신용카드 고객 Segment 세분화 프로젝트

멋쟁이사자처럼 14조

이수형 | 김성철 | 심기열 | 이태윤

Contents

chapter1

프로젝트 개요

01 프로젝트 개요

배경 요약 및 목표

- 신용카드 고객 데이터를 분석하여 고객 등급 분류 모델 개발

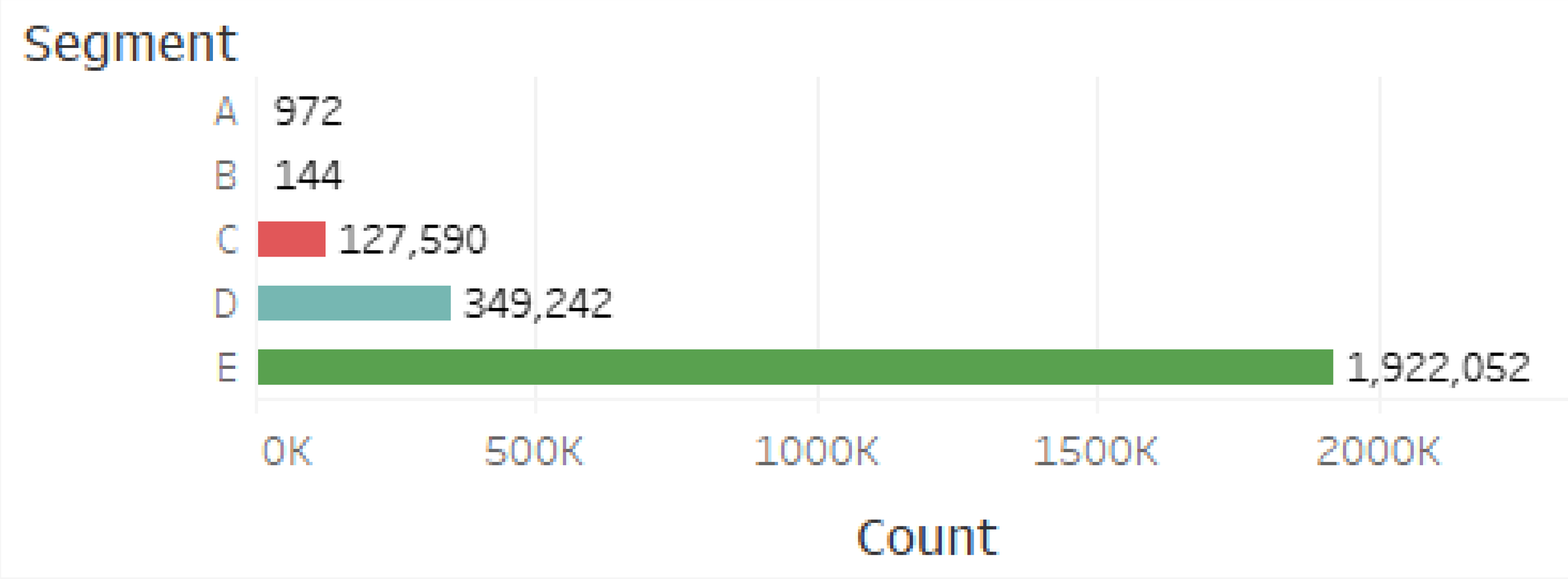
기대 효과

- 그룹별 맞춤형 마케팅 전략 수립
- 신용 리스크 및 이상 거래 탐지 가능
- 경영진용 인사이트 제공
- 월/분기별 Segment 성과 추적 → 전략 조정 용이

