Research Project Proposal

Research Project Title:

Customer Segmentation Using Deep Learning for Enhanced Banking Insights

Group Members:

A005 - Arpan Sakpal

A014 - Himansi Raval

Professor Name:

Dr. Yogesh Naik

BUSINESS PROBLEM

Customer segmentation is the process of dividing customers into groups based on shared characteristics to better understand and serve their needs. Despite its importance, traditional customer segmentation approaches face significant limitations, particularly in the banking sector:

- **Limited Understanding of Diverse Needs:** Traditional rule-based models rely on basic demographic and transactional data. These models fail to account for the complex needs of individual customers and the diversity in their spending patterns. This results in a one-size-fits-all approach that overlooks unique customer characteristics and behaviors.

- **Inability to Capture Complex Behaviors:** Transaction descriptions often contain rich, unstructured data, such as vendor names, product categories, or transaction notes, which are not leveraged effectively. Traditional models struggle to process this unstructured data, missing critical insights into nuanced spending behaviors.

- **Inefficient Targeting:** Broad and generic segmentation leads to poor targeting, where customers receive irrelevant marketing offers. This decreases customer satisfaction, increases churn rates, and damages the perception of the bank's personalized services.

- **Missed Opportunities:**
    - Banks fail to extract actionable insights from transaction descriptions, losing the ability to identify niche spending categories such as "luxury goods," "subscriptions," or "frequent travel."
    - High-value customers, such as those with consistent high-value transactions or niche interests, often remain undetected due to the lack of granular segmentation.
    - Seasonal trends or event-driven spending patterns, such as increased travel during holidays, are overlooked, limiting the effectiveness of timely offers and promotions.
    - Cross-selling and up-selling opportunities are missed, such as recommending relevant loan products to customers frequently purchasing high-value goods or premium credit cards for frequent travelers.
    - The inability to capitalize on these opportunities not only results in immediate revenue losses but also hampers the bank's potential to build loyalty and long-term relationships with its customers.

- **Manual and Inefficient Processes:** Traditional models depend heavily on manual feature engineering to identify patterns in customer data. This approach is time-consuming, labor-intensive, and unsuitable for handling the scale and complexity of modern transactional datasets. With the rise in digital transactions, the sheer volume of data renders manual processes ineffective and prone to oversight.

CONTEXT OF THE DATA SCIENCE PROJECT

This research project focuses on leveraging transaction descriptions to analyze customer spending behaviors for segmentation. Transaction descriptions are a rich, yet often underutilized, source of information that can provide deep insights into customer preferences, habits, and spending patterns. By analyzing these descriptions, banks can unlock previously hidden patterns and significantly enhance their segmentation strategies. **Key Focus Areas are:**

1. **Automated Categorization of Transactions:**
    - Transaction descriptions often contain valuable information about the nature of a customer's spending. For example, a transaction labeled "IRCTC" can indicate travel-related expenses, while "Zomato" points to food delivery services. By employing Natural Language Processing (NLP) techniques, the project aims to extract keywords, classify them, and associate them with relevant spending categories (e.g., "Travel," "Food & Dining," or "Shopping").
    - Automation ensures that this categorization is accurate, scalable, and removes the need for labor-intensive manual processes.
2. **Segmentation Model Development:**
    - Unlike traditional models that rely on static, rule-based segmentation, this project focuses on developing a deep learning-based model to analyze transaction data and create adaptable customer segments.
    - The model will group customers based on patterns derived from the data, ensuring the segmentation reflects detailed and meaningful behaviors over time.
3. **Enhanced Personalization and Insights:**
    - Transaction descriptions offer insights into individual spending habits, such as frequent purchases of groceries, recurring subscriptions, or occasional luxury spending. By analyzing these patterns:
        - Banks can personalize their services to better match customer preferences (e.g., offering cashback on travel expenses for frequent travelers).
        - Emerging trends, such as increased spending on online education or sustainability-focused products, can be identified and acted upon proactively.
        - High-value customers with niche interests, such as those frequently purchasing luxury items, can be targeted with premium banking services.

