

RAG를 이용한 마케팅 활용에 대한 연구 : 신용카드 할부 예측을 중심으로*†

한 선 의(제1저자)

서강대학교 일반대학원 경영학과 Business Analytics(석사)

이 군 희(교신저자)

서강대학교 경영학과(교수)

Research on the Application of RAG in Marketing : Prediction of Credit Card Installment

Han, Seon Ui(First Author)

Department of Business Administration, Sogang University (Master)

Lee, Gun Hee(Corresponding Author)

Department of Business Administration, Sogang University (Professor)

Abstract

This study explores the potential applications of RAG based LLM in predicting installment purchases, a major revenue source for credit card companies. The research compared and analyzed the performance of various predictive models using both large and small datasets, with particular emphasis on validating the

* 이 연구는 AI-Hub 데이터를 기반으로 수행된 연구임

† 본고는 한선이의 2025년도 석사학위 논문 「RAG를 이용한 마케팅 활용에 대한 연구 -신용카드 할부 예측을 중심으로-」의 일부를 발췌하여 재작성한 것임

effectiveness of RAG-based LLM. Analysis results showed that the Random Forest model achieved the highest prediction accuracy in large-scale samples. Conversely, in small-scale samples, RAG-based LLM demonstrated superior predictive performance, particularly proving its high potential in analyzing new customer characteristics and developing personalized marketing strategies based on these insights. This study suggests that RAG-based LLM can serve as an effective tool that maintains comparable performance to existing predictive models while providing additional insights for marketing strategy development. Furthermore, the study derived practical implications by presenting the possibility of automating the "RAG - Analysis - Strategy Development" process using AI Agents.

Keywords : RAG (Retrieval-Augmented Generation), LLM (Large Language Model), Credit Card Installment, Logistic Regression, Tree-Based Ensemble Model

접수일(2025년 02월 17일), 수정일(2025년 03월 10일), 게재확정일(2025년 03월 25일)

I. 서론

신용카드는 국내에서 가장 보편적인 지급수단으로 자리 잡고 있으며, 비용 부담에도 불구하고 편리성과 높은 수용성으로 인해 가장 높은 사용자 만족도를 보인다(금융위원회, 2024). 2023년 12월 기준 국내 신용카드 발급 수는 약 1억 2,980만 개에 달하며, 연간 이용 금액은 약 941조 원으로 추산된다. 이 중 신용카드 할부 이용 금액은 약 162조 원으로 전체 이용 금액의 17.2%를 차지하고 있다(여신금융협회, 2024).

신용카드사의 주요 수익원은 할부 및 연체 이자와 가맹점 수수료가이다. 그러나 가맹점 수수료 적격비용 재산정 제도의 도입으로 수수료 수익의 비중은 지속적으로 감소하고 있다. 현재 우대 가맹점은 연 매출 30억 원 이하인 가맹점으로, 전체의 96%에 달하며, 이는 제도 도입 당시 연 매출 2억 원 이하, 전체 가맹점의 68%였던 것에 비해 크게 증가했다. 반면, 할부 결제의 규모는 꾸준히 증가하여 2024년 상반기에만 약 1조 7,000억 원의 수익을 창출했다(금융감독원 금융통계 정보시스템, 2024). 이는 할

부 서비스가 신용카드사의 핵심 수익원임과 동시에 소비자의 소비 패턴과 재무적 여건을 반영하는 중요한 지표로서, 고객 가치 평가와 개인화된 서비스 제공에 핵심적인 정보를 제공함을 시사한다.

최근 인공지능 기술의 급속한 발전은 다양한 산업 분야에서 새로운 기회를 창출하고 있다. 특히 대규모 언어 모델(Large Language Model, LLM)의 등장으로 실제 산업 현장에서의 활용 가능성이 크게 주목받고 있다. 그러나 기존 LLM은 새로운 지식의 업데이트와 수정의 어려움, 예측에 대한 인사이트 제공의 한계, 환각(Hallucination) 현상 등 세 가지 주요 한계점을 가지고 있다(Lewis et al., 2020). 이러한 한계를 극복하기 위해 생성형 AI와 검색 기반 정보를 결합한 RAG(Retrieval-Augmented Generation) 기술이 대안으로 부상하고 있다. RAG는 최신 데이터의 실시간 반영이 가능하고 기업 내부 데이터를 안전하게 활용할 수 있는 환경을 제공하며, 특히 환각 현상을 최소화할 수 있는 효과적인 방법으로 평가받고 있다.

본 연구는 신용카드 할부 서비스의 전략적 중요성과 데이터 기반 마케팅 전략의 필요성에 주목하여, RAG 기반의 할부 이용 예측 모델을 개발하고자 한다. 이를 위해 인구통계학적 특성, 경제적 특성, 소비 및 생활 패턴, 신용정보, 업종별 카드 소비 금액, 업종별 관심도 등 다양한 데이터를 활용한다. 연구에서는 전통적인 통계 모델과 머신러닝 기반의 앙상블 모델, Logistic Regression 수식 기반 LLM, 그리고 RAG 기반 LLM의 성능을 비교 분석하며, 특히 RAG 기반 LLM을 활용한 할부 이용 가능성 예측을 넘어 효과적인 고객 세분화와 맞춤형 마케팅 전략 도출에 주안점을 둔다.

II. 선행 연구 및 이론적 배경

2.1 신용카드 관련 선행 연구

신용카드 이용 행태와 연체 예측에 관한 연구는 오랫동안 활발히 진행되어 왔다. 본 연구는 이러한 기존 연구들과 차별화하여 신용카드사의 수익성 관점에서 고객의 할부 이용 예측에 초점을 맞추고 있다. 관련 선행 연구는 크게 (1) 신용카드 이용 행태 분석, (2) 신용카드 데이터 기반 예측 모델링, (3) 실제 마케팅 활용 사례의 세 가지 영역으로 구분할 수 있다.

신용카드 이용 행태 연구는 주로 고객의 소비 패턴과 카드 사용 방식이 신용카드사의 수익에 미치는 영향을 분석하는 데 중점을 두었다. 이동규와 신민수(2023)는 휴먼

고객 예측을 위해 소비 금액, 이용 건수, 현금서비스 등의 기본적인 소비 정보와 함께 라이프스타일, 앱 이용 등 부가적인 행태 정보를 활용하였다. 특히 과거 소비 패턴이 고객 행동 예측의 핵심 변수임을 입증하였으며, 라이프스타일이 소비자의 가치관과 구매 행동에 직접적 영향을 미친다는 점을 강조하였다. 이를 통해 고객 세그먼트별 차별화된 상품과 서비스 개발의 필요성이 제기되었다(박만석과 김동준, 1995). 김보영과 박미량(2015)은 여가 서비스의 관계 혜택이 고객을 유치하는 데 긍정적인 영향을 미친다는 것을 입증하였다.

신용카드 데이터 기반 예측 모델링 연구는 주로 부도 예측 등 리스크 관리에 초점을 맞추었다. 장동호(2012)의 연구는 소득/자산 수준이 낮을수록, 연령이 낮을수록 신용카드 대출 의존도가 높아짐을 밝혔다. 또한 교육 수준, 가구 규모, 결혼 상태, 직업, 주거 형태, 성별 등이 신용카드 채무에 유의미한 영향을 미치는 것으로 나타났다. Capon(1982)은 예측의 정확성뿐만 아니라 결과에 대한 명확한 해석이 가능한 변수 선정의 중요성을 강조하였다. 이에 따라 기존의 인구통계학적 변수에 자녀 수, 부양가족 수, 주택 및 부동산 보유 여부 등 경제·생활 관련 변수를 추가하여 모델의 설명력을 제고하려는 시도가 이어졌다. 또한 Arram et al(2023)은 다양한 머신러닝 알고리즘을 활용하여 신용카드 연체 예측 모델을 개발하였으며, 그중 다층 퍼셉트론(MLP) 신경망이 우수한 성능을 보였다고 보고하였다. 정원석 등(2024)은 신용카드 거래 데이터를 기반으로 매출을 예측하였으며, 딥러닝 기반 시계열 모델이 소비 패턴 분석과 매출 예측에 효과적임을 입증하였다. 이러한 연구들은 신용카드사의 리스크 관리 및 마케팅 전략 수립에 데이터 기반 접근이 중요함을 시사한다.

