데이터베이스(002) 기말 프로젝트 보고서

은행 정기예금 마케팅 데이터 분석을 통한 마케팅 대상 추천 시스템 구축

데이터사이언스학과 23011816 황유리

2025.06.22

목차

1. 프로젝트 개요

A. 주제 선정 이유

최근 돈 관리에 관심이 생겨, 여러 은행의 각종 마케팅과 관련 자료들을 접했습니다. 은행은 고객을 대상으로 다양한 금융상품 마케팅을 진행하지만, 실제로 정기예금은 이자율 감소 등으로 인해 신규 가입률이 낮은 편입니다. 이러한 상황에서 불특정 다수에게 일괄적으로 마케팅을 실시할 경우, 불필요한 비용이 발생하고 효율성도 떨어집니다. 이에 따라 데이터 기반으로 고객 특성을 분석하고, 가입 가능성이 높은 대상을 선별해 마케팅 효율을 높일 수 있는 시스템의 필요성을 느꼈습니다. 또한 점점 데이터사이언스 분야에서 고객 행동 예측과 타겟 마케팅이 중요한 주제로 부상하고 있고, 본 데이터베이스 수업에서도 패밀리 레스토랑 실습과 같이 고객 특징에 대해 분석하는 활동을 많이 배웠으므로 실무적으로도 매우 의미 있는 프로젝트라고 판단했습니다.

B. 프로젝트 목적

본 프로젝트의 목적은 은행 정기예금 마케팅 데이터를 체계적으로 분석하고, SQL 기반으로 정규화 된 데이터베이스를 구축하여 효율적으로 데이터를 관리하는 것입니다. 나아가, 분석 결과를 바탕으로 실제 마케팅 대상 추천 프로그램을 구현하는 것을 목표로 삼았습니다. 이를 통해 데이터 기반 의사결정의 실제 과정을 경험하고, 실질적인 마케팅 전략 수립에 도움이 되는 인사이트를 도출하고자 했습니다.

C. 활용 데이터

분석에 사용한 데이터는 UCI Machine Learning Repository에서 제공하는 Bank Marketing 데이터 셋[1]입니다. 이 데이터는 약 4만 5천 명의 은행 고객 정보를 포함하고 있으며, 고객의 나이, 직업, 결혼여부, 잔액, 대출여부, 연락 방식, 캠페인 횟수, 이전 캠페인 결과, 정기예금 가입여부 등 다양한 변수를 담고 있습니다. 데이터는 CSV 파일 형태로 제공되며, 범주형, 수치형, 이진형 변수가 혼합되어 있습니다.

2. 데이터베이스 설계

A. 원본 CSV

원본 CSV 파일은 총 17개의 컬럼으로 구성되어 있습니다. 주요 변수로는 고객의 인구통계학적 정보(나이, 직업, 결혼 여부, 교육 수준 등), 금융 정보(잔액, 대출여부 등), 마케팅 캠페인 이력(연락 방식, 연락 월, 캠페인 횟수, 이전 결과 등), 그리고 최종적으로 정기예금 가입여부가 포함되어 있습니다.

아래는 17개의 피처에 대한 설명입니다:

피처명	설명	데이터 타입	예시
age	고객 나이	정수	30
job	직업	문자열(범주형)	'admin.'
marital	결혼 여부	문자열(범주형)	'married'
education	교육 수준	문자열(범주형)	'secondary'
default	신용불량 여부	'yes' / 'no'	'no'
balance	은행 잔액	정수	1500
housing	주택 대출 여부	'yes' / 'no'	'yes'
loan	개인 대출 여부	'yes' / 'no'	'no'
contact	연락 수단	문자열(범주형)	'cellular'
day	마지막 연락 일	정수	5
month	마지막 연락 월	문자열	'may'
duration	통화 시간(초)	정수	120
campaign	캠페인 동안 연락 횟수	정수	2
pdays	이전 캠페인 이후 경과 일수	정수	999
previous	과거 캠페인 접촉 횟수	정수	0
poutcome	이전 캠페인 결과	문자열(범주형)	'failure'
y	정기예금 가입 여부	'yes' / 'no'	'no'