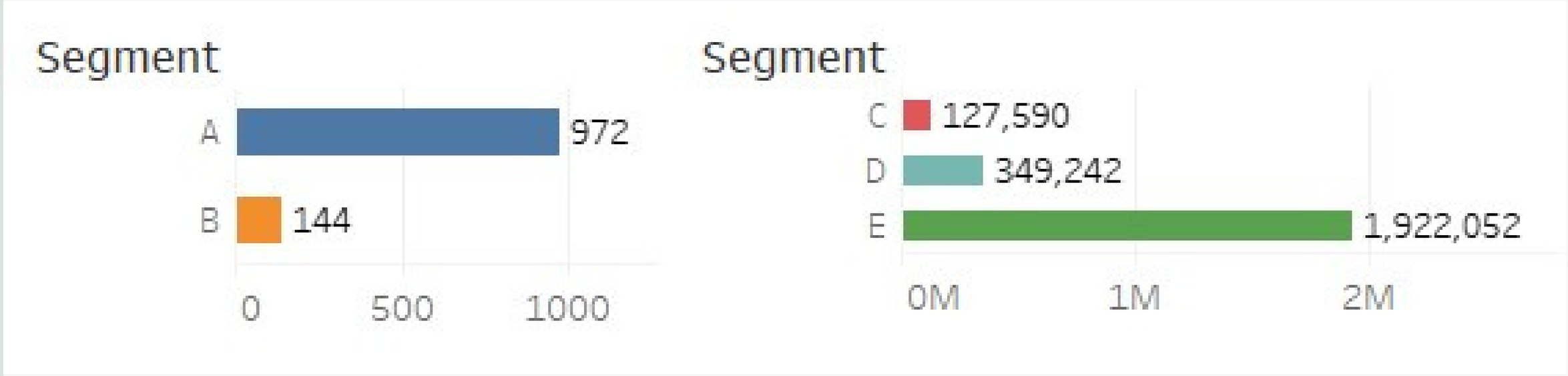
chapter2

EDA

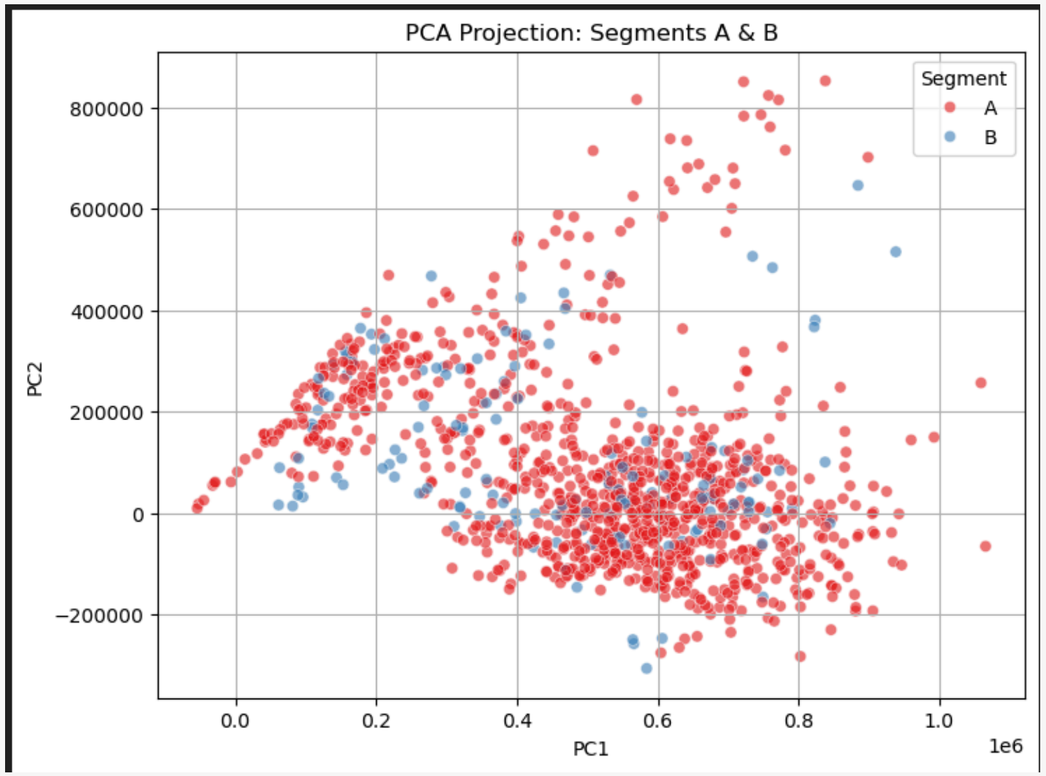
- **Target data 분포**

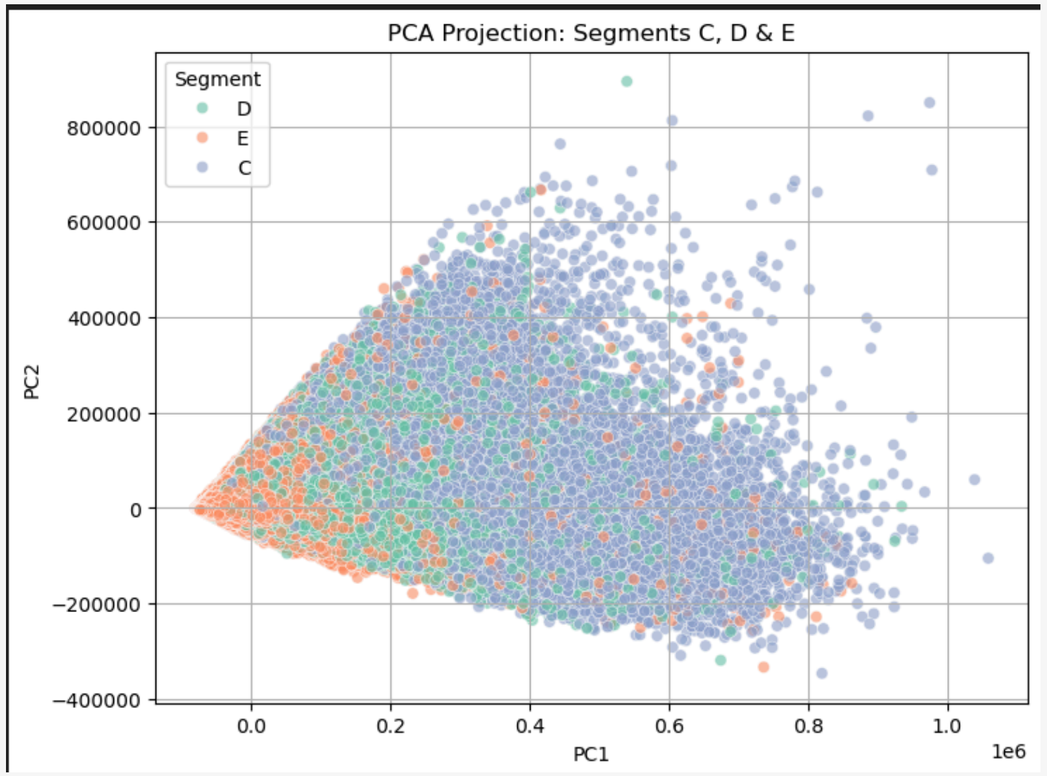


- **Target data 분포**



소수 클래스인 Segment A, B를 잘 예측하는 모델 구축 필요





소수 클래스인 Segment A, B를 잘 예측하는 모델 구축

chapter3

Feature Engineering

전처리 방식

1.데이터 유형에 따른 결측치 처리

- 범주형 데이터일 경우 기타, 수치형 데이터일 경우 평균으로 처리

2.데이터 구성에 따른 변수 처리

- 컬럼이 고유값 1개로만 이루어져 있거나, 결측치 비율이 90% 이상인 경우, 삭제

상관관계 분석

```python
for col in df.columns:
    if col == target_col:
        continue

    x = df[col]

    if pd.api.types.is_numeric_dtype(x):
        # 수치형 → Spearman
        coef, _ = spearmanr(x, target, nan_policy='omit')
        results[col] = coef

    elif x.nunique() == 2:
        # 이진 범주형 → Point-biserial (Label Encoding 필요)
        x_encoded = LabelEncoder().fit_transform(x)
        results[col] = point_biserial(target, x_encoded)

    else:
        # 다중 범주형 → Cramér's V
        results[col] = cramers_v(x, target)
```

- 수치형 → Spearman
- 이진범주형 → Point-biserial
- 다중범주형 → Cramer's V

```python
# 상관계수 계산 결과가 correlations에 들어있다고 가정
filtered = correlations.abs()
filtered = filtered[filtered > 0.4].sort_values(ascending=False)
print(filtered)
```

