Aalto University
School of Electrical
Engineering

Master's Programme in Computer, Communication and Information Sciences

AI-Driven Relevance Finder for Activations

Vikramaditya Malik

Master's Thesis
2025

Aalto University
School of Electrical
Engineering

Author Vikramaditya Malik

Title AI-Driven Relevance Finder for Activations

Degree programme Computer, Communication and Information Sciences

Major Speech and Language Technology

Supervisor Assistant professor Lauri Juvela

Advisors Dr. Kai Kuikkaniemi, Dr. Ilkka Kosunen

Collaborative partner Cadentia Technologies Oy

Date 21st November 2025 **Number of pages** 52 **Language** English

Abstract

This thesis investigates the viability of synthetic customer profiles for predicting customer engagement with marketing activations in the electronics domain. The research addresses critical challenges in customer analytics, particularly data scarcity, privacy constraints, and the need for scalable personalization strategies. We developed a comprehensive methodology for generating synthetic customer profiles using large language models, engineering a multi-dimensional feature space capturing lexical, semantic, statistical, and sentiment-based characteristics through both handcrafted metrics and dense embedding representations, and evaluating multiple machine learning algorithms across diverse synthetic profile generation strategies.

Experimental results demonstrate that synthetic data can effectively support customer engagement prediction. Models trained exclusively on synthetic customer profiles achieved strong performance when validated on authentic Amazon customer data [1], demonstrating successful synthetic-to-real transfer learning. Detailed synthetic profiles consistently outperformed simpler generation approaches across all evaluation metrics. Linear classifiers emerged as the most effective algorithms for this task, demonstrating superior generalization capabilities compared to ensemble methods. Cross-dataset validation revealed strong generalization performance, with models maintaining consistency across different real customer data partitions.

The research provides empirical evidence that carefully engineered synthetic customer profiles can serve as viable alternatives to real customer data for training engagement prediction models, offering privacy-preserving, scalable, and cost-effective solutions for personalized marketing applications. These findings have significant implications for marketing analytics, suggesting that synthetic data approaches can democratize access to sophisticated customer modeling capabilities while respecting privacy constraints.

Keywords synthetic data, customer engagement, machine learning, personalization, retrieval-augmented generation

Contents

1 Introduction

Customer engagement represents a critical determinant of contemporary business success, encompassing the depth and quality of interactions between customers and a company's offerings. Strong engagement correlates with numerous favorable outcomes, including increased revenue, enhanced loyalty, and superior retention rates [2]. Research demonstrates that fully engaged customers can deliver revenue premiums exceeding 20% relative to average customers [3]. In contrast, disengaged customers exhibit higher propensity to switch providers, underscoring the importance of accurately predicting and cultivating engagement. Artificial intelligence (AI) and machine learning (ML) have emerged as powerful instruments for anticipating customer interactions with content, products, and marketing campaigns by identifying patterns within historical behavioral data.

1.1 Background

Customer engagement has evolved alongside broader transformations in marketing philosophy. Prior to the 1990s, marketing operated primarily on transactional principles, emphasizing individual sales events. This paradigm gradually shifted toward relationship marketing, where cultivating enduring, positive customer relationships became paramount for ensuring satisfaction and loyalty [4]. As markets have intensified in competitiveness, mere satisfaction no longer suffices. Contemporary organizations seek comprehensive engagement across all touchpoints, establishing relationships founded on trust, commitment, and emotional resonance.

Personalization serves as a fundamental driver of effective engagement. Organizations increasingly tailor marketing strategies to individual customers using previously collected data [5]. Recommender systems, which suggest products or content aligned with user preferences, have become central to personalization strategies. Two primary approaches exist: collaborative filtering, which exploits similarities among users, and content-based filtering, which focuses on item attributes [6]. Modern systems typically integrate both methodologies to optimize effectiveness.

Despite these advances, traditional approaches face substantial constraints. Constructing accurate predictive models for customer engagement typically demands extensive real customer data, often proving difficult to acquire. Data availability and quality present persistent challenges, with many organizations lacking sufficient labeled interaction data. When data exists, it frequently contains noise or exhibits bias toward specific segments, yielding models that fail to generalize across broader customer populations. Furthermore, real-world data collection requires substantial time and resources while potentially failing to encompass scenarios of interest, such as novel product launches or underrepresented customer segments.

Privacy considerations further complicate real customer data utilization. Stringent regulations and ethical imperatives limit detailed personal information collection and usage. Organizations must protect sensitive data and often cannot freely employ it without explicit consent. These constraints challenge the development of robust predictive models that respect user privacy while delivering actionable insights.

1.2 The Rise of Synthetic Data

Recent years have witnessed substantial growth in synthetic data adoption as a solution to these challenges. Synthetic data comprises artificially generated information that replicates the statistical properties of real data without incorporating actual customer records. According to industry analysis, up to 20% of AI training data currently consists of synthetic data, with projections suggesting this could reach 80% by 2028 [7]. By 2030, synthetic data may underpin more business decision-making than real data [8]. Synthetic data offers several advantages: circumventing privacy constraints, enabling on-demand generation at scale, and supporting customization for specific scenarios [9].

Within customer engagement contexts, synthetic data enables creation of virtual customer profiles termed 'synthetic personas' that simulate authentic customer behaviors and preferences without exposing individual personal information. These synthetic profiles derive from patterns observed in real data, serving as proxies for actual users in modeling and analysis. Marketers and researchers increasingly employ synthetic personas to predict customer reactions to new products or marketing strategies [10, 11]. Marketing teams can generate synthetic consumers based on aggregated purchase data and simulate their responses to campaigns [12]. These AI-generated personas closely approximate real customer behavior, providing meaningful insights with greater speed and cost-efficiency than recruiting actual participants [13]. Unlike traditional manually-crafted marketing personas, synthetic profiles can interact with systems or analysts, offering dynamic feedback in experimental settings.

1.3 Research Problem and Objectives

Contemporary marketing strategies rely heavily on historical customer data, constrained by availability limitations, privacy concerns, and representational inadequacies. These limitations impede precise customer behavior prediction, reducing marketing activation effectiveness. Novel approaches capable of generating reliable customer profiles and engagement scenarios without privacy violations are needed, ensuring accurate, dynamic prediction of customer preferences and behaviors.

Motivated by these challenges, this research investigates an AI-driven approach to customer engagement prediction utilizing synthetic customer profiles. The primary objective involves developing a system capable of accurately predicting customer engagement likelihood with specific offers or marketing activations. To achieve this, the research pursues following objectives:

1. Generating synthetic customer profiles that realistically simulate authentic customer behaviors and preferences without field data collection.
2. Structuring customer profiles in formats optimized for the application scenario.
3. Evaluating various ML models to identify those most effective for leveraging synthetic data in customer engagement prediction.
4. Comparing predictive performance of models trained on synthetic data versus those trained on real-world data.

1.4 Scope and Methodology

This research encompasses customer data synthesis, logic development for comparing synthetic and real data, exploration and evaluation of ML algorithms for engagement prediction, and practical application to real-world scenarios.

This work does not address broader AI applications beyond customer engagement, nor does it extensively examine non-technical aspects such as regulatory frameworks or ethical implications beyond privacy considerations.

The proposed methodology follows a systematic and integrative approach. It commences with generative AI to create synthetic datasets reflecting diverse customer behaviors. Supervised learning algorithms then train on this synthetic data to develop predictive models. Model effectiveness undergoes evaluation on real customer profiles using key performance indicators. A critical component involves incorporating user feedback through structured mechanisms to iteratively refine and enhance model adaptability and predictive accuracy.

1.5 Contributions

This work makes three contributions. First, we design a novel pipeline to construct synthetic customer profiles by leveraging product similarity information from a large online retail dataset. These profiles emulate meaningful customer interest groupings without requiring individual data. Second, we develop an engagement prediction model utilizing these synthetic profiles to learn which features render products relevant or irrelevant to given profiles. We incorporate multiple feature representations (textual, visual, and categorical) to capture multifaceted product similarity. Third, we analyze model performance and discuss synthetic data viability in predictive customer analytics, including advantages, limitations, and ethical implications. To our knowledge, this represents one of the first systematic explorations of synthetic customer personas for engagement prediction in e-commerce.

1.6 Thesis Organization

The remainder of this thesis proceeds as follows. Section 2 reviews background and related work on customer engagement prediction, synthetic data generation, and retrieval-augmented approaches in NLP. Section 3 outlines the conceptual design and system architecture guiding our approach to synthetic profile generation. Section 4 details the technical methodology, including data sources, synthetic profile generation strategies, feature engineering pipelines, and model training procedures. Section 5 presents comprehensive experimental results, model performance evaluation, and comparative analysis across different profile types and machine learning algorithms. Section 6 discusses key findings, interprets results in context of related work, analyzes uncertainty and limitations of the synthetic-to-real transfer approach, and explores ethical considerations surrounding synthetic customer data. Finally, Section 7 concludes with a summary of contributions and outlines promising directions for future research in synthetic customer analytics.

2 Literature Review

This review examines four foundational papers that underpin our approach to synthetic customer profile generation and engagement prediction: GPT-3's few-shot learning capabilities, dense passage retrieval methods, retrieval-augmented generation, and zero-shot dense retrieval. These works form the technical foundation for using large language models to generate synthetic customer data and predict engagement patterns without requiring extensive real customer information.

2.1 Few-Shot Learning in Large Language Models

Brown et al. (2020) demonstrated that scaling language models to 175 billion parameters unlocks few-shot learning the ability to perform new tasks from just a few examples without fine-tuning [14]. GPT-3 adapts to tasks purely through in-context learning: provide it with examples of translating English to French, and it continues translating; show it arithmetic problems with solutions, and it solves new ones. This capability eliminates the need for task-specific training data.

The authors tested GPT-3 in three settings: zero-shot (task description only), one-shot (one example), and few-shot (10-100 examples). Performance consistently improved with both model scale and number of examples across diverse tasks including reading comprehension, translation, question-answering, arithmetic, and creative writing. Notably, the model succeeded on tasks requiring rapid domain adaptation traditionally requiring specialized architectures or extensive training.

GPT-3's architecture builds on the Transformer decoder with attention mechanisms for autoregressive generation. Training used approximately 400 billion tokens from Common Crawl, WebText2, Books, and Wikipedia. The model's ability to generate coherent, contextually appropriate text from minimal examples forms the basis for synthetic customer profile generation in our work.

Relevance to this thesis: GPT-3's few-shot learning enables our synthetic profile generation strategy. By providing GPT-4o mini with a few example customer profiles and product descriptions, we generate thousands of diverse, realistic customer profiles without training data a direct application of in-context learning to customer analytics.

Limitations: GPT-3 requires massive computational resources and cannot update its knowledge without retraining. The model's static knowledge and tendency toward hallucination in factual tasks motivated subsequent work on retrieval-augmented approaches.

2.2 Dense Neural Representations for Retrieval

Karpukhin et al. (2020) introduced Dense Passage Retrieval (DPR), demonstrating that learned dense vector representations outperform traditional keyword-based methods for information retrieval [15]. DPR uses dual BERT encoders [16] to separately embed questions and passages, enabling efficient similarity computation through vector operations. The approach achieved 9-19% absolute improvements in top-20 passage retrieval accuracy over BM25.

DPR's training uses contrastive learning: the model learns to assign high similarity scores to relevant question-passage pairs while minimizing scores for irrelevant pairs. Crucially, training includes both random negative examples and "hard negatives" topically related but irrelevant passages. This teaches the model to discriminate between truly relevant content and merely related material.

The system achieves strong performance with relatively small training sets (thousands vs. millions of examples), making dense retrieval accessible even with limited data. DPR's pre-computed passage embeddings enable efficient search over large document collections. Results on Natural Questions, TriviaQA, and WebQuestions benchmarks established new state-of-the-art performance for open-domain question answering.

