

# SAS Viya Trial

## 모델 개발 가이드

데이터 사이언티스트 Task



# Intro

# 데이터 및 AI 라이프사이클: 모델 개발

Futurum Group의 최신 연구에 따르면, **SAS Viya**는 데이터 및 AI 팀의 **생산성을 4.6배 증가**시키는 것으로 나타났습니다.

데이터 및 AI 라이프사이클의 두 번째 단계는 **모델 개발**입니다. **데이터 사이언티스트**는 데이터 엔지니어가 준비한 데이터를 **탐색**하고 이를 바탕으로 **이탈 가능성을 예측하는 모델을 구축**합니다.

본 가이드는 **SAS Viya** 환경에서 **데이터 사이언티스트**가 개발을 수행한 **과정**을 단계별로 안내합니다.

# 데이터 사이언티스트

데이터 탐색 및 변환

모델 개발, 최적화, 검증 및 문서화

# 주요 작업 항목

1. 시각적 탐색 및 인사이트 도출
2. 시각적 탐색 - 증강 분석 (Augmented Analytics)
3. 이상치 (Outlier) 탐지
4. 모델 프로토타입 생성
5. 모델 구축
6. 모델 비교 (Model Comparisons)
7. 모델 해석 (Explainability)
8. 모델 리포트
9. 파이프라인 비교 (Pipeline Comparisons)
10. 모델 등록 (Model Registration)
11. 프로젝트 인사이트 보고서 - 문서화
12. 프로젝트 공유 - 읽기/쓰기 권한 설정



# 1. 시각적 탐색 및 인사이트 도출

# 변수 유형별 구별

수치형 변수 (Numeric Variable), 측정 변수 (Measure Variable), 범주형 변수 (Categorical Variable)

| 구분                               | 설명                           | 주요 특징          | 예시            |
|----------------------------------|------------------------------|----------------|---------------|
| 수치형 변수<br>(Numeric Variable)     | 데이터 타입이 숫자인 모든 변수            | 연산 가능 여부 관계 없음 | 우편번호, 주민번호    |
| 측정 변수<br>(Measure Variable)      | 수치형 값으로, 합계, 평균 등 연산이 가능한 변수 | 연산 가능          | 매출, 수입        |
| 범주형 변수<br>(Categorical Variable) | 특정 그룹이나 범주를 나타내는 변수          | 연산 불가, 그룹 구분용  | 성별, 국가, 고객 등급 |

수치형 변수는 연산 가능 여부에 따라 측정 변수 또는 범주형 변수로 구분할 수 있습니다.

SAS Viya에서는 변수를 '측정(Measure)'과 '범주(Category)'로 구분하여 데이터 탐색과 분석을 수행합니다.

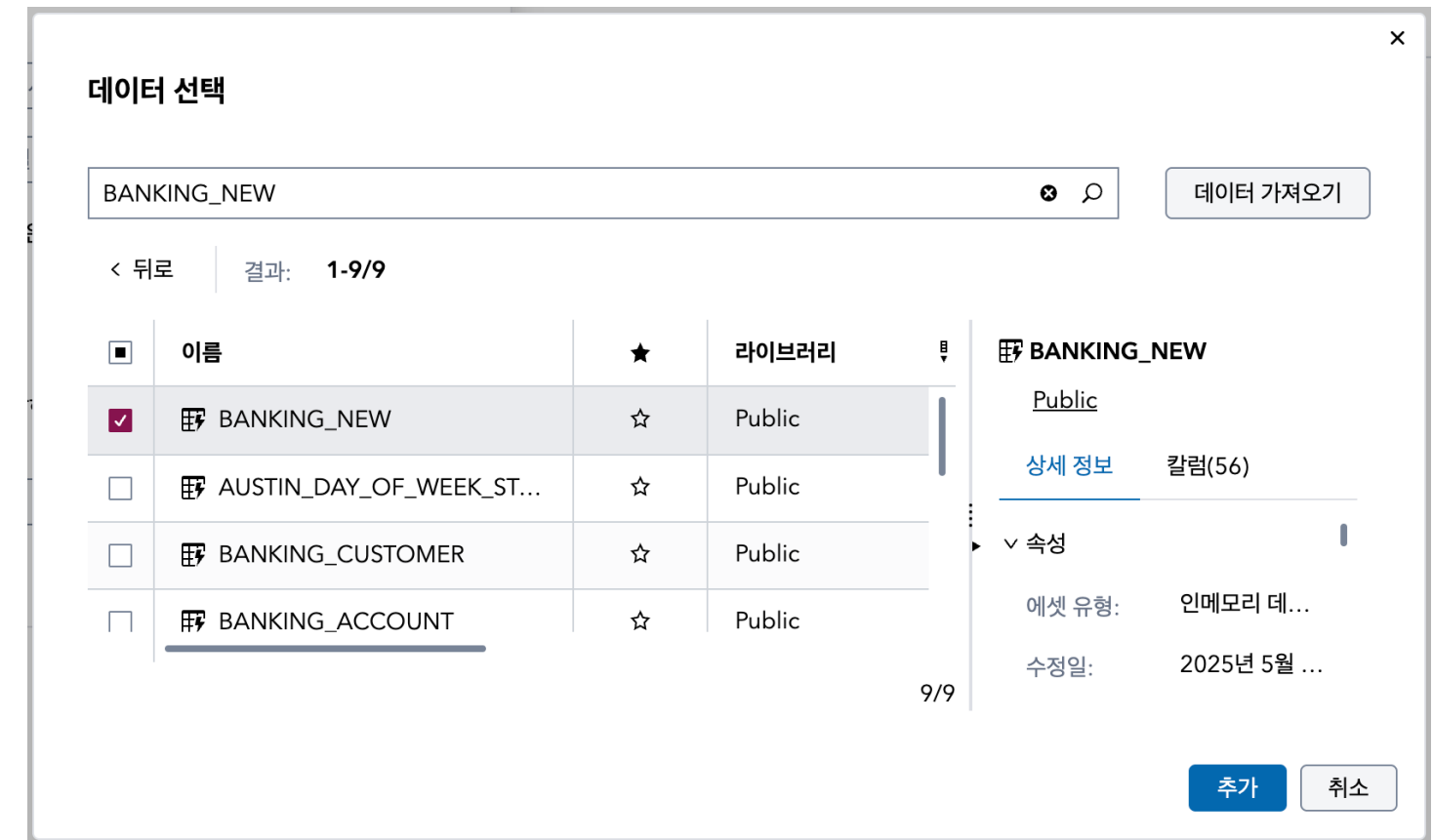
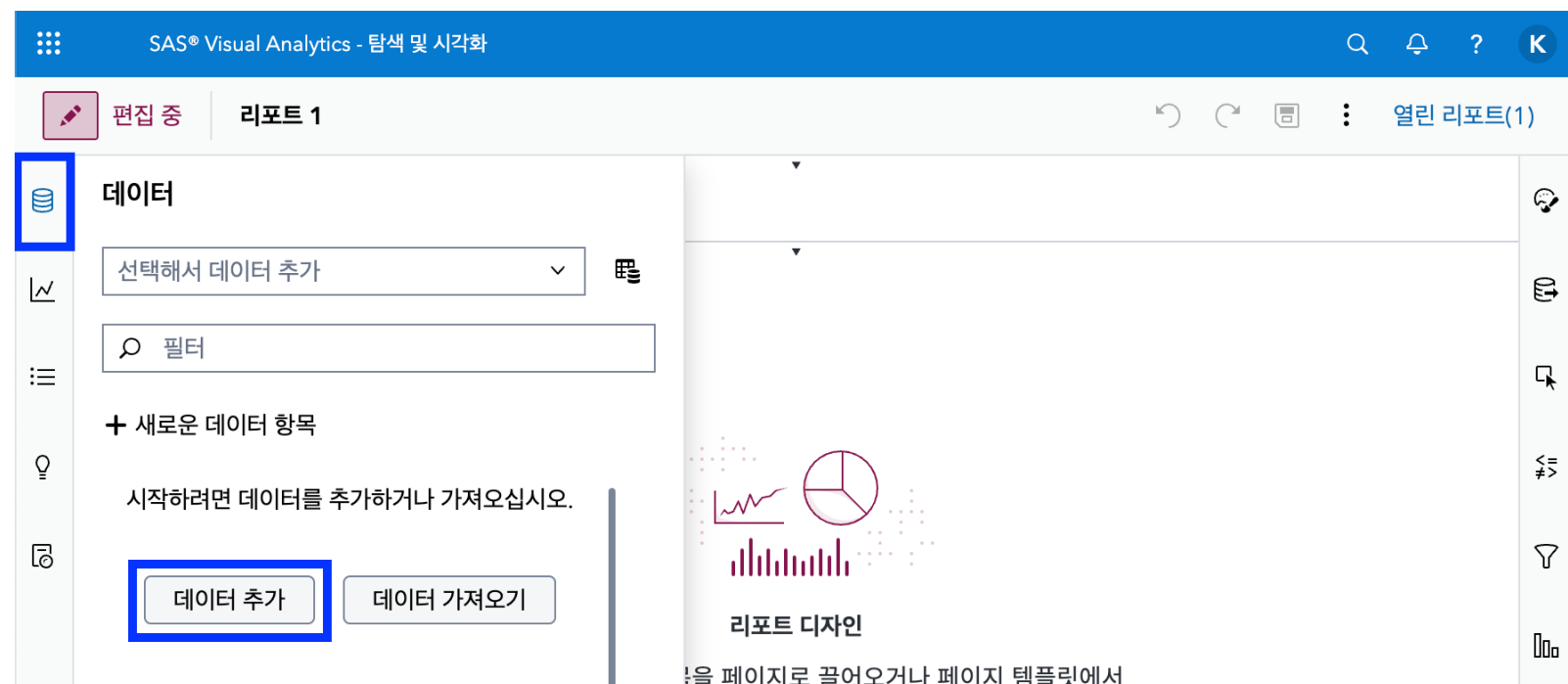
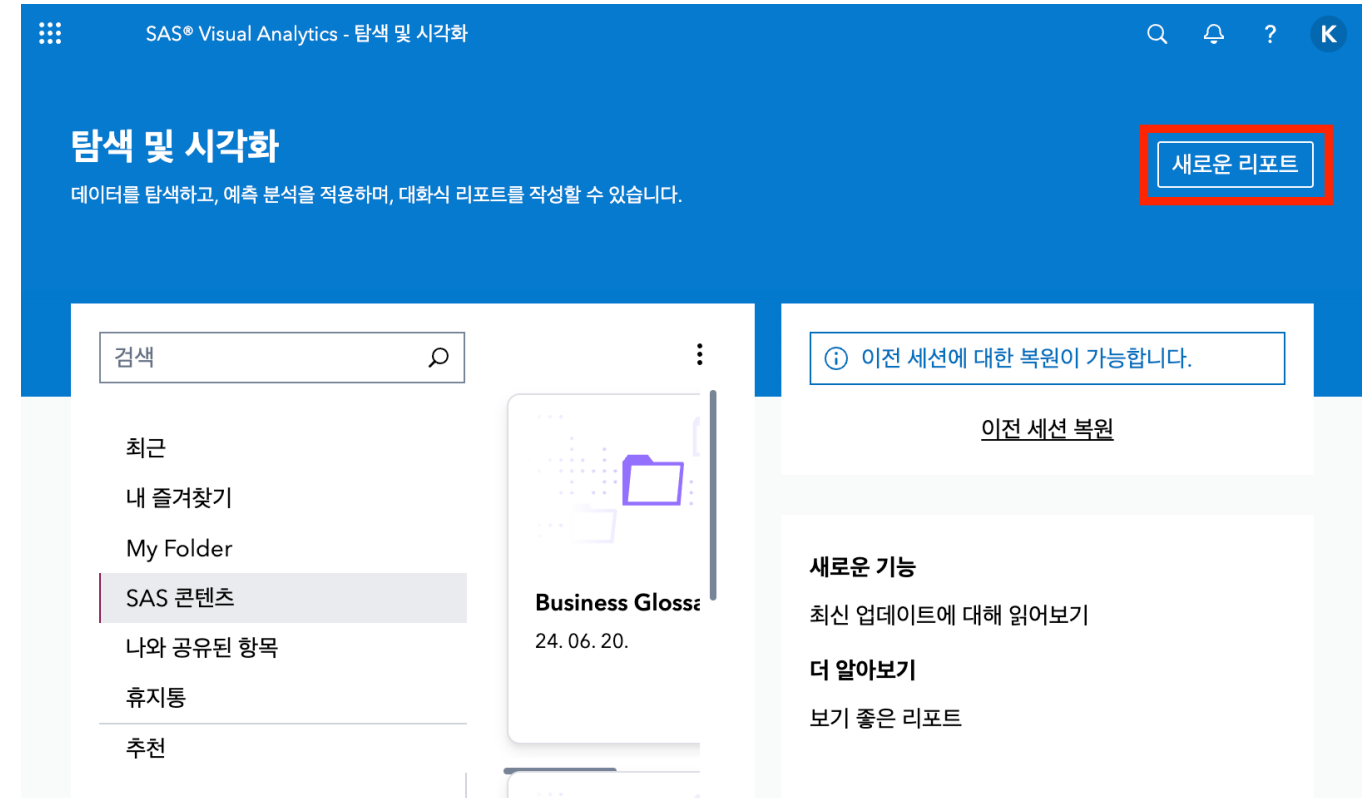
# 1. 시각적 탐색 및 인사이트 도출

## 자동 생성 및 GUI 기반 탐색

응용 프로그램 메뉴 > 탐색 및 시각화를 선택합니다.  
선택할 시 화면은 SAS Visual Analytics로 전환되며, 데이터 탐색, 예측 분석 적용 및 대화식 리포트를 작성할 수 있습니다.

이제 새로운 리포트를 클릭합니다. 데이터 > 데이터 추가를 클릭할 시 팝업 창이 뜨며, 불러올 데이터를 선택할 수 있습니다. 검색창을 활용하여 BANKING\_NEW 데이터 세트를 불러옵니다.

이렇게 불러온 데이터는 범주형(categorical) 과 측정(measure) 변수로 자동 구분됩니다.