실제 마케팅 활용 사례 연구에서는 신용카드사들의 데이터 기반 마케팅 전략 수립 과정을 살펴볼 수 있다. KB국민카드는 인구통계학적 정보, 업종별 카드 사용 데이터, 마케팅 반응 데이터, 소비 상권 데이터를 통합적으로 분석하여 마케팅 전략을 수립하였다(이유재 등, 2014). 현대카드는 라이프스타일과 사회심리학적 분석을 통해 VIP 고객군의 특성을 파악하고 이에 기반한 신상품 개발로 2010년 VIP 신용카드 시장점유율 1위를 달성하였다(이지훈 등, 2014). 이찬경과 노형봉(2018)은 업종별 이용 금액 데이터가 카드사의 혜택 구조와 마케팅 전략 수립에 핵심적인 정보를 제공할 수 있음을 입증하였다.

이러한 선행 연구들은 신용카드 이용 예측과 마케팅 전략 수립에 있어 다양한 변수들의 중요성을 시사하며, 본 연구의 분석 프레임워크 구축에 중요한 이론적 기반을 제공한다.

2.2 Logistic Regression

Logistic Regression은 독립 변수의 선형 결합을 이용하여 종속 변수를 예측하는 확률 추론 통계 기법이다. 특히 종속 변수가 두 개인 이진 분류 문제에서 주로 사용되는 통계적 모형이다. 각 독립변수가 종속변수에 미치는 영향을 해석할 수 있는 장점이 있으며 이를 통해 변수의 중요도를 정량적으로 파악할 수 있다. Logistic Regression 모형의 수식은 다음과 같다.

$$Y = p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (1)$$

여기서 Y 는 종속변수, X_n 는 독립변수, β_n 는 회귀계수를 의미한다. 본 연구에서는 신용카드 할부 이용 여부를 종속 변수로 설정하고, 다양한 독립 변수를 바탕으로 할부 이용 여부를 예측하였다.

2.3 결정 트리 기반 앙상블 기법

앙상블(Ensemble) 기법이란 여러 개의 개별 모델을 결합하여 하나의 모델을 구성하는 머신러닝 기법으로, 앙상블 기법의 목적은 예측 성능과 안정성을 높이는 데 있다. 앙상블 기법은 크게 배깅(Bagging)과 부스팅(Boosting) 기법으로 구분된다. 배깅은 Bootstrap Aggregating의 약자로, 여러 개의 Bootstrap 학습 데이터셋을 사용하여 개별 의사결정 트리를 생성한 후, 그 예측 결과를 결합하는 방법이다(Breiman, 1996). 반면, 부스팅은 여러 개의 의사결정 트리를 순차적으로 학습하여, 선행 트리의 오류를 보정하며 최종 성능을 개선하는 방법이다(Freund and Schapire, 1997). 본 연구에서는 대표적인 앙상블 모형인 Random Forest, XGBoost, LightGBM를 사용하였다.

2.4 Random Forest

Random Forest는 다수의 결정 트리를 생성하고, 각 트리의 예측 결과를 결합하여 최종 예측을 수행하는 배깅 기반의 앙상블 기법이다. 각 트리는 독립적으로 생성되며, 동일한 분포를 갖는 랜덤 벡터의 값에 따라 생성된다. 배깅은 모든 예측 변수 중 최적의 분할을 선택하는 방식으로 트리를 형성하는 반면, Random Forest는 각 노드에

서 예측 변수들을 임의로 추출하고, 추출된 변수 내에서 최적의 분할을 만드는 과정을 추가로 포함한다. 이러한 임의성을 관측치뿐만 아니라 변수에도 적용하는 절차는 결정 트리의 주요 단점인 과적합(over-fitting) 문제를 완화하고 예측의 정확도를 높이는 데 도움을 준다(Breiman, 2001).

2.5 XGBoost

Chen and Guestrin(2016)에 의해 개발된 XGBoost(Extreme Gradient Boosting)는 GBDT(Gradient Boosting Decision Tree)를 개선하고 확장한 모델이다. 기존의 GBDT는 결정 트리를 직렬적으로 결합하여 데이터를 학습하는 방식으로 대용량의 데이터를 학습하기에 적합하지 않지만, XGBoost는 병렬 처리를 통해 트리 생성 속도를 크게 향상하며, 대규모 데이터에서도 빠르게 학습할 수 있다. 또한 L1 및 L2 정규화 항을 목표 함수에 추가하여 과적합을 방지하고, 학습률(Learning Rate)을 활용하여 각 단계에서 모델의 기여도를 조정함으로써 모델이 지나치게 복잡해지는 것을 방지한다.

2.6 LightGBM

LightGBM은 Ke et al(2017)에 의해 제안된 모델로, GBDT 기반의 모형으로 기존의 GBDT 모델과 달리 리프 중심(leaf-wise) 방식을 사용하여 모델의 정확도를 높인다. LightGBM은 부스팅 과정에서 기여도가 높은 데이터 샘플을 우선 사용하고, 기여도가 낮은 샘플은 무작위로 선택하는 Gradient-based One-Side Sampling(Goss) 알고리즘과 희소 데이터(sparse data)에서 상호 배타적인 feature를 묶어 차원을 줄이는 Exclusive Feature Bundling(EFB) 알고리즘을 사용한다. 이러한 방식을 사용하여 데이터셋의 크기를 줄이고, 수행 속도를 큰 폭으로 높였다.

2.7 대규모 언어 모델

LLM은 대규모 데이터를 학습하여 자연어를 이해하고 생성할 수 있는 딥러닝 기반 모델이다. LLM은 트랜스포머(Transformer) 아키텍처를 기반으로 구축되며, 어텐션 메커니즘(Attention Mechanism)을 사용하여 데이터를 처리한다. LLM은 방대한 텍스트 데이터를 기반으로 훈련되며, 다양한 자연어 처리(Natural language processing, NLP) 작업에서 뛰어난 성능을 보인다. 이러한 모델은 텍스트 생성, 번역, 질문 응답 등 다양한 분야에서 응용된다.

최근 금융 분야에서도 LLM을 활용한 연구가 활발히 진행되고 있으며, 고객의 금융 관련 의도를 이해하고 맞춤형 서비스를 제공하는 데 LLM이 중요한 역할을 하고 있다 (Srivastava, 2024). 그러나 LLM은 보안 문제, 환각 현상, 데이터 편향, 최신 정보 부족 등의 한계점을 가지고 있으며, 특히 숫자 계산이나 특정 도메인에 특화된 질문에 대해서는 부족한 성능을 보인다. 이러한 한계를 극복하기 위해 외부 지식을 활용할 수 있는 RAG 기법이 함께 연구되고 있다. 대표적인 LLM으로는 GPT, LLaMA, Claude, Gemini, Mistral 등이 있으며, 본 연구에서는 오픈소스 기반 모델인 LLaMA를 활용하였다.

2.8 LLaMA

LLaMA(Large Language Model Meta AI)는 Meta에서 개발한 오픈소스 대규모 언어 모델로, 다국어 지원, 코드 작성, 추론, 도구 사용 등을 포함한 다양한 기능을 지원한다. 본 연구에서 활용한 LLaMA3 모델의 경우 80억(8B), 700억(70B), 4,050억(405B) 파라미터로 구성된 다양한 크기의 모델을 제공하며, 약 15조 개의 토큰으로 구성된 대규모 데이터셋을 기반으로 학습되었다(Dubey et al., 2024). 또한 안정적인 학습을 위해 Dense Transformer 아키텍처를 채택하여 구축되었다. LLaMA는 오픈소스로 제공되어 연구 및 상업적 목적으로 자유롭게 사용이 가능하다는 장점이 있다.

특히, 최근 연구에서는 LLaMA 모델을 금융 분야에 특화하여 활용하는 사례가 나타나고 있다. Iacovides et al(2024)이 개발한 FinLlama는 금융 감성 분석을 수행하는 모델로 금융 뉴스의 감성 밸런스를 분류하고 그 강도를 정량화하여 알고리즘 트레이딩에 활용할 수 있도록 설계되었다. 이러한 연구는 LLaMA 모델이 특정 도메인에 특화된 지식 학습이 가능함을 보여준다. 또한 LLaMA3 모델은 금융 분야의 도메인 지식을 효과적으로 학습하며, 긴 문장에 비해 짧은 문장에서 더 나은 성능을 보이는 것으로 확인되었다(Mai et al., 2024).