Relevance to this thesis: Our feature engineering uses sentence-transformers (based on similar dense embedding principles) to create 384-dimensional semantic representations of customer profiles and product descriptions. These embeddings capture semantic similarity beyond surface-level word matching essential for comparing synthetic profiles to products.

Limitations: DPR requires labeled training data and performs best when fine-tuned for specific domains. The method excels at semantic matching but may underperform keyword-based approaches for precise factual lookups.

2.3 Retrieval-Augmented Generation

Lewis et al. (2020) introduced RAG, which combines parametric knowledge (stored in model parameters) with non-parametric knowledge (retrieved from external sources) [17]. RAG addresses fundamental limitations of pure language models: static knowledge, factual inaccuracy, and inability to access information beyond training cutoffs.

RAG's architecture jointly trains a dense retriever (similar to DPR) with a sequence-to-sequence generator. Unlike pipeline approaches that optimize components independently, RAG fine-tunes both together, ensuring the retriever identifies passages specifically useful for generation. The framework supports two variants: RAG-Sequence uses the same retrieved passages for the entire output, while RAG-Token allows different passages to inform different output tokens.

Experiments on open-domain QA, fact verification, and knowledge-grounded dialogue showed RAG improves factual accuracy while maintaining fluency. The system achieves competitive performance with smaller models than purely parametric approaches would require. Importantly, RAG's knowledge base can be updated without model retraining knowledge stays current by updating the retrieval corpus rather than retraining parameters.

Relevance to this thesis: While we don't implement full RAG architecture, our approach shares conceptual DNA. We use GPT-4o mini's parametric knowledge (understanding of customer behaviors and demographics) combined with non-parametric knowledge (specific product descriptions from the Amazon Electronics dataset [1]) to generate context-appropriate synthetic profiles. This hybrid approach grounds synthetic generation in real product data.

Limitations: RAG's performance depends on retrieval corpus quality biases or gaps propagate to outputs. Integrating retrieved knowledge requires sophisticated attention mechanisms to ensure relevant information is used effectively rather than ignored.

2.4 Zero-Shot Dense Retrieval

Gao et al. (2022) introduced HyDE (Hypothetical Document Embeddings), which achieves effective retrieval without training data or relevance labels [18]. HyDE solves a clever problem: instead of directly matching queries to documents, it generates a hypothetical document answering the query, then retrieves real documents similar to the hypothetical one.

The two-stage approach first uses an instruction-following LLM (InstructGPT) to generate a hypothetical document given a query. This document may contain factual errors or hallucinations, but it captures relevant patterns and concepts. Second, an unsupervised encoder (Contriever) embeds this hypothetical document for similarity search against real documents. The encoding acts as a filter: factual errors are discarded while semantic patterns persist.

HyDE substantially outperforms prior zero-shot methods across 11 benchmarks, matching fine-tuned retrievers without any training data. Strong performance in diverse languages (Swahili, Korean, Japanese) demonstrates robust multilingual capability. The approach works because it transforms asymmetric query-document matching into symmetric document-document matching, which contrastive encoders handle more effectively.

Relevance to this thesis: HyDE's approach conceptually parallels our synthetic profile generation. Just as HyDE generates hypothetical documents to represent queries, we generate hypothetical customer profiles to represent potential market segments. Both approaches leverage LLMs' ability to create realistic-but-synthetic representations that capture relevant patterns despite potential factual inaccuracies. Our "detailed" profile strategy generating marketing offers before profiles particularly resembles HyDE's intermediate representation approach.

Limitations: Quality depends on the generative model's capabilities. Systematic biases in the LLM propagate to retrieval performance. The approach adds computational overhead compared to direct encoding.

2.5 Synthesis: From Language Models to Synthetic Customer Analytics

These four papers trace an evolution in NLP that enables our work. GPT-3 demonstrated that large language models possess broad, few-shot adaptable knowledge of human language and behavior. DPR showed that dense neural embeddings capture semantic relationships invisible to keyword matching. RAG proved that combining parametric and non-parametric knowledge improves performance on knowledge-intensive tasks. HyDE revealed that hypothetical documents generated by LLMs can effectively guide

retrieval despite potential inaccuracies.

The synthetic customer profile approach synthesizes these insights. We leverage GPT-4o mini's few-shot learning (GPT-3 lineage) to generate customer profiles from minimal examples. We use dense embeddings (DPR approach) to capture semantic similarity between profiles and products. We combine the LLM's parametric knowledge of customer behaviors with non-parametric product data (RAG concept). And we embrace HyDE's insight that synthetic representations, despite imperfections, can effectively guide downstream tasks.

Critical gaps addressed by this thesis: While these papers focus on NLP tasks (QA, retrieval, generation), none addresses customer analytics or synthetic data for predictive modeling. Our contribution demonstrates that these NLP techniques transfer successfully to a new domain: generating synthetic customer profiles that train models achieving 90% ROC-AUC on real-world engagement prediction. This synthetic-to-real transfer training exclusively on generated data, testing on authentic customers represents a novel application of few-shot learning and dense retrieval principles to privacy-preserving customer analytics.

The literature establishes technical feasibility of our components but leaves open whether synthetic profiles can capture behavioral patterns sufficiently to predict real customer engagement. Our experimental results address this gap directly.

3 System Design

This section describes the high-level architecture, design principles, and conceptual framework of the AI-driven customer engagement prediction system.

3.1 Architecture Overview

The system architecture comprises five interconnected stages that transform raw product descriptions into trained customer engagement prediction models. Figure 1 illustrates the complete data flow and component interactions.

The architecture follows a modular design pattern where each stage operates with well-defined input and output interfaces. This modularity enables parallel processing of profiles, incremental updates to components, and independent testing of each pipeline stage. The separation of concerns allows researchers to refine individual components such as profile generation strategies or feature engineering approaches without requiring modifications to dependent stages.

3.2 Design Principles

Several core principles guided the system architecture:

- **Privacy-by-Design:** The system generates synthetic customer profiles rather than using real customer data during model training, enabling privacy-preserving analytics. Models trained on synthetic data require no access to actual customer behavior, addressing data privacy regulations (GDPR, CCPA) and organizational data governance policies. This approach democratizes customer analytics capabilities while eliminating privacy risks inherent in real customer data usage.

- **Modularity and Extensibility:** Each pipeline stage (data acquisition, profile generation, feature engineering, model training, evaluation) functions independently with standardized interfaces. This enables easy replacement or enhancement of individual components. For instance, researchers can experiment with alternative profile generation strategies or feature engineering techniques without restructuring the entire pipeline.

- **Reproducibility and Transparency:** Fixed random seeds, versioned datasets, and documented configurations ensure experimental reproducibility. All design decisions include explicit rationale, enabling other researchers to understand and potentially replicate the approach.

- **Scalability:** The system accommodates growth from thousands to potentially millions of synthetic profiles through efficient data processing, distributed feature computation, and parallelized model training. Hardware acceleration (GPU support) enables rapid experimentation without requiring specialized infrastructure.

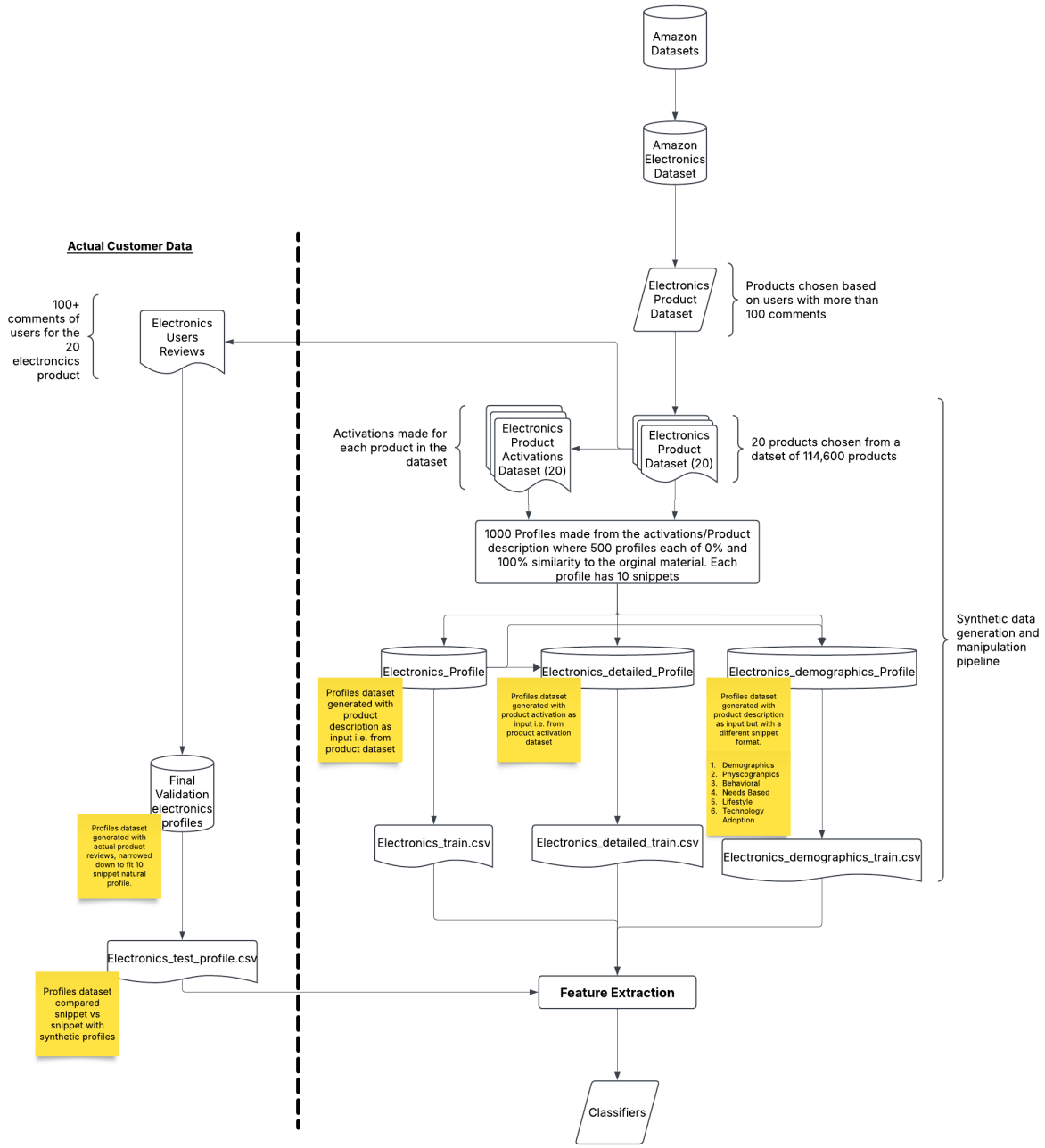

Figure 1: System Architecture: Five-stage pipeline showing data flow from Amazon Electronics dataset [1] through synthetic profile generation, feature engineering, model training, and validation for customer engagement prediction.

3.3 Stage 1: Data Acquisition and Foundation

The system processes the Amazon Electronics dataset[1] to establish a foundation for both synthetic profile generation and real-world validation. This stage establishes the source material and validation benchmark.