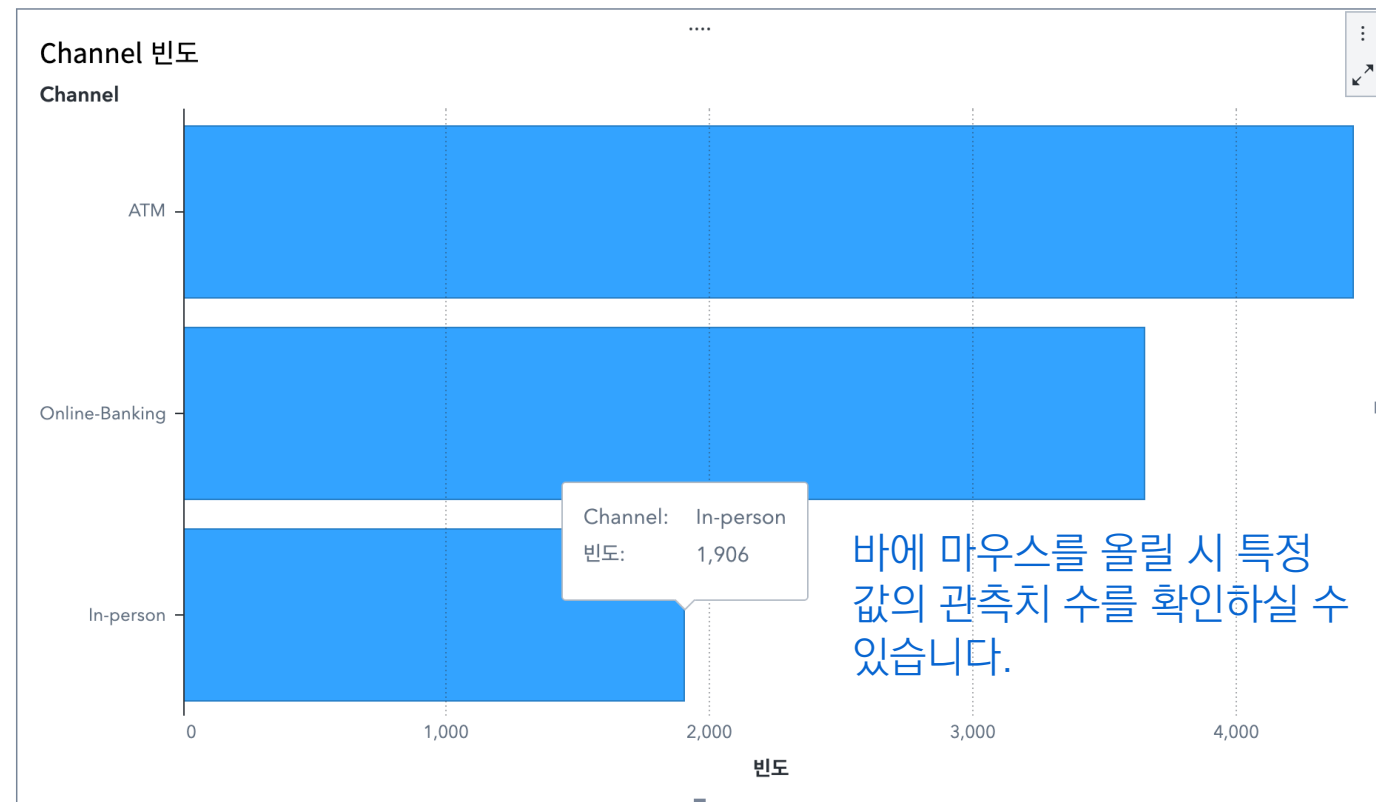


# 1. 시각적 탐색 및 인사이트 도출

## 자동 분포 시각화 (Automatic Distributions)

데이터를 불러왔을 시 **데이터의 범주형 변수와 측정 변수를 나열한 리스트가 제공됩니다.** 이제 범주형 변수와 측정 변수를 각각 하나씩 선택하여 기본적인 시각화를 진행해봅시다.

왼쪽 데이터 탭에서 **Channel** 변수를 더블 클릭하면 변수의 **막대 그래프**가 자동 생성되며, 각 범주의 빈도값이 표시됩니다.



데이터

BANKING\_NEW

필터

+ 새로운 데이터 항목

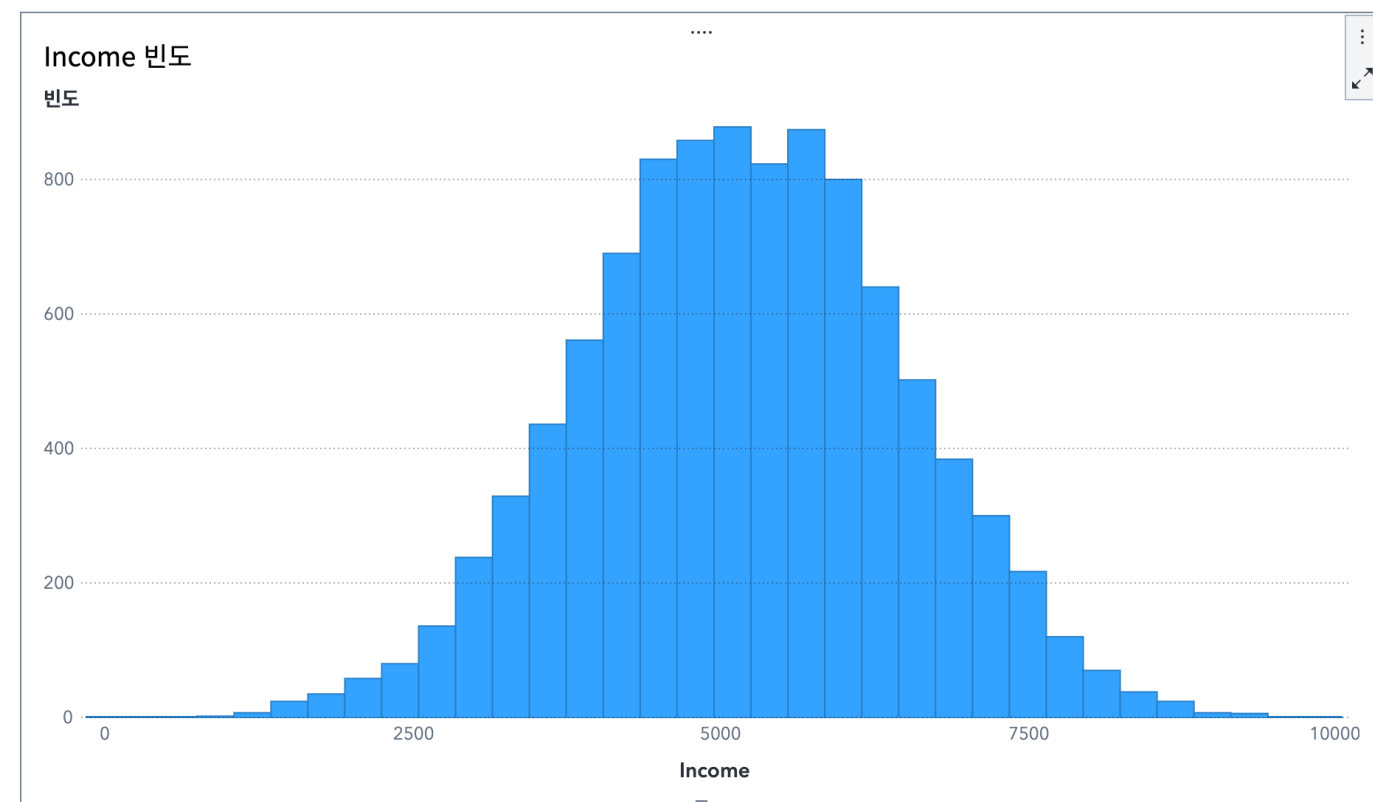
범주

- Channel - 3
- Churn - 2
- Cross\_selling - 2
- Customer\_sentiment - 5

측도

- 빈도
- Age
- Amount\_avg
- Balance\_avg
- Behavior\_segment
- Churn\_num
- Complaints\_num
- Creditcards\_num
- Creditcardusage\_avg
- Creditcardusage\_num

+ 버튼을 클릭해 새 페이지를 추가하고 **Income** 변수를 더블 클릭하면 자동으로 **히스토그램**이 생성됩니다.



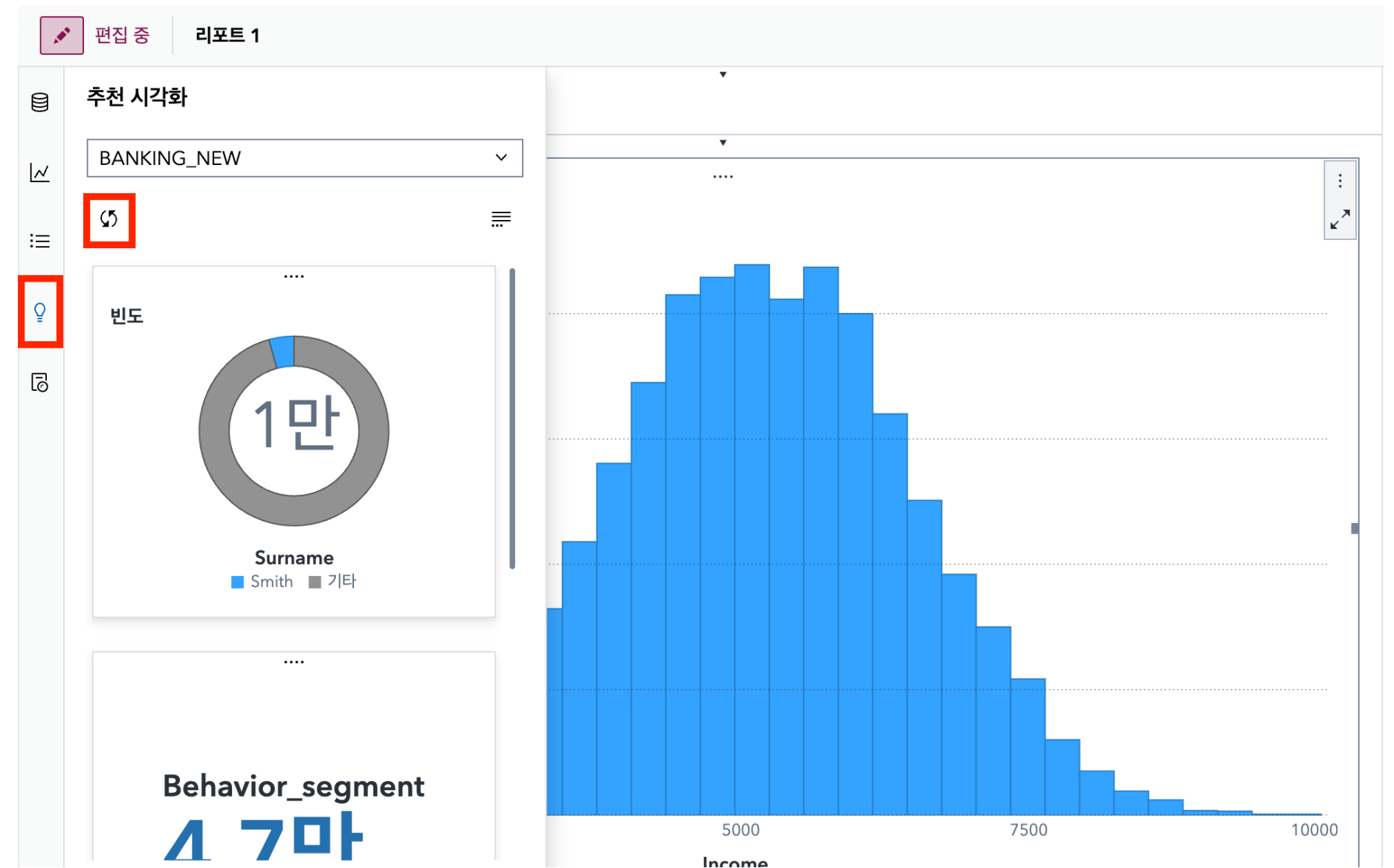
# 1. 시각적 탐색 및 인사이트 도출

## 자동 그래프 추천 기능 (Automatic Graph Suggestions)

SAS Viya는 AI 기반 추천을 통해 데이터 속 숨겨진 패턴과 인사이트를 빠르고 효율적으로 제시합니다.

화면 좌측의 추천 시각화 탭을 클릭하면 자동 생성된 그래프 추천 목록을 확인할 수 있습니다.

더 많은 추천을 보고 싶다면, 상단의 새로고침 아이콘을 클릭하시면 됩니다.

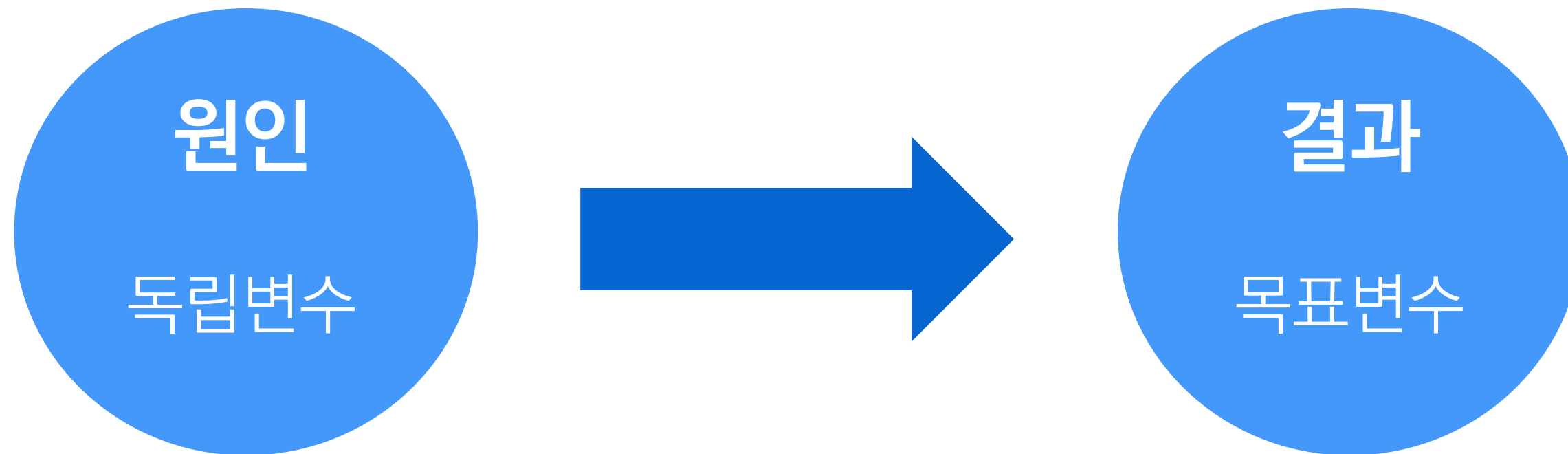


## 2. 시각적 탐색 – 증강 분석 (Augmented Analytics)

# 변수 유형별 정의

## 독립변수, 목표변수

| 구분                             | 설명                              | 다른 표현             |
|--------------------------------|---------------------------------|-------------------|
| 독립변수<br>(Independent Variable) | 예측하려는 목표값(종속 변수)에 영향을 주는 입력 데이터 | 입력변수, 설명변수, 피처 변수 |
| 목표변수<br>(Target Variable)      | 예측을 하려는 목표값 또는 출력값              | 종속변수, 결과변수        |



