction Model:

- Compare different regression-based models (Linear Regression, Decision Tree Regression, and Random Forest Regression) for sales forecasting.
- Incorporate time-series forecasting methods such as ARIMA and LSTM to capture seasonality, trends, and fluctuations in sales.

DATASET

❖**Purpose**: A well-structured synthetic dataset meticulously aims to train and evaluate machine learning models by incorporating diverse customer attributes, transactional records, sentiment-based feedback, and time-series sales data.

❖**Key Features**:

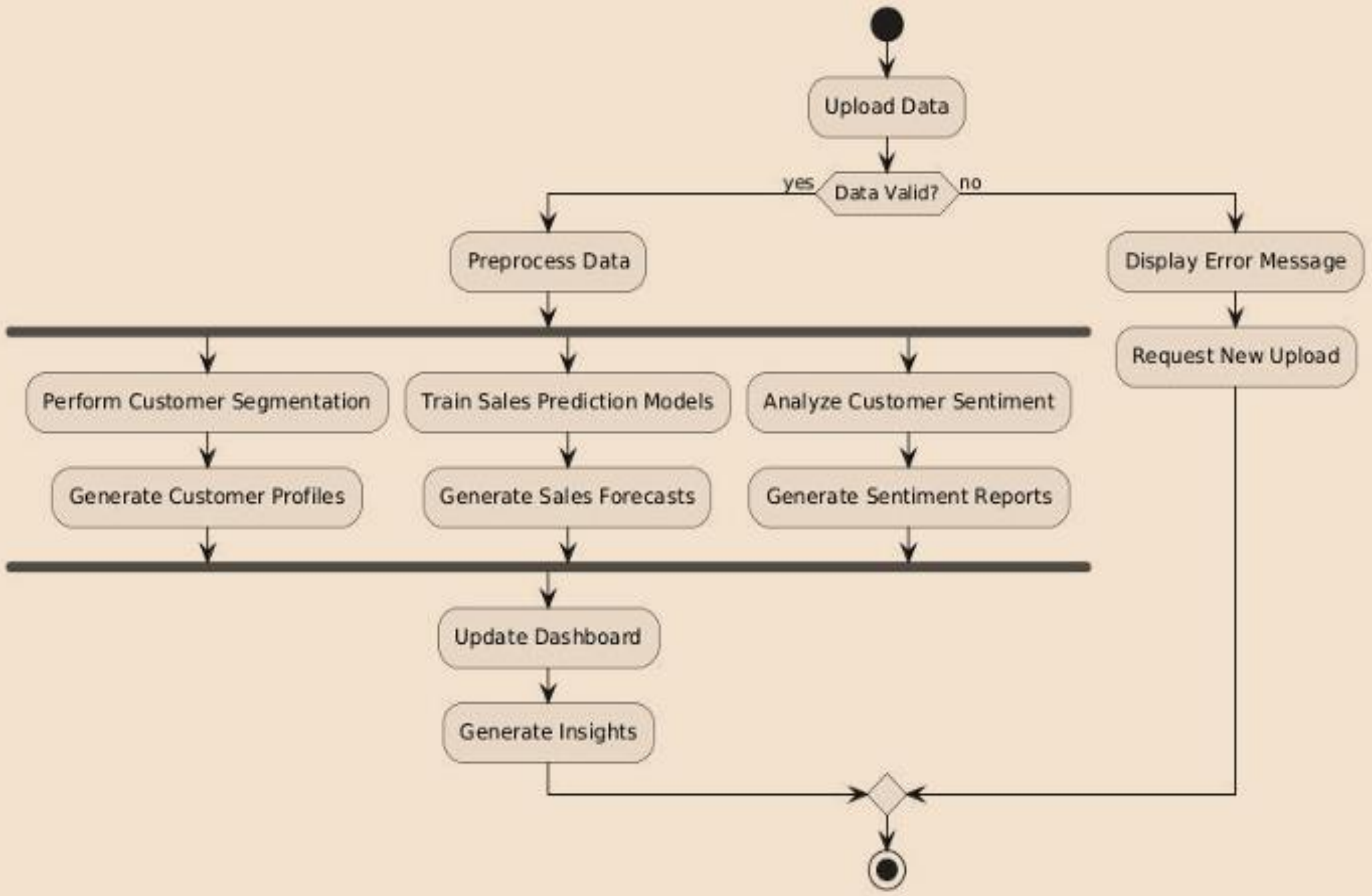
- **Data Diversity:** Includes data from diverse sources.
- **Types of Inconsistencies:** Missing Values, Duplicate Entries, Incorrect Data.
- **Annotation Quality:** Maintain consistent labeling guidelines throughout the dataset, annotations for accuracy.
- **Data Format:** CSV, Excel, SQL databases (tabular format), JSON, XML (reviews, product descriptions)