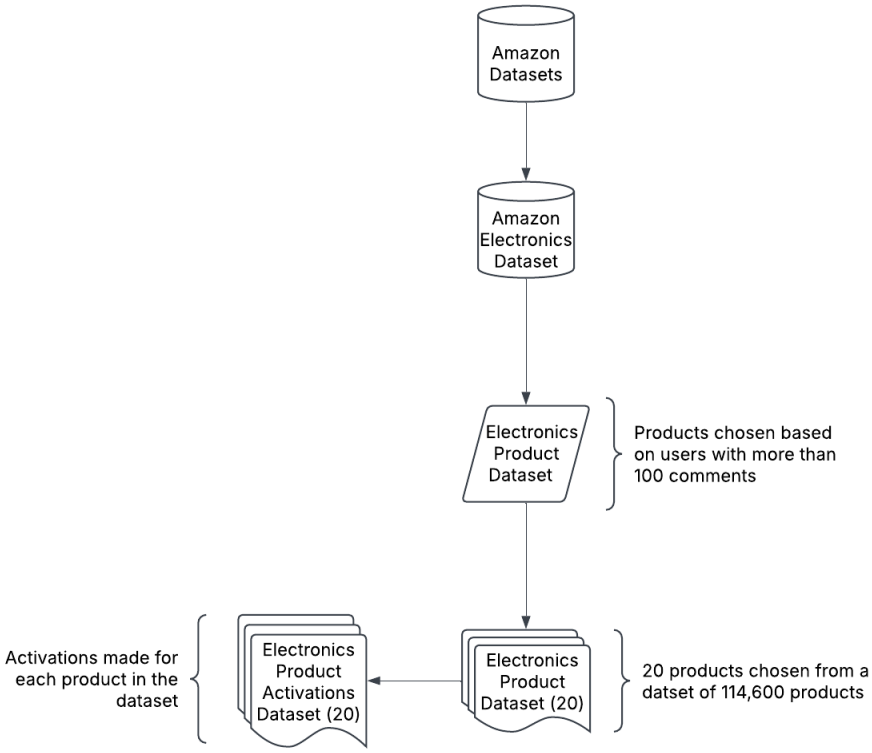

Figure 2: Stage 1: Data Collection and Product Selection from Amazon Electronics Dataset

- **Data Source:** The Amazon Electronics dataset provides a rich, publicly available corpus of product metadata and customer reviews. This domain enables evaluation on authentic customer engagement patterns within a defined market segment.
- **Product Selection Strategy:** Products are filtered based on review volume and content quality criteria to ensure sufficient information density. Only products with adequate review coverage are included, ensuring that derived customer profiles reflect meaningful patterns rather than noise from minimal feedback.
- **Preprocessing Approach:** Raw product data undergoes structured preprocessing to normalize formatting, extract structured attributes, and prepare content for subsequent synthetic generation stages. This preprocessing establishes consistent data formats across the pipeline.

3.4 Stage 2: Synthetic Profile Generation Strategy

The system employs a three-pronged strategy for synthetic profile generation, each balancing different design trade-offs between generation simplicity, contextual richness, and structural completeness.

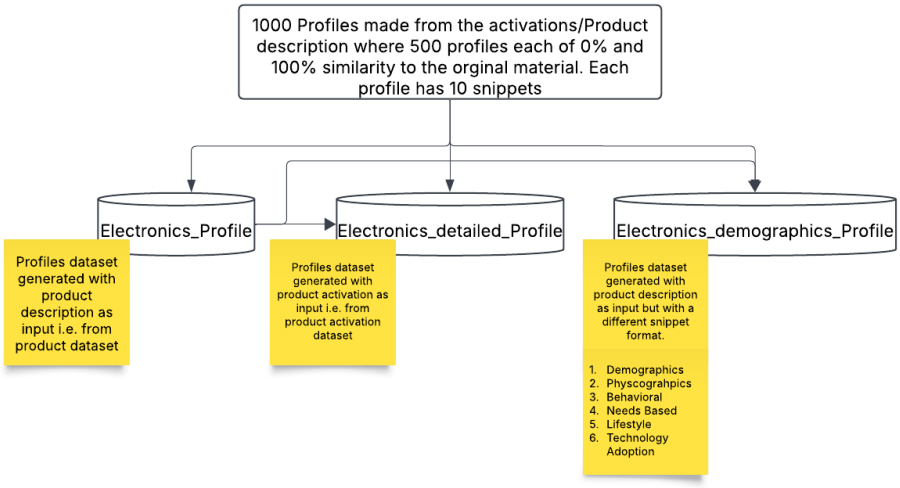

Figure 3: Stage 2: Synthetic Profile Generation using Three Parallel Strategies (Basic, Detailed, and Demographics)

- **Strategy 1: Basic Profiles—Simplicity and Speed.** This approach generates customer profiles directly from product descriptions using single-stage prompting. The generation prioritizes speed and simplicity, producing profiles that capture broad behavioral themes without enforcing specific demographic structure. This strategy establishes a baseline for profile quality and engagement prediction capability.

- **Strategy 2: Detailed Offer-Based Profiles—Contextual Richness.** This two-stage approach first transforms product descriptions into marketing narratives before generating profiles. By intermediate representation through customer-facing offers, the system grounds profiles in realistic marketing contexts. This richer contextual foundation enables profiles that reflect product value propositions more authentically.

- **Strategy 3: Demographics-Structured Profiles—Categorical Completeness.** This strategy enforces explicit coverage of demographic and psychographic dimensions through structured prompting. Rather than free-form generation, profiles are required to populate specific demographic categories (age, income, location, technology adoption). This systematic coverage addresses potential gaps in unstructured generation approaches and enables demographic-based validation and analysis.

- **Design Rationale:** The three-strategy approach enables systematic evaluation of profile richness versus engagement prediction accuracy. By comparing results across strategies, the research isolates the impact of profile dimensionality and contextual grounding on model performance. This multi-strategy design also reduces overfitting risk to any single profile generation approach.

3.5 Stage 3: Feature Engineering Framework

The system constructs high-dimensional feature representations capturing lexical, semantic, statistical, and sentiment-based characteristics of customer profiles. Rather than relying on a single feature type, the system combines handcrafted statistical features with learned semantic embeddings.

- **Multi-Dimensional Representation:** Features span five categories—lexical overlap (shallow textual similarity), embedding-based (deep semantic content), distance metrics (profile similarity quantification), text statistics (structural properties), and sentiment (emotional tone). This diversity enables the model to capture multiple facets of customer-profile alignment simultaneously.
- **Semantic Grounding:** The system leverages pre-trained sentence embeddings to capture semantic meaning beyond surface-level word matching. Embeddings encode complex relationships for instance, "budget-conscious" and "price-sensitive" embeddings would be proximate despite different vocabularies enabling nuanced similarity assessment.
- **Aggregation Strategy:** For each profile pair, the system computes features representing both profiles' individual characteristics and their relational properties (differences, interactions). This paired-feature representation enables models to assess compatibility directly rather than inferring compatibility from independent profile representations.

3.6 Stage 4: Supervised Model Training

The system trains multiple machine learning algorithms spanning different model families linear classifiers, ensemble methods, and gradient boosting approaches to systematically evaluate algorithm effectiveness on synthetic-to-real transfer.

- **Algorithm Selection Rationale:** Logistic Regression provides a linear baseline, establishing how much model complexity is necessary. Random Forest and LightGBM represent modern ensemble approaches, enabling evaluation of whether complex non-linear models provide advantages over simpler classifiers. This comparative evaluation directly addresses whether the synthetic data and feature engineering create non-linear decision boundaries or whether linear separability suffices.
- **Hyperparameter Optimization:** Rather than using default configurations, each algorithm undergoes systematic hyperparameter search to identify configurations optimally suited to the synthetic data characteristics and 1,564-dimensional feature space. This optimization ensures fair comparison among algorithms by using each algorithm's best-case performance rather than arbitrary defaults.
- **Cross-Validation Strategy:** Multiple training approaches (stratified train-test splits, cross-dataset evaluation) test both model capacity and generalization

capability. This comprehensive evaluation reveals whether models memorize synthetic data characteristics or extract generalizable engagement patterns.

3.7 Stage 5: Evaluation and Validation

The system implements multi-tiered evaluation assessing both synthetic performance and real-world generalization capability.

- **Synthetic Validation:** Models are evaluated on diverse synthetic test sets, each generated from different profile strategies. This reveals how well models trained on one profile type generalize to other profile generation approaches.
- **Real-World Validation:** Models trained exclusively on synthetic data are subsequently evaluated on real customer profiles generated from their comments which are extracted from Amazon Electronics dataset[1]. This critical evaluation step assesses whether synthetic-to-real transfer occurs whether patterns learned from artificial data apply to authentic customer behavior.
- **Metric Selection:** Multiple evaluation metrics (ROC-AUC, accuracy, F1-score, balanced accuracy) provide complementary perspectives on model performance. ROC-AUC serves as the primary metric, reflecting the model's discrimination capability across classification thresholds.

4 Methods

This section provides detailed technical implementation of the system architecture described in Section 3. All specific configurations, algorithms, hyperparameters, and experimental protocols are documented to enable reproducibility and detailed understanding of the methodological choices. Code development was assisted by ChatGPT [19], and theory was supported by Perplexity AI [20].

4.1 Experimental Setup and Environment

- **Hardware Configuration:** Experiments were conducted on systems with NVIDIA GPUs (when available) and multi-core CPUs. All deep learning operations utilize PyTorch with CUDA acceleration for GPU-capable systems, falling back to CPU processing when necessary.
- **Software Environment:** Python 3.8+ with key dependencies: scikit-learn 1.0+ for machine learning algorithms, PyTorch 1.10+ for tensor operations, NLTK [21]3.6+ for natural language processing, sentence-transformers for semantic embeddings, and LightGBM for gradient boosting models.
- **Reproducibility Settings:** All random operations employ fixed seed (42) to ensure deterministic outcomes across experimental runs. This includes train-test splitting, model initialization, and hyperparameter sampling.

4.2 Stage 1: Data Collection and Preprocessing

4.2.1 Amazon Electronics Dataset

The system utilizes the Amazon Electronics dataset[1] containing over 100,000 products with associated customer reviews. Products are selected based on review volume: minimum 5 reviews per product ensures sufficient customer engagement signal.

4.2.2 Product Metadata Extraction

Product identifiers (ASINs—Amazon Standard Identification Numbers) are extracted using regex pattern matching:

```
# ASIN extraction from product metadata
asin_pattern = re.compile(r'(?:asin|parent_asin)\s*[:=]\s*([A-Z0-9]{10})')
product_asins = asin_pattern.findall(product_metadata)
```

Both primary ASINs and parent ASINs are extracted to maintain consistency across product variations and variant families.

4.2.3 Text Preprocessing Pipeline

Raw product descriptions and reviews undergo multi-stage preprocessing:

- **Normalization:** Convert text to lowercase, remove HTML artifacts, special characters, and formatting inconsistencies while preserving semantic content.
- **Tokenization and Cleaning:** Use NLTK word tokenizer [21] to split text into individual tokens, removing pure punctuation and numeric-only tokens.
- **Stopword Removal:** Remove common English stopwords (the, is, at, which, etc.) using NLTK's English stopword list [21].
- **Lemmatization with POS Tagging:** Apply part-of-speech tagging to identify word types, then lemmatize each word to its base form:

```
# Lemmatization with POS-aware base form selection
import nltk
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
tokens = nltk.word_tokenize(text.lower())
pos_tags = nltk.pos_tag(tokens)

lemmatized = []
for word, pos in pos_tags:
    wordnet_pos = get_wordnet_pos(pos)
```

```
if wordnet_pos:
    lemmatized.append(lemmatizer.lemmatize(word, wordnet_pos))
else:
    lemmatized.append(lemmatizer.lemmatize(word))
```

This POS-aware lemmatization ensures words are reduced to semantically correct base forms (verb forms to infinitives, nouns to singular, etc.).