2.9 RAG (Retrieval-Augmented Generation)

RAG(Retrieval-Augmented Generation) Lewis et al(2020)에 의해 고안된 기법으로 검색 시스템을 통해 관련 정보를 검색하고, 이를 바탕으로 텍스트를 생성하는 방법이다. 이 기술은 LLM의 한계인 환각 현상을 줄이고, 외부 지식 베이스를 활용하여 사실에 기반한 응답을 생성할 수 있다는 장점이 있다. RAG는 대규모 언어 모델의 성능을 강화하기 위해 설계되었으며, 대규모 문서 데이터베이스에서 필요한 정보를 검색하여 보다 정확하고 상세한 답변을 생성할 수 있게 한다.

RAG는 크게 검색기(Retriever)와 생성기(Generator)로 구성되어 있다. 검색기는 저장된 벡터 데이터베이스에서 입력된 질문과 관련된 문서를 검색하는 과정이다. 이 단계는 질문에 가장 적합한 정보를 신속하게 찾아내는 것이 목표이며, RAG의 성과 직결되는 매우 중요한 과정이다. 생성기는 앞서 검색된 문서를 기반으로 대규모 언어 모델을 활용하여 응답을 생성하는 과정이다. 이 단계는 언어 모델의 능력을 최대한 활용하여 사용자의 질문에 대한 최종 답변을 생성한다.

최근 금융 분야에서는 RAG를 활용하여 LLM의 성능을 개선하는 연구가 활발하게 이루어지고 있다. Srivastava(2024)는 금융 분야의 텍스트 분류 작업에서 RAG 기법을 적용한 소규모 오픈소스 LLM이 대규모 폐쇄형 LLM보다 우수한 성능을 보인다는 것을 확인하였다. 이는 금융 데이터의 최신성을 유지하면서도, 외부 문서를 적극적으로 활용하여 모델의 정확성을 높일 수 있다는 점에서 의미가 있다.

RAG는 기존 LLM의 한계였던 보안 측면에서 강점이 있다. 외부 데이터베이스에서 실시간으로 정보를 검색하기 때문에, 기밀성이 높은 데이터를 LLM 내부에 저장하지 않아도 된다. 특히 기업 내부 문서나 개인 정보와 같은 민감한 정보를 모델에 학습시키지 않고 사용하여 보안성을 강화할 수 있다.

본 연구에서는 LangChain을 활용하여 RAG 프로세스를 구축하였다. Langchain은 언어 모델 응용 애플리케이션 개발을 위한 오픈소스 프레임워크로 복잡한 과정을 단순화하여 개발 시간을 줄이고, 다양한 데이터 소스와의 호환성을 확보할 수 있도록 설계되었다(Topsakal and Akinci, 2023). 이를 통해 보다 효율적인 RAG 기반 시스템을 구축하고, 금융 도메인에 특화된 맞춤형 분석을 진행할 수 있도록 설계하였다.

2.10 벡터 스토어

벡터 스토어(Vector Store)는 텍스트, 이미지, 음성 등 비정형 데이터를 벡터 형식으로 변환하여 저장한 데이터베이스로 이를 기반으로 유사성 검색을 수행하는 데 사용된다. 사용자가 입력한 쿼리를 벡터화한 뒤, 벡터 스토어에 저장되어 있는 벡터와의 유사성을 코사인 유사도 및 유클리디안 거리와 같은 수학적 거리를 계산하여 가장 연관성이 높은 데이터를 반환한다. 벡터 스토어의 경우 빠른 검색 속도를 제공하며, 시스템의 성능 저하 없이 대규모 데이터를 관리할 수 있도록 확장성(Scalability)을 제공한다. 또한 벡터 스토어는 의미적으로 유사한 단락 검색을 지원하기에 키워드에 의존하는 검색 형태에서 벗어나 더욱 정확한 검색을 가능하게 한다. 벡터 스토어는 RAG의 검색 기능과 직접적으로 연결되어 있으며, 응답 시간과 정확성이 큰 영향을 미쳐 벡터 스토어에 저장하는 단계가 굉장히 중요하다.

벡터 스토어에는 대표적으로 Chroma, FAISS, Pinecone 등이 있으며, 로컬 및 클라

우드 환경 지원, 유사성 검색 방식에서 차이가 있다. 본 연구에서는 로컬에서 사용이 가능하며, RAG 시스템 개발에 최적화되어 있어 LLM 기반 작업에 적합한 ChromaDB를 사용하였다.

Ⅲ. 연구 모형 및 분석

3.1 데이터 설명

본 연구는 한국지능정보사회진흥원(National Information Society Agency, NIA)의 지원을 받아 구축된 금융 합성 데이터를 사용하였다. 해당 데이터는 AI-Hub에서 제공되었으며, 연구 목적으로 사용할 수 있는 데이터이다(AI-Hub, n.d.). 본 연구에서 활용한 데이터는 SK텔레콤, 신한카드, KCB가 협력하여 구축한 GranData 사업의 일환으로 생성된 가명 정보 결합 금융 합성 데이터로, 금융 분야의 인공지능 학습과 데이터 분석을 지원하기 위해 구축되었다. 이 데이터는 실제 금융기관에서 활용되고 있는 데이터의 특성을 인공지능 알고리즘으로 학습하여 생성한 데이터로 소비자의 트렌드를 파악하고 예측 모델을 구축하는 데 강점이 있다.

본 연구에 사용된 데이터는 개인 단위의 가명 결합 데이터로, 서로 다른 개인정보 처리자가 보유하고 있는 개인 정보를 가명 처리하여 결합한 데이터이다(가명정보지원 플랫폼, 2024). 결합 데이터를 활용하며 발생할 수 있는 개인 정보 노출 위험 및 데이터 불균형 등과 같은 여러 한계점을 극복하고자 합성 기술을 적용하였다. 합성 데이터는 원본 데이터와 통계적 특성이 유사하여, 실제 원본 데이터 분석 결과와 유사한 결과를 얻을 수 있도록 재현한 가상의 데이터이다. 합성 데이터는 민감한 정보가 포함된 경우 법적 제약 내에서 안전하게 사용할 수 있으며, 비율 효율성과 AI 모델 성능 향상에 있어 장점을 가진다.

해당 데이터는 생성적 적대 신경망(Generative Adversarial Network, GAN) 알고리즘과 R의 Synthpop 라이브러리를 활용하여 데이터를 합성하였다. 분석 데이터는 2021년 1분기부터 2022년 4분기까지 총 2년간 수집되었으며, 분기별 33만 명에 대한 738개의 변수를 포함하고 있다. 데이터에는 SK텔레콤의 전국 단위의 유동 인구 데이터 및 관심사 관련 정보, 신한카드의 결제 정보 및 소비 정보, KCB의 지역별 신용, 소득, 자산, 부채 등 신용과 관련한 정보들이 포함되어 있다. 본 연구는 할부를 예측하기 위해 2022년 전년 말 기준 3개월 내 신용카드 할부 총이용 금액 변수를 기준으로 할부를 사용하면 '1', 할부를 사용하지 않으면 '0'으로 분류하여 분석을 진행하였다.

3.1 데이터 전처리 및 최종 고려 변수

본 연구에서는 738개의 변수로 구성된 방대한 데이터셋을 사용하여 분석을 진행하였다. 모든 변수를 활용할 경우 해석의 명확성이 떨어지고, 과적합의 가능성이 높아질 수 있다. 따라서 이를 방지하고 분석의 효율성을 위해 선행 연구에서 주로 사용된 핵심 변수를 기반으로 분석을 진행하였다. 최종적으로 고려된 변수들은 1장에서 검토한 인구통계학적 변수, 경제적 특성 변수, 소비 및 생활 관련 변수, 신용도 관련 변수, 소비 금액 관련 변수이다. 또한 기존 연구에서 다루지 않았던 관심도 변수를 새롭게 추가하여 신용카드 사용을 촉진할 수 있는 관심사를 분석하고자 하였다. 관심도 변수는 고객이 특정 업종의 관심도를 반영하며, 이를 통해 폭넓은 마케팅 전략을 수립하고자 한다.