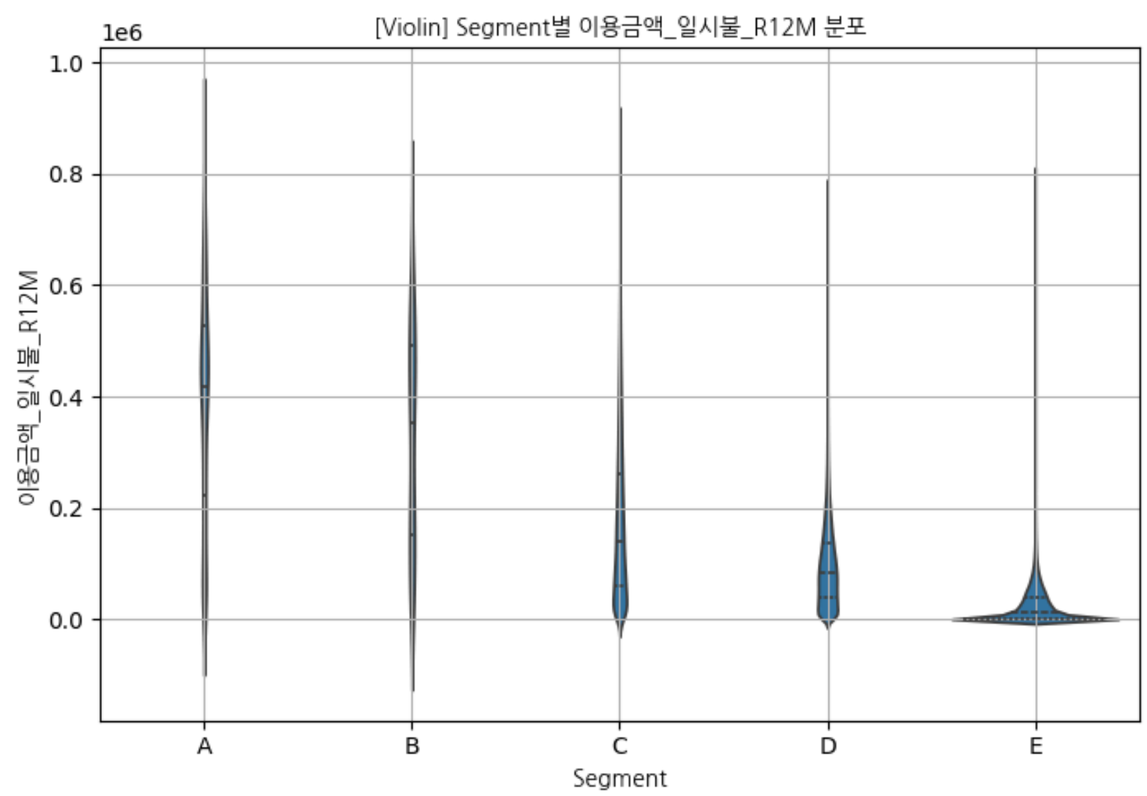
- 변수 유형에 따라 각기 다른 상관계수 사용을 통해 상관계수 도출
- target과의 상관계수 > 0.4인 변수들로 선정

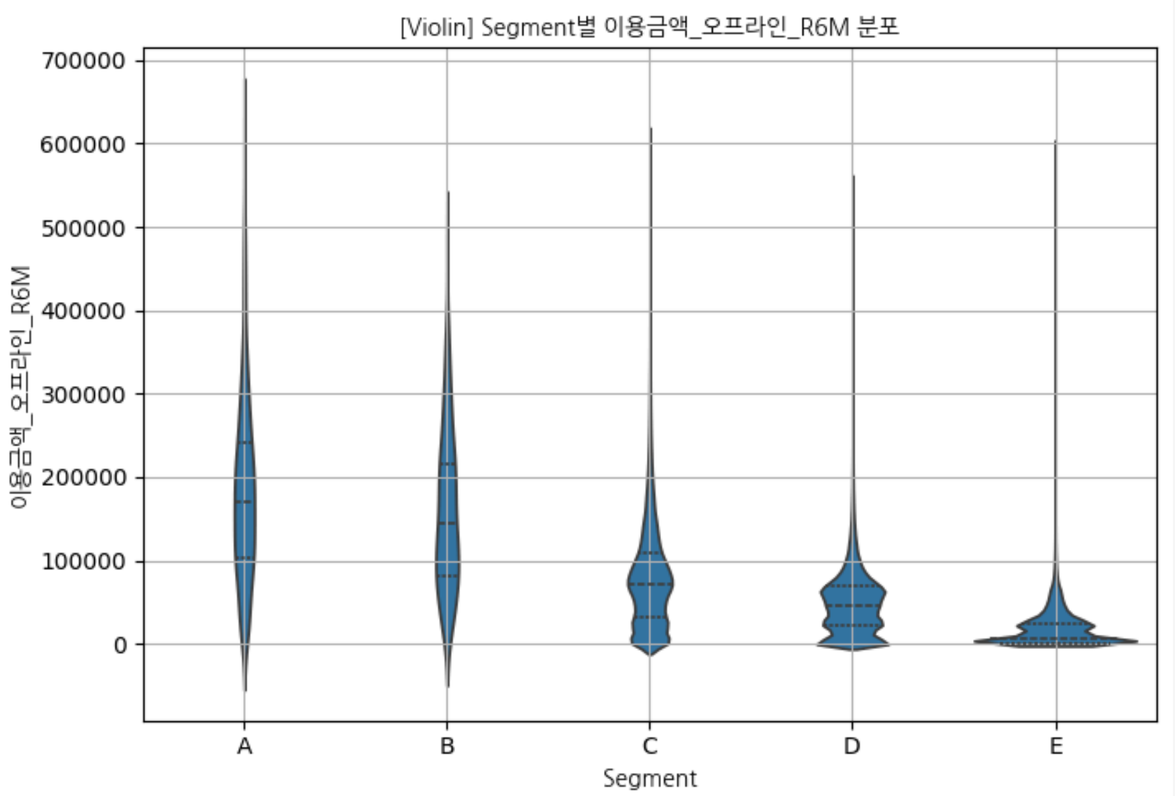
다중공선성 제거(VIF)

```python
# 상관계수로 선택된 변수만 사용
X = df

# 상수항 추가 (intercept term)
X = add_constant(X)

# VIF 계산
vif_df = pd.DataFrame()
vif_df["변수명"] = X.columns
vif_df["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
```