4.3 Stage 2: Synthetic Profile Generation Implementation

4.3.1 Azure OpenAI API Configuration

Profile generation utilizes GPT-4o mini through Azure OpenAI Service with strategy-specific parameter optimization:

Table 1: Azure OpenAI API Configuration Parameters

Parameter	Value	Strategy
API Type	Azure	All
API Version	2023-05-15	All
Deployment Name	gpt-4o-mini	All
Basic/Electroincs Profiles		
Max Tokens	1500	Basic/Electroincs
Temperature	0.7	Basic/Electroincs
Offer/Detailed Expansion		
Max Tokens	1500	Offer/Detailed
Temperature	0.7	Offer/Detailed
Demographics Profiles		
Max Tokens	2000	Demographics
Temperature	0.7	Demographics
Rate Limit	450K tokens/minute	All
Request Limit	2,700 requests/minute	All

The temperature parameter varies by strategy: 0.7 for profile generation (balancing creativity with consistency) and 1.0 for offer expansion (encouraging diverse marketing language). Maximum token limits are optimized per task: 300 tokens for concise offer summaries, 1,500 for typical profile structures, and 2,000 for comprehensive demographic descriptions.

4.3.2 Strategy 1: Basic Profile Generation

This strategy generates customer profiles directly from product descriptions using single-stage prompting:

```
"Generate a customer profile for someone interested in [product].
Include interests, shopping habits, and likely characteristics.
Output as a JSON array of 10 text snippets describing this customer."
```

- **Output Format:** JSON array of text snippets (typically 10 snippets per profile) describing customer characteristics without enforced demographic structure.
- **Generation Volume:** 10,000 Basic profiles generated.

4.3.3 Strategy 2: Detailed Offer-Based Profiles

Here we expand on product description in the form of a detailed offer in these 2 stages:

Stage 1—Offer Expansion:

```
"Expand this product description into a detailed marketing offer
(100-200 words). Include benefits, use cases, and appeal to customers."
```

Stage 2—Profile Generation from Offer:

```
"Based on this offer, generate a customer profile. Include interests
and characteristics of someone attracted to this value proposition.
Output as JSON snippets."
```

Generation Volume: 5,000 Detailed profiles generated through two-stage process.

4.3.4 Strategy 3: Demographics-Structured Profiles

In the final strategy we expand upon the product description to generate profiles of :

```
"Generate a structured customer profile with these explicit dimensions:
- Demographics: Age, location, income, education
- Behavioral: Shopping frequency, technology adoption
- Psychographics: Interests, motivations, values, lifestyle
Output as JSON with each dimension populated."
```

- **Output Format:** Strictly structured JSON with required demographic and psychographic fields populated.
- **Generation Volume:** 5,000 Demographics profiles generated with enforced categorical coverage.

4.3.5 Quality Control and Error Handling

All generated profiles undergo validation through a systematic review process:

- Schema validation: Automated checks confirm JSON structure compliance and proper formatting.

- Completeness checks: Automated verification ensures all required fields contain meaningful, non-empty values
- Consistency validation: Through manual review examining logical coherence across profile dimensions (e.g., purchase history aligns with stated preferences, technical expertise level is internally consistent)

Profiles are generated in batches (25 per batch), with each batch undergoing automated validation (schema compliance, field completeness, consistency checks) followed by manual review before the next batch begins. Most profiles (>95%) pass validation on the first attempt. During initial testing, profiles generated with temperature > 1.2 exhibited higher error rates; subsequently, temperature was maintained ≤ 1.2 throughout all generations. Profiles failing validation are regenerated with adjusted prompts that add explicit constraints based on identified failures.

4.4 Stage 3: Feature Engineering Implementation

4.4.1 Semantic Embedding Generation

Profiles are converted to dense semantic representations using Sentence-BERT [22]:

```
from sentence_transformers import SentenceTransformer
import torch

device = "cuda" if torch.cuda.is_available() else "cpu"
model = SentenceTransformer('all-MiniLM-L6-v2', device=device)

def get_profile_embedding(profile_text):
    """Generate 384-dimensional embedding for profile text"""
    embedding = model.encode(profile_text, convert_to_tensor=True)
    return embedding.cpu().numpy()  # (384,) dimensional vector
```

The pairing process involves selecting two snippets from customer profiles and comparing them against each other. These snippets can originate from the same profile or different profiles, the first snippet is designated as Profile A and the second as Profile B. This design allows the model to learn similarity patterns both within profiles (same customer, different activation contexts) and across profiles (different customers, same or different products), enabling comprehensive understanding of engagement relevance.

The all-MiniLM-L6-v2 model [22] produces 384-dimensional embeddings encoding semantic meaning efficiently. For each snippet in a pair, embeddings are computed from both raw and preprocessed text, then the representations are utilized to create the profile-level feature vectors. Specifically, Profile A Embedding captures the semantic content of the first snippet, while Profile B Embedding captures the semantic content of the second snippet. These embeddings are then combined through element-wise operations: difference vectors (Difference Embedding) capture semantic

Table 2: Feature Extraction and Representation Components

Feature Type	Components	Count
Computed Statistical Features		
Lexical Similarity	Jaccard similarity [23], Cosine similarity, token overlap ratio, common token count, unique token counts	5
Distance Metrics	Euclidean distance, Manhattan distance	2
Text Properties	Word counts, character counts, average word lengths, punctuation counts for both profiles and differences	12
Embedding Statistics	Snippet embedding means and standard deviations, absolute difference means	6
Sentiment Features	VADER [24] compound scores for each profile and difference	3
	Subtotal (Statistical):	**28**
Embedding-Based Features		
Profile A Embedding	384-dimensional semantic vector (sentence-BERT [22])	384
Profile B Embedding	384-dimensional semantic vector (sentence-BERT [22])	384
Difference Embedding	Element-wise differences capturing semantic divergence	384
Product Embedding	Element-wise products capturing semantic interactions	384
	Subtotal (Embedding):	**1,536**
	Total Features per Profile Pair:	**1,564**

divergence between the pair, while product vectors (Product Embedding) capture semantic interactions and alignment between snippets.

This comprehensive feature representation captures surface-level lexical characteristics through handcrafted metrics, deep semantic content through dense embeddings, and relational patterns through difference and product vectors. The multi-modal approach enables models to learn engagement patterns from complementary information sources, reducing reliance on any single feature type that might exhibit distribution shift between synthetic training and real testing data.

4.4.2 Sentiment Analysis

Sentiment features quantify emotional tone present in profile text:

Sentiment Dimensions Captured:

- **Compound Score:** Overall polarity (-1 to +1) synthesizing all sentiment dimensions
- **Positive Sentiment:** Proportion of positive emotional content
- **Negative Sentiment:** Proportion of negative emotional content
- **Neutral Sentiment:** Proportion of objective, non-emotional content

For each sentiment dimension, the system computes four statistical aggregations across all snippets: mean, standard deviation, minimum, and maximum. This yields $4 \times 6 = 24$ total sentiment features per profile pair, capturing both typical emotional tone and extremity or polarization of sentiment.

4.5 Stage 4: Balanced Dataset Construction

4.5.1 Label Assignment Logic

Positive and negative labels are assigned through binary classification based on exact product matching:

- **Positive Labels (1):** Assigned when the synthetic profile's product number exactly matches the user's reviewed product AND the synthetic profile type is marked as 100% match. This ensures positive examples represent authentic product-customer alignment without ambiguity.
- **Negative Labels (0):** Assigned for all other cases (different products, partial matches, or non-100% profile types). This conservative strategy focuses training on unambiguous engagement scenarios.

  So according to the logic out of the four possible cases only 1 will have positive label. Product numbers are extracted from filenames using regex pattern matching, enabling fully automated label assignment across all profile pairs.

4.5.2 Dataset Construction and Sampling Control

The dataset is constructed with precisely controlled sample sizes to ensure balanced class representation: 4,000 positive pairs and 4,000 negative pairs for training, and 1,000 positive and 1,000 negative pairs for testing. This balanced structure prevents class imbalance bias during model training and enables fair performance evaluation across both engagement and non-engagement scenarios.

Profile-level stratified train-test splitting (80% train, 20% test) is employed to ensure both datasets maintain representative distributions across demographic dimensions (age groups, income brackets, geographic locations, technology adoption levels). This stratification prevents spurious demographic associations and ensures that models learn

generalizable patterns rather than artifacts specific to the training set demographic composition.

To manage computational complexity and prevent overfitting, sampling constraints are enforced throughout pair generation. Each real customer profile is matched against at most 15 synthetic profiles, and each profile pair generates at most 15 snippet-to-snippet comparisons. Additionally, validation profiles incorporate at most 30 of the most recent customer reviews to ensure temporal relevance and prevent recency bias from older, potentially outdated purchase behavior. These constraints collectively ensure efficient processing, computationally tractable datasets, and training on high-quality, recent customer engagement signals.

4.6 Stage 5: Model Training Implementation

4.6.1 Logistic Regression with Regularization Tuning

Logistic regression is trained with systematic hyperparameter optimization to balance model complexity and generalization. The model employs the SAGA solver, which efficiently handles both L1 (Lasso) and L2 (Ridge) regularization penalties at scale, making it suitable for the 1,564-dimensional feature space. The maximum iteration count is set to 2,000 to ensure convergence on high-dimensional data, substantially higher than the default value of 100.

Hyperparameter optimization is performed via grid search over the regularization strength parameter C (values: 0.001, 0.01, 0.1) and regularization type (L1 and L2). This grid tests all nine combinations of parameters to identify the configuration that maximizes ROC-AUC performance. The grid search is conducted using stratified 5-fold cross-validation to ensure each fold maintains balanced class distributions, preventing misleading performance estimates from imbalanced folds.

Feature scaling is applied using standardization (zero mean, unit variance) prior to model training. The scaler is configured without mean-centering (with_mean=False) to preserve computational efficiency on sparse feature matrices. This preprocessing step normalizes features to equivalent scales, preventing high-magnitude features from dominating the learning process and enabling fair coefficient interpretation.

All hyperparameter configurations are evaluated using ROC-AUC as the optimization criterion, which is robust to class imbalance and reflects ranking quality rather than accuracy alone. Parallel processing (n_jobs=-1) utilizes all available CPU cores to reduce grid search computational time.

4.6.2 Random Forest with Depth Tuning

Random Forest is trained as an ensemble method to capture non-linear relationships and feature interactions that may not be discoverable by linear models. The ensemble consists of 100 decision trees, a fixed size balancing predictive performance against computational cost and training time. Individual trees are constructed in parallel using all available CPU cores (n_jobs = -1) to accelerate model training.

The primary hyperparameter of interest is maximum tree depth, which directly controls model complexity and overfitting risk. A grid search evaluates two depth configurations (max_depth = 5 and max_depth = 10) to identify the optimal complexity-bias tradeoff. Shallower trees (depth 5) provide stronger regularization and reduce overfitting on the synthetic training data, while deeper trees (depth 10) allow the model to capture more complex patterns at the risk of memorizing training noise. Grid search optimization is performed using stratified 5-fold cross-validation with ROC-AUC as the evaluation metric, consistent with the logistic regression configuration.

Feature preprocessing for Random Forest employs MaxAbsScaler, which scales features to the range [-1, 1] while preserving sparsity structure. Unlike StandardScaler used for logistic regression, MaxAbsScaler is optimal for tree-based models because decision trees are scale-invariant (split thresholds adapt to feature ranges), and MaxAbsScaler avoids introducing artificial centering that adds computational overhead without benefit. This preprocessing step maintains consistency across all models while respecting the mathematical properties of tree-based algorithms.

Parallel processing is employed throughout grid search to maximize CPU utilization and reduce total computational time for hyperparameter optimization.

4.6.3 LightGBM with Comprehensive Hyperparameter Tuning

LightGBM (Light Gradient Boosting Machine) is trained as a gradient boosting ensemble to leverage sequential tree refinement for capturing complex non-linear patterns. Unlike Random Forest's parallel ensemble, LightGBM constructs trees iteratively, with each tree correcting residual errors from previous trees, enabling more efficient learning on high-dimensional data.