분석을 진행하기에 앞서 데이터에 대한 결측치 비중을 확인하여 데이터 전처리 작업을 수행하였다. 결측치의 비중이 95% 이상인 변수는 분석에서 제외하였다. 또한 결측치 비율이 3% 미만인 '소득대비카드소비금액비중' 변수는 중앙값으로 대체하였으며, 결측치의 비율이 70% 미만인 '최초차량구매경과월수'와 '생애최초주택구입경과월수' 변수는 차량과 주택을 구입하지 않은 경우로 가정하고 결측치를 0으로 대체하였다. 또한 '신용카드충한도금액' 변수의 경우 '신용카드건수' 값이 0인 경우에는 결측치를 0으로 처리하였고, 나머지 결측치는 중앙값으로 대체하였다. 또한 명목형 변수는 One-Hot Encoding 방식으로 변환하여 분석을 진행하였다.

본 연구에서는 전체 데이터를 Large Sample과 Small Sample로 구분하여 각각 분석을 진행하였다. Large Sample은 전체 고객 데이터에서 전처리 및 Under Sampling을 적용하여 데이터 불균형 문제를 해결한 후, 최종적으로 33만 건의 데이터 중 약 27만 5천 건의 데이터를 선별하였다. 이렇게 선별된 데이터는 통계 및 앙상블 모델 분석을 위해 Train 데이터와 Test 데이터를 7:3의 비율로 분할하여 활용하였다.

Small Sample은 LLM 적용 및 마케팅 인사이트 도출을 위해 특정 세그먼트에서 600건의 데이터를 추출하여 별도로 구성하였다. Small Sample을 구축하기 위해 특정 세그먼트 선정 및 이를 활용한 분석 과정은 본 논문의 제4장에서 추가로 논의될 예정이다. 데이터 분할 및 활용한 모델에 대한 자세한 정보는 <표 1>에 정리하였다.

<표 1> Sample 크기별 데이터 및 활용 모델

	Train 데이터	Test 데이터	활용 모델
Large Sample	192,616건	82,550건	Logistic Regression
			결정 트리 기반 앙상블 기법
Small Sample	500건	100건	Logistic Regression
			결정 트리 기반 앙상블 기법
			Logistic Regression 수식 기반 LLM
			RAG 기반 LLM

최종적으로 고려한 변수는 <표 2>에 제시된 바와 같이 크게 인구통계학적 변수, 경제적 특성 변수, 소비 및 생활 관련 변수, 신용도 관련 변수, 소비 금액 관련 변수, 관심도 관련 변수, 타겟 변수로 나뉜다.

<표 2> 최종 고려 변수

구분	변수명
인구통계학적 변수	성별
	연령대 (만)
	직업군
	거주지시도
경제적 특성 변수	전년도 추정 연소득
	전년도 순자산평가금액
	추정가구원수_전년말기준
	연매출추정금액
	소득대비카드소비금액비중_전년말기준
소비 및 생활 관련 변수	하이엔드대상자코드
	라이프스태이지
	VIP카드소지여부
	최초차량구매경과월수
	주택보유건수
	자가거주여부
	주택보유건수_가구기준
	카드소비금액
	생애최초주택구입경과월수
사업체건수	

신용도 관련 변수	3개월내신용카드일시불총이용금액(해지포함)_전년말기준
	신용카드건수(미해지)_전년말기준
	신용카드총한도금액(활성카드)(미해지)_전년말기준
	최초카드개설일로부터경과일수(활성카드)(미해지)_전년말기준
	신용카드총단기카드대출(CA)_20년말기준이용잔액유효한도소진율(2+)(활성카드)(해지포함)_전년말기준
	3개월전대출건수(미해지)_전년말기준
소비 금액 관련 변수	연체건수(3개월내발생)(해제포함)_전년말기준
	최근3개월_백화점_이용금액
	최근3개월_스포츠/문화/레저_이용금액
	최근3개월_의료_이용금액
	최근3개월_주유_이용금액
최근3개월_배달앱_이용금액	

구분	변수명
관심도 관련 변수	최근3개월_금융_결제_관심도
	최근3개월_금융_보험_관심도
	최근3개월_금융_암호화폐_관심도
	최근3개월_금융_은행/카드_관심도
	최근3개월_금융_자산관리_관심도
	최근3개월_금융_증권_관심도
	최근3개월_생활_건강/운동/의료_관심도
	최근3개월_생활_배달/식음료_관심도
	최근3개월_생활_부동산_관심도
	최근3개월_생활_이사/인테리어_관심도
	최근3개월_생활_카라이프_관심도
	최근3개월_생활_키즈_관심도
	최근3개월_생활_뷰티_관심도
	최근3개월_쇼핑_백화점/대형마트/면세점_관심도
	최근3개월_쇼핑_소셜커머스_관심도
	최근3개월_쇼핑_오픈마켓_관심도
	최근3개월_쇼핑_패션/뷰티_관심도
	최근3개월_쇼핑_해외직구_관심도
	최근3개월_레저_스포츠_관심도
	최근3개월_레저_여행_관심도
최근3개월_레저_호텔/숙박_관심도	
최근3개월_엔터_sVOD_관심도	
최근3개월_엔터_동영상/방송_관심도	
최근3개월_엔터_라디오_관심도	
최근3개월_엔터_웹툰/애니_관심도	
최근3개월_엔터_음악_관심도	
타겟변수	할부이용여부

3.2 평가 지표

본 연구에서는 예측 모델을 평가하기 위해 오차 행렬(Confusion Matrix)에서 계산할 수 있는 성능 지표를 사용하였다. 오차 행렬은 다음과 같은 형태로 나타낼 수 있다.

<표 3> 오차 행렬(Confusion Matrix)

Confusion Matrix		Predicted Class	
		Negative (0)	Positive (1)
Actual Class	Negative (0)	True Negative (TN)	False Positive (FP)
	Positive (1)	False Negative (FN)	True Positive (TP)

<표 3>에서 TN는 실제 음성인 경우를 음성으로 올바르게 예측한 경우, FN는 실제 양성인 경우를 음성으로 잘못 예측한 경우를 의미한다. FP는 실제로 음성인 경우를 양성으로 잘못 예측한 경우, TP는 실제 양성인 경우를 양성으로 올바르게 예측한 경우이다. 본 연구에서는 예측 모델의 성능을 평가하기 위해 오차 행렬에서 계산할 수 있는 metric을 사용하였다.

3.2.1 정확도, 정밀도, 재현율

정확도(Accuracy)는 전체 데이터에서 모델이 올바르게 예측한 비율이다. 전체적인 예측 성능을 파악할 수 있지만, 데이터가 불균형한 경우에는 모델의 성능이 왜곡될 수 있다. 정확도의 수식은 다음과 같다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

정밀도(Precision)는 양성으로 예측한 샘플 중에서 실제로 양성인 샘플의 비율을 나타낸다. 정밀도는 잘못된 양성 예측이 큰 리스크를 초래하는 경우 중요한 지표로 활용된다. 정밀도의 수식은 다음과 같다.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

재현율(Recall)은 실제 양성인 샘플 중에서 모델이 양성으로 올바르게 예측한 비율

을 의미한다. 재현율은 양성 샘플을 최소화하는 것이 중요한 상황에서 중요하게 활용된다. 재현율의 수식은 다음과 같다.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

3.2.2 F1-Score

F1-Score는 정밀도와 재현율의 조화 평균으로, 두 지표 간의 균형을 평가하는 데 사용된다. 정밀도는 양성으로 예측한 샘플 중에 실제로 양성인 샘플의 비율을 나타내며, 재현율은 실제 양성인 샘플 중에서 모델이 양성으로 올바르게 예측한 비율을 의미한다. F1-Score의 수식은 다음과 같다.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

F1-Score는 정밀도와 재현율의 가중 평균처럼 작동하며, 정밀도와 재현율 모두 높은 경우에는 F1-Score도 높아지지만, 두 지표 중 하나라도 낮으면 F1-Score는 낮아진다. 이는 두 가지의 지표 모두 좋은 성능을 유지해야 함을 의미한다.