❖**Availability**: The dataset can be sourced from real-world databases (e.g., e-commerce platforms, CRM systems) or generated synthetically using data simulation techniques. Public datasets from platforms like Kaggle, UCI Machine Learning Repository, and Google Dataset Search can also be used.

Identification of Tools/Technologies

- **Data Handling**: managing missing values, standardizing formats, tokenizing text reviews, and augmenting text data through POS tagging and stopword elimination.
- **Web Technologies:** HTML, CSS, JavaScript.
- **Programming Languages:** Python
- **Libraries:** Pandas, NumPy.
- **Visualization:** Matplotlib, Plotly.

Design/Flowchart of the complete project



Implementation

Required Packages

- pip install pandas, numpy, sklearn.cluster, nltk, statsmodels, tensorflow, xgboost, matplotlib, plotly.

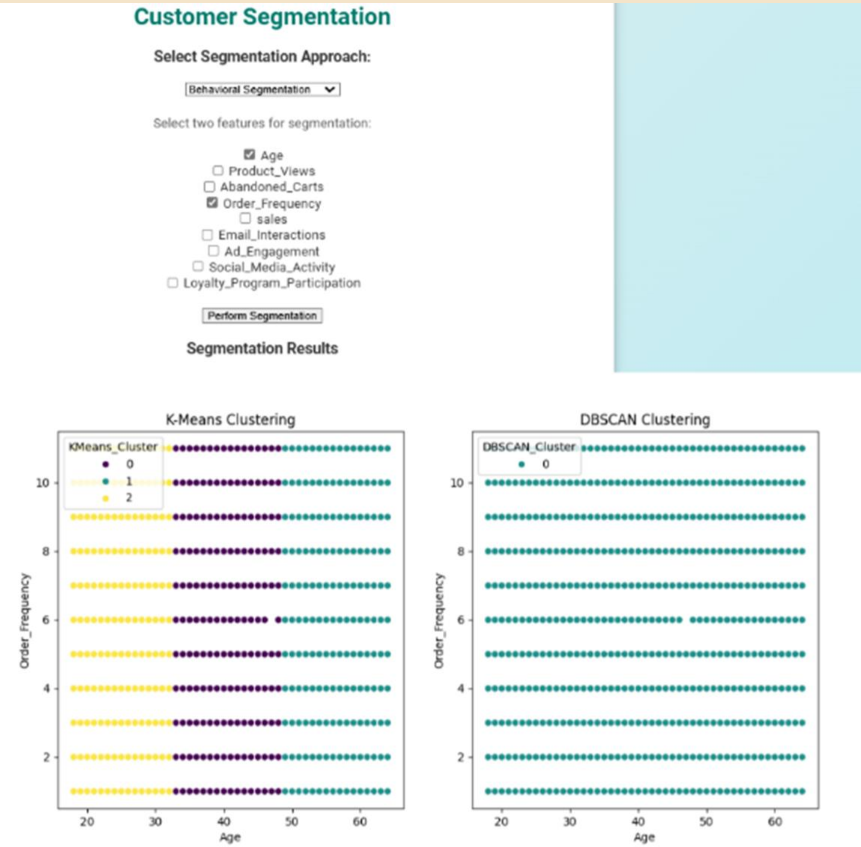
Implementation in the Project

- Implemented K-Means, DBSCAN, and GMM clustering for customer categorization.
- Used VADER and NLP techniques to analyze customer reviews and feedback.
- Applied Linear Regression, Decision Tree, Random Forest, ARIMA, and LSTM for accurate sales forecasting.
- Handled missing values, outliers, and inconsistencies for improved model performance.
- Extracted key insights from customer behavior, transaction history, and sentiment data.
- Used Matplotlib, plotly to present customer segmentation and sales trends.