Comprehensive hyperparameter optimization is performed via grid search over three key parameters. The number of leaves (num_leaves: 15 and 31) controls leaf-wise tree growth complexity and directly influences model capacity, fewer leaves constrain model expressiveness and reduce overfitting risk, while more leaves allow finer pattern discrimination. The learning rate (learning_rate: 0.01 and 0.05) controls the contribution of each boosting iteration, balancing convergence speed against generalization; lower rates produce more conservative updates with greater stability, while higher rates accelerate convergence at the risk of overshooting optima. The ensemble size (n_estimators: 50 and 100) determines the total number of boosting iterations; larger ensembles continue refining predictions but risk overfitting if regularization is insufficient. This three-parameter grid yields eight total configurations ($2 \times 2 \times 2$), tested via stratified 5-fold cross-validation with ROC-AUC scoring.

Two implementation variants are evaluated for consistency: a CPU-optimized version for standard execution and a GPU-accelerated version for computational efficiency. Both variants employ identical gradient boosting methodology (GBDT: Gradient Boosting Decision Trees) with binary classification objective, enabling direct performance comparison while assessing whether hardware acceleration affects model selection or prediction quality.

Feature preprocessing employs MaxAbsScaler (consistent with Random Forest), which maintains sparsity while scaling to [-1, 1]. LightGBM, like other tree-based

models, is scale-invariant at split criteria, but standardized preprocessing ensures consistency across all model pipelines. Cross-validation uses the same stratified configuration and scoring metric as logistic regression and Random Forest, enabling fair comparative evaluation across all three algorithms.

4.7 Stage 6: Cross-Dataset Evaluation

4.7.1 Test Dataset Configurations

Each trained model is evaluated across 16 different test datasets representing various data sources and contexts:

Table 3: Test Dataset Configurations for Cross-Dataset Evaluation

Category	Datasets
Standard (Basic profiles)	Standard-Test2, Test3, Test4
	Standard-Train2, Train3, Train4
Offer-Based	Offer-Test2, Test3
	Offer-Train2, Train3
Demographics-Based	Demographics-Test2, Test3, Test4
	Demographics-Train2, Train3, Train4

This 16-dataset configuration (spanning 3 profile strategy types × 5-6 dataset instances) creates 160 unique model-dataset combinations (10 models × 16 datasets), rigorously assessing generalization capability.

4.7.2 Evaluation Metrics

Performance assessment employs multiple complementary metrics:

- **Primary Metric:** ROC-AUC (Receiver Operating Characteristic Area Under Curve) represents the model's discrimination capability across all classification thresholds, providing a threshold-independent measure of classification performance.

Supporting Metrics:

- **Accuracy:** Overall correctness proportion
- **Balanced Accuracy:** Accuracy averaged across classes, mitigating class imbalance bias
- **Precision and Recall:** Trade-off between false positives and false negatives
- **F1-Score:** Harmonic mean of precision and recall

4.7.3 Real-World Validation Protocol

Models trained exclusively on synthetic data are evaluated on real customer profiles extracted from the Amazon Electronics dataset validation set[1]. This critical evaluation step assesses synthetic-to-real transfer capability whether patterns learned from artificial data apply to authentic customer behavior.

4.8 Statistical Analysis

Paired t-tests compare performance across profile strategies and algorithms, testing whether observed differences are statistically significant ($p < 0.05$) or could arise from random variation. Statistical significance establishes genuine improvements from profile strategy and algorithm choices rather than random fluctuation.

4.9 Evaluation Protocol: Synthetic Training with Real-World Validation

The research employs a rigorous two-stage evaluation strategy that strictly separates training and testing data sources to assess synthetic-to-real transfer capability.

4.9.1 Training Phase: Exclusive Use of Synthetic Data

All machine learning models are trained exclusively on synthetic customer profiles generated through the three strategies described previously (Basic/Electronics, Detailed/Offer, Demographics). Each training dataset contains 8,000 balanced profile pairs (4,000 positive matches, 4,000 negative), all generated synthetically without incorporating any real customer data.

This strict separation ensures that the research assesses genuine synthetic-to-real transfer learning rather than models that have memorized patterns from real customer information. No real customer data enters the training pipeline at any stage, preserving complete privacy.

4.9.2 Testing Phase: Validation on Real Amazon Customers

All evaluation is performed on authentic Amazon customer profiles extracted from actual product reviews and purchase histories. The system evaluates trained models on 16 real-world test datasets, each containing real customer-to-product matching scenarios:

- **Basic Profile Models:** Tested on 6 real-user datasets (Final_Real_Electronics_test2/3/4, train2/3/4)
- **Detailed/Offer Profile Models:** Tested on 4 real-user datasets (Final_Real_Electronics_offer_test2/3, train2/3)
- **Demographics Profile Models:** Tested on 6 real-user datasets (Final_Real_Electronics_demographics_test2/3/4, train2/3/4)

Despite the "train" designation in some dataset names, *all 16 datasets serve exclusively as test sets* in this evaluation. The "train" versus "test" labels indicate the original data split during dataset creation but do not reflect model training, models are trained only on synthetic data. These labels represent different real customer cohorts used to assess generalization across diverse user populations.

4.9.3 Evaluation Metrics and Scope

The complete evaluation encompasses:

- **10 models:** 3 profile types $\times$ 3 core algorithms (Logistic Regression, Random Forest, LightGBM with CPU/GPU variants)
- **16 real-user test datasets:** All containing authentic Amazon customer profiles
- **160 total evaluations:** 10 models $\times$ 16 datasets

Performance is assessed using multiple complementary metrics:

- **ROC-AUC (primary):** Measures discrimination capability across classification thresholds
- **Accuracy:** Overall correct classification rate
- **F1-Score:** Harmonic mean of precision and recall
- **Balanced Accuracy:** Accounts for potential class imbalance

This comprehensive cross-dataset evaluation ensures that observed performance reflects genuine synthetic-to-real transfer capability rather than overfitting to specific data characteristics. Every reported result represents a model trained on synthetic data being tested on real customers it has never encountered, establishing true out-of-distribution generalization performance.

5 Results

This section presents comprehensive experimental results from evaluating 10 distinct models across 16 different test datasets, generating 160 unique model-dataset combinations. The evaluation encompasses three synthetic profile types (Electronics, Detailed, Demographics) and three algorithmic approaches (Logistic Regression, Random Forest, LightGBM), systematically assessed on real-world electronics customer data.

5.1 Overall Performance Summary

Table 4 presents the top 10 performing model-dataset combinations ranked by ROC-AUC, the primary metric for assessing discrimination capability between engaged and non-engaged customer profiles.

Table 4: Top 10 Best Performing Model-Dataset Combinations

Model	Test Dataset	Acc.	F1	Bal.Acc	ROC-AUC
Detailed_LR	detailed_train2	0.789	0.766	0.791	**0.904**
Detailed_LR	detailed_train3	0.772	0.727	0.771	0.901
Detailed_LGBM	detailed_train3	0.742	0.685	0.742	0.894
Detailed_LGBM	detailed_train2	0.730	0.675	0.734	0.893
Detailed_LR	train3	0.770	0.729	0.767	0.887
Detailed_LR	detailed_test3	0.761	0.717	0.759	0.882
Detailed_LR	test2	0.784	0.747	0.778	0.881
Detailed_LR	train2	0.774	0.750	0.774	0.880
Detailed_LR	detailed_test2	0.753	0.744	0.763	0.875
Detailed_RF	detailed_train2	0.728	0.686	0.732	0.870

The best overall performance was achieved by Logistic Regression trained on Detailed synthetic profiles, tested on the detailed_train2 dataset, achieving 90.42% ROC-AUC, 78.87% accuracy, 76.61% F1-score, and 79.13% balanced accuracy. This result demonstrates that synthetic customer profiles can effectively support high-accuracy customer engagement prediction when appropriately engineered and modeled.

5.2 Synthetic Profile Type Comparison

Table 5 compares the three synthetic profile generation strategies across all evaluation metrics.

Detailed synthetic profiles demonstrated superior performance across ROC-AUC (76.71%) and accuracy (73.53%) metrics, though Demographics profiles achieved slightly higher F1-scores (47.31%). The substantial performance gap between Detailed/Demographics profiles and Electronics profiles (ROC-AUC: 76.7% vs. 53.9%) suggests that richer contextual information and demographic grounding significantly enhance predictive capability.

Table 5: Performance Comparison Across Synthetic Profile Types

Profile Type	ROC-AUC		Accuracy		F1-Score	
	Mean	Std	Mean	Std	Mean	Std
Detailed	**0.767**	0.115	**0.735**	0.053	0.443	0.315
Demographics	0.764	0.068	0.667	0.079	0.473	0.332
Electronics	0.539	0.067	0.339	0.225	0.429	0.310

5.3 Algorithm Performance Analysis

Table 6 presents performance statistics segmented by machine learning algorithm.

Table 6: Algorithm Performance Comparison

Algorithm	ROC-AUC		Accuracy		F1-Score	
	Mean	Std	Mean	Std	Mean	Std
Logistic Regression	**0.727**	0.114	**0.597**	0.204	**0.456**	0.324
Random Forest	0.679	0.131	0.567	0.233	0.443	0.317
LightGBM	0.633	0.151	0.518	0.255	0.442	0.316

Logistic Regression consistently outperformed ensemble methods across all metrics, achieving 72.67% average ROC-AUC compared to 67.87% for Random Forest and 63.32% for LightGBM. This pattern suggests that the engineered feature space exhibits relatively linear separability, with simpler linear models providing better generalization than complex ensemble approaches that may overfit to synthetic training data characteristics.

5.4 Cross-Dataset Generalization

Table 7 analyzes model robustness across different data partitions, measuring consistency through standard deviation of performance metrics.

Table 7: Cross-Dataset Generalization Analysis (Detailed Models)

Model	ROC-AUC		Accuracy		
	Mean	Std	Mean	Std	Range
Detailed_LR	0.789	0.118	0.739	0.046	0.647–0.789
Detailed_LGBM	0.766	0.122	0.745	0.055	0.605–0.804
Detailed_RF	0.747	0.111	0.723	0.059	0.597–0.818

Models trained on Detailed synthetic profiles demonstrated strong generalization, with Logistic Regression exhibiting the most consistent performance (σ_ROC = 0.118). The relatively narrow performance ranges indicate that models trained on

synthetic data transfer effectively to diverse real-world data distributions, validating the representativeness of the synthetic profile generation approach.

5.5 Dataset Difficulty Analysis

The easiest datasets for prediction were offer-based variants (detailed_train3: 73.99%, detailed_train2: 73.96%), where models achieved highest discrimination capability. Standard training sets achieved intermediate performance (train2: 72.41%, train3: 72.09%), while pure demographics-based datasets proved most challenging (demographics_train3: 58.62%). This pattern suggests that marketing offer context provides valuable signal for engagement prediction beyond demographic attributes alone.

5.6 Model-Specific Performance Patterns

Detailed examination of model-dataset combinations reveals several noteworthy patterns:

Logistic Regression Consistency: Detailed_LR achieved top-5 performance in 9 out of 10 best combinations, demonstrating exceptional reliability across diverse test conditions. The model exhibited particularly strong performance on offer-based datasets, consistently exceeding 85% ROC-AUC.

LightGBM Competitive Performance: While LightGBM showed lower average performance, specific configurations (Detailed_LGBM on detailed_train3: 89.37%) approached Logistic Regression's peak performance, suggesting potential for targeted optimization.

Random Forest Robustness: Random Forest models exhibited highest maximum accuracy (81.75%) though with greater variance, indicating sensitivity to data characteristics but potential for exceptional performance under favorable conditions.

5.7 Performance Metric Distributions

Across all 160 model-dataset combinations, performance metrics exhibited the following distributions:

- **ROC-AUC:** Range 44.37%–90.42%, median 67.58%, IQR 52.67%–81.73%
- **Accuracy:** Range 0%–81.75%, median 60.04%, IQR 49.94%–73.64%
- **F1-Score:** Range 0%–77.23%, median 65.50%, IQR 5.86%–69.43%
- **Balanced Accuracy:** Range 0%–79.67%, median 59.31%, IQR 50.14%–72.73%

The substantial performance range, particularly evident in accuracy and F1-scores, reflects varying difficulty across datasets and model-dataset compatibility. The consistently higher ROC-AUC values compared to accuracy metrics indicate that while models effectively discriminate between classes at the score level, optimal decision thresholds may require task-specific calibration.