B. 정규화

효율적인 데이터 관리와 중복 최소화를 위해 데이터 정규화를 적용했습니다.

i. 1차 정규화(1NF): 반복 속성 제거

1차 정규화(1NF) 과정에서는 각 셀이 단일 값만을 갖도록 구조를 개선했

습니다. 원본 데이터에서 직접적으로 반복되는 속성은 없었지만, job이나 education과 같은 범주형 속성들이 문자열 형태로 저장되어 있었습니다. 이를 해결하기 위해 각 범주형 속성을 별도의 코드 테이블로 분리했습니다. 예를 들어, job 컬럼은 Job 테이블을 만들어 job_id와 job_name으로 구성하여 관리하도록 했습니다.

ii. 2차 정규화(2NF): 부분 종속성 제거

2차 정규화(2NF) 단계에서는 부분 종속성 문제를 해결했습니다. 초기에는 모든 정보가 하나의 테이블에 포함되어 있어, 고객의 기본 정보(나이, 직업 등)가 캠페인마다 중복되어 저장되는 문제가 있었습니다. 이를 해결하기 위해 고객 정보를 담는 Customer 테이블과 캠페인 기록을 담는 Campaign 테이블로 분리했습니다. 이렇게 함으로써 한 고객이 여러 번의 마케팅 캠페인에 참여한 경우에도 고객 정보의 중복 없이 효율적으로 관리할 수 있게 되었습니다.

iii. 3차 정규화(3NF): 이행적 종속성 제거

3차 정규화(3NF) 과정에서는 이행적 종속성을 제거했습니다. education, contact, poutcome과 같은 속성들은 고객이나 캠페인에 종속되지만, 그 자체로도 의미 있는 독립적인 속성이므로 각각을 별도의 코드 테이블로 분리했습니다. 예를 들어, Education 테이블을 만들어 education_id와 education_name으로 구성하고, 이를 외래키로 참조하도록 설계했습니다. contact와 poutcome도 동일한 방식으로 처리하여, 코드 관리의 일관성을 확보하고 향후 새로운 범주가 추가될 때의 확장성도 고려했습니다.

iv. 주요 제약조건

데이터베이스 설계 과정에서 데이터 무결성과 일관성을 확보하기 위해 여러 제약조건을 설정했습니다.

먼저 각 테이블의 기본키는 AUTO_INCREMENT 속성을 부여하여 자동으로 증가하는 정수값을 사용하도록 했습니다. 이를 통해 새로운 데이터가 삽입될 때마다 고유한 식별자가 자동으로 생성되어, 중복이나 식별자 관리 문제를 방지할 수 있었습니다. 예를 들어, Job 테이블의 job_id, Customer 테이블의 customer_id 등이 이에 해당합니다.

참조 무결성을 위해서는 외래키 제약조건을 활용했습니다. Customer 테이블과 Campaign 테이블 간에는 customer_id를 통해 연결하고, 각 범주형 속성 테이블들(Job, Education, MaritalStatus, Contact, Poutcome)과는 해당하는 ID 컬럼(job_id, education_id 등)을 외래키로 설정했습니다. 이

렇게 함으로써 존재하지 않는 코드값이 입력되는 것을 방지하고, 데이터 간의 일관성을 유지할 수 있었습니다.
또한 각 코드 테이블의 name 컬럼에는 UNIQUE 제약조건을 설정하여 동일한 값이 중복으로 저장되는 것을 방지했습니다. 이를 통해 'admin.', 'management' 등의 직업 코드가 중복되어 저장되지 않도록 관리할 수 있었습니다.