# 2. 시각적 탐색 - 증강 분석 (Augmented Analytics)

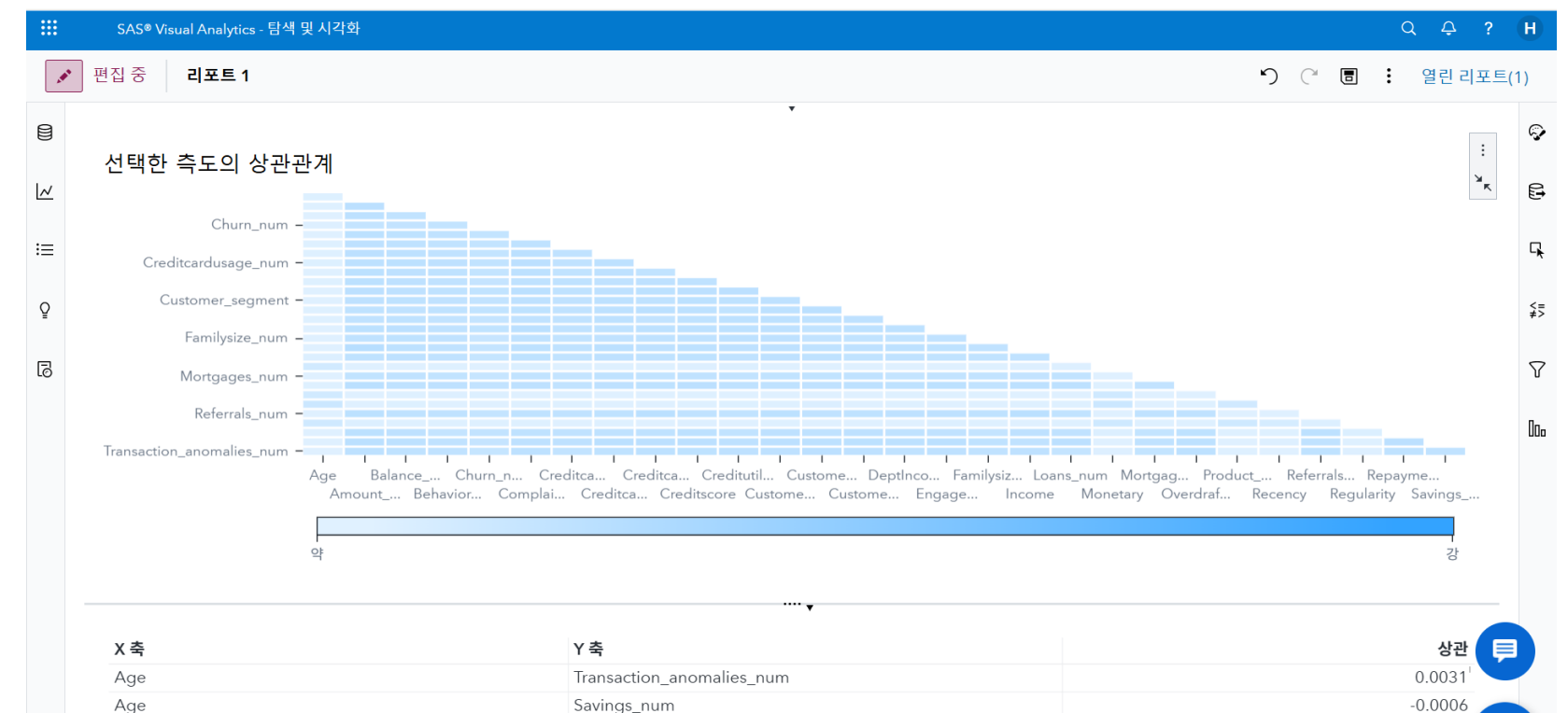
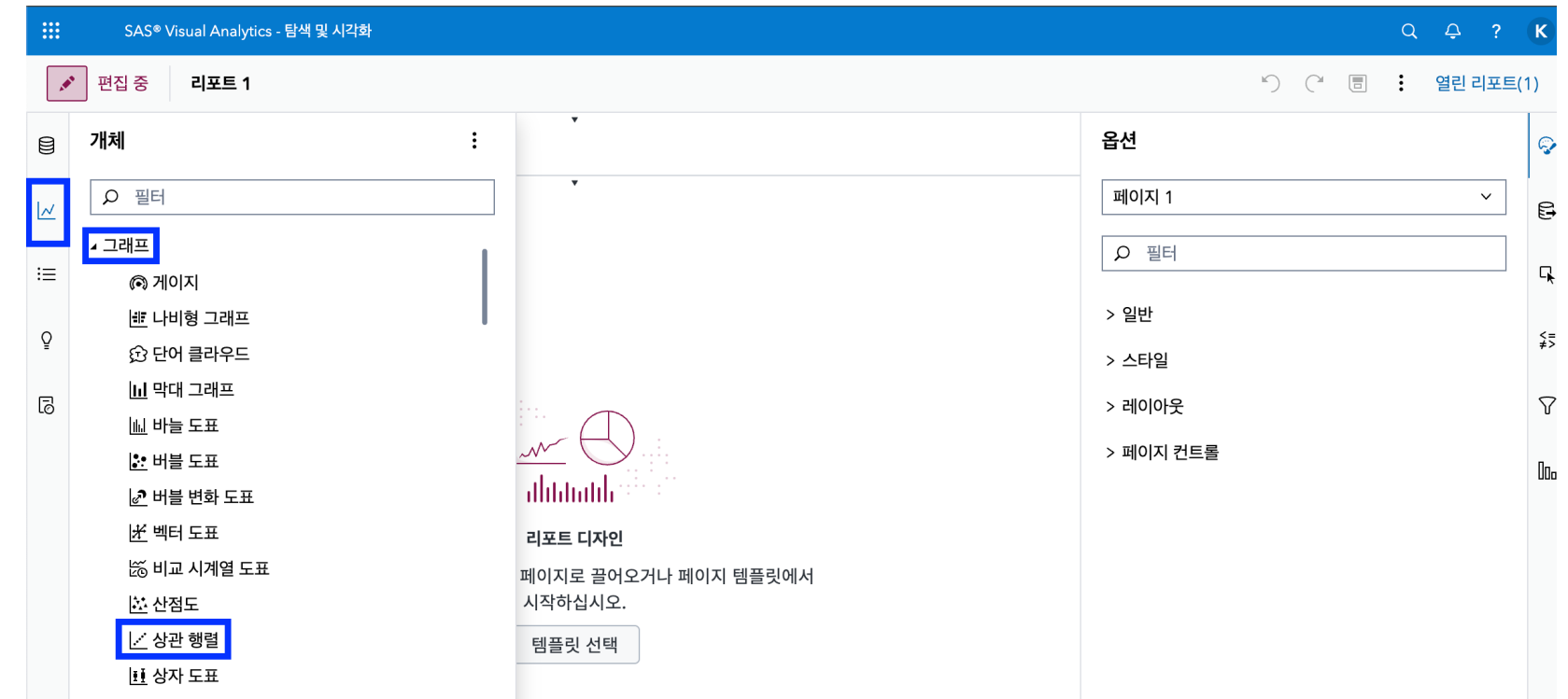
## 상관 행렬 그래프 (Correlation Matrix)

데이터 사이언티스트는 독립 변수와 목표(타겟) 변수 간의 패턴과 상관관계를 이해하기 위해 기초 그래프 시각화를 수행합니다.

시각화는 수동 생성 또는 자동 생성 방식 모두 가능합니다. 우선, 수동으로 그래프를 만드는 방법을 살펴봅시다.

새 페이지를 열고, 왼쪽에서 **개체** 탭을 선택하면 가능한 **그래프 및 모델 프로토타이핑 방법 목록**을 볼 수 있습니다. 본 프로젝트에선 예시로 **상관 행렬 그래프**를 선택해보겠습니다. 데이터 항목을 할당하고 모든 측정 변수를 추가합니다.

**상관 행렬 그래프**에서 마우스 우클릭 > **뷰 최대화**를 선택하면 하단에 **상관계수 값**이 표시되며 상관계수 항목을 클릭하면 **오름/내림차순으로 정렬** 가능합니다.

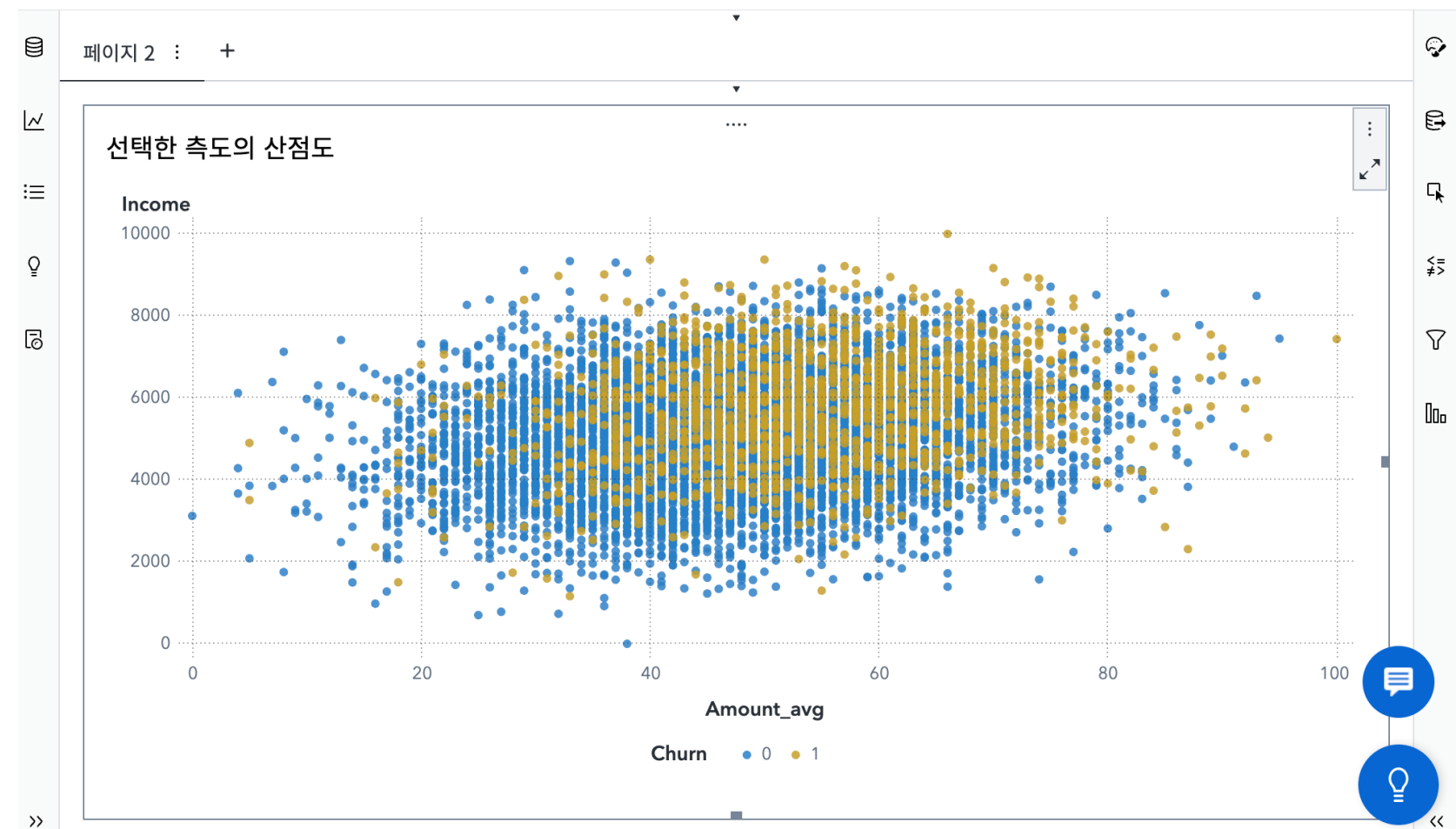
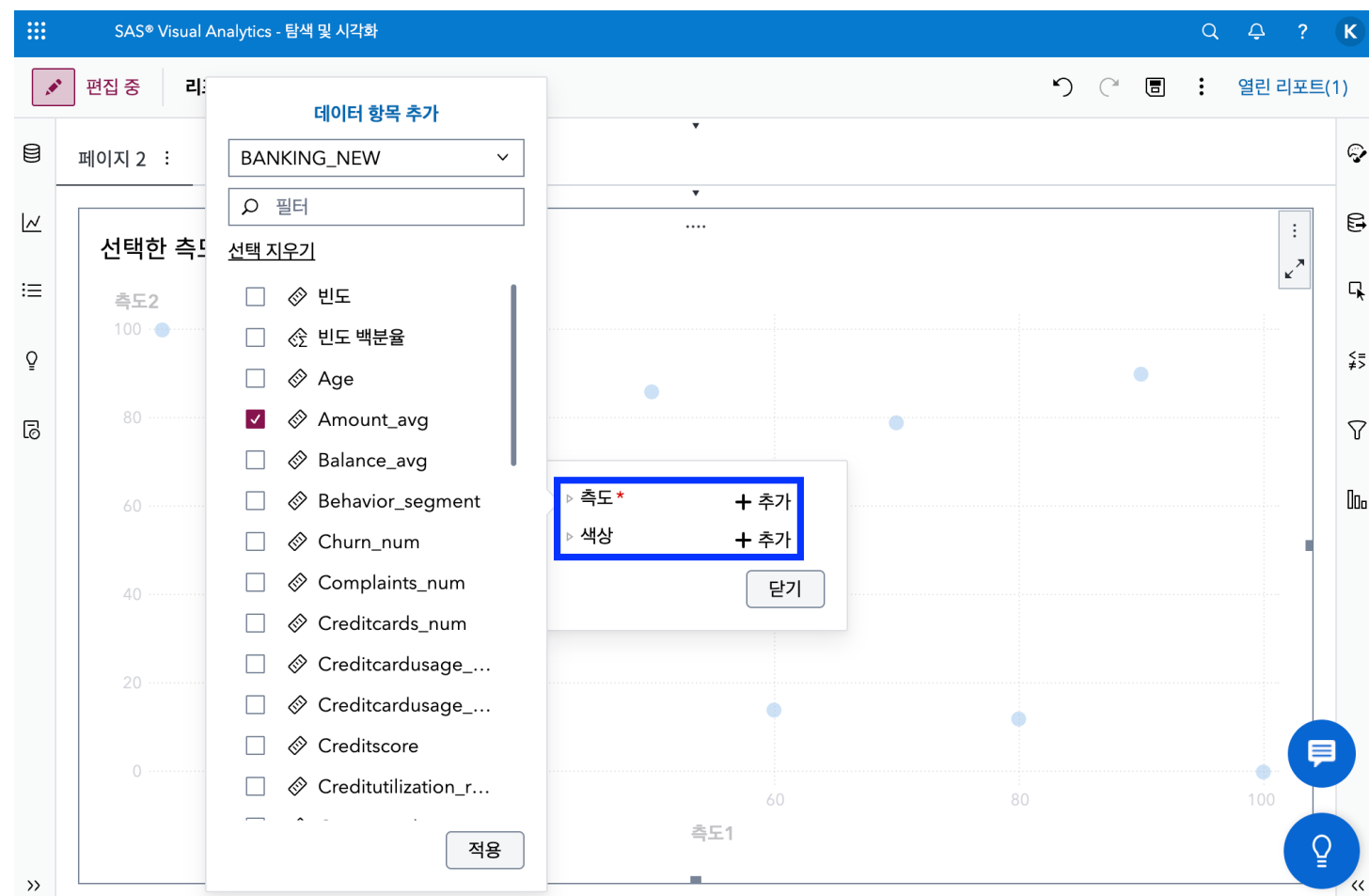


# 2. 시각적 탐색 – 증강 분석 (Augmented Analytics)

## 산점도 (Scatter Plots)

이번에는 새 페이지를 열고 산점도를 만들어 봅시다. 개체 탭에서 ‘산점도’을 선택합니다.