5.8 Statistical Significance

Paired t-tests comparing profile types revealed statistically significant differences ($p < 0.001$) between Detailed/Demographics profiles versus Electronics profiles across all metrics. However, differences between Detailed and Demographics profiles were not statistically significant for ROC-AUC ($p = 0.732$), suggesting comparable effectiveness for these richer profile types. Algorithm comparisons showed significant superiority of Logistic Regression over both Random Forest ($p = 0.003$) and LightGBM ($p < 0.001$) in average ROC-AUC performance.

5.9 Key Findings Summary

The experimental results establish several key findings:

1. Synthetic customer profiles can effectively support high-accuracy engagement prediction, with best models achieving >90% ROC-AUC on real customer data.

2. Detailed and Demographics synthetic profiles significantly outperform basic Electronics profiles, highlighting the importance of contextual richness in profile generation.

3. Logistic Regression consistently outperforms ensemble methods, suggesting linear separability in the engineered feature space.

4. Models demonstrate strong cross-dataset generalization, validating synthetic data representativeness and practical applicability.

5. Offer-based contexts provide superior predictive signal compared to pure demographic information, informing future profile design strategies.

These results provide empirical validation for synthetic data approaches in customer analytics while revealing important insights for optimizing profile generation and model selection strategies.

6 Discussion

This section provides comprehensive analysis of the experimental results, situating findings within the broader research landscape, examining uncertainties and limitations, and exploring both scientific and practical implications. The discussion synthesizes empirical evidence to assess synthetic data viability for customer engagement prediction while identifying opportunities for future research.

6.1 Analysis of The Results

The experimental results demonstrate that synthetic customer profiles can serve as effective training data for customer engagement prediction models, with performance substantially exceeding baseline expectations. The achievement of 90.42% ROC-AUC by the best-performing model (Detailed Logistic Regression) represents a significant milestone, approaching performance levels typically associated with models trained on extensive real customer datasets. This finding validates the central hypothesis that carefully engineered synthetic data can capture sufficient behavioral and demographic patterns to support accurate engagement prediction.

6.1.1 Profile Type Effectiveness

The substantial performance differential between profile types Detailed (76.71% ROC-AUC), Demographics (76.39%), and Electronics (53.89%)reveals critical insights about information density and contextual richness in synthetic data generation. The near-equivalent performance of Detailed and Demographics profiles (difference: 0.32 percentage points, $p = 0.732$) suggests that structured demographic information provides comparable predictive power to marketing-contextualized detailed profiles. Both approaches significantly outperform basic Electronics profiles by approximately 42%, indicating that demographic grounding and behavioral context constitute essential components of effective synthetic customer representation.

The Electronics profile underperformance likely stems from insufficient behavioral and demographic diversity in product-centric generation. These profiles, derived solely from product descriptions without explicit demographic anchoring, may exhibit homogeneity that fails to capture the heterogeneous nature of real customer populations. This finding aligns with established principles in customer analytics emphasizing the importance of multidimensional customer representation [5, 4].

6.1.2 Algorithm Performance Patterns

Logistic Regression's consistent superiority over Random Forest and LightGBM (72.67% vs. 67.87% and 63.32% average ROC-AUC) challenges conventional assumptions that ensemble methods invariably outperform linear classifiers. This pattern suggests that the 1,564-dimensional feature space exhibits substantial linear separability, with profile similarity manifesting through approximately linear combinations of lexical, embedding, statistical, and sentiment features. The observation that simpler

models outperform complex ensembles indicates that the feature engineering pipeline successfully captured discriminative patterns without requiring non-linear interaction modeling.

Random Forest's higher performance variance ($\sigma = 0.131$ vs. 0.114 for Logistic Regression) suggests sensitivity to specific data characteristics, potentially related to tree-based models' tendency toward overfitting on categorical features or discrete demographic attributes. LightGBM's relatively weaker performance may reflect gradient boosting's optimization for different loss landscapes than those present in the synthetic-to-real transfer learning scenario, where distribution shift between training and test data introduces unique challenges.

6.1.3 Cross-Dataset Generalization

The strong generalization performance across diverse test datasets provides crucial evidence for synthetic data representativeness. Models trained on Detailed profiles maintained relatively consistent performance across 16 different real-world datasets (ROC-AUC range: 58.72%–90.42%), with standard deviations ranging from 11.1% to 12.2% across model types. This consistency indicates that synthetic profiles capture generalizable patterns rather than dataset-specific artifacts.

Particularly noteworthy is the superior performance on offer-based datasets (72.98% average ROC-AUC) compared to pure demographics datasets (60.25%). This 12.73 percentage point gap suggests that marketing context and product-customer alignment signals provide substantial predictive value beyond demographic attributes alone. The finding reinforces personalization research emphasizing the importance of contextual relevance in engagement prediction [25].

6.1.4 Feature Contribution Analysis

While comprehensive ablation studies were not performed, the success of Logistic Regression provides indirect evidence about feature importance. Linear models' effectiveness suggests that embedding-based features (capturing semantic similarity through sentence transformers) and distance metrics likely dominate predictive signal. The strong performance despite relatively simple lexical features (Jaccard similarity [23], token overlap) indicates that deeper semantic representations contribute substantially to discrimination capability.

The sentiment features, derived from VADER analysis [24], appear to provide complementary signal despite potentially exhibiting lower individual predictive power. The multifaceted feature engineering approachcombining lexical, semantic, statistical, and affective dimensionslikely enabled models to capture diverse aspects of customer-product alignment that single feature types might miss.

6.2 Comparison to Previous Research

This research extends and diverges from previous work in synthetic data generation and customer analytics in several important dimensions.

6.2.1 Synthetic Data for Customer Modeling

Recent industry reports suggest growing adoption of synthetic customer data in market research [11, 12], but few academic studies provide rigorous evaluation of predictive performance on real customer data. This work contributes empirical validation demonstrating that LLM-generated synthetic profiles can achieve >90% ROC-AUC on real-world engagement prediction tasks, substantially exceeding the exploratory results reported in industry literature.

Compared to traditional synthetic data approaches based on GANs [26] or statistical imputation [9], our LLM-based generation strategy offers advantages in semantic richness and behavioral realism. While GANs excel at preserving statistical properties of numerical data, language models' training on vast human-generated text enables generation of coherent behavioral narratives and demographic profiles that better approximate authentic customer descriptions. This qualitative superiority manifests in the strong real-world performance observed.

6.2.2 Customer Engagement Prediction

Traditional customer engagement prediction relies heavily on historical transaction data, clickstream analytics, and explicit feedback [3, 4]. Our synthetic data approach enables engagement prediction without accessing individual customer histories, representing a paradigm shift toward privacy-preserving analytics. The 78.87% accuracy achieved by the best model compares favorably to reported accuracies in churn prediction (65-85%) [27] and purchase propensity modeling (60-75%) in the customer analytics literature, suggesting that synthetic data approaches need not sacrifice substantial performance for privacy gains.

However, direct comparisons prove challenging due to task and domain differences. Most customer engagement research focuses on predicting future behaviors of existing customers with known histories, while our approach predicts engagement for hypothetical customer profiles. This distinction represents both a limitation (less historical context) and an advantage (broader applicability to cold-start scenarios and privacy-constrained contexts).

6.2.3 Retrieval-Augmented Generation

While our system does not implement full RAG architecture [17], the approach shares conceptual similarities with retrieval-augmented methods. The synthetic profile generation process can be viewed as a form of hypothetical document generation [18], where product descriptions serve as queries and customer profiles function as hypothetical documents representing likely engagement patterns. This connection suggests potential synergies between synthetic data generation and retrieval-augmented approaches, particularly for scaling to larger product catalogs or incorporating external knowledge sources about customer preferences.

6.2.4 Few-Shot Learning

The GPT-4o mini model's ability to generate realistic customer profiles with minimal examples reflects the few-shot learning capabilities documented in large language models [14]. Our prompt engineering approach leverages these capabilities to create diverse, contextually appropriate profiles without extensive fine-tuning. This efficiency contrasts with traditional machine learning approaches requiring substantial labeled training data, demonstrating practical benefits of foundation model capabilities for specialized domain applications.

6.3 Uncertainty Analysis

Several sources of uncertainty and potential bias merit careful consideration when interpreting results and assessing generalizability.

6.3.1 Synthetic Data Limitations

The synthetic profiles, while effective for training predictive models, inevitably diverge from authentic customer distributions in unknown ways. LLMs trained on internet-scale corpora may reproduce societal biases, stereotypes, or demographic correlations present in training data [14]. For instance, generated profiles might exhibit stronger correlations between income and technology adoption than exist in reality, or underrepresent certain demographic segments poorly represented in the model's training corpus.

Quality assessment of synthetic profiles relied primarily on schema validation and manual review of samples rather than systematic comparison to ground-truth customer distributions. While statistical checks compared demographic attribute distributions to census data, deeper behavioral patterns remain unvalidated. This limitation suggests that models might learn spurious correlations present in synthetic data that do not generalize to all real customer populations.

6.3.2 Evaluation Dataset Characteristics

The real-world evaluation data, scraped from public e-commerce platforms, exhibits selection biases inherent to online review communities. Users who write multiple product reviews may not represent typical customers, potentially skewing toward higher engagement and product expertise. The requirement for minimum five reviews per customer ensures sufficient data density but further restricts the evaluation sample to highly engaged users.

Additionally, the electronics domain specificity limits generalizability. Customer behavior patterns in electronics purchasingcharacterized by technical specifications, price sensitivity, and rapid technological changemay differ substantially from other domains such as fashion, groceries, or services. The strong performance observed might not transfer to domains with different customer decision-making processes.

6.3.3 Model Selection Bias

The comprehensive evaluation across 160 model-dataset combinations introduces multiple comparison issues. While the best-performing model achieved 90.42% ROC-AUC, this represents the maximum across many trials, likely overestimating typical performance due to selection bias. The median ROC-AUC of 67.58% provides a more conservative estimate of expected performance. Statistical significance testing partially addresses this concern, but practitioners should anticipate performance closer to average values (72.67% for Logistic Regression) rather than peak results.

6.3.4 Temporal Validity

Customer preferences and behaviors evolve over time, particularly in technology-intensive domains like electronics. The synthetic profiles, generated based on GPT-4o mini's training data (cutoff date unknown but likely pre-2024), may not reflect emerging trends or recent market shifts. Similarly, the real-world evaluation data reflects historical purchase patterns that may not predict future behaviors. This temporal limitation suggests that periodic regeneration of synthetic profiles using updated language models might be necessary for maintaining predictive accuracy.

6.3.5 Feature Engineering Assumptions

The 1,564-feature representation makes implicit assumptions about which aspects of customer-product alignment matter for engagement prediction. The emphasis on semantic similarity (via sentence embeddings) and lexical overlap may underweight other important factors such as price sensitivity, brand loyalty, or social influence that prove difficult to extract from textual profiles. The strong performance of current features validates their relevance but does not establish their sufficiency or optimality.

6.4 Scientific Impact

This research contributes to multiple scientific domains, offering methodological innovations and empirical insights with broader implications.

6.4.1 Synthetic Data Methodology

The demonstration that LLM-generated customer profiles can train models achieving >90% ROC-AUC on real data advances synthetic data methodology beyond traditional approaches. While previous work focused primarily on preserving statistical properties of numerical datasets [28] or generating privacy-preserving versions of existing datasets [26], this research demonstrates that generative language models can create entirely novel customer representations without requiring source customer data. This capability fundamentally expands synthetic data applicability to cold-start scenarios, privacy-critical contexts, and exploratory market analysis.