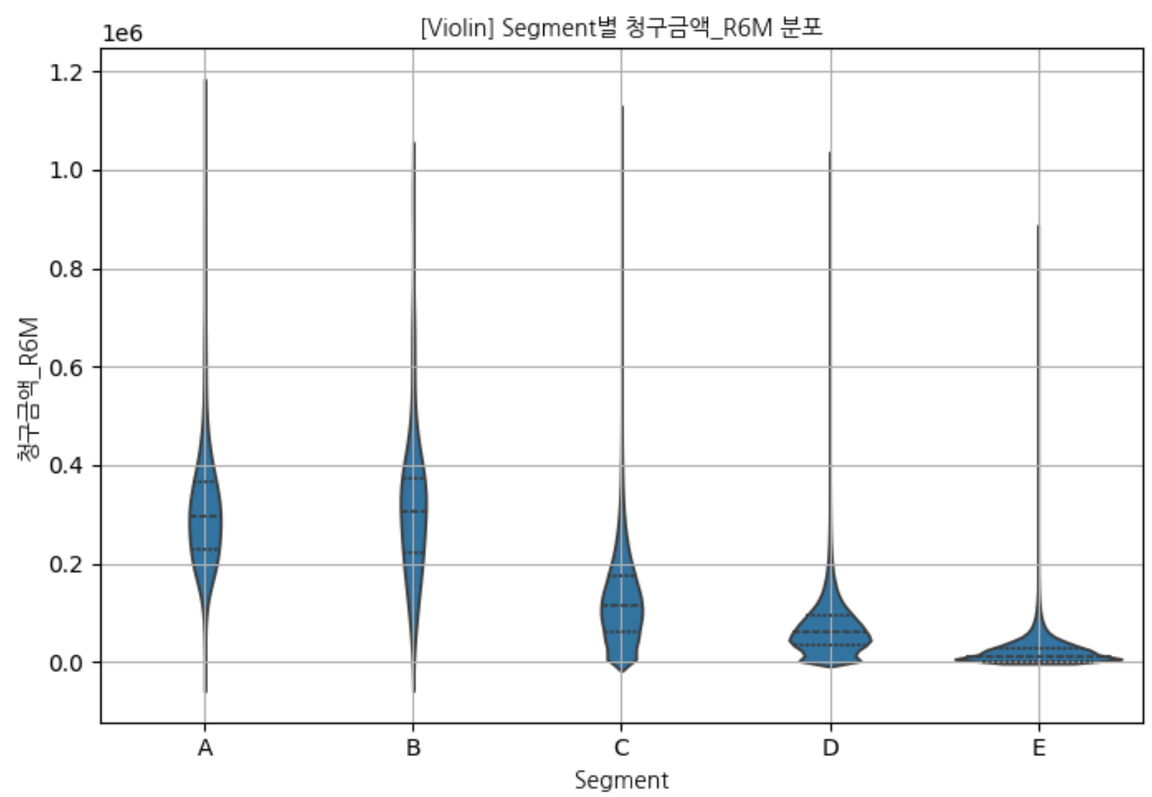
- VIF 계수가 높은 순대로 제거해 최종적으로 변수들이 26 이하의 VIF계수를 가짐으로써 다중공선성 해소