3.2.3 Area Under the ROC Curve(AUC)

AUC는 ROC 곡선 아래의 면적을 의미하며, 모델의 양성과 음성을 구분하는 능력을 평가하는 지표이다. ROC 곡선은 FPR(False Positive Rate)과 TPR(True Positive Rate)을 각각 x축과 y축으로 놓은 그래프로, FPR과 TPR의 수식은 다음과 같다.

$$FPR = \frac{FP}{FP+TN} \quad (6)$$

$$TPR = \frac{TP}{TP+FN} \quad (7)$$

FPR이란 위양성률로 음성인 케이스에 대해 양성으로 잘못 예측한 비율로 1-특이도(Specificity)로 계산된다. TPR이란 양성률로 양성인 케이스에 대해 양성으로 올바르게 예측한 비율로 민감도(Sensitivity)와 같은 의미가 있다. ROC 곡선이 좌측 상단에 붙어있을수록 좋은 이진 분류의 성능을 보이며, AUC 역시 1에 가까울수록 좋은

성능을 보인다고 평가된다.

3.3 분석 결과 (Large Sample)

본 연구에서는 27만 5천 건의 대규모 표본(Large Sample)을 활용하여 신용카드 할부 이용 예측 모델의 성능을 분석하였다. 로지스틱 회귀모형의 경우, 단계적 변수 선택 방법(Stepwise Method)을 통해 최적의 예측 변수를 선별하였으며, Random Forest, XGBoost, LightGBM 등 앙상블 모델에서는 전체 변수를 활용하여 예측 성능을 극대화하였다.

분석 결과를 살펴보면, Random Forest 모델이 모든 평가 지표에서 가장 우수한 성능을 보였다. 구체적으로 정확도 0.8682, 정밀도 0.8400, 재현율 0.9116, F1-Score 0.8743, AUC 0.9451을 기록했으며, 특히 할부 이용 고객에 대한 예측 정확도가 높게 나타났다.

반면, Logistic Regression 모델은 정확도 0.7383, 정밀도 0.7348, 재현율 0.7506, F1-score 0.7426, AUC Score 0.8214로 상대적으로 저조한 성능을 보였다. 이는 신용카드 할부 사용에 내재된 복잡한 비선형 관계를 모델이 충분히 포착하지 못한 것으로 해석된다.

XGBoost 모델은 정확도 0.7763, 정밀도 0.7541, 재현율 0.8238, F1-score 0.7874, AUC Score 0.8675를 기록하여 Random Forest에 이어 두 번째로 높은 성능을 보였다. 다만, 상대적으로 낮은 정밀도(0.7541)는 할부 이용 예측의 정확성에 일부 한계가 있음을 시사한다.

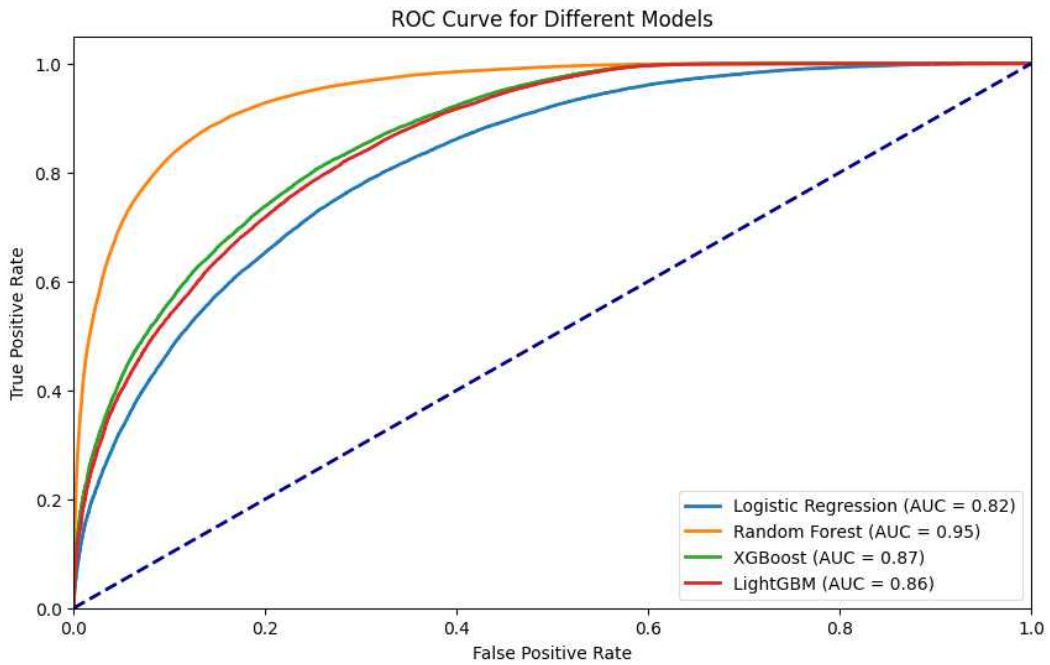
LightGBM 모델의 경우 정확도 0.7688, 정밀도 0.7418, 재현율 0.8288, F1-score 0.7829, AUC Score 0.8595로 XGBoost와 유사한 수준의 성능을 보였다. 그러나 낮은 재현율(0.7080)은 실제 할부 미이용 고객에 대한 오분류 가능성이 있음을 나타낸다.

각 모델의 상세한 성능 지표와 ROC Curve는 <표 4>와 <그림 1>에 제시되어 있다. 전반적으로 앙상블 모델들이 전통적인 Logistic Regression에 비해 우수한 예측 성능을 보였으며, 특히 Random Forest 모델이 가장 안정적이고 높은 예측력을 보여주었다.

<표 4> 모델별 성능 비교(Large Sample)

	Logistic Regression	Random Forest	XGBoost	LightGBM
Accuracy	0.7383	0.8682	0.7763	0.7688
Precision	0.7348	0.8400	0.7541	0.7418
Recall	0.7506	0.9116	0.8238	0.8288
F1-score	0.7426	0.8743	0.7874	0.7829
AUC Score	0.8214	0.9451	0.8675	0.8595

<그림 1> 모델별 ROC Curve (Large Sample)



IV. LLM 모델을 활용한 할부 이용 예측

본 장에서는 신용카드 할부 이용 예측을 위해 Logistic Regression 수식 기반 모델과 RAG 기반 LLM 모델을 활용하였다. Logistic Regression 수식 기반 모델은 LLM 프롬프트에 회귀 수식을 직접 입력하는 방식을 채택하였으며, RAG 기반 LLM 모델은

기존 고객 데이터를 바탕으로 벡터 스토어를 구축하여 새로운 고객의 할부 이용을 예측하고자 하였다. 또한 기존 통계 및 앙상블 모델과의 성능 비교 분석도 함께 진행하였다.

RAG는 대규모 데이터에서 질문과 관련된 소수의 문서만을 선택적으로 검색하는 구조적 특성을 가진다. 이러한 특성을 고려하여 본 연구는 소규모 데이터(Small Sample) 기반의 LLM 적용 방안을 모색하였다. 이는 신용카드사가 새로운 상품이나 마케팅 캠페인 도입 전 특정 고객군 대상의 수요 예측이나 VIP 고객군 맞춤 상품 개발 등에 활용될 수 있다.

타겟 고객군 선정을 위한 탐색적 데이터 분석(EDA) 결과, 40~54세 연령대가 할부 이용 빈도와 평균 사용량에서 가장 높은 수치를 보였으며, 성별로는 여성의 할부 이용량이 상대적으로 높게 나타났다(<표 5> 참조).

그러나 40~54세 여성 고객층은 이미 할부 서비스를 적극적으로 활용하는 그룹으로 추가적인 마케팅 개입 효과가 제한적일 가능성이 있다. 반면, 35~39세 여성 고객층은 연령상 40세 이상 그룹과 연속성이 있으면서도, 아직 할부 이용률이 상대적으로 낮아 분석을 진행하는 데 있어 가장 적합한 집단으로 판단된다. 더불어 직장과 가정을 병행하는 특성상 생활비, 육아, 가전제품 등의 고정비 지출에 할부 서비스를 활용할 가능성이 높은 것으로 판단되었다. 따라서 본 연구에서는 35~39세 여성 중 전년도 추정 연 소득 상위 33% 그룹을 최종 분석 대상으로 선정하였다.