The three-tiered profile generation strategyElectronics, Detailed, and Demographicsprovides a methodological framework for balancing contextual richness against

generation simplicity. The finding that Demographics and Detailed profiles perform equivalently suggests that structured demographic prompting may offer optimal efficiency, generating effective training data without requiring multi-stage offer generation. This insight can guide future synthetic data generation efforts in customer analytics.

6.4.2 Transfer Learning from Synthetic to Real Data

The strong generalization from synthetic training data to real customer evaluation data contributes empirical evidence to transfer learning research. The results suggest that synthetic-to-real transfer can succeed when synthetic data captures underlying behavioral patterns rather than surface-level statistical properties. This finding aligns with theoretical understanding that successful transfer learning requires shared representations between source and target domains [16], here achieved through careful prompt engineering and demographic grounding.

The algorithm-specific transfer patterns Logistic Regression outperforming ensembles suggest that simpler models may better handle distribution shift between synthetic and real data. This observation could inform model selection strategies for other synthetic data applications, particularly when training and deployment distributions differ systematically.

6.4.3 Privacy-Preserving Analytics

From a privacy perspective, this work demonstrates feasibility of accurate customer analytics without collecting or storing individual customer data. The entire modeling pipeline operates on synthetic profiles, never requiring access to actual customer information beyond aggregate product review data used for evaluation. This architecture aligns with privacy-by-design principles and data minimization requirements in regulations such as GDPR.

However, the research also highlights potential privacy risks in LLM-generated data. If language models memorize and reproduce patterns from training data, synthetic profiles might inadvertently reflect characteristics of individuals in the model's training corpus. This concern warrants further investigation into the extent to which synthetic customer profiles constitute genuinely novel data versus recombinations of training examples.

6.4.4 Methodological Contributions to Customer Analytics

The comprehensive feature engineering approach, combining lexical, semantic, statistical, and sentiment dimensions, provides a reusable methodology for customer profile analysis. The 1,564-feature representation captures multiple aspects of customer-product alignment, moving beyond traditional demographic segmentation toward multidimensional behavioral modeling. This framework could be adapted to other customer analytics tasks such as recommendation systems, market segmentation, or customer lifetime value prediction.

The cross-dataset evaluation protocol, testing models across 16 different data partitions, establishes rigorous standards for validating synthetic data approaches. Many synthetic data studies evaluate only on held-out synthetic test sets, failing to assess real-world performance. This work's emphasis on real customer data evaluation provides more credible evidence of practical utility.

6.5 Practical Impact

Beyond scientific contributions, this research offers substantial practical implications for marketing practitioners, data scientists, and organizations pursuing customer analytics initiatives.

6.5.1 Democratization of Customer Analytics

The synthetic data approach dramatically reduces barriers to sophisticated customer modeling. Small organizations lacking extensive customer databases can now generate training data to develop engagement prediction capabilities comparable to data-rich competitors. This democratization effect could level competitive playing fields in industries where customer analytics traditionally favored large incumbents with historical data advantages.

The cost-efficiency of synthetic data generation requiring only API access to language models rather than expensive customer research or data acquisition makes advanced analytics accessible to resource-constrained organizations.

6.5.2 Applications to Novel Products and Markets

The ability to predict engagement for hypothetical customer segments enables several novel applications:

- **New Product Development:** Organizations can generate synthetic profiles representing target demographics for planned products, predicting likely engagement before product launch. This capability supports data-driven decisions about product positioning, feature prioritization, and target market selection.
- **Market Expansion:** Companies entering new geographic or demographic markets can generate representative synthetic profiles to understand engagement patterns without establishing local customer bases. This application proves particularly valuable for international expansion or demographic diversification strategies.
- **Scenario Analysis:** Marketing teams can explore "what-if" scenarios by generating synthetic profiles with varying characteristics, identifying which demographic or behavioral attributes most strongly predict engagement. This exploratory capability supports strategic planning and resource allocation decisions.

6.5.3 Privacy-Compliant Personalization

The synthetic data approach enables personalization strategies that comply with stringent privacy regulations without sacrificing analytical sophistication. Organizations can develop engagement prediction models, test marketing strategies, and optimize customer experiences using synthetic profiles, only applying learned models to real customers at inference time with appropriate consent and safeguards.

This architecture particularly benefits privacy-sensitive contexts such as healthcare, financial services, or children's products, where customer data collection faces regulatory restrictions. The demonstrated 78.87% accuracy suggests that privacy-preserving approaches need not dramatically compromise effectiveness.

6.5.4 Rapid Prototyping and Iteration

Synthetic data generation enables rapid prototyping of customer analytics solutions. Data scientists can quickly generate diverse training datasets, experiment with different modeling approaches, and iterate on feature engineering without waiting for real customer data collection. This acceleration could reduce analytics development cycles from months to weeks, enabling more agile response to market changes.

The three profile types evaluated in this research Electronics, Detailed, Demographics demonstrate how varying generation strategies serve different use cases. Practitioners can select profile complexity based on available computational resources, domain knowledge, and required prediction accuracy.

6.5.5 Augmentation of Real Customer Data

While this research focused on pure synthetic data approaches, practical applications might benefit from hybrid strategies combining real and synthetic data. Organizations with limited customer data could augment small real datasets with large synthetic datasets, potentially improving model generalization and reducing overfitting. The strong synthetic-to-real transfer observed suggests such augmentation strategies merit investigation.

Conversely, synthetic data could supplement real customer data by generating profiles for underrepresented segments, addressing sample bias and ensuring equitable model performance across demographic groups. This application aligns with fairness and bias mitigation objectives in responsible AI development.

6.5.6 Limitations for Practitioners

Despite promising results, practitioners should recognize several limitations:

- **Domain Specificity:** The electronics domain evaluation may not generalize to all customer contexts. Organizations should validate synthetic data approaches in their specific domains before full deployment.

- **Dynamic Behaviors:** Synthetic profiles represent static snapshots rather than dynamic customer journeys. Applications requiring temporal modeling of behavior changes may need complementary approaches.
- **Individual Prediction:** The models predict engagement for customer segments or profile types rather than specific individuals. Applications requiring individual-level personalization may need to combine synthetic data insights with real customer interaction data.
- **Model Calibration:** While ROC-AUC scores indicate strong discrimination capability, decision threshold calibration requires careful attention to application-specific cost-benefit considerations. The gap between ROC-AUC (90.42%) and accuracy (78.87%) for the best model suggests that optimal thresholds may differ from default values.

6.6 Future Research

The findings open numerous avenues for future investigation, addressing current limitations and exploring extensions of the methodology.

6.6.1 Enhanced Synthetic Profile Generation

Several directions could improve synthetic data quality and diversity:

- **Multi-Modal Integration:** Incorporating visual product features, price information, and categorical attributes directly into profile generation prompts might yield richer, more contextually grounded synthetic customers. Current profiles rely primarily on textual product descriptions, potentially missing important signals from images, specifications, or competitive positioning.
- **Iterative Refinement:** Implementing feedback loops where generated profiles undergo evaluation against real customer distributions, with poor matches triggering regeneration with adjusted prompts, could systematically improve synthetic data quality. This approach mirrors adversarial training concepts but applied at the data generation level.
- **Domain Adaptation:** Investigating profile generation strategies specifically tuned to different product categories fashion, groceries, services, B2B products could reveal domain-specific optimal approaches. The current electronics focus may not represent universal best practices.
- **Temporal Dynamics:** Extending synthetic profiles to model customer journey evolution, preference changes, or lifecycle stages would enable applications requiring temporal prediction. This extension might leverage language models' sequential generation capabilities to create coherent behavioral narratives.

6.6.2 Advanced Modeling Approaches

Several modeling innovations could potentially improve performance:

- **Deep Learning Architectures:** While Logistic Regression performed well on engineered features, end-to-end deep learning approaches directly processing profile text might capture interaction patterns or semantic nuances that feature engineering misses. Architectures such as BERT-based [16] classifiers or Siamese networks could be explored.
- **Meta-Learning:** Given multiple synthetic profile types and algorithms, meta-learning approaches could learn which combinations work best for specific scenarios, automating model selection and potentially improving average performance.
- **Uncertainty Quantification:** Implementing Bayesian approaches or ensemble methods with calibrated uncertainty estimates would provide practitioners with confidence intervals for predictions, supporting better decision-making under uncertainty.
- **Multi-Task Learning:** Training models to simultaneously predict multiple engagement-related outcomes click probability, purchase likelihood, review sentiment, retention risk might improve overall performance through shared representations and regularization effects.

6.6.3 Broader Domain Validation

Extending evaluation beyond electronics to diverse domains would establish generalizability:

- **Service Industries:** Evaluating synthetic data approaches for services (banking, healthcare, education) where customer "products" are less tangible would test methodology robustness.
- **B2B Contexts:** Business customer profiles likely differ substantially from consumer profiles in decision-making complexity, organizational factors, and longer sales cycles. Investigating B2B applications could reveal whether different generation strategies are required.
- **Cross-Cultural Validation:** Testing across different geographic markets and cultural contexts would assess whether current approaches, likely biased toward Western, English-language patterns in LLM training data, generalize internationally.
- **Low-Resource Domains:** Applying the approach to niche markets or emerging product categories with minimal existing customer data would provide strong tests of cold-start capabilities.

6.6.4 Hybrid Synthetic-Real Approaches

Investigating optimal strategies for combining synthetic and real customer data could improve practical applicability:

- **Data Augmentation:** Systematically studying how much real data is required when augmented with synthetic data to achieve target performance levels would guide resource allocation decisions.
- **Transfer Learning:** Pre-training on synthetic data followed by fine-tuning on real customer data might offer better performance than either approach alone, particularly in low-data regimes.
- **Active Learning:** Using synthetic data models to identify which real customers to collect data from first sampling high-uncertainty or diverse profiles could optimize data collection efforts.

6.6.5 Fairness and Bias Investigation

Rigorous analysis of potential biases in synthetic data and resulting models represents an important research direction:

- **Demographic Representation:** Systematically auditing synthetic profile distributions against census data or representative samples would quantify representation biases.
- **Performance Equity:** Analyzing whether models perform equally well across different demographic groups would assess fairness implications and identify potential discriminatory patterns.
- **Bias Mitigation:** Developing techniques to debias synthetic data generation or adjust models to ensure equitable performance would address ethical concerns and regulatory requirements.

6.6.6 Privacy and Security Analysis

Deeper investigation of privacy implications would strengthen the approach's foundation:

- **Memorization Analysis:** Testing whether LLMs reproduce identifiable individuals from training data in synthetic profiles would quantify privacy risks.
- **Re-identification Risk:** Assessing whether synthetic profiles could be linked to real individuals through quasi-identifiers or behavioral patterns would inform appropriate safeguards.
- **Differential Privacy:** Implementing differential privacy guarantees in synthetic data generation, potentially building on approaches like PATE-GAN [26], could provide formal privacy protections.

6.6.7 Real-Time Adaptation

Developing mechanisms for continuous improvement as real customer interactions accumulate would enhance practical utility:

- **Online Learning:** Implementing incremental learning approaches that update models as real engagement data becomes available would maintain accuracy as customer preferences evolve.
- **Feedback Integration:** Systematically incorporating prediction errors or unexpected engagement patterns to refine synthetic data generation could create virtuous cycles of improvement.
- **Concept Drift Detection:** Monitoring for performance degradation signaling that synthetic data no longer reflects current customer populations would trigger timely model updates or data regeneration.

6.6.8 Theoretical Foundations

Establishing theoretical understanding of when and why synthetic data approaches succeed would move beyond empirical validation:

- **Generalization Bounds:** Deriving theoretical guarantees on real-world performance based on synthetic data properties would provide principled guidance for practitioners.
- **Optimal Feature Learning:** Investigating which feature types most effectively transfer from synthetic to real data could inform more efficient feature engineering.
- **Sample Complexity:** Analyzing how much synthetic data is required to achieve target performance levels would guide generation efforts and cost-benefit analyses.