Goals and Impact of the Project:

- **Improved Customer Experience:** Personalized services and offers, tailored to individual preferences, enhance customer satisfaction and loyalty.
- **Increased Revenue Opportunities:** By identifying cross-selling and up-selling opportunities (e.g., travel insurance for frequent travelers or premium cards for high-spenders), banks can significantly boost their revenue streams.
- **Scalable and Efficient Processes:** Automating the categorization of transaction descriptions eliminates inefficiencies and ensures that the system is scalable to handle large datasets and increasing transaction volumes.

LITERATURE REVIEW

1) *A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases* - [LINK]

Author(s): Miguel Alves Gomes, Tobias Meisen
Source: Springer (2023)

- This paper provides an in-depth review of customer segmentation methods used in e-commerce and highlights the progression from traditional rule-based approaches to modern machine learning-driven techniques.
- It outlines various clustering algorithms (e.g., k-means, DBSCAN) and advanced techniques like deep learning, which improve segmentation accuracy and adaptability.
- The authors emphasize the importance of dynamic segmentation models capable of processing large-scale and diverse customer data.
- The paper identifies key challenges in customer segmentation, such as difficulty in handling high-dimensional data, lack of scalability, and inability to adapt to real-time changes in consumer behavior.
- Case studies are presented to showcase how advanced segmentation has led to improved personalized targeting and customer retention in e-commerce.

Limitations:
- The study primarily focuses on e-commerce data and does not extensively explore segmentation in other domains like banking, where transaction data plays a crucial role.
- It does not address the challenges of integrating unstructured data, such as transaction descriptions, into segmentation models.

Relevance to Current Research:
- Reinforces the need for scalable and dynamic segmentation models.
- Suggests that clustering and deep learning methods can be adapted to the banking domain for improved transaction-based segmentation.

2) *Sentiment Analysis of Consumer Feedback and Its Impact on Business Strategies by Machine Learning* - [LINK]

Author(s): Gupta Mohit, Prof. Dr. Brune Philipp, Prof. Dr. Faußer Stefan
Source: HNU Publications

- This paper focuses on how consumer feedback can be analyzed using machine learning for strategic business decisions.
- Various methods, including supervised learning models (e.g., support vector machines, logistic regression) and unsupervised techniques (e.g., topic modeling), are explored for understanding sentiment in text data.
- The authors stress that unstructured data like consumer feedback can reveal deeper behavioral insights, such as customer satisfaction levels, recurring complaints, and preferences.

- The findings show that businesses leveraging sentiment analysis can significantly improve customer satisfaction by addressing pain points and tailoring their offerings.

Limitations:

- The paper is limited to sentiment analysis and does not directly address customer segmentation.
- It primarily discusses feedback and review-based data, which may differ in structure and complexity from transaction descriptions.

Relevance to Current Research:

- Highlights the importance of extracting meaningful insights from unstructured data, such as transaction descriptions.
- Suggests that NLP techniques used in sentiment analysis can be adapted for keyword extraction and categorization in banking transaction data.

3) *Improve Profiling Bank Customer's Behavior Using Machine Learning-* LINK

Author(s): Emad Abd Elaziz Dawood, Essamedean Elfakhrany, Fahima A. Maghraby
Source: IEEE Xplore (2019)

- This paper investigates how machine learning techniques can improve the profiling of banking customers by analyzing transactional and demographic data.
- Traditional clustering algorithms like k-means, DBSCAN, and hierarchical clustering are applied to segment customers based on their spending behaviors.
- The study emphasizes the importance of accurately profiling customers to offer tailored banking products and services.
- The authors discuss the challenges of handling imbalanced datasets and propose preprocessing methods to address this issue, ensuring better segmentation outcomes.
- Results demonstrate that machine learning-based segmentation improves the identification of high-value customers and their specific needs.

Limitations:

- The study does not explore the use of advanced models, such as deep learning or NLP, for handling unstructured or textual data.
- It primarily addresses clustering techniques without discussing how to integrate real-time data into customer profiling.

Relevance to Current Research:

- Validates the use of clustering techniques for customer segmentation in the banking sector.
- Points to the need for incorporating advanced methods, such as deep learning, to address the limitations of traditional clustering models.