측도 속성에는 ‘Amount avg’ 와 ‘Income’ 변수를 추가하고, 색상속성에는 목표변수인 ‘Churn’을 추가합니다 (Churn 변수는 고객 이탈을 의미하며, 누가 이탈할지를 예측하는 것이 목표이기 때문에 모델의 목표변수로 설정됩니다).



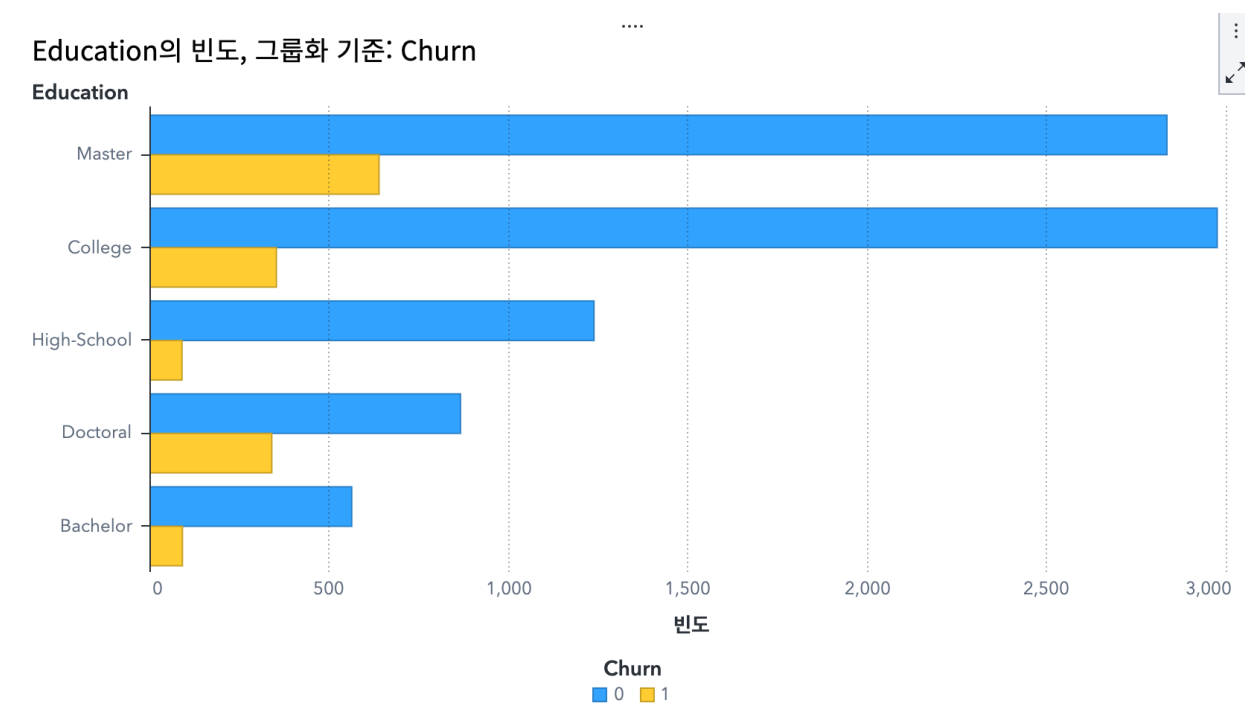
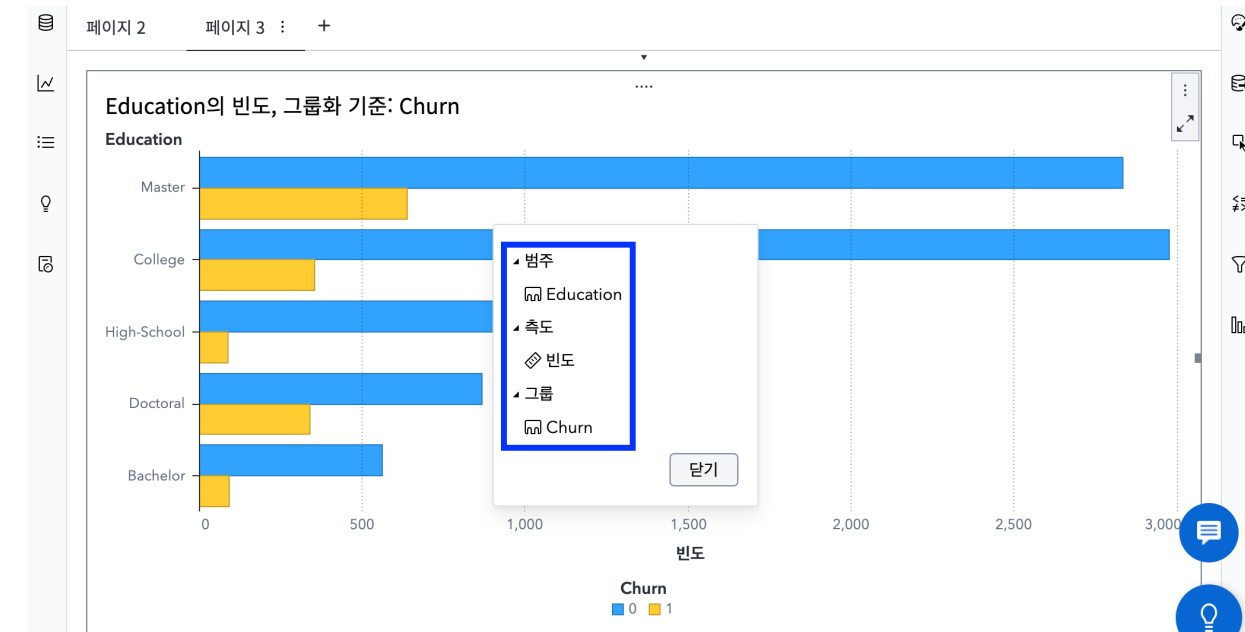
# 2. 시각적 탐색 – 증강 분석 (Augmented Analytics)

## 막대 그래프 (Nested Graphs: 특정 그래프 내에 또 다른 그래프를 포함하는 것을 의미)

SAS Visual Analytics을 이용하여 모델링 단계에서 편향 분석 평가를 자동으로 수행할 수 있지만, 시각적 탐색 단계에서 수동으로 확인하는 것도 가능합니다.

이번 단계에서는 시각적 탐색을 통해 수동으로 편향 여부를 점검해보겠습니다. 'Education' 변수를 범주로, 'Churn' 변수를 그룹 기준으로 지정하여, 교육 수준별 이탈 여부에 따른 고객 수를 보여주는 막대 그래프를 생성합니다.

이 그래프를 통해 각 교육 수준별로 이탈 여부에 따라 고객 수가 어떻게 분포되어 있는지를 비교할 수 있으며, 이를 통해 데이터에 잠재적인 편향이 존재하는지를 시각적으로 판단할 수 있습니다.



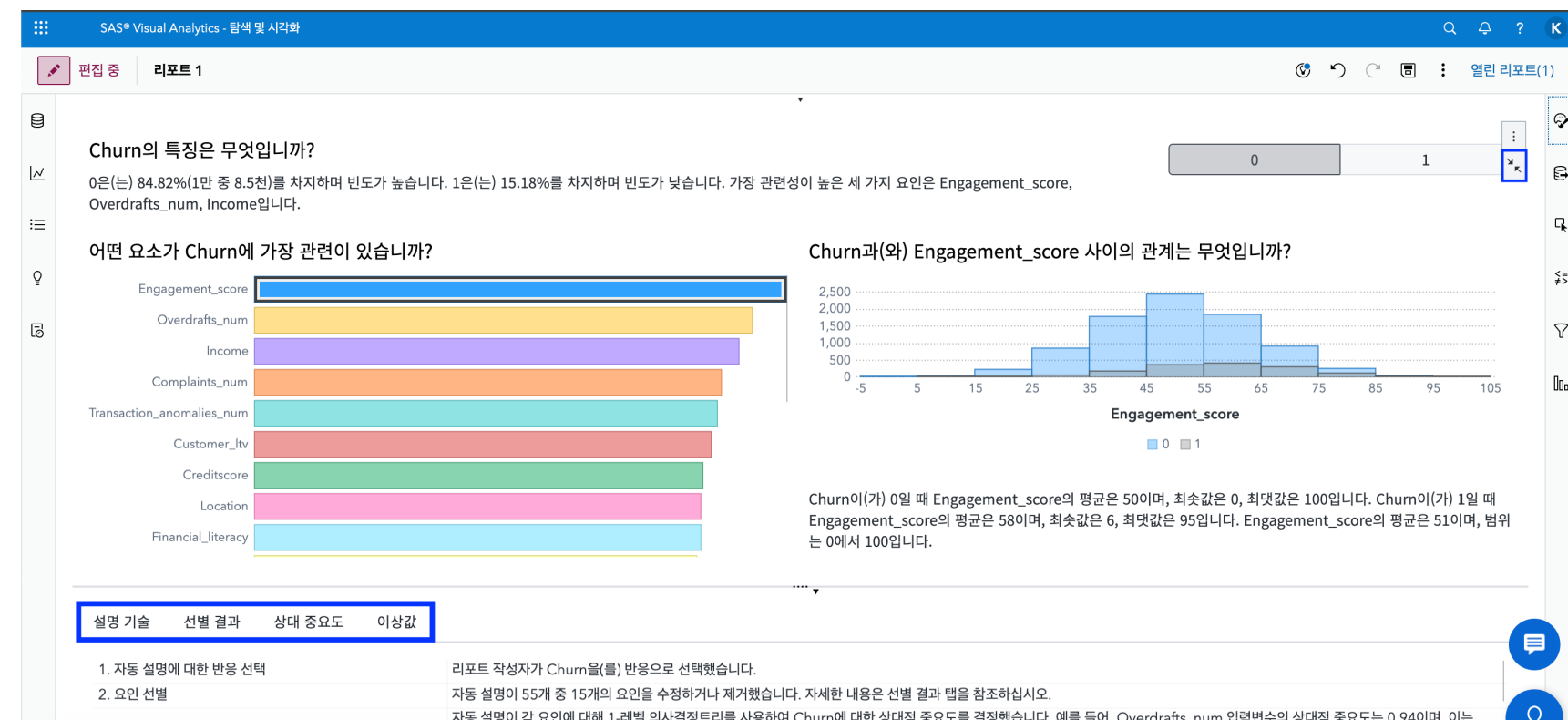
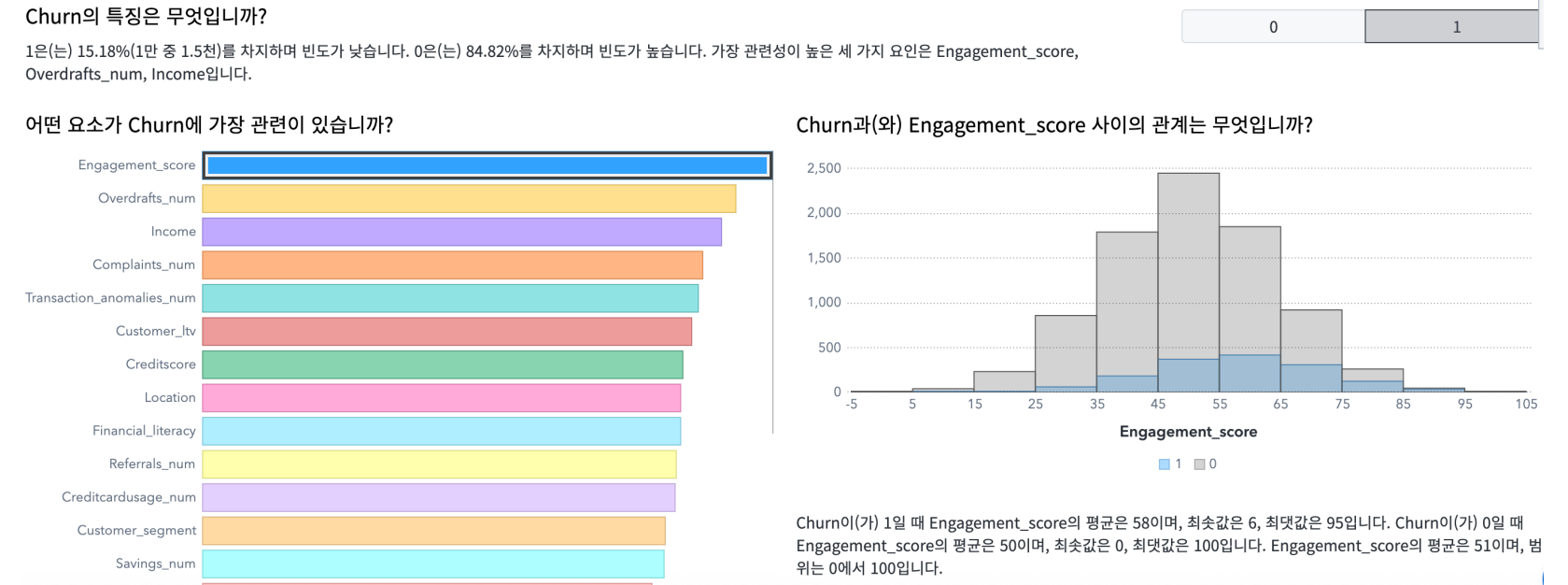
# 2. 시각적 탐색 – 증강 분석 (Augmented Analytics)

## 자동화된 탐색 및 시각화 (Automated Exploration & Visualization)

모델링 단계 전, 데이터에서 발견한 초기 인사이트를 제시하고 논의하는 과정은 중요합니다.