<표 5> 성별, 연령대별 할부 이용 고객 및 평균 금액

SEX	AGE	count	mean	SEX	AGE	count	mean
남성	20	519	540,676	여성	20	1,044	1,104,578
	25	6,035	7,289,803		25	6,274	7,613,014
	30	7,460	11,567,957		30	8,918	13,650,207
	35	8,719	15,726,651		35	10,069	17,785,039
	40	10,749	21,468,288		40	13,965	27,055,069
	45	11,135	21,641,804		45	14,527	28,495,892
	50	12,195	25,718,035		50	14,448	30,126,076
	55	9,803	20,575,910		55	11,788	23,690,231
	60	9,018	18,218,411		60	10,742	19,768,587
65	6,465	11,073,837	65	6,761	11,315,594		
70	6,591	9,178,565	70	5,192	7,408,473		

실증 분석을 위해 타겟 집단에서 무작위로 선정된 500명의 훈련 데이터와 100명의 테스트 데이터를 활용하였다. 모든 모델링 과정에서 할부 이용/미이용 고객 비율을 1:1로 조정하여 데이터 불균형 문제를 해결하였다. LLM 기반 모델의 경우 Huggingface의 meta-llama/Meta-Llama-3-70B-Instruct 모델을 활용하였다.

본 연구의 주요 목적은 두 가지로 요약된다. 첫째, LLM 기반 모델들의 이진 분류

예측에 대한 실질적 적용 가능성을 평가하는 것이다. 둘째, 잠재적 할부 이용 가능성이 높은 고객을 식별하여 신용카드사의 세그먼트별 맞춤형 상품 및 서비스 제공을 위한 기반을 마련하는 것이다. 이를 통해 세그먼트별 예측 정확도를 높이고, 그룹별 특성에 기반한 효과적인 마케팅 전략 수립의 가능성을 제시하고자 한다.

4.1 Logistic Regression 수식 기반 LLM

본 연구에서는 Logistic Regression에서 유의미하게 도출된 변수를 중심으로 활용하였다. 앙상블 모델의 경우에는 트리 기반의 비선형 모델로, 변수의 중요도를 도출할 수 있으나 변수 간 방향성을 제공하지 않는다. 이러한 특성상 LLM 기반의 예측에는 적합하지 않다고 판단하였다. 타겟 고객층 내의 Logistic Regression 식을 도출하기 위해 타겟 고객층 내의 500명을 대상으로 단계적 변수 선택 방법을 실행하였다.

최종 선택된 변수의 회귀계수를 입력 프롬프트에 명시적으로 제시하여 전체 회귀식을 기반으로 신용카드 할부 이용 여부를 예측할 수 있도록 설정하였다. 이는 모든 변수의 영향력을 반영하여, 예측의 정확도를 높이는 데 중점을 둔 방식이다. 입력 프롬프트 템플릿은 <그림 2>에 제시되어 있다.

<그림 2> 입력 프롬프트 형식 (Logistic Regression 수식 기반 모델)

```

""" You are an expert in installment prediction analysis.
Based on the following customer information and the logistic regression model
provided, predict whether the customer is likely to use an installment plan (1: Yes, 0:
No).

Logistic Regression Model:

로지스틱 회귀분석 식 예시
Intercept: -1.0960
3개월전대출건수(미해지)_전년말기준 : 0.2578

answer sample :
Prediction: (1: Yes, 0: No)
Log(odds) Calculate Conclusion :

#Question:
{question}

#Answer: """
    
```

4.2 RAG 기반 LLM

본 연구는 기존 고객 데이터를 활용하여 새로운 고객의 신용카드 할부 이용 여부를 예측하고자 하였다. 이러한 분석을 진행하기 위해 앞서, 활용한 할부 이용 여부 정보를 포함한 500명의 데이터를 벡터 스토어인 ChromaDB에 저장하였다. 또한, 데이터 불균형 문제를 해소하기 위해 할부 이용 고객과 미이용 고객을 1:1 비율로 조정하였다.

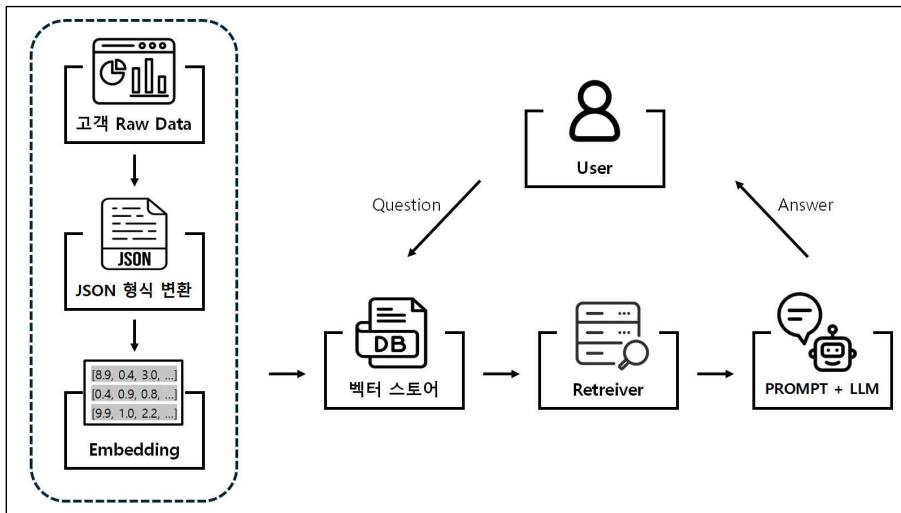
RAG를 진행하기 위해 Pandas DataFrame에 저장된 고객 데이터를 JSON 형식으로 변환하고 이를 Langchain의 Document 객체로 재구성한 뒤, ChromaDB에 저장하는 과정을 수행하였다. 데이터를 JSON 형식으로 변환한 주된 이유는 직관성, 간결성, 자원 효율성 때문이다(Nurseitov et al., 2009). JSON 형식 키-값 쌍으로 구성되어 있어 불필요한 태그 없이 간결한 데이터 표현이 가능하며, 높은 가독성을 제공한다.

이 과정에서 각 행의 데이터를 JSON 형식의 객체로 변환하고 Langchain을 통해

Document 객체로 재구성하여 page_content에 문자열로 저장하였다. 동시에, 고객 식별 코드와 같은 주요 변수는 metadata로 포함하여 문서 간의 구분과 검색 효율성을 높였다. 이후 OpenAI의 기본 임베딩 모델인 text-embedding-ada-002 모델을 사용하여 생성된 데이터를 벡터 형식으로 변환하여 ChromaDB에 저장하였다.

예측 과정에서는 새로운 고객 데이터를 JSON 형식으로 입력받아, 기존 벡터 스토어에서 가장 유사한 고객 데이터를 검색하였다. 검색된 고객 데이터에는 할부 이용 여부 변수가 함께 저장되어 있어, 이 정보를 바탕으로 평균을 계산하여 새로운 고객의 할부 이용 여부를 예측하였다. LLaMA3 모델은 입력 및 출력 토큰 수의 제한(8,192 토큰)으로 인해 k=5로 설정하여, 상위 5개의 정보만을 검색하도록 설정하였다. RAG 기반 LLM 모델 구축 과정 및 분석 과정은 <그림 3>과 같이 진행되며, 새로운 고객 데이터와 기존 고객 데이터를 연계하여 LLM 기반의 예측 실효성을 입증하고자 한다.

<그림 3> RAG 기반 LLM 모델 분석 과정 도식화



RAG를 통해 검색된 정보 중 3명 이상이 할부를 이용하는 경우 1(할부 이용)로, 2명 미만인 경우 0(할부 미이용)으로 예측할 수 있도록 프롬프트를 구성하였다. 프롬프트의 {context}에는 벡터 스토어에서 검색된 데이터가 입력되며, {question}에는 새롭게 예측하고자 하는 고객의 데이터가 입력된다. 분석을 진행하기 위해 입력한 프롬프트는 <그림4>에 제시하였다.

<그림 4> 입력 프롬프트 형식 (RAG 기반 모델)

```

"""You are an expert in installment prediction analysis.
Based on the following customer information, predict whether the customer is likely to
use an installment plan (1: Yes, 0: No).

Instructions:
Use the variable INS_STATUS to determine past installment usage. If INS_STATUS
= 1, the customer used an installment plan.
Calculate Probability as count(INS_STATUS = 1) / 5.

Provide a concise explanation with the calculated Probability and the predicted
INS_STATUS.

answer sample :
INS_STATUS :
Probability :

#Context:
{context}

#Question:
{question}

#Answer:"""

```

4.3 분석 결과 (Small Sample)

Small Sample 데이터 분석 결과, RAG 기반 LLM 모델이 F1-Score 0.7048과 재현율 0.7400으로 가장 균형 잡힌 예측 성능을 보여주었다. 다만, 정확도와 정밀도 측면에서는 LightGBM이 더 우수한 성능을 나타내어, 성능 지표의 우선순위에 따라 최적 모델의 선택이 달라질 수 있음을 시사한다.