v. 데이터 삽입 방법

실제 데이터 삽입 과정에서는 pandas와 pymysql 라이브러리를 활용하여 효율적으로 데이터를 전처리하고 데이터베이스에 저장했습니다.
먼저 UCI 데이터셋을 pandas DataFrame으로 불러온 후, 각 범주형 컬럼의 고유값들을 추출하여 해당하는 코드 테이블에 삽입했습니다. 이 과정에서 INSERT IGNORE 구문을 사용하여 동일한 값이 중복으로 삽입되는 것을 방지했습니다.
결측값 처리에 있어서는 원본 데이터에서 빈 값이나 null 값이 발견되는 경우, 일관되게 'unknown'으로 대체하여 처리했습니다. 이를 통해 데이터 분석 과정에서 결측값으로 인한 오류를 방지하고, 향후 분석에서도 일관된 기준을 적용할 수 있었습니다.
데이터 삽입 순서는 참조 무결성을 고려하여 부모 테이블부터 자식 테이블 순으로 진행했습니다. 먼저 각 코드 테이블(Job, Education, MaritalStatus 등)에 데이터를 삽입한 후, Customer 테이블에 고객 정보를 저장하고, 마지막으로 Campaign 테이블에 캠페인 이력을 저장하는 방식으로 구성했습니다. 이 과정에서 pandas의 to_sql() 메서드를 활용하여 대량의 데이터를 한 번에 효율적으로 삽입할 수 있었습니다.

C. 최종 테이블 구조

최종적으로 데이터베이스는 8개의 테이블(원본 테이블 1개, 정규화 테이블 7개)로 구성되었으며, 각 테이블 간 외래키 제약조건을 통해 데이터 무결성을 확보했습니다. 이를 통해 SQL 쿼리로 다양한 분석을 손쉽게 수행할 수 있게 되었습니다.

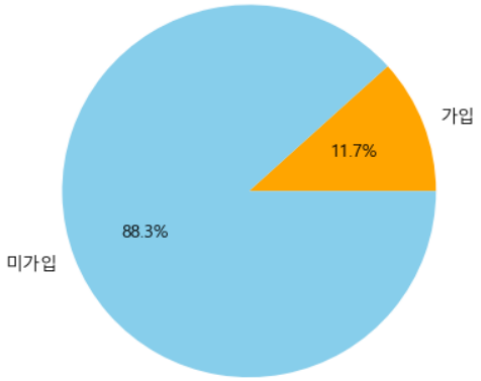
3. SQL 분석 및 시각화

다양한 SQL 구문과 pandas의 시각화 기능을 활용하여 데이터 분석을 진행하였습니다. 본 보고서에는 결과물 이미지만 첨부하였으며, 코드는 별도로 제출하는 파이썬 파일을 통해 확인할 수 있습니다.

A. 전체 고객 가입률

	전체고객수	가입자수	전체가입률
0	45211	5289.0	11.7

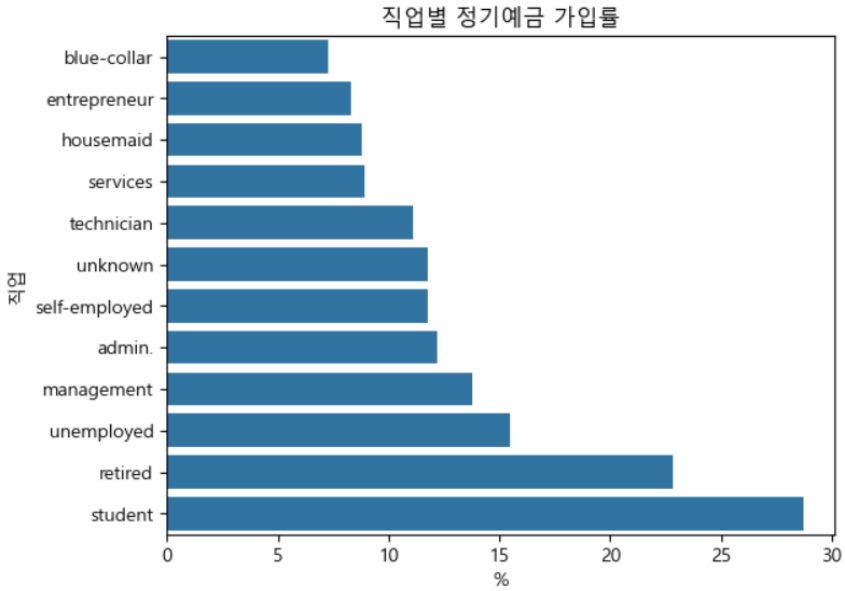
정기예금 가입률 비율



SQL 쿼리를 이용해 전체 고객 중 정기예금에 실제로 가입한 비율을 분석했습니다. 특히 Campaign 테이블은 1:N 관계성을 가지므로, 서브쿼리를 이용하여 분석을 진행하였습니다. 분석 결과 전체 고객 수는 약 45,211명이며, 이 중 5,289명이 정기예금에 가입해 가입률은 약 11.7%로 나타났습니다. 이는 실제 마케팅의 성공률이 매우 낮다는 점을 보여줍니다.