The reviewed papers collectively highlight the shift towards machine learning and deep learning models for dynamic, scalable customer segmentation. They emphasize the importance of leveraging unstructured data, such as transaction descriptions, to extract deeper behavioral insights. These advancements enable personalized targeting, improved customer retention, and enhanced revenue opportunities across domains like e-commerce and banking.

RESEARCH GAP

In the evolving financial sector, customer segmentation is key to delivering personalized banking experiences. Traditional segmentation approaches rely heavily on demographic and transactional data but fail to leverage the rich, unstructured information within transaction descriptions. This project aims to address these limitations by applying advanced techniques like Natural Language Processing (NLP) and Deep Learning to analyze transaction descriptions. By doing so, we seek to uncover hidden patterns in customer behavior, enabling banks to enhance personalization, improve targeting, and unlock new revenue opportunities. The existing body of research highlights the effectiveness of machine learning and deep learning techniques for customer segmentation in e-commerce and other domains. However, several gaps remain, particularly in the context of banking:

Lack of Focus on Transaction Descriptions:

While previous studies have explored demographic and transactional data, they fail to utilize the rich, unstructured data contained in transaction descriptions. This data has the potential to unlock deeper insights into customer behavior and preferences.

Limited Application of Advanced NLP Techniques in Banking:

Although Natural Language Processing (NLP) is extensively used for sentiment analysis and topic modeling, its application in categorizing banking transaction descriptions remains underexplored. Current models fail to identify specific spending patterns or categories like luxury goods, subscriptions, or seasonal trends.

Static and Generic Segmentation Models:

Most existing studies rely on static segmentation techniques, such as rule-based or clustering methods, which lack adaptability to evolving customer behaviors. There is a need for dynamic and scalable models that can process large volumes of real-time transactional data.

Challenges in Personalization:

Despite advancements in machine learning, current banking models do not effectively leverage segmentation to offer hyper-personalized services. This results in missed opportunities for cross-selling and up-selling.

Integration of Deep Learning Models in Banking:

While clustering algorithms like k-means have been widely used, the potential of deep learning models, particularly in the context of banking, is not fully realized. Models such as autoencoders or recurrent neural networks (RNNs) could provide more granular insights into spending behaviors.

RESEARCH DESIGN

This research adopts a **Quantitative Design** to analyze and understand customer spending behaviors using advanced analytical techniques. The process is structured into several phases:

1. **Data Understanding and Preprocessing**
   The research begins with gathering and understanding multiple datasets, including:

- Publicly available transactional data sourced from platforms like Kaggle.
- Synthetic datasets, which simulate diverse transaction categories and enhance the scope of analysis.
- Personal transactional records, which act as a testing set to validate the developed models.

The preprocessing phase focuses on:

- Cleaning and normalizing data to address issues such as missing values, inconsistent formats, and irrelevant attributes.
- Handling transaction descriptions using Natural Language Processing (NLP) techniques, including tokenization, stemming, and stop-word removal.
- Reducing noise and extracting meaningful patterns to ensure high-quality data for analysis.
- **Algorithms Used:** Standard preprocessing libraries such as NLTK and spaCy for NLP-based cleaning and normalization, and deep learning-based text preprocessing using LSTMs.

This phase ensures the datasets are consistent, reliable, and ready for feature extraction and analysis.

2. **Feature Engineering and Representation:**
   The focus is on extracting meaningful information from transaction descriptions using advanced Natural Language Processing (NLP) techniques. The key steps in this phase involve:

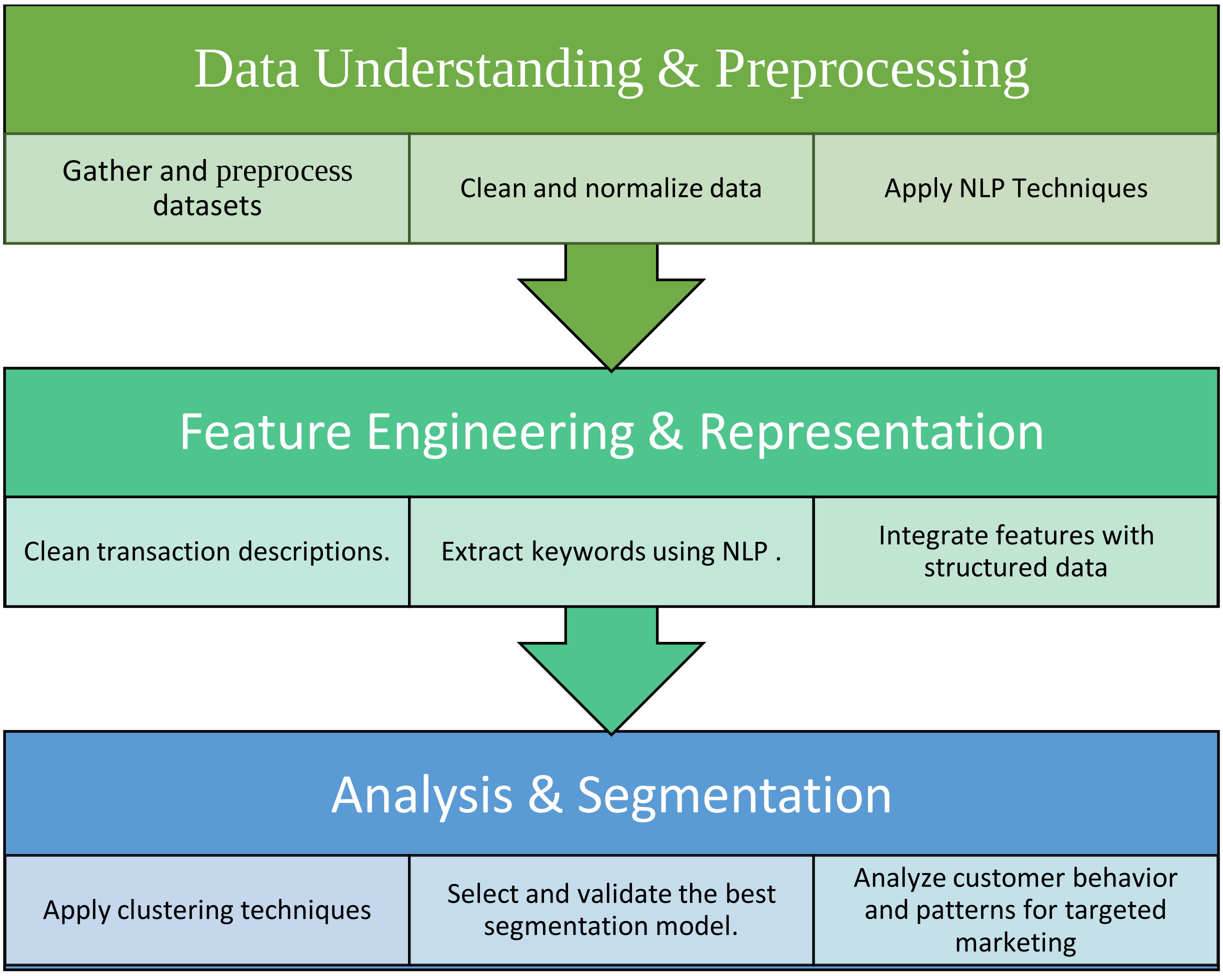
- Data Transformation: The transaction description column is cleaned and preprocessed to remove any irrelevant content, such as stop words or special characters.
- Keyword Extraction: NLP techniques, such as Named Entity Recognition (NER), are used to identify specific keywords in the text, such as hotel names, brands (e.g., Starbucks), and locations (e.g., IRCTC).
- Embedding Representation: Deep learning models like Word2Vec or BERT are employed to convert the transaction descriptions into word embeddings, capturing semantic relationships between keywords. This will allow the model to better understand context and improve extraction accuracy.
- Feature Integration: The extracted keywords and word embeddings are combined with structured features (such as transaction amount, frequency, and customer demographics) to build a robust feature set that can be used for downstream analysis or model training.
- **Algorithms Used:** Named Entity Recognition (NER) with spaCy, BERT embeddings using Transformers library, and deep learning-based autoencoders for feature extraction.

This process ensures the transaction descriptions are leveraged effectively to extract meaningful insights and support customer segmentation or other behavioral analysis tasks.