기존 모델들의 경우 Large Sample 대비 전반적인 성능 저하가 관찰되었다. 전통적인 로지스틱 회귀모델은 Small Sample에서도 앙상블 모델들에 비해 상대적으로 낮은 성능을 보였다. 앙상블 모델 중에서는 Random Forest와 LightGBM이 각각 정확도 0.6900, 0.7000과 F1-Score 0.6869, 0.6739로 비교적 우수한 성능을 기록했다.

Logistic Regression 수식을 프롬프트로 입력한 LLM 모델의 경우, 기존 Logistic Regression 모델보다 오히려 낮은 성능을 보였다. 특히 재현율은 0.7200으로 높았으

나, 정밀도가 0.5070으로 현저히 낮아 할부 미이용 고객에 대한 오분류가 많았다. 이는 모델이 할부 이용 가능성을 과대 추정하는 경향이 있음을 시사한다.

이러한 분석 결과들은(<표 6> 참조) Small Sample 환경에서 RAG 기반 LLM 모델이 가장 실용적이고 효과적인 예측 도구에 대한 가능성을 보여준다. 특히 소규모 데이터셋을 활용한 예측 시나리오에서 RAG 기반 LLM의 활용이 더욱 유용할 것으로 판단된다.

<표 6> 모델별 성능 비교(Small Sample)

	기존 활용 모델				LLM 활용	
	Logistic Regression	Random Forest	XGBoost	LightGBM	PROMPT	RAG
Accuracy	0.5700	0.6900	0.6200	0.7000	0.5100	0.6900
Precision	0.5660	0.6939	0.7000	0.7381	0.5070	0.6727
Recall	0.6000	0.6800	0.4200	0.6200	0.7200	0.7400
F1-score	0.5825	0.6869	0.5250	0.6739	0.5950	0.7048

4.4 RAG 기반 LLM을 활용한 마케팅 인사이트 도출

본 장에서는 RAG 기반 LLM을 활용한 고객 할부 이용 예측을 수행하고, 이를 바탕으로 추가적인 인사이트 도출 과정을 진행하였다. 본 연구에서는 단순히 입력된 고객의 특성을 파악하는 것이 아닌, 입력된 고객과 유사한 상위 5명의 고객을 함께 분석하는 접근법을 적용하였다. 이를 통해 개별 고객의 특성을 더욱 직관적으로 해석할 수 있으며, 고객 맞춤형 전략을 수립하는 데 활용할 수 있다.

특히 LLM이 고객 데이터를 더욱 정확하게 해석할 수 있도록 입력 프롬프트에는 직업군, 라이프스타일 등 명목형 변수에 대한 설명을 추가하였다. 이를 통해 LLM이 변수의 의미를 올바르게 해석하고, 고객들의 특성을 적절하게 파악할 수 있도록 설정하였다.

분석 과정에서는 입력된 고객과 유사성이 높은 상위 5명의 고객을 검색한 후, 이들과의 공통된 특성을 요약할 수 있도록 입력 프롬프트를 구성하였다. 이 접근 방식은 단순히 개별 고객의 데이터를 해석하는 것보다, 유사 고객과의 비교를 통해 더욱 신뢰성 있는 인사이트를 도출할 수 있도록 돕는다. 예를 들어, 특정 고객이 백화점/대형마트/면세점에 높은 관심이 있을 때, 유사한 고객들도 동일한 관심도 패턴을 보이는지 확인함으로써 소비 성향의 일관성을 파악할 수 있다. 이러한 분석 결과는 마케팅 전

략 수립에 있어 더욱 실질적인 의사결정을 지원할 수 있다. 입력 프롬프트는 <그림 5>에 제시하였다.

<그림 5> 입력 프롬프트 형식 (RAG 기반 모델-인사이트)

```

"""You are an expert in extracting insights from data.
I will provide the information of one customer and five other customers who are
similar to this customer.
Your task is to analyze these customers and identify their common characteristics.

Review the data provided for the main customer and the five similar customers.
Identify and explain at least three key common traits or patterns shared by these
customers.
Provide actionable insights or recommendations based on your analysis.

직업군, 라이프스타일 등 명목형 변수에 대한 설명

answer sample :
- Simillar CUST_ID :
- Analysis Result :
1. JB_TP :
2. LIF_STG :
3. HOME_ADM :

#Context:
{context}

#Question:
{question}

#Answer:"""
    
```

분석 결과 <그림 6>과 같이 입력된 프롬프트 기반으로 LLM이 응답을 생성하였으며, 민감한 정보 보호를 위해 고객 식별 ID는 제외하여 결과를 제시하였다. 동시에 고객 식별 ID를 포함한 결과를 도출하여 실제 데이터와 분석 결과가 적합하게 진행되었는지 검증하였다. 분석 결과의 쉬운 이해를 돕기 위해 영어를 한글로 번역하여 <그림 6>에 제시하였다.

<그림 6> 답변 예시 (RAG 기반 모델-인사이트)

Similar CUST_ID:

유사한 고객의 ID

분석 결과:

1. JB_TP: 모든 고객의 직업 유형은 420으로, 일반적인 급여생활자를 나타냅니다.
2. LIF_STG: 대부분의 고객은 생활의 2단계 또는 3단계에 속해 있으며, 이는 유아자녀가 있는 가족 또는 청소년 자녀가 있는 가족을 포함합니다.
3. HOME_ADM: 모든 고객은 41 또는 42의 HOME_ADM 값을 가지고 있습니다.
4. OWN_LIV_YN: 모든 고객은 3의 값을 가지고 있으며, 이는 비자가임을 나타냅니다.
5. R3M_ITRT_*: R3M_ITRT_*에서 가장 빈번한 값은 1로, 이러한 고객들은 다양한 금융, 생활, 여가 및 엔터테인먼트 서비스에 관심이 있는 것으로 나타났습니다.

실천 가능한 인사이트 또는 권장 사항:

이러한 공통된 특성을 바탕으로, 이들 고객은 유아 자녀를 둔 중산층 가족일 가능성이 높습니다. 이들은 생활에 필요한 금융 상품과 서비스에 관심이 있을 수 있습니다. 예를 들어, 신용카드, 개인 대출, 투자 기회 등이 있습니다. 또한, 가족 및 생활과 관련된 보험, 교육, 엔터테인먼트 등의 상품과 서비스에 관심이 있을 수 있습니다. 마케팅 전략은 이러한 인구 통계를 대상으로 디지털 채널, 예를 들어 소셜 미디어와 온라인 광고를 통해 초점을 맞출 수 있습니다.

분석 결과 고객 데이터에서 도출된 주요 특성이 실제 고객의 관심사와 일치함을 확인하였다. 이는 RAG 기반 LLM이 단순한 데이터 분석을 넘어, 고객이 관심을 가지는 요소를 더욱 정교하게 추출하고 해석하는 데 활용될 수 있음을 의미한다. 이를 기반으로 개인화된 마케팅 전략을 수립하거나, 새로운 고객을 기존 데이터에 기반한 세그먼트에 배치하는 데 활용할 수 있을 것으로 기대된다.

특히, 신규 고객의 경우 고객 특성에 맞춘 세분화 과정을 자동화하거나 확장할 수 있는 기초 자료로 활용할 수 있을 것으로 기대된다. 또한, 추후에는 관심이 있는 변수만 선별하여 새로운 벡터 스토어를 구성해 분석을 진행된다면, 토큰 제한의 영향을 최소화하면서 더욱 많은 고객 데이터를 기반으로 분석을 진행할 수 있을 것으로 예상된다.

V. 결 론

5.1 연구 결과 요약 및 시사점

본 연구는 기존 연구에서 상대적으로 주목받지 못했던 신용카드 할부 사용 예측에 초점을 맞추어 실증적 분석을 수행하였다. 특히 전통적인 통계 및 앙상블 모델과 함께 RAG 기반 LLM 예측 모델을 새롭게 도입하여 혁신적인 분석 프레임워크를 제시하였다. 또한 벡터 스토어에 저장된 기존 고객 정보를 RAG 기반 LLM과 연계하여 신규 고객과의 공통 특성을 파악하고 개인화된 마케팅 인사이트를 도출하고자 하였다.