SAS Viya에서는 목표 변수에 대해 포괄적이고 상호작용형 리포트를 자동 생성하여 상당한 시간이 요구되는 작업을 단 몇 분 만에 완성할 수 있습니다.

새 페이지를 열고, 개체 탭 > 분석에서 자동 설명을 선택한 뒤, 'Churn' 변수를 반응변수로 추가합니다.



이 그래프는 변수 중요도(Variable Importance)를 기반으로 목표 변수와 가장 관련 있는 요인들을 보여줍니다.

막대 그래프에서 다른 변수를 클릭하면, 선택한 변수와 목표 변수간의 관계를 나타내는 관련 그래프가 오른쪽에 표시됩니다.

화면 오른쪽 상단의 최대화 버튼을 선택하면

1. 리포트의 전체적인 요약 (설명 기술)
2. 각 변수의 스크리닝 결과 (선별 결과)
3. 목표 변수에 영향을 미치는 변수들의 중요도 (상대 중요도)
4. 정상 패턴에서 벗어난 데이터 이상값

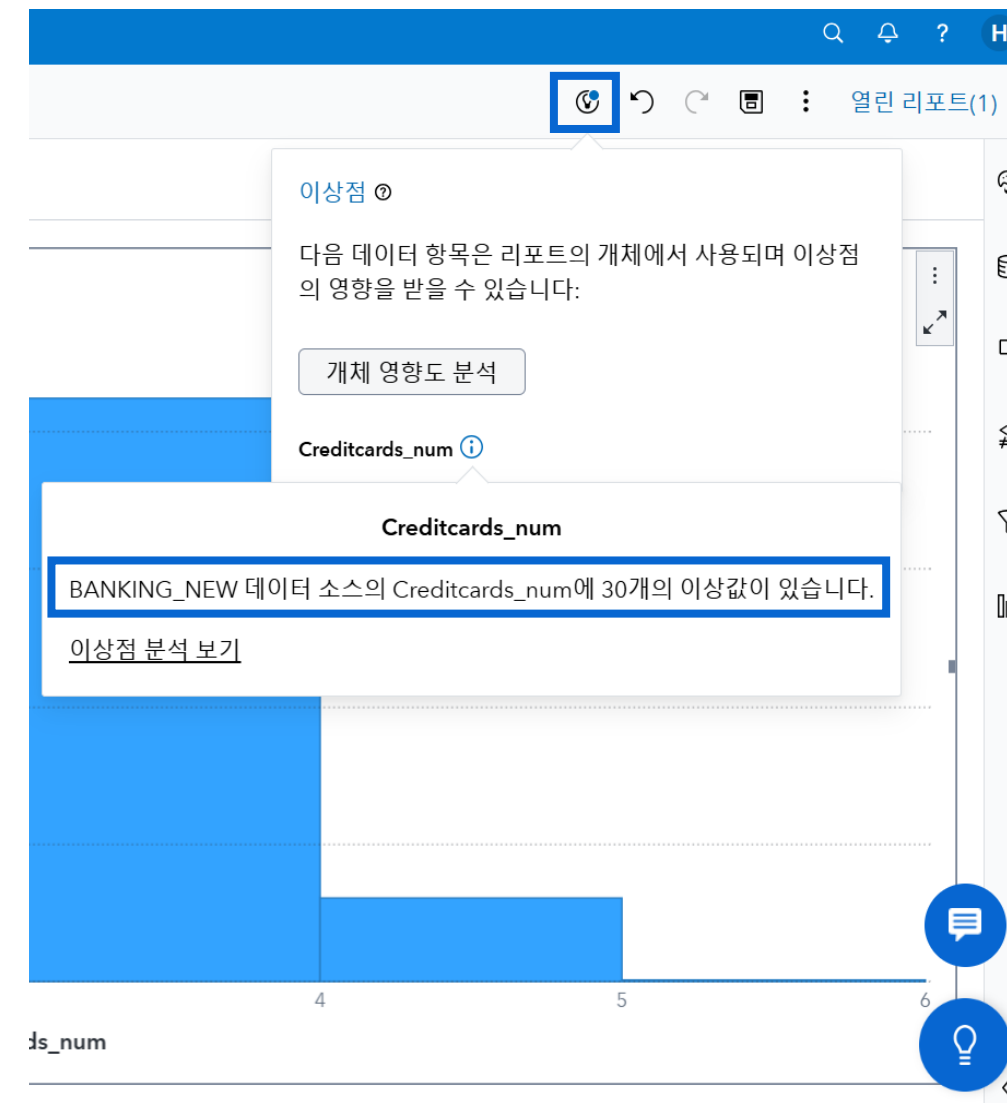
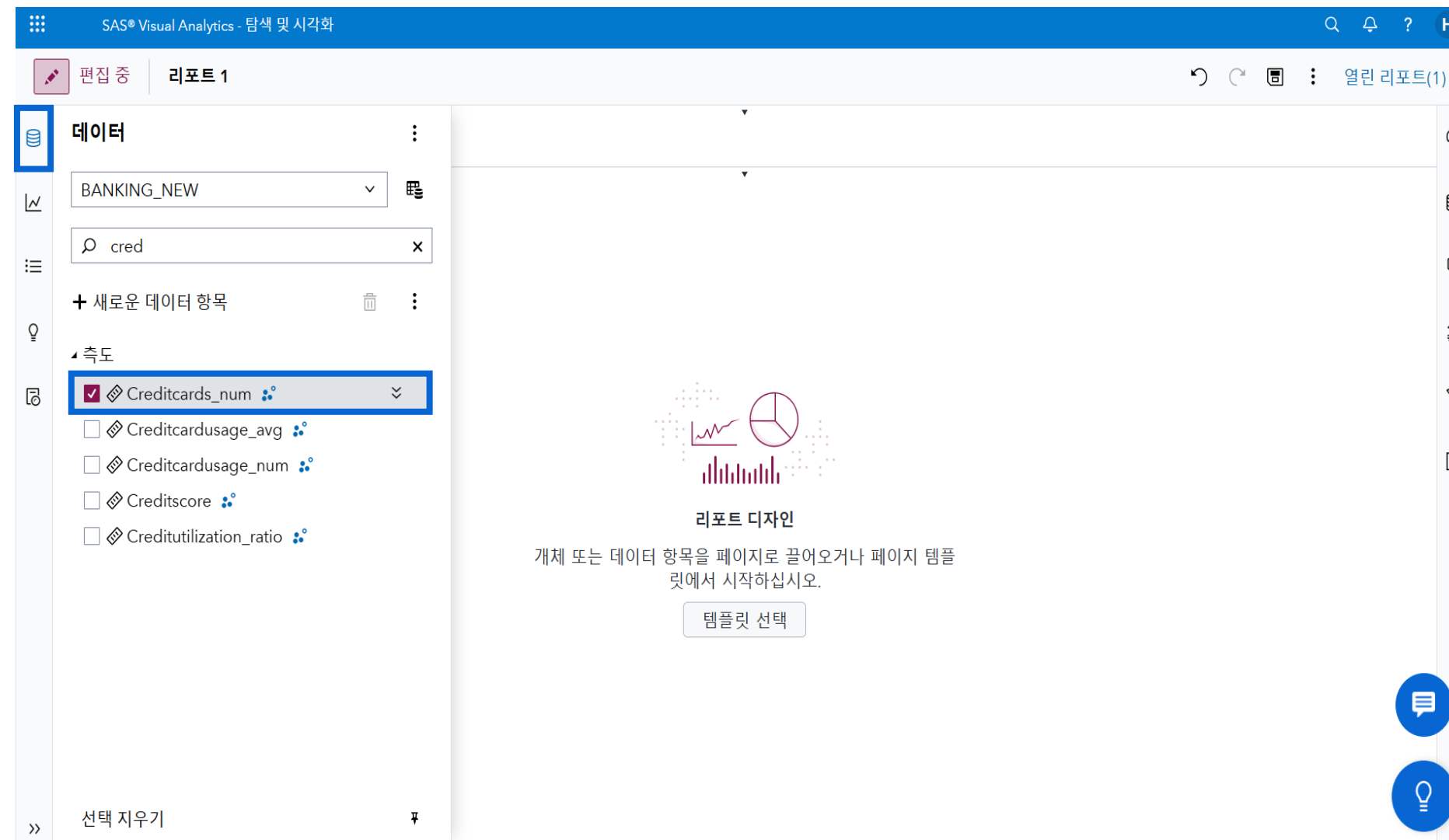
등 추가 정보를 확인할 수 있습니다.

# 3. 이상치 (Outlier) 탐지

# 3. 이상치 (Outlier) 탐지

이제 모델 프로토타입 생성 전에 모델 성능에 부정적인 영향을 줄 수 있는 이상치가 데이터에 포함되어 있는지 확인해보겠습니다.

이상치를 확인하기 위해 새 페이지를 열고, 좌측 데이터 탭에서 'Creditcards\_num' 변수를 더블 클릭합니다. 이후 우측 상단의 전구 아이콘을 클릭하면 SAS Visual Analytics가 자동으로 분석한 결과를 바탕으로 해당 변수에 이상치가 감지되었는지 확인할 수 있습니다.

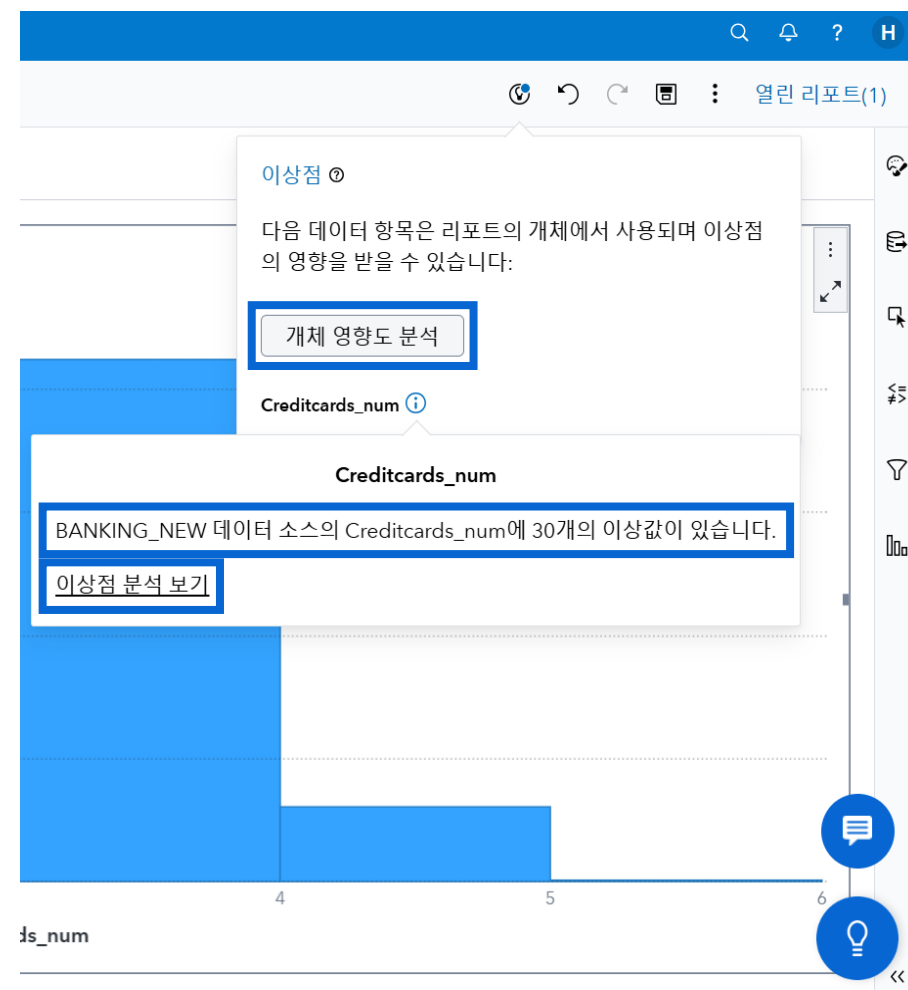


# 3. 이상치 (Outlier) 탐지

이상치를 확인하는 이유는 이상치가 데이터의 합계나 평균, 중간값 등 대표값에 미치는 영향을 사전에 점검하기 위함입니다.

개체 영향도 분석을 클릭하면 해당 변수가 모델에 어떤 영향을 주는지 분석할 수 있으며, 자동으로 이상치 여부도 평가됩니다. 이번 경우에는 이상치가 모델 성능에 큰 영향을 미치지 않는 것으로 나타났습니다. 또한 이상점 분석 보기를 선택하면 이상치가 대표값에 얼마나 영향을 주는지를 보여주는 리포트를 확인할 수 있습니다.

이런 인사이트를 기반으로 데이터 사이언티스트는 이상치를 제거하거나 유지할지 결정할 수 있습니다.



## 이상점

다음 데이터 항목은 리포트의 개체에서 사용되며 이상점의 영향을 받을 수 있습니다:

100% 이상점의 영향을 받는 개체가 없습니다.

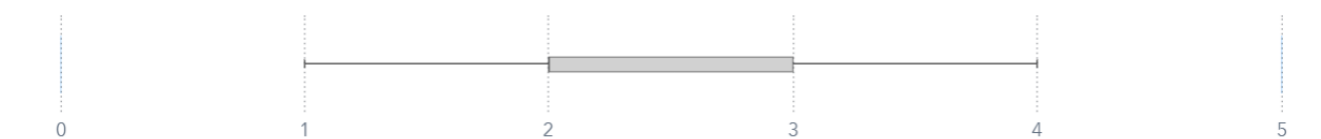
## Creditcards\_num

이상점의 영향을 받는 개체가 없습니다.

## Creditcards\_num의 이상점

Creditcards\_num의 이상값이 있습니까?

Creditcards\_num에 30개의 이상값이 있습니다. 이러한 이상값은 전체 합계, 평균 또는 중위수(를) 5% 이상 변경시키지 않습니다.