B. 직업별 가입률

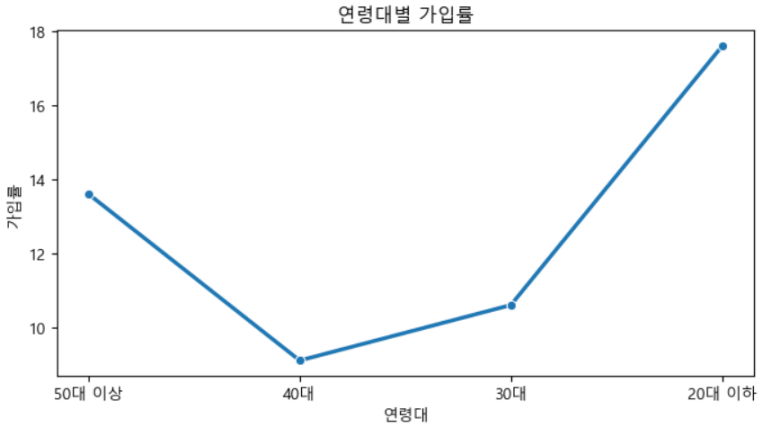
	직업	전체	가입수	가입률
0	student	4690	1345.0	28.7
1	retired	11320	2580.0	22.8
2	unemployed	6515	1010.0	15.5
3	management	47290	6505.0	13.8
4	admin.	25855	3155.0	12.2
5	unknown	1440	170.0	11.8
6	self-employed	7895	935.0	11.8
7	technician	37985	4200.0	11.1
8	services	20770	1845.0	8.9
9	housemaid	6200	545.0	8.8
10	entrepreneur	7435	615.0	8.3
11	blue-collar	48660	3540.0	7.3



직업별로 가입률을 분석한 결과, 학생(28.7%), 은퇴자(22.8%), 실업자(15.5%), 관리자(13.8%) 순으로 가입률이 높았습니다. 반면, 블루칼라(7.3%), 자영업(8.3%), 서비스직(8.9%) 등은 가입률이 낮은 편이었습니다.
이러한 결과는 특정 직업군이 정기예금 상품에 더 관심이 많거나, 마케팅에 더 잘 반응한다는 것을 시사합니다.

C. 연령대별 가입률

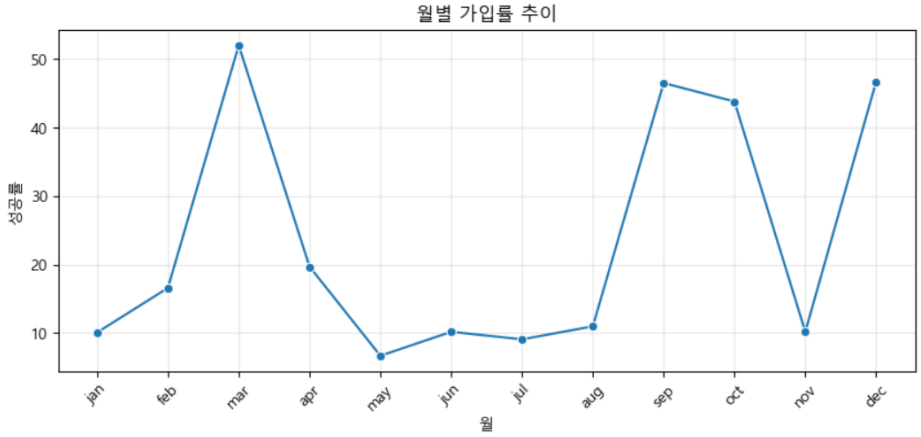
	연령대	고객수	가입자	가입률
0	50대 이상	50970	6925.0	13.6
1	40대	58275	5315.0	9.1
2	30대	90445	9565.0	10.6
3	20대 이하	26365	4640.0	17.6



연령대별로는 50대 이상(13.6%)과 20대 이하(17.6%)에서 가입률이 높게 나타났습니다.