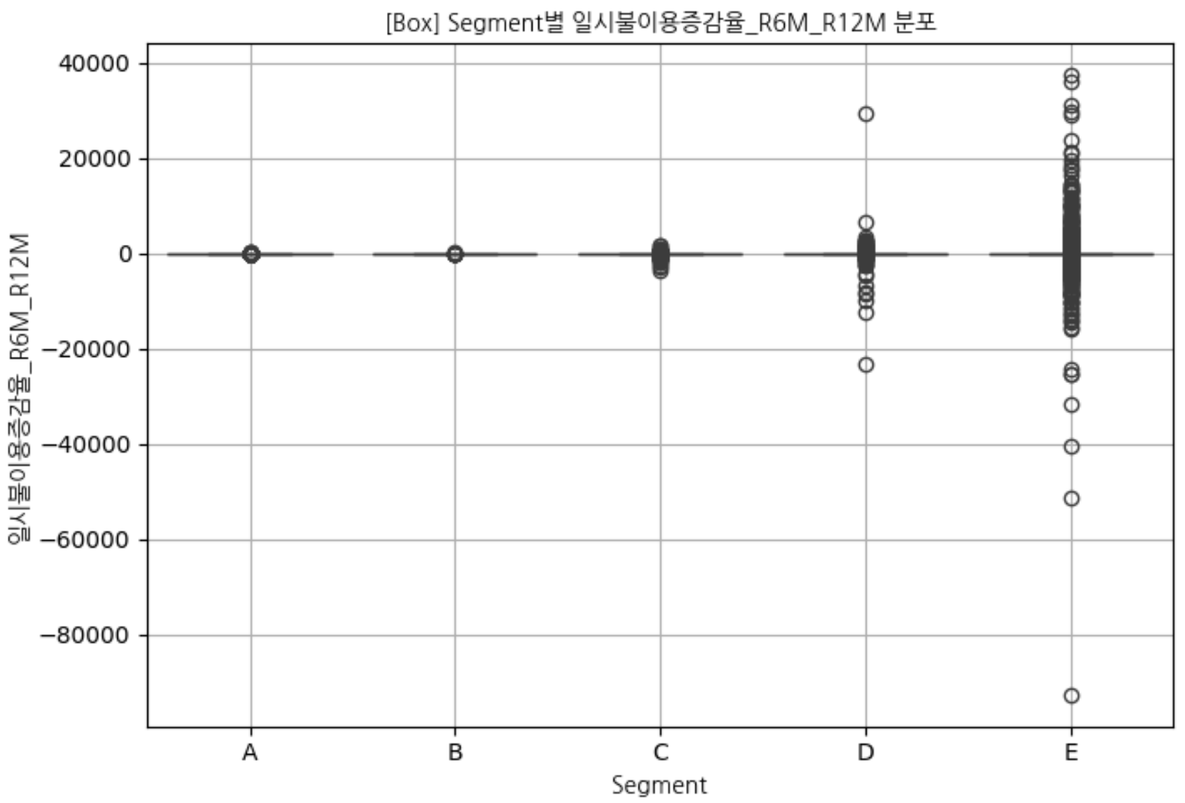
Segment(A/B , C/D/E) 간 column 차이



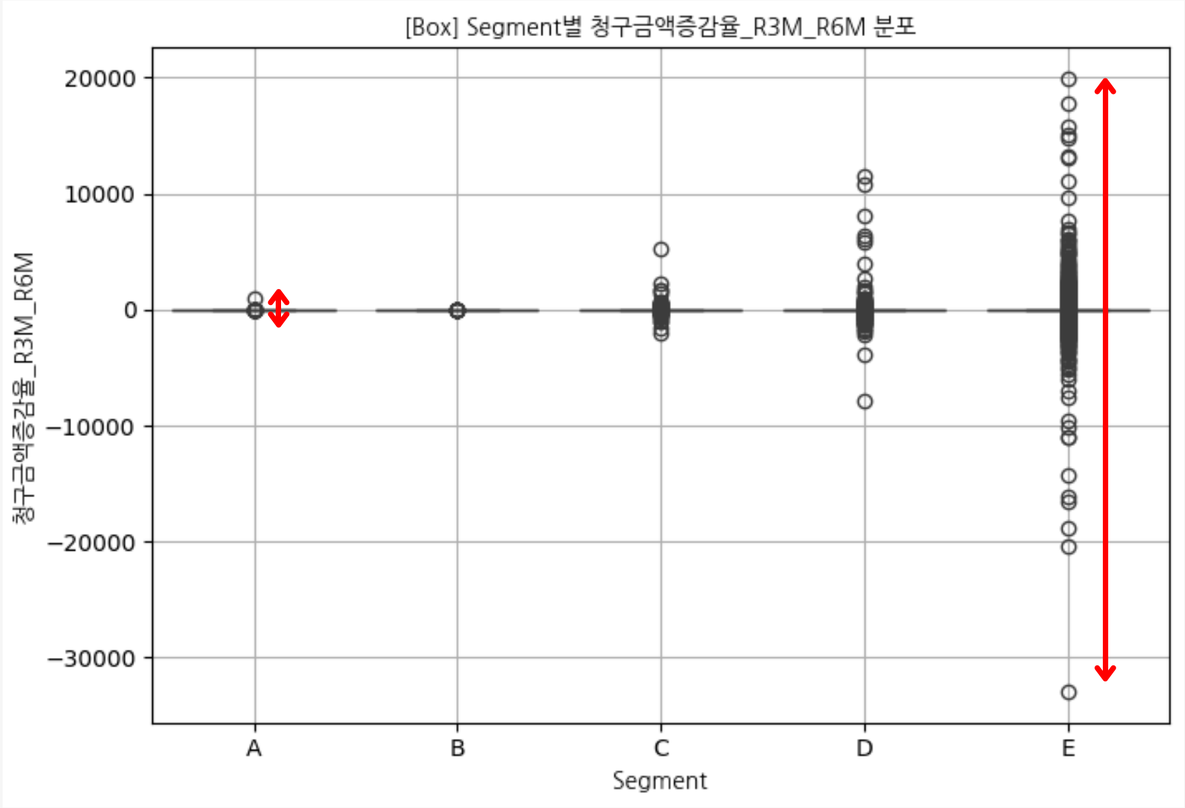


Segment(A/B , C/D/E) 별 차이





Segment(A/B , C/D/E) 별 차이



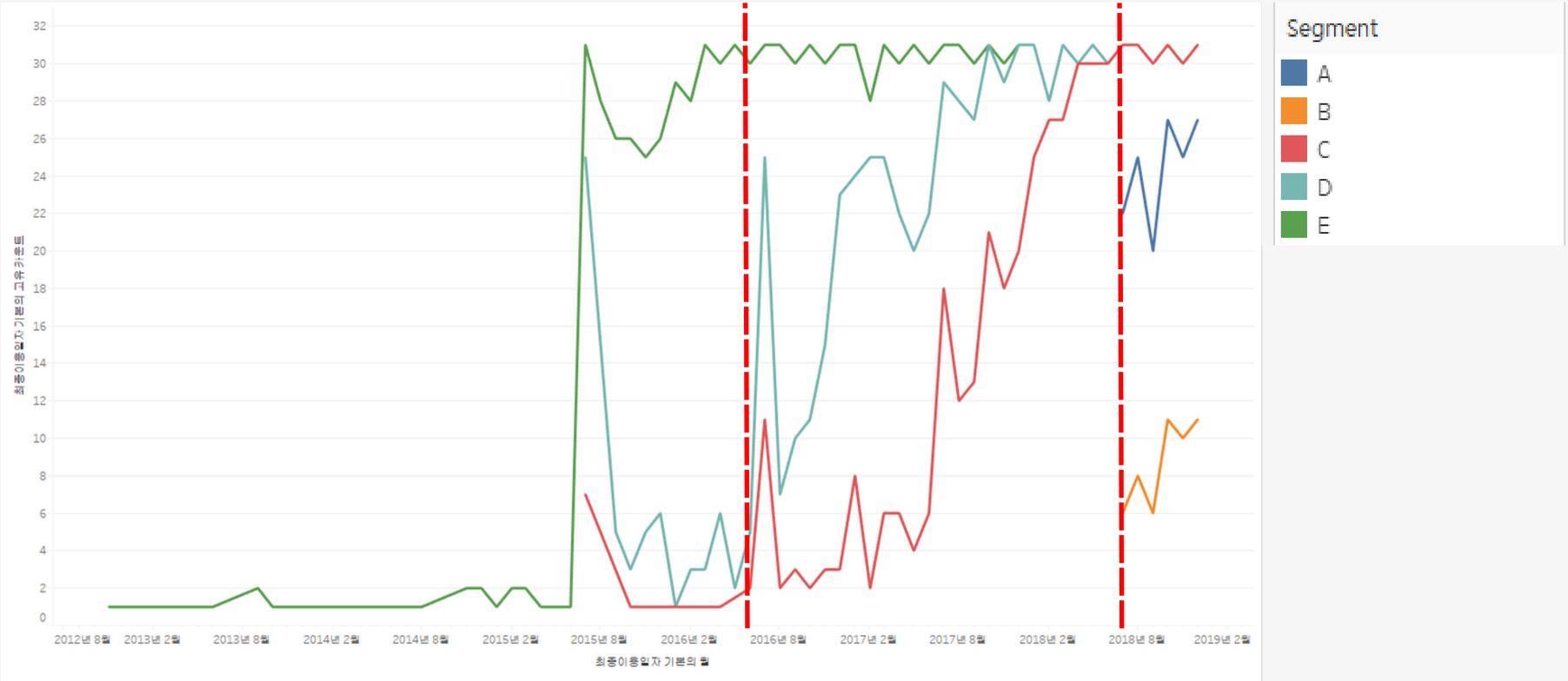
RFM 고객 세분화 분석

1.Recency : 얼마나 최근에 구매했는가
→ ‘최종이용일자_기본’ 컬럼 선정

2. Frequency : 얼마나 자주 구매했는가
→ '이용건수신용_R12M' 컬럼 선정

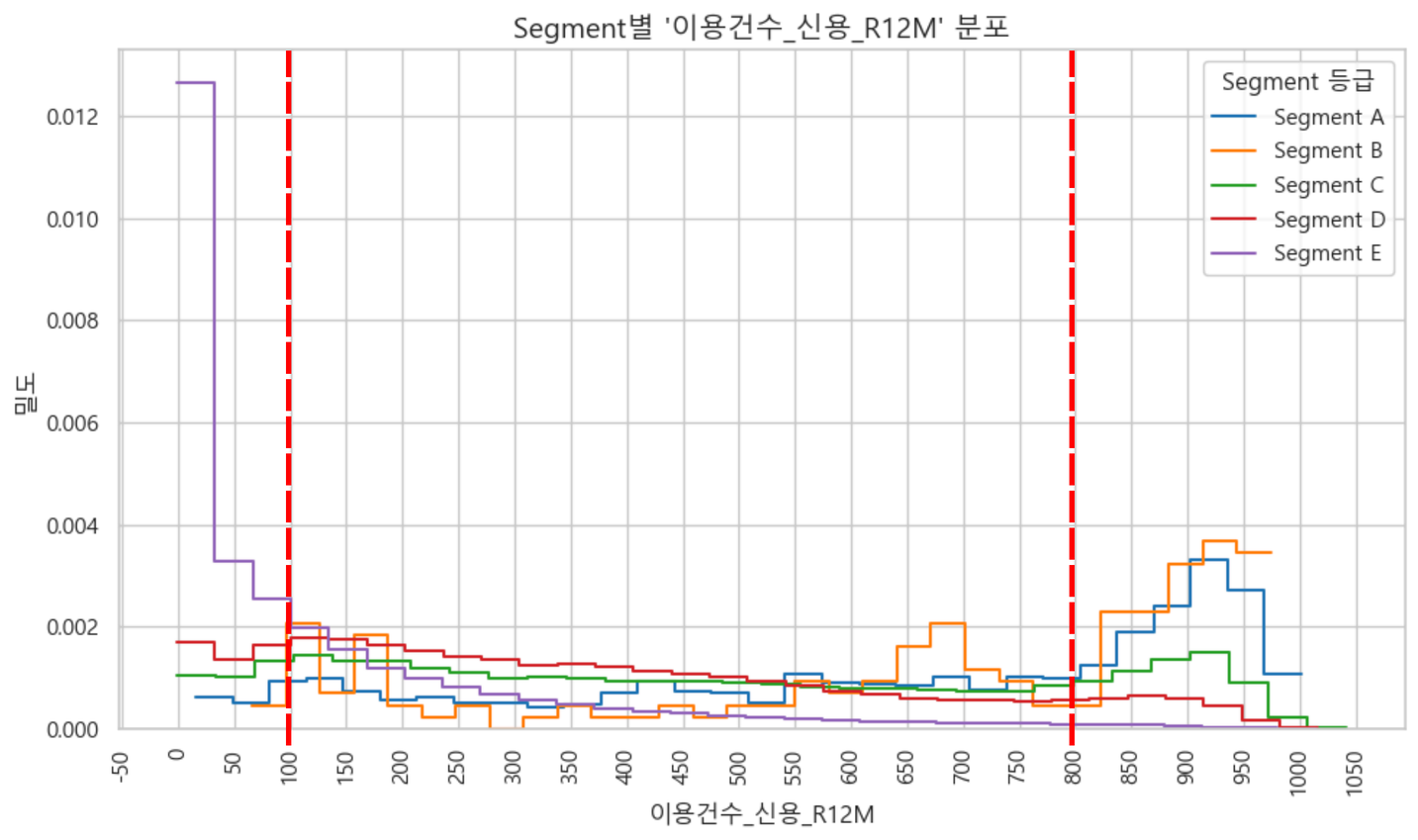