이 이상점의 상세 정보는 무엇입니까?

|   | Creditcards_num | Savings_num | Creditutilization_ratio | Amount_avg | Balance_avg  | Income |
|---|-----------------|-------------|-------------------------|------------|--------------|--------|
| • | 5               | 19          | 52.969984758            | 47         | 547.63975299 | 5113   |
| • | 0               | 17          | 11.546247511            | 36         | 471.9675259  | 5029   |
| • | 0               | 17          | 41.609299478            | 43         | 397.32784684 | 4402   |

닫기

# 4. 모델 프로토타입 생성

# 모델 프로토타입이란

## 모델 프로토타입 (Quick Model Prototyping)

최종 모델을 만들기 전, 어떤 변수들이 유의미한지, 어느 정도 정확도가 나오는지 빠르게 실험해보는 **‘초기 모델’**

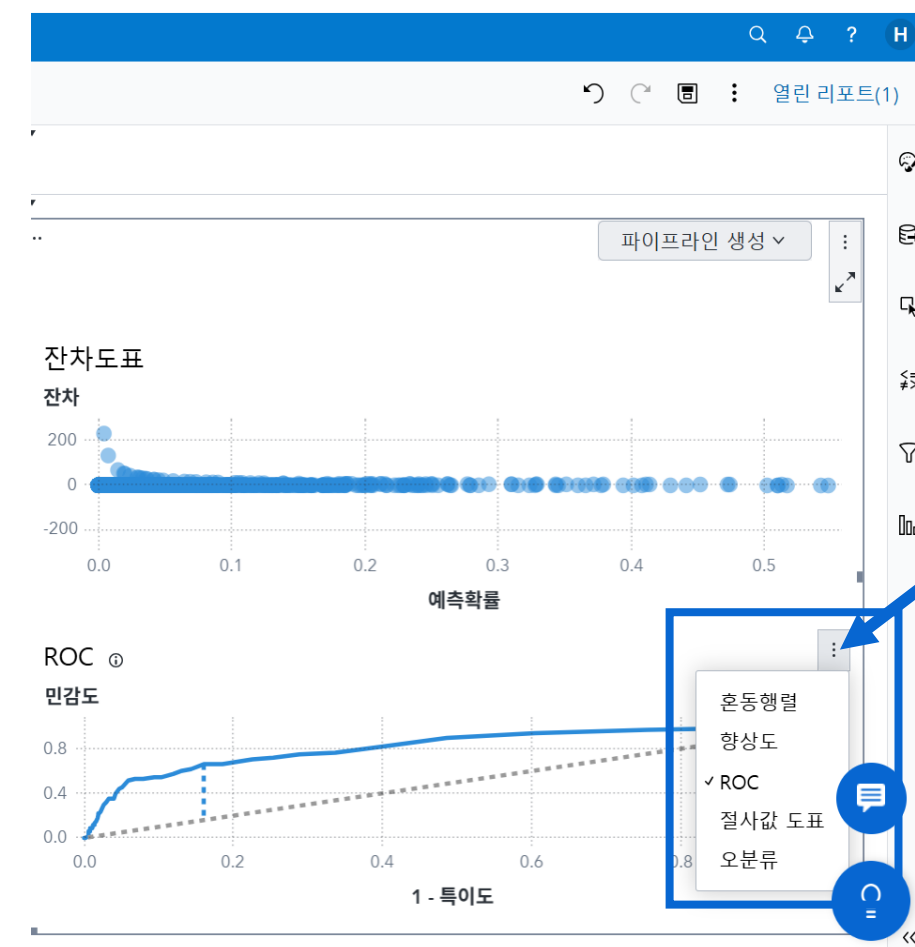
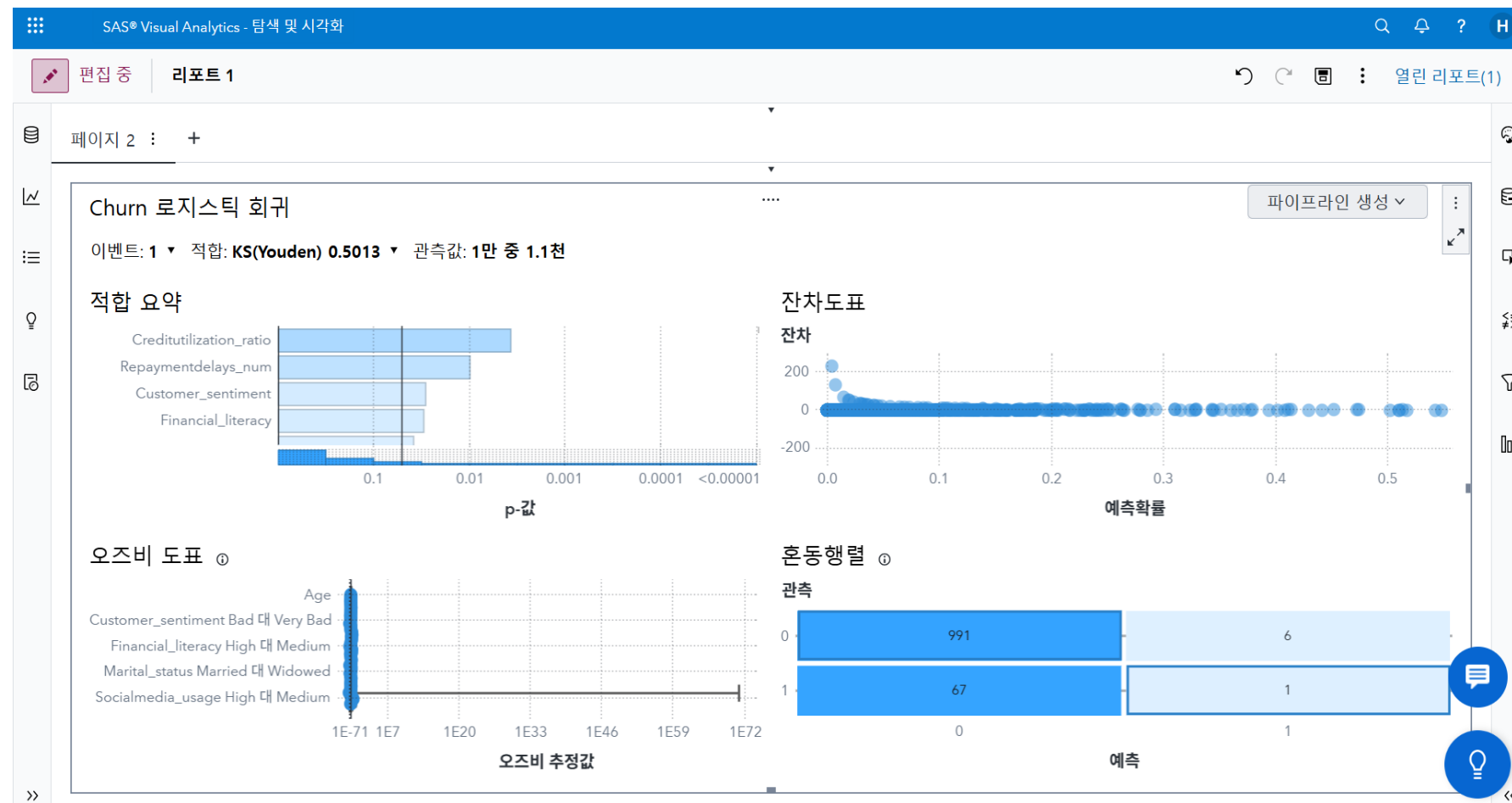
| 목적        | 설명                            |
|-----------|-------------------------------|
| 정확도 점검    | 모델이 목표에 맞는 성능을 낼 수 있을지 빠르게 확인 |
| 변수 중요도 파악 | 어떤 변수가 예측에 가장 큰 영향을 미치는지 파악   |
| 방향 설정     | 모델 개발 전 어떤 접근법이 유망한지 결정       |

# 4. 모델 프로토타입 생성

## 로지스틱 회귀 (Logistic Regression)

모델을 본격적으로 구축하기에 앞서, 빠르게 **프로토타입**을 만들어 **모델의 정확도와 예측에 중요한 변수가 무엇인지 확인**하는 과정이 필요합니다.

**개체** 탭에서 **로지스틱 회귀**를 선택한 뒤, **Churn**을 반응(응답) 변수로 지정합니다. **연속 효과**에는 **모든 변수**를 포함하고, **분류 효과** 변수에서는 **'이름(name)'과 '성(surname)'을 제외한 나머지를** 선택합니다. 이후 우측의 **점 세 개 아이콘(작업 버튼)**을 클릭하여 모델 평가 그래프를 혼동 행렬에서 ROC 차트로 변경하는 것 또한 가능합니다.

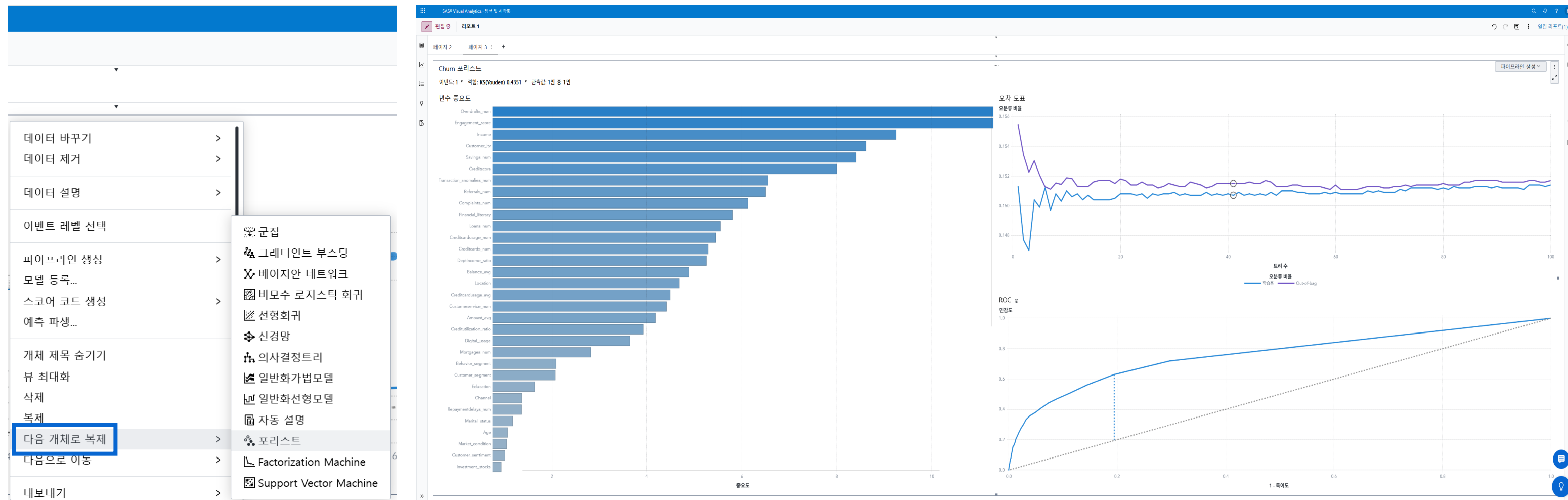


# 4. 모델 프로토타입 생성

## 랜덤 포리스트 (Random Forest)

로지스틱 회귀 페이지를 복제하면 선택한 변수가 모든 모델에 동일하게 적용되며, 다양한 모델의 잠재적 성능을 쉽게 비교할 수 있습니다.

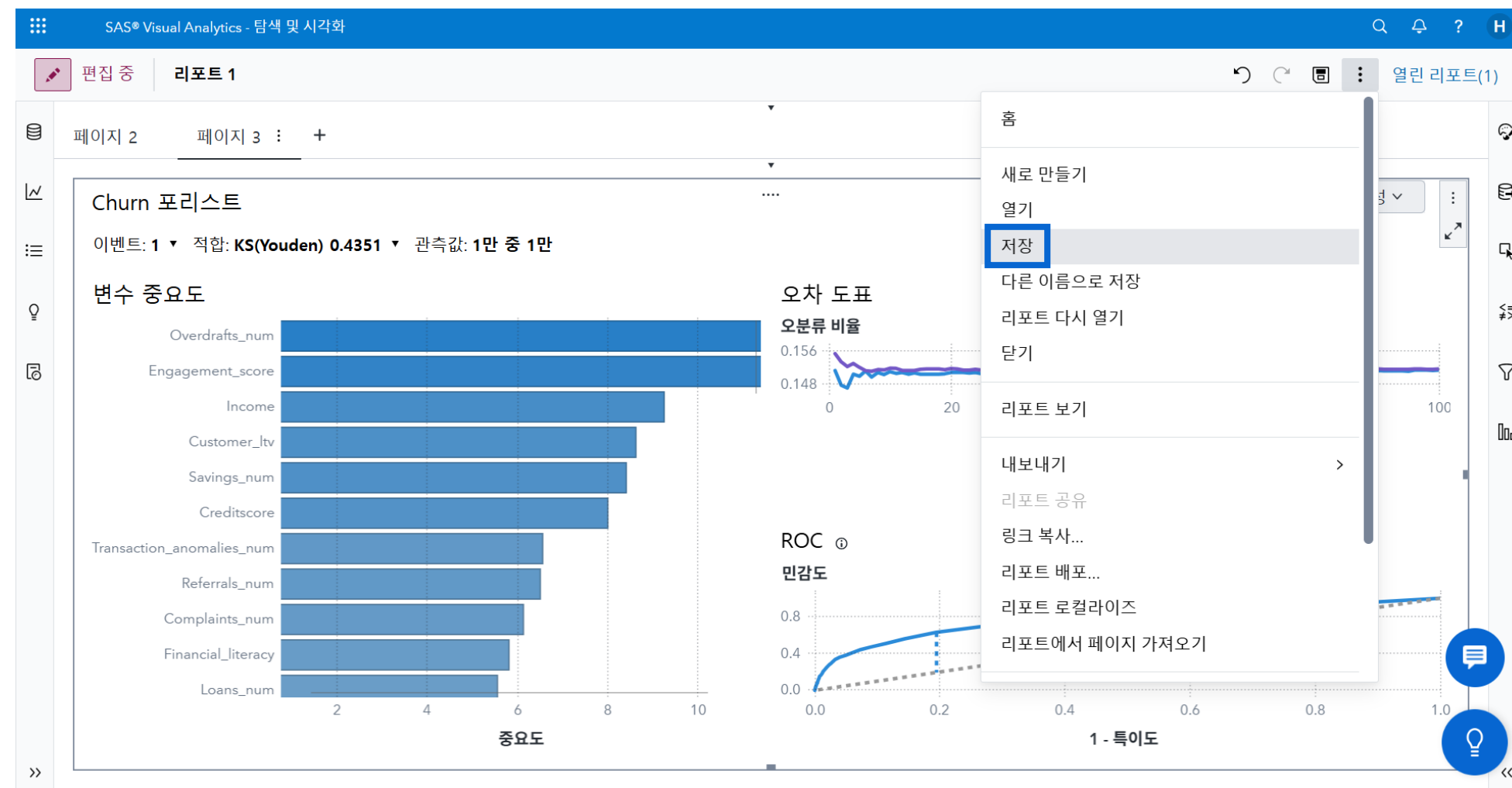
모델 성능 비교를 위해 로지스틱 회귀 템플릿을 마우스 우클릭 > 다음 개체로 복제 > '포리스트'를 선택합니다. 생성된 포리스트 템플릿을 다시 우클릭하여 '다음으로 이동'을 선택하면 새 페이지로 이동되어 독립적으로 확인 할 수 있습니다.