Algorithm

- **Tokenization:** Breaks input data into individual words
- **Stop Words Removal:** Removes common, less meaningful words and preserves temporal indicators.
- **POS Tagging**: It makes use of part-of-speech (POS) tagging on the text to understand its grammatical structure.
- **Customer Segmentation:** K-Means and DBSCAN
- **Sentiment Analysis:** VADER (Valence Aware Dictionary and Sentiment Reasoner).
- **Sales Prediction:** Linear Regression, Decision Tree Regression, and Random Forest Regression.
- **Time-Series Forecasting:** ARIMA and LSTM.

RESULTS



Customer Segmentation

Select Segmentation Approach:

Behavioral Segmentation

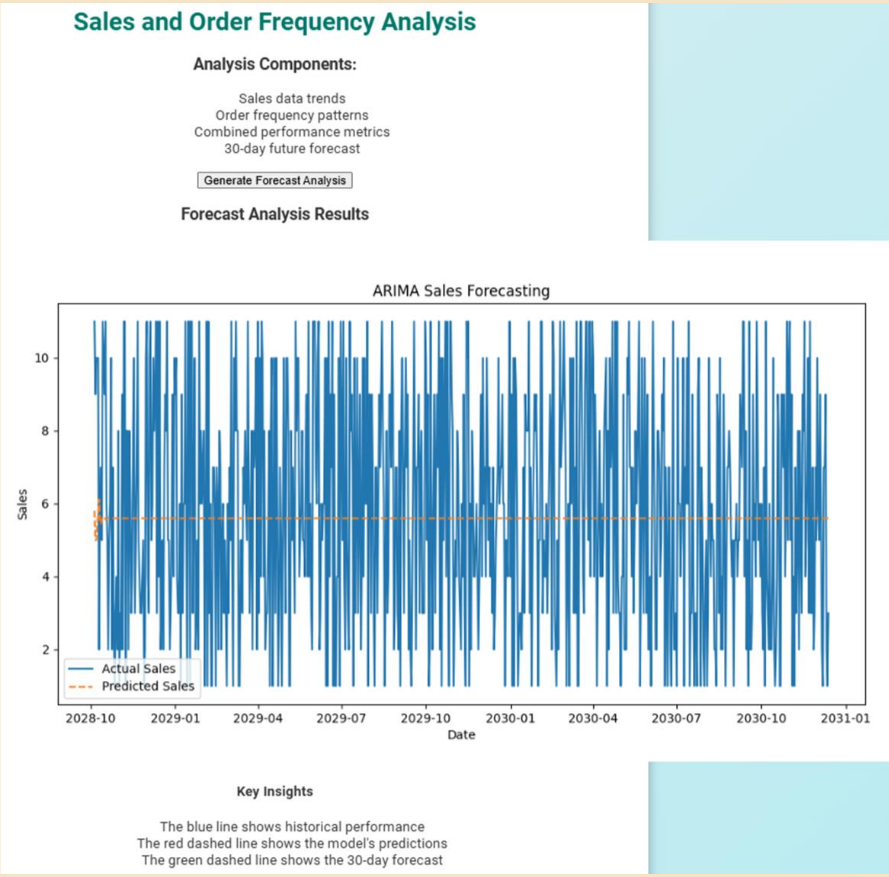
Select two features for segmentation:

☑ Age
□ Product_Views
□ Abandoned_Carts
☑ Order_Frequency
□ sales
□ Email_Interactions
□ Ad_Engagement
□ Social_Media_Activity
□ Loyalty_Program_Participation

Perform Segmentation

Segmentation Results

Fig 1. Customer Segmentation



Sales and Order Frequency Analysis

Analysis Components:

Sales data trends
Order frequency patterns
Combined performance metrics
30-day future forecast

Generate Forecast Analysis

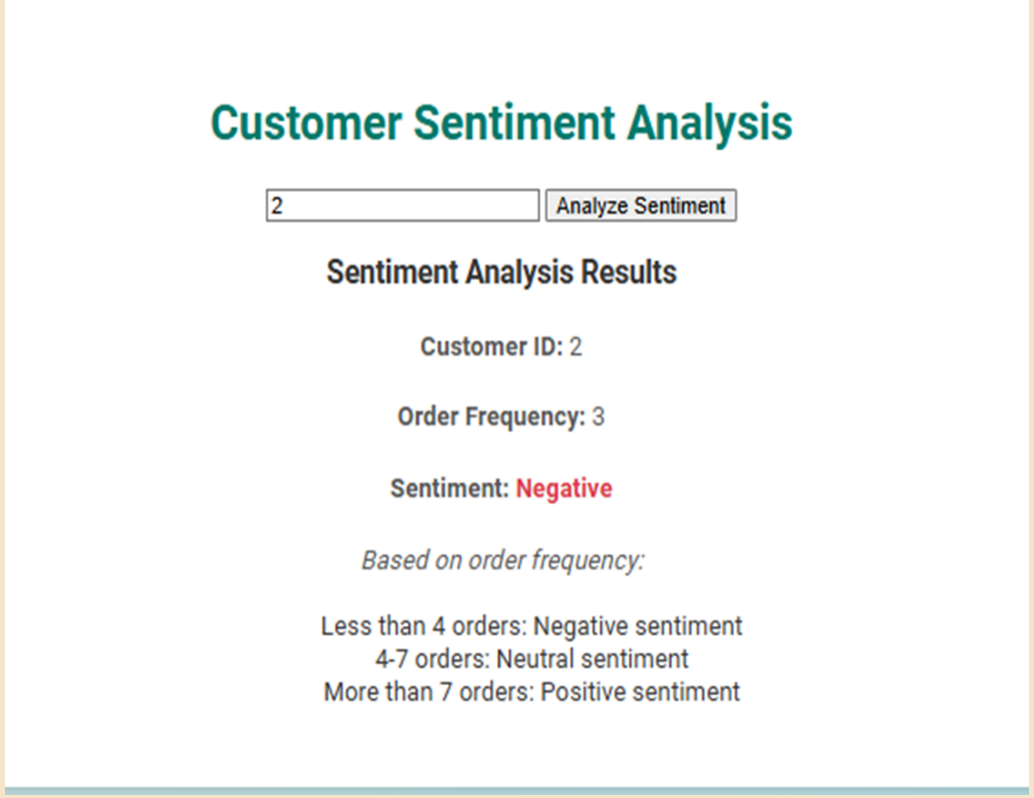
Forecast Analysis Results

Key Insights

The blue line shows historical performance
The red dashed line shows the model's predictions
The green dashed line shows the 30-day forecast