3. **Analysis and Segmentation:**
In this phase, analytical techniques are applied to uncover patterns in customer spending behaviors based on the features created. The focus is on segmenting customers into meaningful groups using both structured and unstructured data, including the transaction descriptions and extracted keywords. Key steps include:

- Segmentation Techniques: Different methods like clustering (e.g., K-means, DBSCAN) are tested to group customers based on similarities.
- Model Selection: The best-performing segmentation model is chosen based on evaluation metrics.
- Behavioral Patterns: Insights from the segmentation provide a deeper understanding of customer preferences, purchasing habits, and potential opportunities for targeted marketing or personalized services.
- Segmentation Validation: The segmentation model is validated using the personal transaction dataset, ensuring its effectiveness in real-world scenarios. This step is crucial to verify the model's reliability and its ability to generalize beyond training data.
- **Algorithms Used:** K-means for clustering, DBSCAN for density-based segmentation, deep learning-based variational autoencoders (VAEs) for representation learning, and Silhouette Score for model evaluation.

This structured approach ensures a comprehensive understanding of customer spending behaviors, leading to valuable insights that can enhance financial decision-making and targeted marketing strategies.



DATA COLLECTION STRATEGY

The data collection strategy for this research combines multiple datasets to ensure a comprehensive and diverse approach to customer segmentation.
By leveraging both *real-world and synthetic data*, the research aims to build a robust model that can handle various customer behaviors and transaction patterns.
The datasets used in this study include :
- two Kaggle datasets, which provide rich transactional and demographic information,
- synthetic data generated using the Faker library to simulate diverse spending behaviors, and
- personal transaction records to validate the model's real-world applicability.
This combination ensures a well-rounded foundation for the segmentation model, allowing for accurate analysis and insights into customer spending behaviors.

1. **Kaggle Datasets**:

    1. **Spending Habits by Category and Item (Kaggle)**: This dataset provides detailed transactional data categorized by spending habits. It includes information like transaction amounts, descriptions, and item categories, which are essential for understanding the types of purchases made by different customer groups. By using this dataset, the model can identify spending behaviors across various categories and determine how customers with similar spending patterns should be grouped. The data's variety and real-world relevance make it an ideal foundation for building segmentation models.
    2. **Customer Segmentation Dataset (Kaggle)**: This dataset focuses on customer demographics, behavioral data, and transactional history. It includes attributes like age, income, geographic location, and spending patterns. It will serve to enrich the customer segmentation process by combining demographic information with transactional data. This will help identify how different customer attributes influence purchasing behavior and enable more tailored segmentation. The rich diversity of customer attributes allows for more nuanced segmentation strategies and helps address the challenges of varied customer needs.

2. **Synthetic Transaction Dataset**:

    - **Simulated Data Generated Using Faker**: To complement the Kaggle datasets and provide additional diversity, a synthetic transactional dataset has been generated using the Faker library. This dataset includes fabricated transaction records, each with corresponding transaction descriptions (e.g., shopping, dining, subscriptions) and transaction categories. The synthetic data simulates a broad range of customer behavior, ensuring that the segmentation model can be trained on diverse types of spending behaviors. The use of synthetic data provides flexibility in expanding the dataset, while also addressing potential concerns regarding data privacy or limited access to real-world data. Furthermore, the synthetic nature of the data allows for controlled testing and model validation without being constrained by privacy or ethical issues related to real customer information.

3. Personal Transaction Record Statements:

- **User’s Own Bank Statement Data**: The personal transaction record statements, which include the user’s own banking transactions, are employed as a testing set to validate the performance. This data includes transaction descriptions, amounts, and merchant names, providing a real-world benchmark for evaluating the model’s effectiveness. By using personal data, the research ensures that the final model performs accurately on data similar to what actual customers experience. Additionally, these records are used to test how well the model can identify unique customer behaviors and segment them accordingly. The authenticity of the personal transaction data ensures that the model is grounded in real-world scenarios, providing more confidence in its predictive power and ability to generalize across different customer types.