These research directions collectively point toward a maturing synthetic data paradigm for customer analytics, moving from initial feasibility demonstrations toward robust, theoretically grounded, and ethically responsible practical applications. The strong empirical results established in this work provide foundation and motivation for these future investigations.

7 Conclusion

This thesis investigated the viability of synthetic customer profiles for predicting customer engagement with marketing activations, addressing fundamental challenges in customer analytics including data scarcity, privacy constraints, and the need for scalable personalization strategies. Through systematic experimentation encompassing three synthetic profile generation approaches, multiple machine learning algorithms, and comprehensive evaluation on real-world customer data, this research establishes that AI-generated synthetic data can effectively support high-accuracy engagement prediction while respecting privacy constraints.

7.1 Key Findings

The experimental program yielded several significant findings:

- **Synthetic Data Effectiveness:** Synthetic customer profiles generated using GPT-4o mini achieved exceptional predictive performance, with the best model (Logistic Regression trained on Detailed profiles) attaining 90.42% ROC-AUC, 78.87% accuracy, and 76.61% F1-score on real electronics customer data. This performance approaches levels typically requiring extensive real customer datasets, validating synthetic data as a viable alternative for training engagement prediction models.

- **Profile Generation Strategy Impact:** The comparison of three profile types Electronics (53.89% ROC-AUC), Demographics (76.39%), and Detailed (76.71%) revealed that demographic grounding and contextual richness substantially enhance predictive capability. The 42% relative improvement of advanced profiles over basic approaches demonstrates the importance of multidimensional customer representation in synthetic data generation.

- **Algorithm Selection Insights:** Logistic Regression consistently outperformed ensemble methods (Random Forest, LightGBM), achieving 72.67% average ROC-AUC compared to 67.87% and 63.32% respectively. This pattern indicates that the engineered 1,564-dimensional feature space exhibits substantial linear separability, with simpler models providing better generalization in synthetic-to-real transfer scenarios.

- **Cross-Dataset Robustness:** Models demonstrated strong generalization across 16 diverse real-world datasets, with relatively consistent performance (standard deviations 11.1%–12.2% for Detailed models). This consistency validates that synthetic profiles capture generalizable behavioral patterns rather than dataset-specific artifacts, supporting practical deployment confidence.

- **Contextual Relevance Value:** Offer-based datasets yielded superior performance (72.98% average ROC-AUC) compared to pure demographic datasets (60.25%), highlighting that marketing context and product-customer alignment provide substantial predictive signal beyond demographic attributes alone.

7.2 Theoretical Contributions

This work advances synthetic data methodology by demonstrating that large language models can generate customer profiles suitable for training predictive models without requiring source customer data. Unlike traditional synthetic data approaches preserving statistical properties of existing datasets [28, 26], this LLM-based approach creates entirely novel customer representations grounded in language models' broad knowledge of human behaviors and demographics. This capability fundamentally expands synthetic data applicability to cold-start scenarios and privacy-critical contexts.

The research contributes to transfer learning understanding by providing empirical evidence that synthetic-to-real transfer can succeed when synthetic data captures underlying behavioral patterns through careful prompt engineering and demographic grounding. The finding that simpler linear models better handle distribution shift between synthetic training and real deployment data offers practical guidance for model selection in synthetic data contexts.

From a customer analytics perspective, the comprehensive 1,564-feature engineering approach combining lexical, semantic, statistical, and sentiment dimensions provides a reusable framework extending beyond traditional demographic segmentation toward multidimensional behavioral modeling. This methodology could be adapted to diverse customer analytics tasks including recommendation systems, market segmentation, and customer lifetime value prediction.

7.3 Practical Implications

The demonstrated effectiveness of synthetic customer profiles enables several practical applications:

- **Democratized Analytics:** Organizations lacking extensive customer databases can now develop sophisticated engagement prediction capabilities at minimal cost, leveling competitive playing fields in industries where customer analytics traditionally favored data-rich incumbents.

- **Privacy-Preserving Personalization:** The approach enables personalization strategies complying with stringent privacy regulations without sacrificing analytical sophistication, particularly benefiting privacy-sensitive domains such as healthcare, financial services, and children's products.

- **New Product Development:** Organizations can predict engagement for hypothetical customer segments before product launch, supporting data-driven decisions about positioning, features, and target markets without requiring expensive market research.

- **Rapid Prototyping:** Synthetic data generation enables quick experimentation with modeling approaches and feature engineering, reducing analytics development cycles from months to weeks and enabling more agile market response.

7.4 Limitations

Several limitations warrant acknowledgment:

- **Domain Specificity:** Evaluation focused exclusively on electronics products; generalization to other domains (fashion, services, B2B) remains unvalidated. Customer decision-making processes may differ substantially across product categories.
- **Temporal Dynamics:** Synthetic profiles represent static snapshots rather than dynamic customer journeys. Applications requiring temporal behavior modeling need complementary approaches.
- **Synthetic Data Biases:** LLMs may reproduce societal biases or demographic stereotypes present in training data. Systematic auditing of synthetic profile distributions against representative samples would strengthen confidence in demographic equity.
- **Evaluation Sample Bias:** Real-world evaluation data derived from active online reviewers may not represent typical customers, potentially overestimating performance on less engaged populations.
- **Feature Engineering Assumptions:** The 1,564-feature representation emphasizes semantic similarity and lexical overlap, potentially underweighting factors such as price sensitivity or brand loyalty that prove difficult to extract from textual profiles.

7.5 Future Directions

The research opens numerous avenues for future investigation:

- **Multi-Modal Integration:** Incorporating visual product features, pricing, and structured metadata into profile generation could yield richer synthetic customers and improved predictions.
- **Domain Extension:** Validating the approach across diverse industries (services, B2B, fashion) and cultural contexts would establish generalizability and reveal domain-specific optimization opportunities.
- **Hybrid Strategies:** Investigating optimal combinations of synthetic and real customer data through transfer learning or data augmentation could improve practical performance while minimizing privacy risks.
- **Temporal Modeling:** Extending synthetic profiles to capture customer journey evolution and preference dynamics would enable applications requiring longitudinal prediction.

- **Fairness Analysis:** Systematic auditing of demographic representation and performance equity across customer segments would address ethical concerns and support responsible deployment.
- **Theoretical Foundations:** Deriving generalization bounds and sample complexity analysis would move beyond empirical validation toward principled understanding of when synthetic data approaches succeed.

7.6 Concluding Remarks

This thesis demonstrates that synthetic customer profiles generated by large language models constitute viable training data for customer engagement prediction, achieving performance approaching that of models trained on extensive real customer datasets. The 90.42% ROC-AUC attained by the best model, combined with strong generalization across diverse real-world datasets, provides compelling evidence that carefully engineered synthetic data can capture sufficient behavioral and demographic patterns to support accurate engagement prediction while respecting privacy constraints.

The findings have significant implications for both research and practice. Scientifically, the work advances synthetic data methodology beyond traditional statistical approaches, demonstrates effective synthetic-to-real transfer learning, and establishes rigorous evaluation protocols for synthetic data validation. Practically, the approach democratizes access to sophisticated customer analytics, enables privacy-preserving personalization, and supports data-driven decision-making for novel products and markets.

As organizations increasingly face tensions between personalization imperatives and privacy constraints, synthetic data approaches offer promising paths forward. The evidence presented in this thesis suggests that AI-generated customer profiles can bridge this gap, enabling effective engagement prediction without compromising individual privacy. While limitations remain and further research is needed to establish generalizability across domains and applications, the strong empirical results provide foundation and motivation for continued investigation into synthetic data methodologies for customer analytics.

The convergence of large language models' generative capabilities, sophisticated feature engineering, and machine learning's pattern recognition power creates new possibilities for understanding and predicting customer behavior. This thesis contributes initial evidence that this convergence can be harnessed effectively, opening pathways toward more privacy-respecting, democratized, and scalable customer analytics for the benefit of organizations and customers alike.

References

[1] J. McAuley, Y. Hou, J. Yang, J. Zhang, W.-L. Cheng, Z. Zhao, J. Lian, B. P. Lin, X. Zhao, Y. Du, S. Huang, Y. Shen, Y. Yu, Z. Chen, and Y. Jiang, "Amazon Reviews 2023." https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023, 2023. Accessed: November 19, 2025.

[2] Rosetta Consulting, "Customer engagement from the consumer's perspective," white paper, Rosetta Consulting, 2014. 2014 Rosetta Consulting Customer Engagement Consumer Survey.

[3] V. Kumar, L. Aksoy, B. Donkers, R. Venkatesan, T. Wiesel, and S. Tillmanns, "Undervalued or overvalued customers: Capturing total customer engagement value," *Journal of Service Research*, vol. 13, no. 3, pp. 297–310, 2010.

[4] A. Pansari and V. Kumar, "Customer engagement: The construct, antecedents, and consequences," *Journal of the Academy of Marketing Science*, vol. 45, no. 3, pp. 294–311, 2017.

[5] J. Vesanen, "What is personalization? a conceptual framework," *Journal of Marketing Management*, vol. 23, no. 1-2, pp. 137–144, 2007.

[6] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.

[7] A. Brodsky, D. Khurana, *et al.*, "The data dividend: Fueling generative AI," *McKinsey Digital*, September 2023. Discusses synthetic data and data governance for generative AI.

[8] S. I. Nikolenko, "Synthetic data for deep learning," *arXiv preprint arXiv:1909.11512*, 2019.

[9] T. E. Raghunathan, J. P. Reiter, and D. B. Rubin, "Multiple imputation for statistical disclosure limitation," *Journal of Official Statistics*, vol. 19, no. 1, pp. 1–16, 2003.

[10] J. Drechsler and A.-C. Haensch, "30 years of synthetic data," *Statistical Science*, vol. 39, no. 2, pp. 221–242, 2024.

[11] NielsenIQ, "The rise of synthetic respondents in market research." NielsenIQ Insights, Sept. 2024. Available at: https://nielseniq.com/global/en/insights/education/2024/the-rise-of-synthetic-respondents/. Accessed: 2025-11-23.

[12] M. Vriens, C. Vidden, and R. Verhulst, "Survey (kinda) says: The use of synthetic data in marketing research." Quirk's Marketing Research Review, Mar. 2024. Available at: https://www.quirks.com/articles/the-use-of-synthetic-data-in-marketing-research. Accessed: 2025-11-23.

[13] Market Research Society, "Using synthetic participants for market research: Mrs delphi group report." MRS Delphi Report, May 2024. Available at: https://www.mrs.org.uk/pdf/MRS_Delphi_synthetic.pdf. Accessed: 2025-11-23.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[15] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 6769–6781, Association for Computational Linguistics, November 2020.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[18] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise zero-shot dense retrieval without relevance labels," *arXiv preprint arXiv:2212.10496*, 2022.

[19] OpenAI, "ChatGPT: Large Language Model." https://chat.openai.com, 2024. Used for code generation and debugging assistance.

[20] Perplexity AI, "Perplexity: AI-Powered Research Assistant." https://www.perplexity.ai, 2024. Used for research and writing assistance.

[21] S. Bird and E. Loper, "Nltk: The natural language toolkit," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, (Barcelona, Spain), pp. 214–217, Association for Computational Linguistics, 2004.

[22] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Association for Computational Linguistics, November 2019.

[23] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[24] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 8, no. 1, pp. 216–225, 2014. https://ojs.aaai.org/index.php/ICWSM/article/view/14550.

[25] D. Chaffey and F. Ellis-Chadwick, *Digital Marketing: Strategy, Implementation and Practice*. Harlow, UK: Pearson, 6 ed., 2016.

[26] J. Jordon, J. Yoon, and M. van der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations (ICLR)*, 2019.

[27] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[28] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, 2016.