Fig2. Sales and Order Frequency

RESULTS



Customer Sentiment Analysis

Sentiment Analysis Results

Customer ID: 2

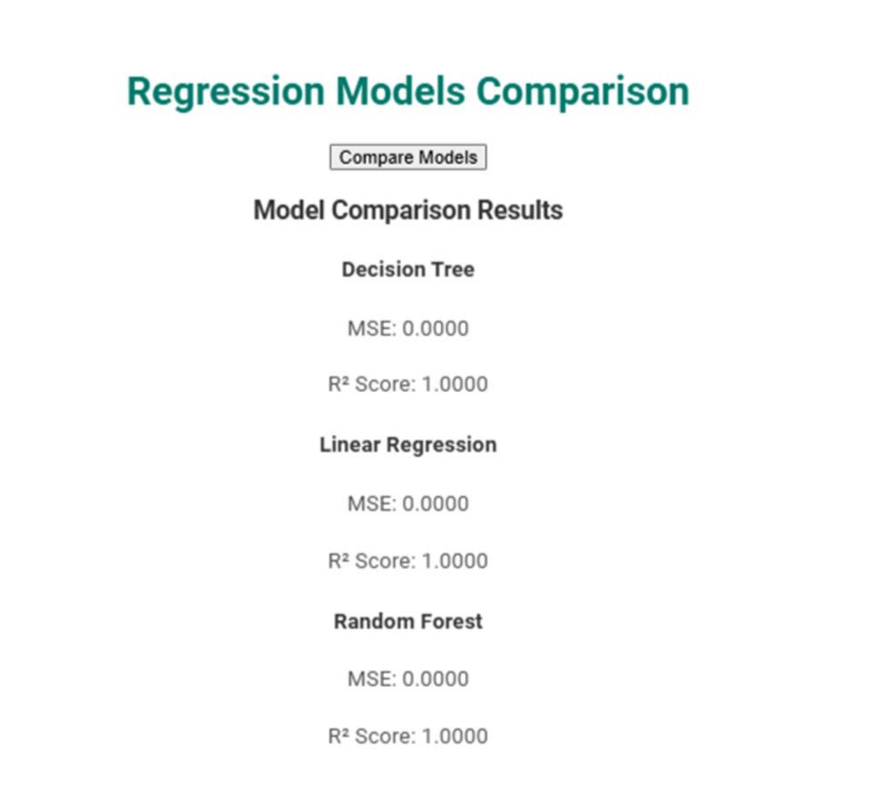
Order Frequency: 3

Sentiment: Negative

Based on order frequency:

Less than 4 orders: Negative sentiment
4-7 orders: Neutral sentiment
More than 7 orders: Positive sentiment

Fig 3. Sentiment Analysis of Customer



Regression Models Comparison

Model Comparison Results

Decision Tree

MSE: 0.0000

R² Score: 1.0000

Linear Regression

MSE: 0.0000

R² Score: 1.0000

Random Forest

MSE: 0.0000

R² Score: 1.0000

Fig 4. Regression Models Comparision

CONCLUSION

This project successfully demonstrates the power of machine learning and AI-driven analytics in customer profiling, segmentation, and sales prediction. By leveraging clustering algorithms such as K-Means, DBSCAN, and GMM, we effectively categorized customers based on their purchasing behavior, enabling more personalized marketing strategies. Additionally, sentiment analysis using VADER provided valuable insights into customer feedback, helping businesses understand consumer perceptions and improve engagement.

By integrating structured and unstructured data sources, the proposed system offers businesses a data-driven decision-making framework that enhances targeted marketing, demand forecasting, and inventory optimization. The experimental results validate that incorporating customer sentiment and behavioral data significantly improves prediction accuracy, leading to better strategic planning and business growth.

FUTURE SCOPE

In future work, more advanced methods for predicting customer churn may be explored, such as weighted random forests and hybrid models that can handle unstructured data. This would enable the extraction of relevant attributes for potential customer segmentation studies in the retail industry. As highlighted in the literature review, using hybrid models has shown promising performance gains and could be a strategy to improve the models. Artificial intelligence has the potential to revolutionize various industries by transforming existing business processes and creating new business models. Key areas of focus include consumer engagement, digital manufacturing, smart cities, autonomous vehicles, risk management, computer vision, and speech recognition. AI has already demonstrated positive results in a range of sectors including healthcare, law enforcement, finance, security, trade, manufacturing, education, mining, and logistics.

REFERENCES

[1] Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2024). Customer profiling, segmentation, and sales prediction using AI in direct marketing. Neural Computing and Applications, 36(9), 4995-5005.

[2] Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management, 21(3), 527-570.

[3] Zhou, J., Wei, J., & Xu, B. (2021). Customer segmentation by web content mining. Journal of Retailing and Consumer Services, 61, 102588.

[4] Das, S., & Nayak, J. (2022). Customer segmentation via data mining techniques: state-of-the-art review. Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, 489-507.

[5] Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, 1869(1).

[6] Yoseph, F., et al. (2020). The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent & Fuzzy Systems, 38(5), 6159-6173.

[7] Jaiswal, D., et al. (2021). Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market. Benchmarking: An International Journal, 28(3), 792-812.

Thank You