TOOLSET

1. **Algorithms and Models**:
    - Deep Learning Models (LSTM, CNN, etc.): These models are being utilized for processing and analyzing transaction descriptions. Long Short-Term Memory (LSTM) models can be effective in understanding sequences of transaction descriptions, which may hold significant predictive value for segmentation. Convolutional Neural Networks (CNNs) may be leveraged to extract patterns from textual data using the concept of feature maps.
    - NLP Models (Word2Vec, GloVe, or BERT,etc): These pre-trained models can be applied to understand the context of transaction descriptions.
    - Clustering Models (K-means, DBSCAN, Agglomerative Clustering): These will help segment customers based on various features extracted from the transaction data, both structured and unstructured.

2. **Toolsets**:
    - Python Libraries (Pandas, NumPy, Scikit-learn): These libraries will be used for data preprocessing, feature engineering, and data analysis. They provide powerful tools for cleaning, transforming, and manipulating transactional data.
    - TensorFlow / Keras: For implementing deep learning models such as LSTM and CNN, TensorFlow (or Keras as a higher-level interface) is used. These libraries offer powerful tools for model creation, training, and optimization. TensorFlow's wide support for distributed computing ensures that the models can scale efficiently.
    - NLTK / SpaCy: Natural Language Processing libraries like NLTK or SpaCy will be employed to tokenize, clean, and preprocess transaction descriptions. These tools will be used for removing noise from the text, extracting keywords, and analyzing customer behavior patterns.

3. **Datasets / Surveys**:
    - Transactional Datasets from Kaggle: These datasets serve as the primary source for training the customer segmentation model. They provide a comprehensive set of transaction records that can be used to detect spending patterns, identify customer preferences, and segment customers accordingly.
    - Synthetic Transactional Data: The synthetic data provides additional variety and complexity to the dataset, helping to test the generalization capability of the model across different types of customers and spending patterns.
    - Personal Transaction Record Statements: The real-world testing set comes from the user's personal transaction records, which will be used to test the model's real-world applicability and ensure it aligns with real customer spending behavior.

DATA ANALYSIS APPROACH

1. **Understanding the Problem**:
   - Focus on customer segmentation using deep learning to address limitations of traditional rule-based models in the banking sector.
   - Leverage transactional descriptions to capture unstructured data and extract deeper behavioral insights.

2. **Data Collection and Preprocessing**:
   - Use Kaggle datasets, synthetic data (via the Faker library), and personal transaction records for a comprehensive dataset.
   - Clean transaction descriptions (remove noise, stop words) and tokenize for NLP processing.
   - Categorize transactions into categories (e.g., Travel, Food, Luxury Goods).

3. **Feature Engineering**:
   - Generate word embeddings for semantic relationships in transaction descriptions.

4. **Model Development**:
   - Use clustering algorithms (e.g., k-means, DBSCAN) for initial segmentation.
   - Build advanced deep learning models (LSTM, Autoencoders) to dynamically refine segmentation.
   - Apply NLP models (e.g., BERT, Word2Vec) to process and analyze unstructured transaction descriptions.

5. **Validation and Testing**:
   - Various metrics can be used to assess the quality of the customer segmentation models. Commonly used metrics include the Silhouette Score, Davies-Bouldin Index ,Inertia,etc.
   - Test models across real-world (Kaggle, personal records) and synthetic datasets for robustness.

6. **Insights and Recommendations**:
   - The analysis focuses on understanding customer groups and identify key customer segments.

TIMETABLE OF THE PROJECT

Phase	Timeline
Research and Data Exploration	Week 1
Data Preprocessing and EDA	Week 2
Model Development	Week 3
Documentation and Final Preparations	Week 4

Week 1: Research and Data Exploration
In the first week, we conducted a literature review and explored various datasets. After evaluating the data, we finalized the most relevant datasets for the project.

Week 2: Data Preprocessing and EDA
We focused on the transaction descriptions, performing extensive Exploratory Data Analysis (EDA) to understand patterns and clean the data. This step was crucial for ensuring data quality and preparing it for modeling.

Week 3: Model Development
We moved toward model development, experimenting with different algorithms. After testing and evaluating the performance, we finalized the model that best fit the requirements.

Week 4: Documentation and Final Preparations
In the final week, we focused on documenting the process, including methodology and results. We also prepared a report for submission and worked on presenting the project through a poster.