3.Monetary : 얼마나 많은 금액을 지출했는가
→ '이용금액_R3M_신용체크' 컬럼 선정

RFM 고객 세분화 분석(Recency)



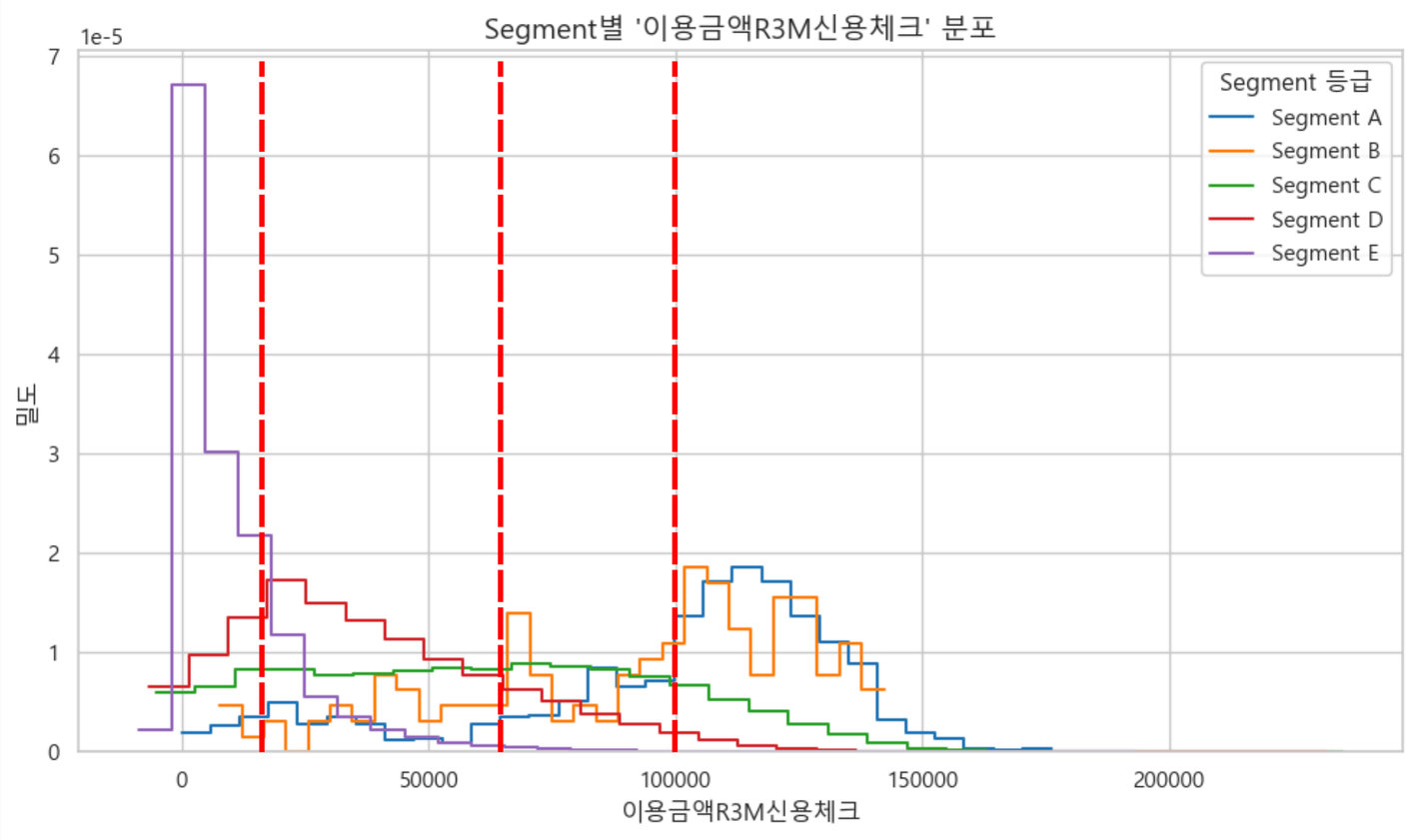
2016-08과 2018-07을 기준으로 very past/past/recent의 3가지 등급을 가진 변수 생성

RFM 고객 세분화 분석(Frequency)



100-800을 기준으로 low/middle/high의 3가지 등급을 가진 변수 생성

RFM 고객 세분화 분석(Monetary)



18000-65000-100000을 기준으로 very low/low/middle/high의 4가지 등급을 가진 변수 생성

파생변수(증감률)