D. 월별 캠페인 성공률 추이

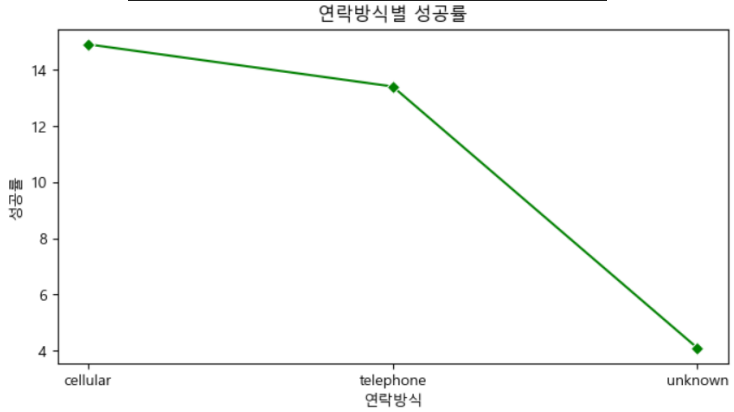
	월	연락수	가입수	성공률
0	jan	7015	710.0	10.1
1	feb	13245	2205.0	16.6
2	mar	2385	1240.0	52.0
3	apr	14660	2885.0	19.7
4	may	68830	4625.0	6.7
5	jun	26705	2730.0	10.2
6	jul	34475	3135.0	9.1
7	aug	31235	3440.0	11.0
8	sep	2895	1345.0	46.5
9	oct	3690	1615.0	43.8
10	nov	19850	2015.0	10.2
11	dec	1070	500.0	46.7



월별로 마케팅 캠페인 성공률을 분석한 결과, 3월(52.0%), 9월(46.5%), 10월(43.8%), 12월(46.7%) 등 특정 시기에 가입률이 급상승하는 경향이 있었습니다. 다만, 연락 건수가 적은 달의 성공률이 높다는 점도 확인되어, 단순히 성공률만으로 마케팅 전략을 세우기보다는 실제 연락 건수와의 균형도 고려해야 함을 알 수 있었습니다.

E. 통화 수단별 가입률

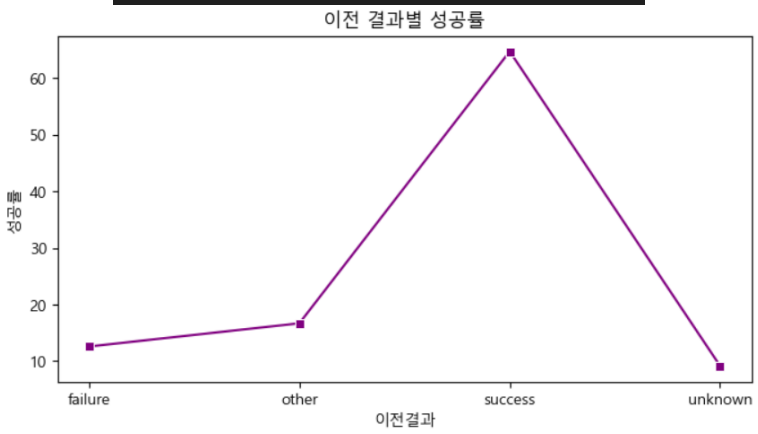
연락방식	전체	가입자	성공률
cellular	146425	21845.0	14.9
elephone	14530	1950.0	13.4
unknown	65100	2650.0	4.1



연락 방식별로는 휴대전화(cellular)를 통한 연락 시 가입률이 14.9%로 가장 높았고, 전화(telephone) 13.4%, 미확인(unknown) 4.1% 순이었습니다. 이는 휴대전화가 가장 효과적인 마케팅 채널임을 보여줍니다.