# 인사이트 저장 및 모델링 단계로 이동

인사이트를 충분히 수집했다면, 이제 리포트를 저장하고 모델링 단계로 넘어갈 차례입니다.

화면 우측 상단의 **저장**을 클릭 시 **탐색 및 시각화**에서 생성한 리포트와 모든 그래프가 저장됩니다. 이후, **공유 및 협업** 탭을 통해 프로젝트의 읽기(Read) 권한을 부여하거나, 추가 분석을 위해 읽기 및 쓰기 (Read and Write) 권한을 부여해 협업이 가능합니다.



The screenshot shows the SAS navigation menu. The menu is organized into several categories: "분석 라이프사이클" (Analysis Lifecycle), "스트리밍 분석" (Streaming Analytics), and "관리" (Management). The "공유 및 협업" (Share and Collaborate) option is highlighted in the "분석 라이프사이클" section. Other options include "정보 애셋 검색", "데이터 관리", "탐색 및 시각화", "모델 생성", "모델 관리", "의사결정 생성", "코드 및 플로우 개발", "스트리밍 분석 관리", "스트리밍 프로젝트 디자인", "사용자 정의 그래프 생성", "테마 관리", "Lineage 탐색", "환경 관리", and "워크플로우 관리".

# 5. 모델 구축

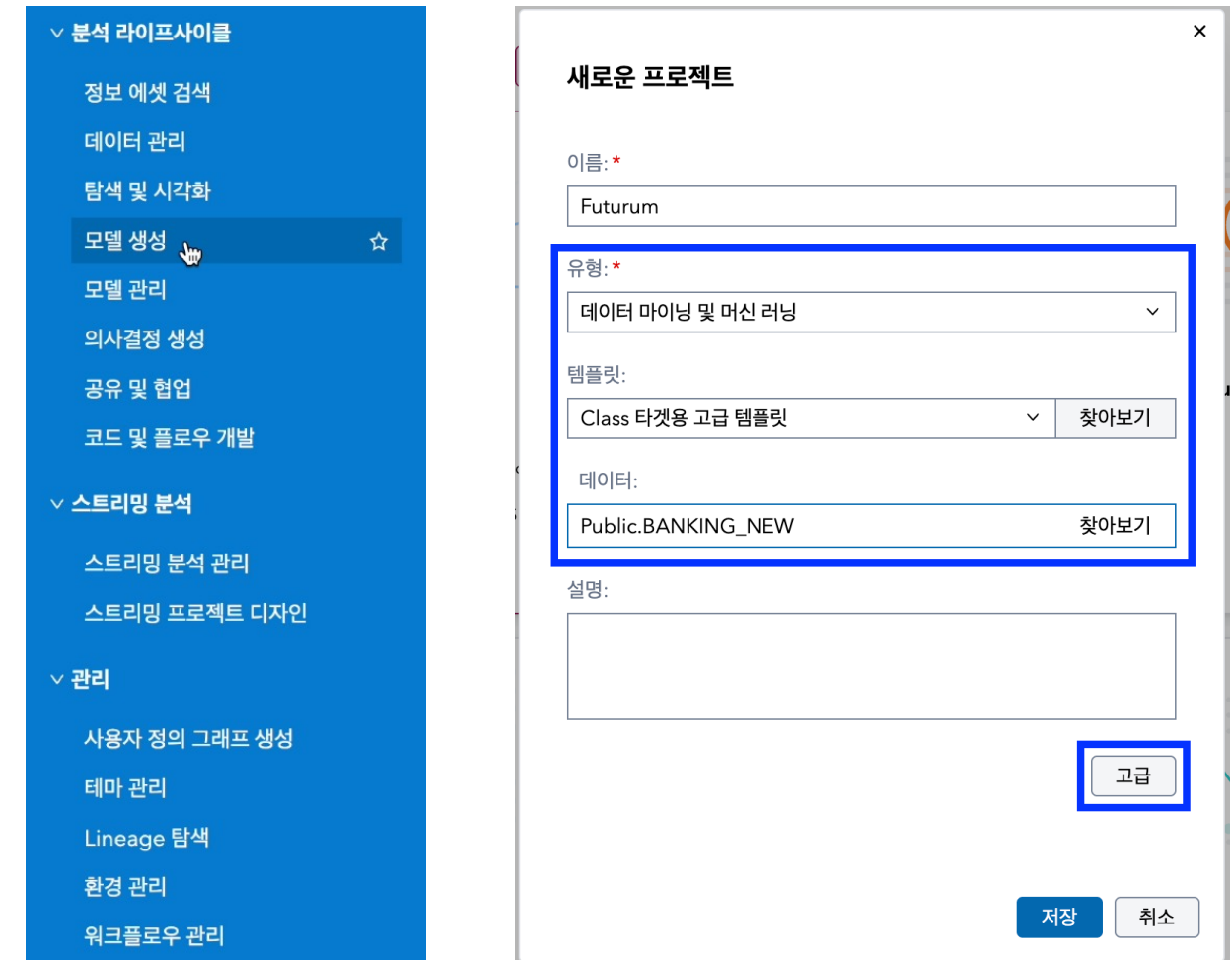
# 5. 모델 구축

## 프로젝트 만들기

데이터 탐색을 마친 후, 좌측 메뉴에서 **모델 생성**으로 이동합니다. 이동할 시 화면은 **SAS Visual Analytics**에서 **SAS Model Studio**로 전환됩니다.

우측 상단에 **새로운 프로젝트**를 클릭하면 새로운 창이 열립니다. 유형은 **'데이터 마이닝 및 머신 러닝'**으로 선택합니다. 템플릿은 **'Class 타겟용 고급 템플릿'**으로 선택합니다. 마지막으로, **BANKING\_NEW** 데이터 세트를 선택합니다.

템플릿의 경우 **SAS**에서 제공하며, **모델링을 위한 파이프라인을 자동 생성**합니다. 사용자가 구성하단의 **고급** 버튼을 클릭하면 **기본 설정을 조정** 할 수 있으며, 데이터 분할, 이벤트 기반 표본추출, 결측치 비율이 임계값을 초과하는 변수 제외 등 기타 구성 설정 조정이 가능합니다.

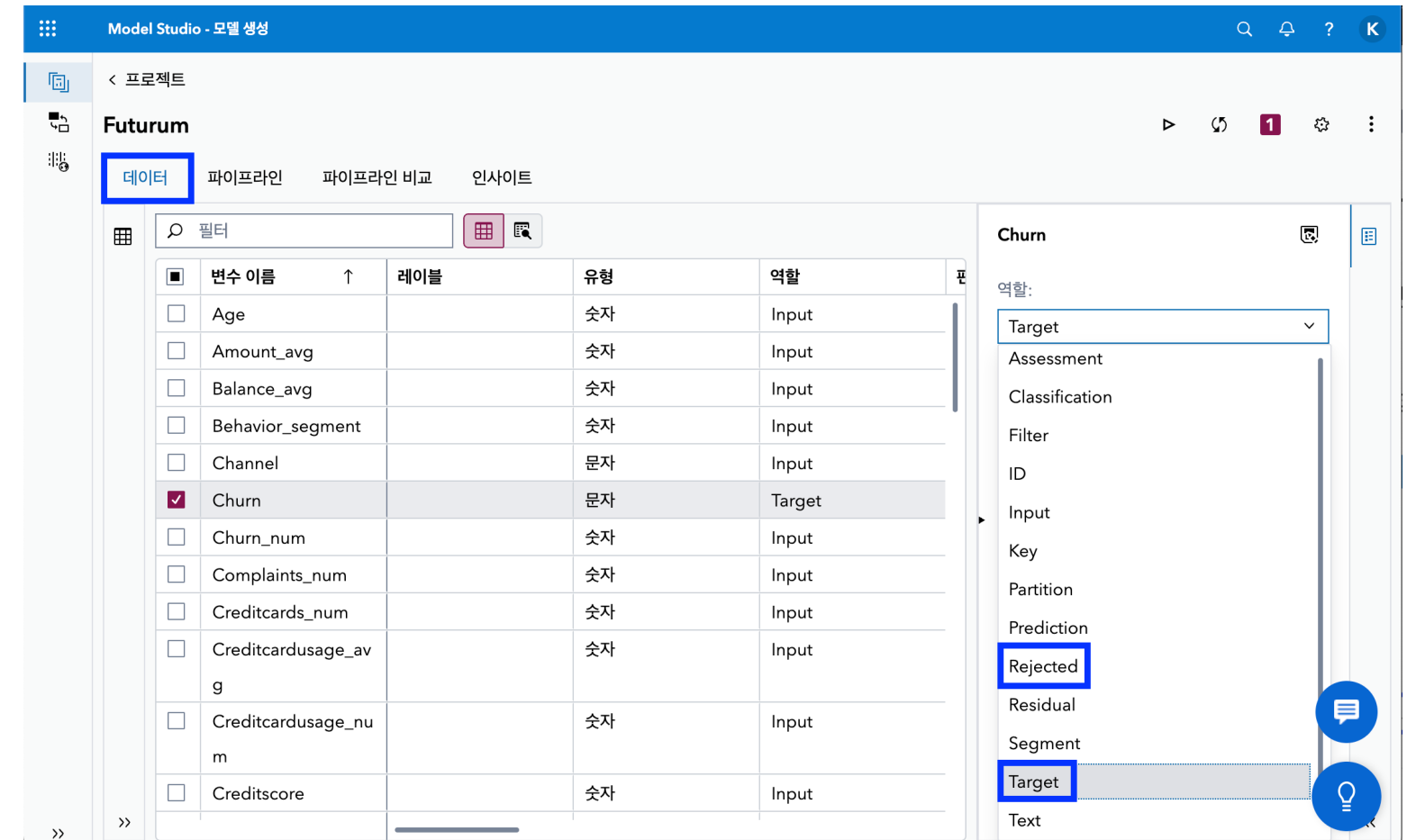


# 5. 모델 구축

## 변수 역할 수정

프로젝트가 생성되었다면 이제 **데이터** 탭에서 **모델링에 사용할 변수를 선택할** 차례입니다. 더 나아가, **데이터** 탭에서 **변수의 역할을 변수의 속성을 고려하여 수정**해보겠습니다.

Churn은 기존 'Input' 역할에서 'Target'으로, Age, Engagement\_score, Gender, Loyalty\_program, Name, Surname의 경우 'Reject'로 변경합니다. 그리고 나머지 변수들은 'Input' 역할을 유지합니다.



# 5. 모델 구축

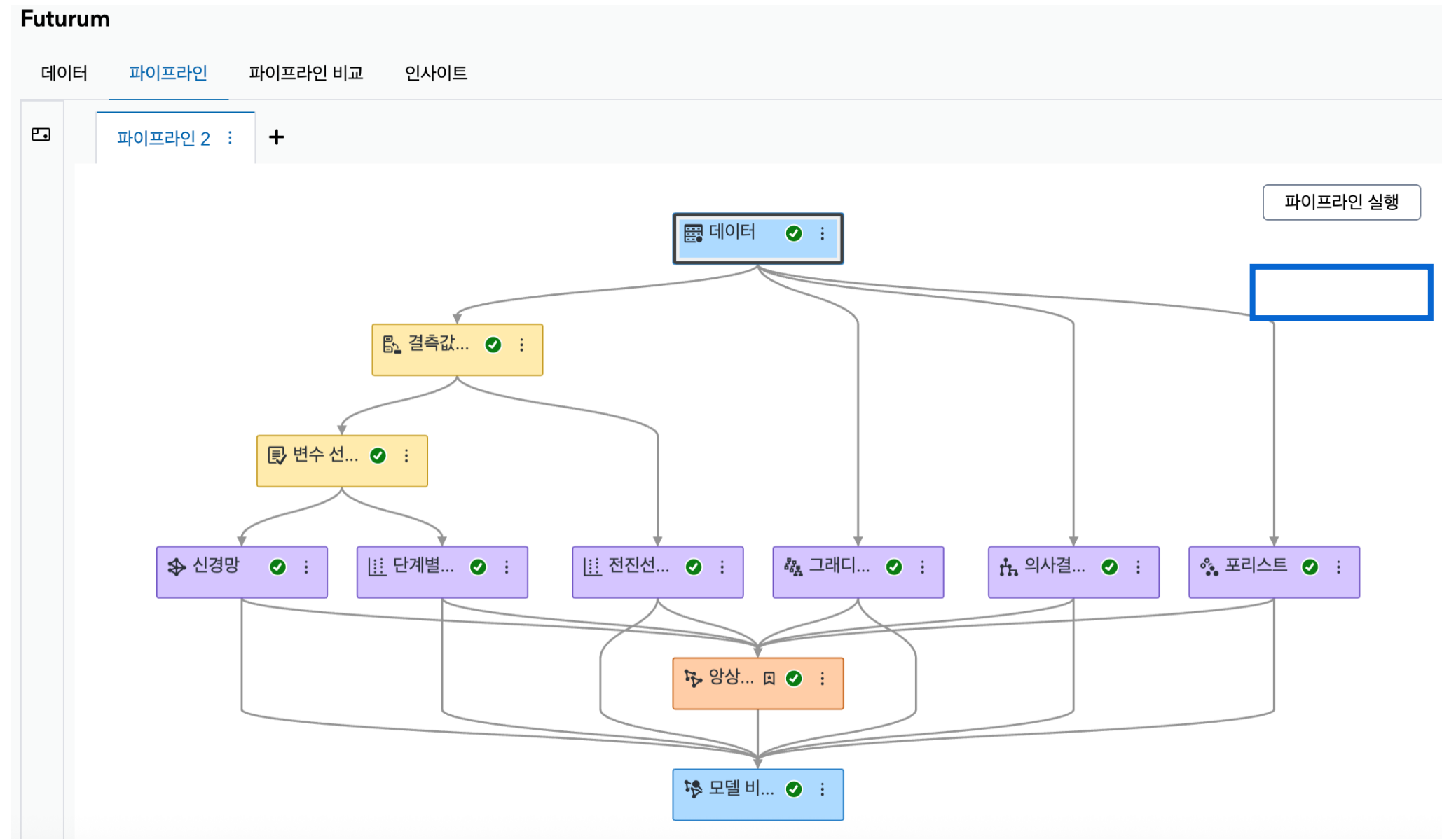
## 파이프라인 실행

이제 **파이프라인을 실행**하기 위해 **파이프라인** 탭으로 넘어가겠습니다.