LIMITATIONS

1. Dependence on Data Quality:

The accuracy of segmentation relies heavily on the quality of transaction data. Missing, incomplete, or inconsistent data (e.g., poorly labeled transaction descriptions) may affect model performance.

2. Unstructured Data Challenges:

Transaction descriptions may vary significantly in format and quality across different banks or merchants, making it challenging to standardize and categorize them effectively.

3. Scalability Issues:

Processing large volumes of transactional data in real-time requires significant computational resources, which may limit scalability for smaller banking institutions.

4. Synthetic Data Limitations:

Synthetic datasets, while useful for training and testing, may not perfectly capture the diversity and nuances of real-world customer behaviors, reducing the model's effectiveness when applied to real-world scenarios.

5. Generalization to Different Regions:

Spending patterns and transaction descriptions may vary based on geographic or cultural factors, which could limit the generalizability of the model to different regions.

6. Complexity of Deep Learning Models:

Deep learning models like LSTMs or Autoencoders are computationally intensive and require substantial training time. Moreover, they may be challenging to interpret, making it difficult to explain segmentation outcomes to stakeholders.

7. Static Segmentation for Evolving Behaviors:

The model may struggle to adapt to rapidly changing spending patterns, such as seasonal trends or sudden shifts (e.g., during a pandemic or economic crisis).

8. Ethical and Privacy Concerns:

Using personal transaction records for segmentation raises privacy concerns, and compliance with data protection laws like GDPR must be ensured.

FUTURE SCOPE

1. **Real-Time Customer Segmentation:**
   Extend the project to enable real-time processing and segmentation of transaction data, allowing banks to provide instant recommendations and offers.

2. **Incorporation of Additional Data Sources:**
   Integrate external data sources like social media activity, loyalty programs, and customer feedback to enrich segmentation and provide deeper insights.

3. **Cross-Bank Implementation:**
   Adapt the model for use across multiple banks, addressing the challenges of varying transaction formats and customer demographics.

4. **Customizable Segmentation Framework:**
   Create a modular framework that banks can customize based on their specific needs, such as focusing on niche customer groups or regional markets.

5. **Scalability Enhancements:**
   Optimize the model for deployment in cloud-based environments, enabling efficient handling of massive transactional datasets for large banks.

6. **Impact Analysis of Recommendations:**
   Include a module to measure the effectiveness of personalized offers and marketing campaigns based on the segmented groups.

7. **Focus on Niche Markets:**
   Explore niche customer segments, such as environmentally conscious customers, luxury spenders, or frequent travelers, and design targeted banking products for these groups.

By addressing the limitations and expanding the project scope, this work can evolve into a comprehensive and adaptable customer segmentation framework that significantly enhances banking insights and customer experiences.

LITERATURE REFERENCE

1) How to Use Machine Learning for Customer Segmentation.(Medium)
https://hrvoje-smolic.medium.com/how-to-use-machine-learning-for-customer-segmentation-49612667301d)

2) Advanced Customer Segmentation with AI in Banking (LinkedIn)
https://www.linkedin.com/pulse/advanced-customer-segmentation-ai-banking-mohammad-arif-gyvxc/

3) Customer Segmentation and Personalization in Banking
https://www.jetir.org/papers/JETIRGD06019.pdf

4) AI Machine Learning Algorithms for Customer Segmentation and Personalization in the Banking / FinTech Industries(LinkedIn)
https://www.linkedin.com/pulse/ai-machine-learning-algorithms-customer-segmentation-nasser-sami-yyupe/

5) Customer Segmentation and Personalization in Banking Services: Investigating the use of Big Data Analytics to Segment Banking Customers based on their Behavior, Demographics, and Preferences, and Leveraging these Insights to Personalize Banking Services and Marketing Campaigns -Sree Sandhya Kona
https://www.ijsr.net/archive/v9i8/SR24522131706.pdf

6) Experimental Analysis on Banking Customer Segmentation using Machine Learning Techniques
https://www.researchgate.net/publication/379927526_Experimental_Analysis_on_Banking_Customer_Segmentation_using_Machine_Learning_Techniques

7) How banks can leverage AI customer segmentation
https://thingsolver.com/blog/customer-segmentation-in-banking/