F. 이전 캠페인 영향 분석

	이전결과	전체	가입자	성공률
0	failure	24505	3090.0	12.6
1	other	9200	1535.0	16.7
2	success	7555	4890.0	64.7
3	unknown	184795	16930.0	9.2



이전 캠페인에 성공한 경험이 있는 고객의 경우, 재가입률이 64.7%에 달해 매우 높았습니다. 반면, 이전 캠페인에서 실패하거나 결과가 미확인인 경우에는 가입률이 각각 12.6%, 9.2%로 낮았습니다. 이는 과거 마케팅 경험이 향후 행동에 큰 영향을 미친다는 점을 의미합니다.

G. 종합 결론 및 타겟 전략 제안

분석 결과, 연령, 직업, 잔액, 이전 캠페인 결과, 연락 방식 등 여러 특성이 정기예금 가입 가능성에 영향을 미치는 것을 확인할 수 있었습니다. 예를 들어, 50세 이상이면서 관리자/은퇴자 직군, 11~12월에 연락을 받고, 이전 캠페인에서 성공 경험이 있는 고객은 가입 가능성이 매우 높았습니다.

따라서, 이러한 특성을 조합해 마케팅 자원을 집중할 타겟군을 선별하는 전략이 효과적임을 제안합니다.

4. 프로그램 구현

본 프로젝트에서는 분석 결과를 바탕으로 마케팅 담당자가 쉽게 활용할 수 있는 추천 프로그램을 개발하였습니다. Python의 Streamlit 라이브러리를 활용하여 웹 기반 인터페이스를 구현하였으며, 사용자는 고객의 나이, 직업, 연락 월, 연락 방식, 이전 캠페인 결과를 입력하면 가입 가능성을 안내받을 수 있습니다.

아래는 '3. SQL 분석 및 시각화'에서 진행한 실제 분석 결과를 반영한 점수화 규칙입니다.

- **직업**:
  학생/은퇴자: +3점
  실업자/관리자/경영직: +2점
  자영업/미확인/기술직: +1점
- **연령대**:
  20대 이하 또는 50대 이상: +2점
  30~39세: +1점
- **연락 월**:
  3월, 9월, 10월, 12월: +2점
- **연락 방식**:
  휴대전화(cellular): +2점
  전화(telephone): +1점
- **이전 캠페인 결과**:
  성공: +4점 (64.7% 재가입률)
  실패: -1점

이러한 규칙으로 계산된 점수에 따라, 아래와 같이 가입 가능성이 분류됩니다.

점수 범위	추천 등급	설명
8점 이상	적극 추천	가입 가능성 매우 높음
5~7점	추천	가입 가능성 높음
3~4점	보통	신중히 고려
2점 이하	비추전	가입 가능성 낮음

이 프로그램을 통해 마케팅 담당자는 실제 데이터 분석 결과와 일치하는 점수 체계를 통해, 5개 핵심 요소를 기반으로 신속한 의사결정이 가능할 것입니다. 특히 1분 이내로 분석이 가능하기 때문에, 업무 효율성이 매우 향상될 것입니다.

아래는 실제 프로그램을 실행하여 사용해본 예시입니다.

5. 마무리

이번 프로젝트를 통해 데이터베이스 설계, SQL 분석, Python 프로그래밍, 그리고 실제 마케팅 전략 수립까지 데이터사이언스의 전 과정을 경험할 수 있었습니다. 단순한 통계 분석을 넘어, 데이터베이스를 직접 구축하고, SQL로 데이터를 가공/분석하는 과정에서 데이터의 구조화와 실용성의 중요성을 다시 한번 느꼈습니다. 또한, 분석 결과를 실제 추천 시스템 프로그램으로 구현함으로써, 데이터 기반 의사결정의 실질적인 효과와 한계를 모두 체감할 수 있었습니다.

향후에는 머신러닝 기반의 예측 모델까지 확장해, 더욱 정교한 마케팅 대상 추천 시스템을 개발해보고 싶습니다.

6. 참고자료

[1] UCI Machine Learning Repository: Bank Marketing Data Set, https://archive.ics.uci.edu/dataset/222/bank+marketing

[2] 세종대학교 2025-1 '데이터베이스(002)' 강의자료