Large Sample 분석에서는 전통적인 접근 방식인 로지스틱 회귀, Random Forest, XGBoost, LightGBM 모델을 활용하였다. 분석 결과, Random Forest 모델이 모든 평가 지표에서 최고 성능을 보였다. 이는 Random Forest가 복잡한 데이터 구조와 변수 간 상호작용을 효과적으로 포착하여 신용카드 할부 이용 예측에 특히 적합함을 시사한다.

Small Sample 분석에서는 기존 모델들과 함께 Logistic Regression 수식 기반 LLM 및 RAG 기반 LLM 모델을 적용하였다. 분석 결과, RAG 기반 LLM 모델이 가장 균형 잡힌 예측 성능을 보여주었다. Logistic Regression 기반 LLM의 경우, 적은 수의 변수에서는 정확한 예측이 가능했으나, 변수가 많아질수록 환각 현상이 발생하여 복잡한 데이터 구조에서의 한계를 드러냈다.

본 연구의 주요 발견은 Small Sample 환경에서 RAG 기반 LLM 모델이 전통적 통계 모델과 대등한 예측 성능을 보이면서도, 추가적인 마케팅 전략 인사이트를 제공할 수 있다는 점이다. 이는 단순한 예측을 넘어 개인화된 마케팅 전략 수립과 고객 세분화를 가능하게 한다. 신용카드사가 고객 특성을 더욱 정확하게 이해하고 이를 바탕으로 전략적 의사결정을 수행할 수 있는 기회를 제공하며, 궁극적으로 실질적인 비즈니스 가치 창출로 이어질 수 있음을 시사한다.

5.2 향후 연구 방향

본 연구의 결과를 바탕으로 RAG 기반 LLM 모델의 성능 향상을 위한 다섯 가지 향후 연구 방향을 제시하고자 한다.

첫째, 대규모 데이터 처리 역량 강화가 필요하다. LLM 모델의 입력 토큰 제한으로 인한 대용량 데이터 처리의 한계를 극복하기 위해, 핵심 변수 선별을 통한 효율적인

벡터 스토어 구성이나 토큰 제한이 적은 LLM 모델의 활용을 고려할 수 있다. 이를 통해 더욱 정교하고 세밀한 분석이 가능할 것으로 기대된다.

둘째, LLM 모델의 예측 편향 최소화가 요구된다. 현재 모델은 높은 재현율에 비해 상대적으로 낮은 정밀도를 보여 할부 미이용 고객에 대한 높은 오분류 가능성이 존재한다. RAG 프로세스에서의 검색 개수(k 값) 최적화나 새로운 데이터 처리 방식의 도입을 통해 예측 정확성을 제고할 수 있을 것이다.

셋째, 신용카드사의 대규모 데이터를 활용한 파인튜닝(Fine-tuning)을 통해 LLM 모델의 최적화가 가능하다. 이를 RAG와 결합하면 예측 성능과 정확도가 향상될 뿐만 아니라, 더욱 정교한 마케팅 전략 수립이 가능할 것으로 전망된다.

넷째, 결측치 처리 방식의 다양화가 필요하다. 본 연구에서는 95% 이상의 결측치를 포함한 변수를 제외하였으나, 범주형 변환 등 다양한 결측치 처리 방식을 도입함으로써 더욱 풍부한 데이터 기반의 심층 분석이 가능할 것이다.

다섯째, 본 연구를 통하여 AI Agent를 활용한 "RAG - 분석 - 전략 도출" 프로세스의 완전 자동화에 대한 가능성이 확인되었다. 이러한 자동화 시스템이 구현되면 기업들은 고객 데이터의 실시간 분석과 이에 기반한 신속하고 정확한 전략적 의사결정이 가능해질 것이다. 이는 궁극적으로 업무 효율성 향상과 고객 만족도 제고에 기여할 것으로 기대된다.

참고문헌

- 가명정보 소개, 2024년 11월 18일, 가명정보지원플랫폼
금융위원회 (2024), *신용카드업 상생·발전을 위한 제도개선 방안*, 금융위원회.
- 김보영, 박미랑 (2015), 신용카드 여가서비스 관계해택이 고객 서비스 만족 및 유지
의도에 미치는 영향에 관한 연구. *서비스경영학회지*, 16(1), 126-147.
- 박만석, 김동준 (1995), 신용카드 이용자의 라이프 스타일과 구매행동 연구, *광고연
구*, 가을호, 231-260.
- 신용카드 소비 분석 데이터, 2024년 10월 21일, AI-Hub
- 연도별 신용카드 이용실적, 2024년 11월 18일, 여신금융협회
- 이동규, 신민수 (2023), 카드산업에서 휴면 고객 예측, *서비스 연구*, 13(2), 99-113.
- 이유재, 이신형, 이종세 (2014), KB국민카드의 마케팅 활동과 빅데이터 활용, *Korea
Business Review*, 18(1), 145-176.
- 이지훈, 홍성태, 강동균 (2014), 현대카드의 마케팅 성공요인: 비즈니스 모델, 상품개
발 그리고 BTL 마케팅, *Korea Business Review*, 18(3), 147-170.
- 이찬경, 노형봉 (2018), 신용카드 부가서비스 요율 분석 방법론에 대한 연구, *품질경
영학회지*, 46(4), 797-819.
- 장동호 (2012), 가계 신용카드 채무의 결정요인 분석: 가구특성을 중심으로, *사회복
지정책*, 39(1), 227-255.
- 정원석, 김도형, 엄영익 (2024), A Study on Sales Prediction Model Based on
BiLSTM-GAT Using Credit Card Transaction Data, *Journal of KIISE*, 51(9),
807-816.
- 2024년 11월 15일, 금융감독원 금융통계시스템
- Arram, A., Ayob, M., Albadr, M. A. A., Sulaiman, A., & Albashish, D. (2023),
Credit card score prediction using machine learning models: A new dataset,
arXiv preprint arXiv:2310.02956.
- Breiman, L. (1996), Bagging predictors. *Machine learning*, 24, 123-140.
- Breiman, L. (2001), Random forests, *Machine Learning*, 45(1), 5-32.
- Capon, N. (1982), Credit scoring systems: A critical analysis, *Journal of
Marketing*, 46(2), 82-91.
- Chen, T., & Guestrin, C. (2016), Xgboost: A scalable tree boosting system,

- Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024), The llama 3 herd of models, *arXiv preprint arXiv:2407.21783*.
- Freund, Y., & Schapire, R. E. (1997), A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences*, 55(1), 119-139.
- Iacovides, G., Konstantinidis, T., Xu, M., & Mandic, D. (2024), FinLlama: LLM-Based Financial Sentiment Analysis for Algorithmic Trading, *Proceedings of the 5th ACM International Conference on AI in Finance*, 134–141.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2017), Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, 3146–3154.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020), Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Mai, Z., Zhang, J., Xu, Z., & Xiao, Z. (2024), Financial sentiment analysis meets llama 3: A comprehensive analysis, *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, 171–175.
- Nurseitov, N., Paulson, M., Reynolds, R., & Izurieta, C. (2009), Comparison of JSON and XML data interchange formats: a case study, *Caine*, 9, 157–162.
- Srivastava, V. (2024), Lending an Ear: How LLMs Hear Your Banking Intentions, *Proceedings of the 5th ACM International Conference on AI in Finance*, 301–309
- Topsakal, O., & Akinci, T. C. (2023), Creating large language model applications utilizing langchain: A primer on developing llm apps fast, *International Conference on Applied Engineering and Natural Sciences*, 1(1), 1050–1056.

*** 저자소개 ***

· **한 선 의(seonuihn@sogang.ac.kr)**

서강대학교 일반대학원에서 경영학과 Business Analytics 전공으로 석사학위를 취득하였다. 주요 관심 분야는 생성형 AI, 금융 서비스, 서비스 경영 등이다.

· **이 군 희(ghlee@sogang.ac.kr)**

University of Missouri에서 통계학 박사 학위를 취득하였으며, 현재 서강대학교 경영학과 교수로 재직 중이다. 주요 관심 분야는 계량 분석, 경영과학, 서비스 사이언스, 금융 서비스, 인공지능 등이다.