파생변수명	=	항목1	연산	항목2
카드이용금액합	=	1순위 카드	+	2순위 카드
카드이용금액차	=	1순위 카드	-	2순위 카드
업종이용금액합	=	1순위 업종	+	2순위 업종 + 3순위 업종
업종 집중도	=	1순위 업종	/	업종이용금액합 + 1
쇼핑합계	=	1순위 쇼핑	+	2순위 쇼핑
쇼핑비중	=	쇼핑합계	/	업종이용금액합 + 1
입금률_B2M	=	정상입금원금_B2M	/	정상청구원금_B2M + 1
입금률_B5M	=	정상입금원금_B5M	/	정상청구원금_B5M + 1
입금률증감_B2M_B5M	=	입금률_B2M	-	입금률_B5M

파생변수(증감률)

$$\text{증감률}\,(\%) = \frac{\text{현재값} - \text{이전값}}{\text{이전값}} \times 100$$

```python
def calc_growth_rate(current, past_avg):
    # 1) 숫자형 변환 (NaN 있으면 그대로 유지)
    current = pd.to_numeric(current, errors='coerce').fillna(0)
    past_avg = pd.to_numeric(past_avg, errors='coerce').fillna(0)

    # 2) 분모가 0인 경우 False, 아니면 True mask
    mask = past_avg != 0

    # 3) 기본값 0 배열 생성
    result = np.zeros(len(current))

    # 4) mask=True인 곳만 계산
    result[mask] = (current[mask] - past_avg[mask]) / past_avg[mask]

    return result
```

- '일시불이용증감률_R6M_R12M'
- '일시불평잔증감률_R3M_R6M'
- '일시불평잔증감률_B0M_R3M'
- '오프라인증감률_R3M_R6M'
- '오프라인증감률_B0M_R3M'
- '청구금액증감률_R3M_R6M'
- '청구금액증감률_B0M_R3M'

Feature Engineering

1.초기 데이터

분석을 시작하는 전체 변수 집합

총 855개 Columns

2. 상관 계수

목표 변수와 유의미한 관계의 변수 선택

기준: 상관계수 > 0.4

35개 Columns

3. VIF

변수 간 다중공선성 문제 해결 → 모델 안정성

기준: VIF > 26 변수 제거

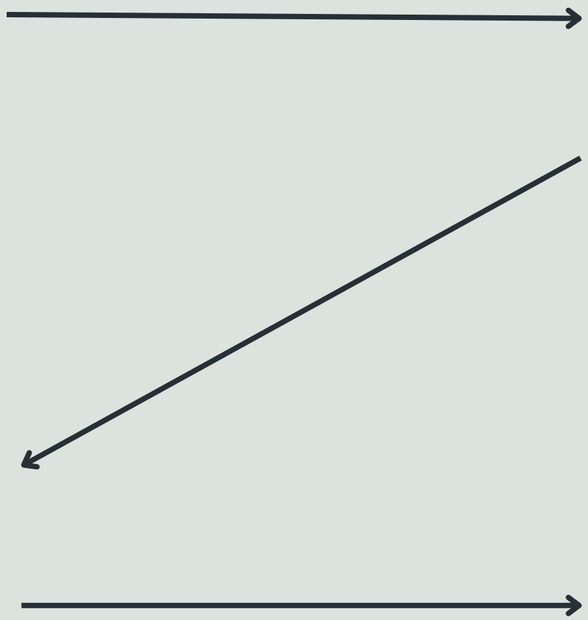
30개 Columns

4.최종 변수 확정

RFM 고객 세분화 방법론 적용
기간에 따른 증감율 적용

파생변수 생성 / 대체

38개 Columns



chapter4

Modeling

모델링 방법론

Method1

A,B/C,D,E 그룹으로 나눠 각각 모델 학습 -> A,B 모델 예측 결과 중 A 또는 B일 확률이 높은 결과를 C,D,E 모델 예측 결과에 덮어씌움 -> 각 예측 결과 토대로 Segment 예측

Method2

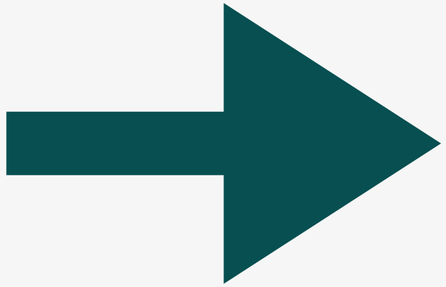
A,B/C,D,E 그룹으로 나눠 각각 모델 학습 -> 각 모델을 통한 Segment의 예측 확률 도출 -> 각 예측 확률을 하나의 행렬로 합산해 최종결과 환산

Method3

데이터를 E와 not E 그룹으로Set1 생성 -> not E 그룹에서 not C,D,CD 그룹으로 Set2 생성 -> A,B만 모아 Set3 생성 ->Set1,2,3로 생성한 model1,2,3를 차례대로 사용해 예측

모델링 최종선택

Method1 public score : 0.8495
Method2 public score : 0.7054
Method3 public score : 0.9024



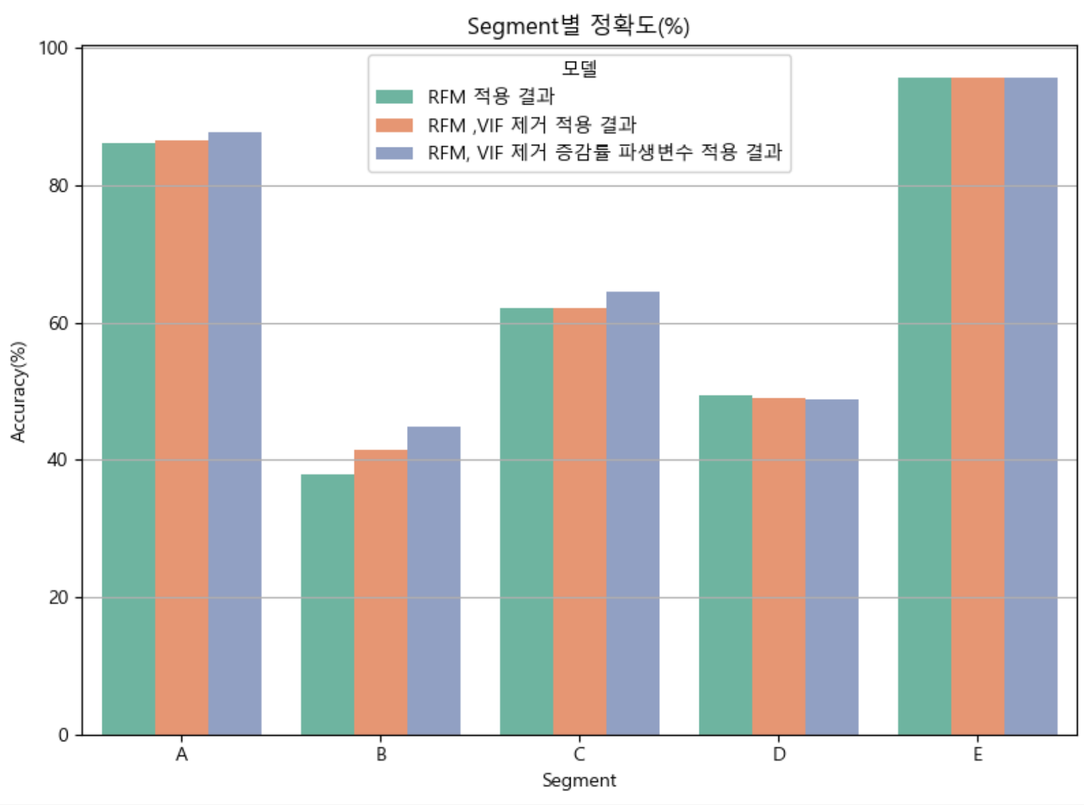
Method3 최종선정

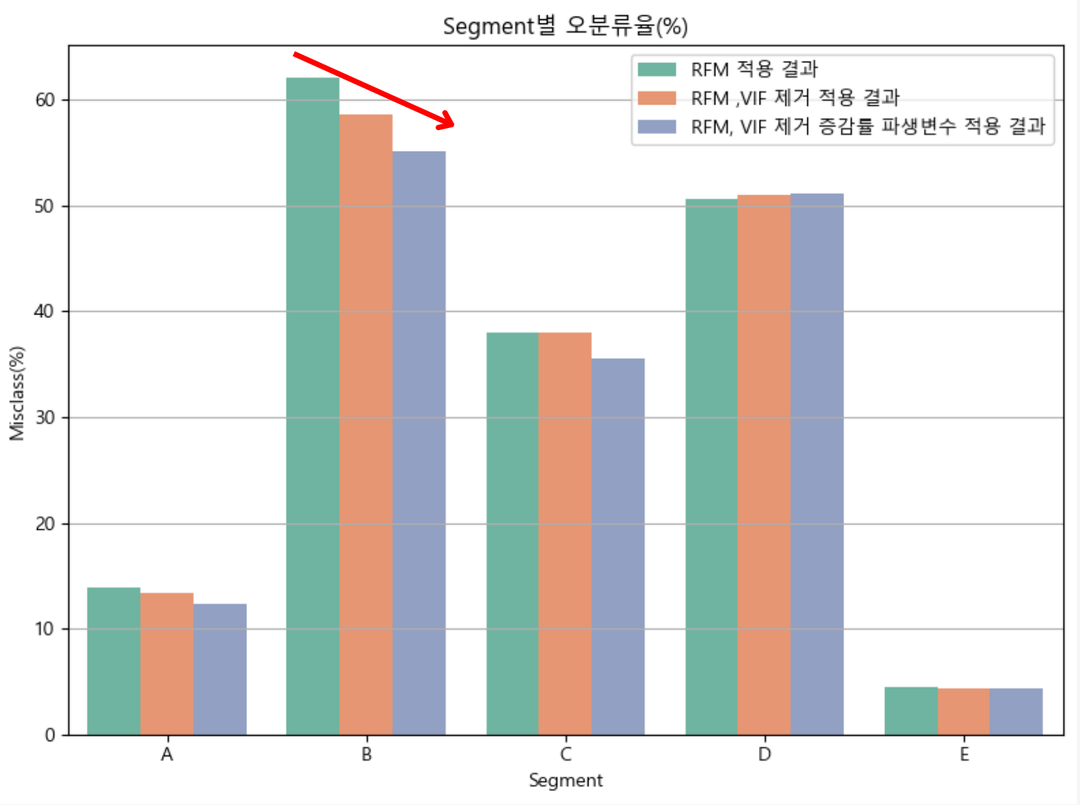
Method3 모델 바탕으로 VIF, 파생변수 기반의 피쳐 변화에 의한 모델 성능 향상 도모

chapter5

모델 성능 비교

모델별 정확도, 오분류율


Segment별 정확도(%)
모델
RFM 적용 결과
RFM ,VIF 제거 적용 결과
RFM, VIF 제거 증감률 파생변수 적용 결과
Accuracy(%)
100
80
60
40
20
0
A
B
C
D
E
Segment


Segment별 오분류율(%)
RFM 적용 결과
RFM ,VIF 제거 적용 결과
RFM, VIF 제거 증감률 파생변수 적용 결과
Misclass(%)
60
50
40
30
20
10
0
A
B
C
D
E
Segment

chapter6

결론 및 시사점

6. 결론

1. E vs Others → C/D vs Others → A vs B, 3단계 모델링

2. RFM 고객 세분화 방법론을 적용한 파생변수 생성

3. 기간에 따른 증감율 파생변수 생성

Segment 소수 클래스(A, B) 예측 성능 향상

상관계수 0.4이상 column

	Segment	Total	Correct	Incorrect	Accuracy(%)	Misclass(%)
0	A	194	167	27	86.08	13.92
1	B	29	10	19	34.48	65.52
2	C	25518	16384	9134	64.21	35.79
3	D	69848	35064	34784	50.20	49.80
4	E	384411	367359	17052	95.56	4.44

상관계수 0.4이상 column + 파생변수 + VIF 제거

	Segment	Total	Correct	Incorrect	Accuracy(%)	Misclass(%)
0	A	194	167	27	86.08	13.92
1	B	29	13	16	44.83	55.17
2	C	25518	16083	9435	63.03	36.97
3	D	69848	34211	35637	48.98	51.02
4	E	384411	367569	16842	95.62	4.38

6. 의의 및 한계

- 불균형 데이터 해결 전략 수행 : auto balanced를 통해 불균형 정도가 심한 데이터에 적합한 방식 채택
- Segment 별 특징이 두드러지는 column 발견

- 불균형 데이터 해결에 집중했음에도 Segment B, C, D 오분류율이 낮지 않음
- 금융 관련 방법론 등 다양한 방법 적용 부족
- 결측치 대체 과정에서 평균, '기타'가 아닌 다른 방식으로의 대체 미실시

6. 향후 연구방향

- 고객 분류, 금융 관련 방법론 등 다양한 방법 적용하여 후속 연구 실시
- 결측치 대체 과정에서 평균, ‘기타’가 아닌 다른 방식으로의 대체 실시
- 모델 다양화한 Ensemble 모델링 실시

chapter7

질의응답

궁금한 사항을 질문해 주세요.

Thank you

감사합니다.