우측 상단에 **파이프라인 실행**을 클릭하면 파이프라인 내 모든 모델을 자동으로 학습 및 비교합니다.

해당 파이프라인은 총 **7개의 모델**을 생성합니다:

1. 신경망
2. 단계별 로지스틱 회귀
3. 전진선택 로지스틱 회귀
4. 그래디언트 부스팅
5. 의사결정트리
6. 포리스트
7. 위 모델들을 결합한 앙상블 모델



위 사진과 같이 파이프라인 실행이 완료되었을 시 노드 이름 옆으로 체크 마크가 새로 생긴 것을 확인하실 수 있습니다.

# 5. 모델 구축

## 사용자 정의 파이프라인 저장 및 공유

데이터 사이언티스트는 모델링 노드를 추가하거나 제거 또는 각 모델의 하이퍼파라미터 및 옵션을 설정하여 자신만의 맞춤형 파이프라인을 생성할 수 있습니다.

**Exchange에 저장**을 클릭시 현재 구성된 파이프라인을 다른 부서에서도 활용할 수 있습니다 (저장된 파이프라인은 **Exchange**에 추가된 것을 확인하실 수 있습니다).

또한, **특정 모델링 노드를 직접 수정하거나 새로 만들고 싶은 경우에는** 수정 후 해당 노드에서 우클릭 > **다른 이름으로 저장** 을 선택하면 사용자 정의 노드로 저장할 수 있습니다.



# 6. 모델 비교 (Model Comparisons)

# 6. 모델 비교

## 모델 결과 및 평가 지표

완성된 파이프라인에서 **모델 노드 우클릭 > 결과**를 선택합니다.  
**펼치기(화살표)**를 누를 시 하면 **계수와 그래프에 대한 구체적 설명이** 제공됩니다.

(예: 단계별 로지스틱 회귀 모델에서는 각 파라미터의 t-value, 추정계수, 모델 선택 요약, 적합도 통계, 스코어 코드 등을 제공합니다.)

The screenshot shows the SAS Model Studio interface for a logistic regression model. On the left, a context menu is open with the '결과' (Results) option highlighted. The main window displays the '모수별 t 값' (t-values for coefficients) bar chart, where the '모수' (Coefficients) are listed on the x-axis and 't 값' (t-values) on the y-axis. Below the chart is a table of selected variables:

| 단계 | 추가된 효과         | 효과 수 | SBC        | 최적 SBC |
|----|----------------|------|------------|--------|
| 0  | Intercept      | 1    | 5,119.2121 | 0      |
| 1  | Overdrafts_num | 2    | 4,889.9359 | 0      |

On the right, there are sections for '모수 추정값' (Coefficient Estimates) and '회귀 적합통계량' (Regression Statistics), including metrics like Intercept, Overdrafts\_num, Location, Channel, Creditscore, and Creditcardu.

이제 **평가** 탭으로 넘어가봅시다. 여기서는 선택한 **모델의 적합도 통계 및 시각화 결과**를 확인할 수 있습니다.

본 페이지에서 총 **3가지의 그래프**를 볼 수 있으며 (누적 향상도, ROC 곡선, 이벤트 분류), 그래프의 오른쪽 상단 ⓘ를 클릭해 더 자세한 설명을 보실 수 있습니다. 또한, **Model Studio**는 **다양한 그래프 옵션**을 제공해 필요에 맞게 활용할 수 있습니다.

The screenshot shows the '평가' (Evaluation) tab in SAS Model Studio. It displays three main visualizations: '향상도 리포트' (Improvement Report) showing cumulative improvement curves, 'ROC 리포트' (ROC Report) showing the ROC curve, and '이벤트 분류' (Event Classification) showing the distribution of predicted classes. A dropdown menu is open over the Improvement Report, listing various metrics: 누적향상도, 향상도, 이득, 반응검출률, 누적반응검출률, 반응률, and 누적반응률. The ROC curve shows a curve above the diagonal line, indicating good model performance. The event classification section shows a bar chart for '백분율 도표' (Percentage Chart).

# 6. 모델 비교

## 파이프라인 모델 비교

모델 결과 창을 종료하고 다시 파이프라인 화면으로 돌아가 '모델 비교' 노드에서 결과를 보기 전 사용자가 **기준 지표** (정확도, KS 등) 및 **사용할 데이터 분할**(train, valid, test)을 설정할 수 있습니다. 본 프로젝트에서는 기본 설정 그대로 유지하겠습니다.

'모델 비교' 노드 우클릭 > 결과를 선택하면 파이프라인 내 여러 모델들의 성능 비교 결과를 확인할 수 있습니다.



Model Studio - 모델 생성

Futurum > "모델 비교" 결과

노드 평가

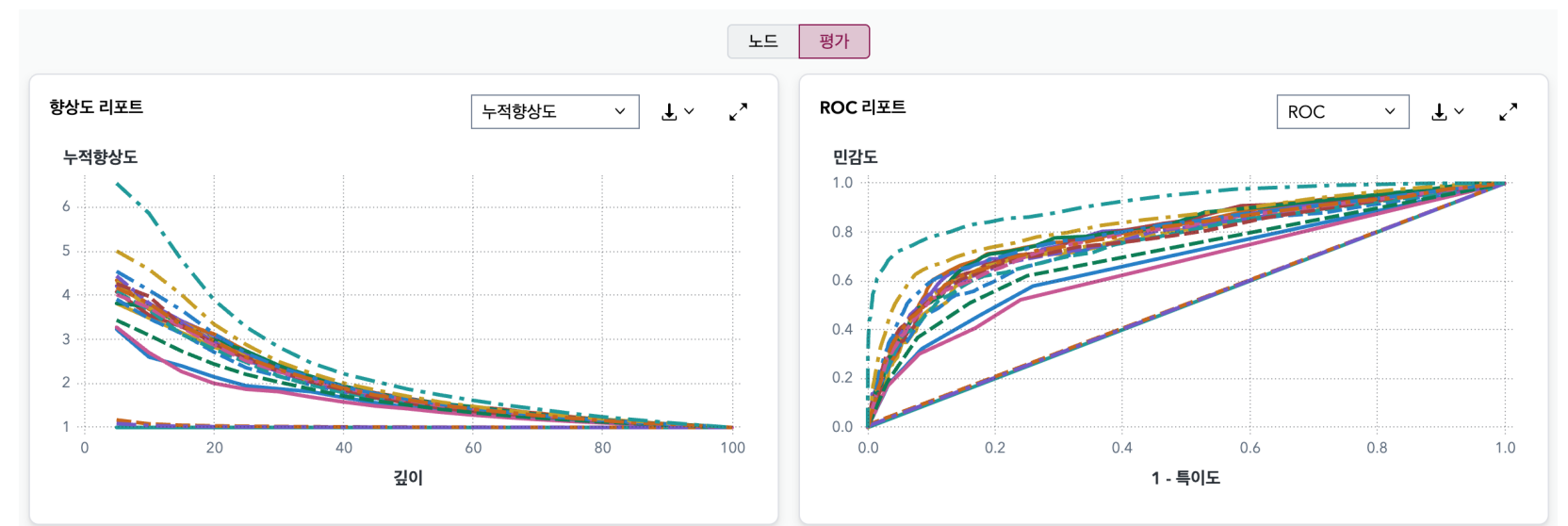
| 챔피언 | 이름          | 알고리즘...   | KS(You... | 정확도    | 평균제곱... | ROC 아... | 누적향상도  | 누적반응... | 절사값    | 데이터 ... | 깊이 |
|-----|-------------|-----------|-----------|--------|---------|----------|--------|---------|--------|---------|----|
| ★   | 양상블         | 양상블       | 0.5230    | 0.8510 | 0.1072  | 0.7993   | 3.6842 | 36.8421 | 0.5000 | TEST    | 1  |
|     | 그래디언트 부스팅   | 그래디언트 부스팅 | 0.3195    | 0.8480 | 0.1259  | 0.6798   | 2.5997 | 25.9975 | 0.5000 | TEST    | 1  |
|     | 신경망         | 신경망       | 0.5037    | 0.8600 | 0.1020  | 0.7934   | 3.8158 | 38.1579 | 0.5000 | TEST    | 1  |
|     | 단계별 로지스틱 회귀 | 로지스틱 회귀   | 0.4711    | 0.8550 | 0.1072  | 0.7784   | 3.4868 | 34.8684 | 0.5000 | TEST    | 1  |
|     | 의사결정트리      | 의사결정트리    | 0         | 0.8480 | 0.1289  | 0.5000   | 1      | 10      | 0.5000 | TEST    | 1  |
|     | 포리스트        | 포리스트      | 0.5183    | 0.8560 | 0.1039  | 0.8018   | 3.7500 | 37.5000 | 0.5000 | TEST    | 1  |
|     | 저지서택 로지     | 로지스틱 회귀   | 0.4805    | 0.8530 | 0.1050  | 0.7915   | 3.5524 | 35.5243 | 0.5000 | TEST    | 1  |

속성

| 속성 이름                  | 속성 값   |
|------------------------|--------|
| selectionCriteriaClass | KS 통계량 |

양상블 모델이 KS지표 기준으로 가장 우수한 성능을 보였다는 것을 알 수 있습니다.

평가 탭에서는 각 모델별로 학습, 검증, 테스트 데이터에 대한 각 모델의 성능 지표의 시각화된 그래프 및 구체적인 적합 통계량 값을 알 수 있습니다.



# 7. 모델 해석 (Explainability)

# 7. 모델 해석 (Explainability)

모델 해석 방법의 분류:

Global Interpretability vs Local Interpretability

## Global Interpretability

전체 데이터 세트를 기반으로 중요한 특성 (feature)과 변수 간 상호작용을 분석

- 전체 패턴과 추세를 보여주는 것을 목표
- “어떤 특성이 모델 예측에 영향을 많이 주는가?”

## Local Interpretability

개별 데이터 포인트에 집중하여 예측을 설명

- 모델이 특정 결과에 도달한 이유를 보여주는 데 도움
- “이 특정 집( $100m^2$ )의 가격은 면적에 따라 선형적으로 변한다.”

# 7. 모델 해석 (Explainability)

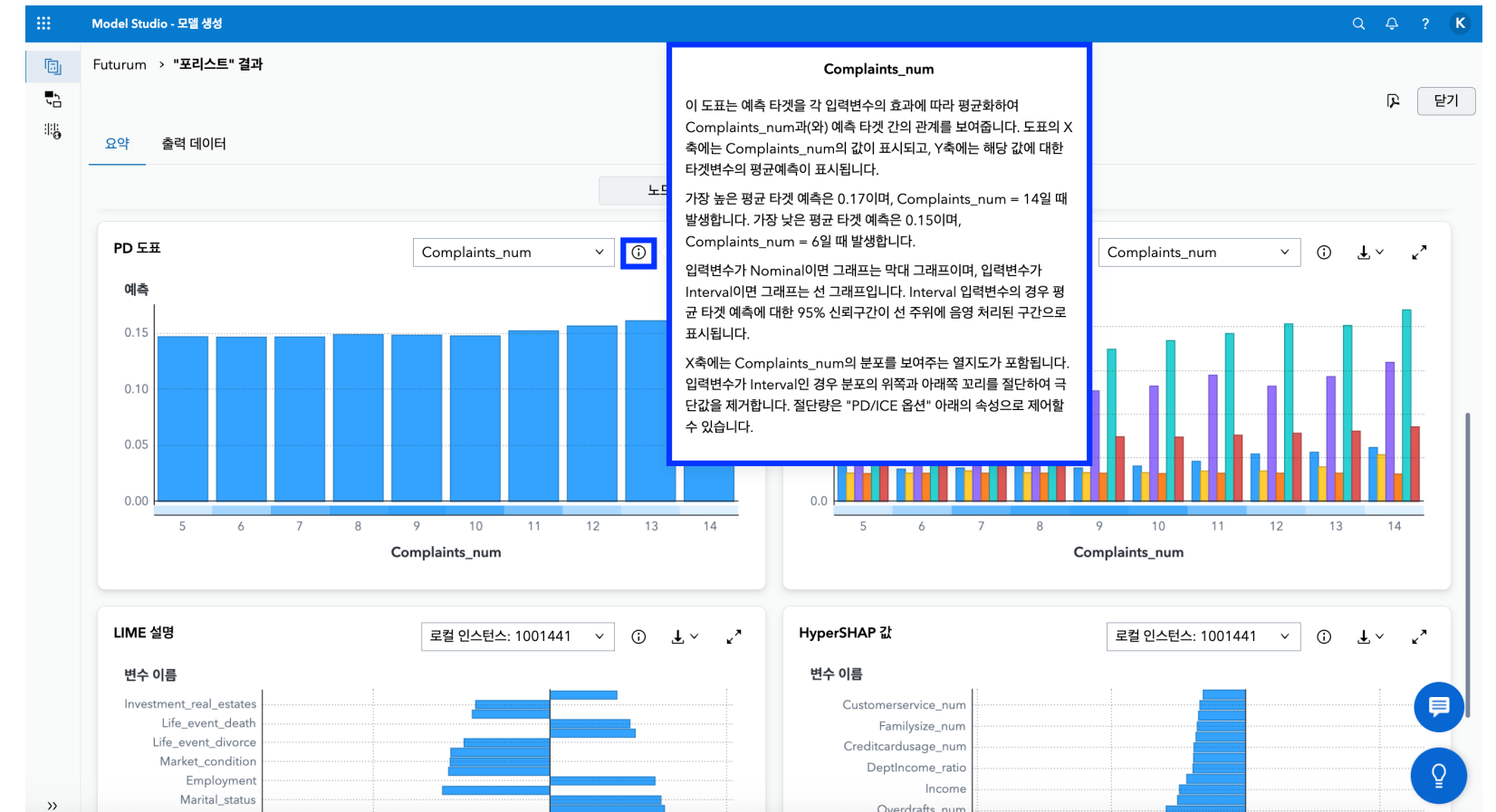
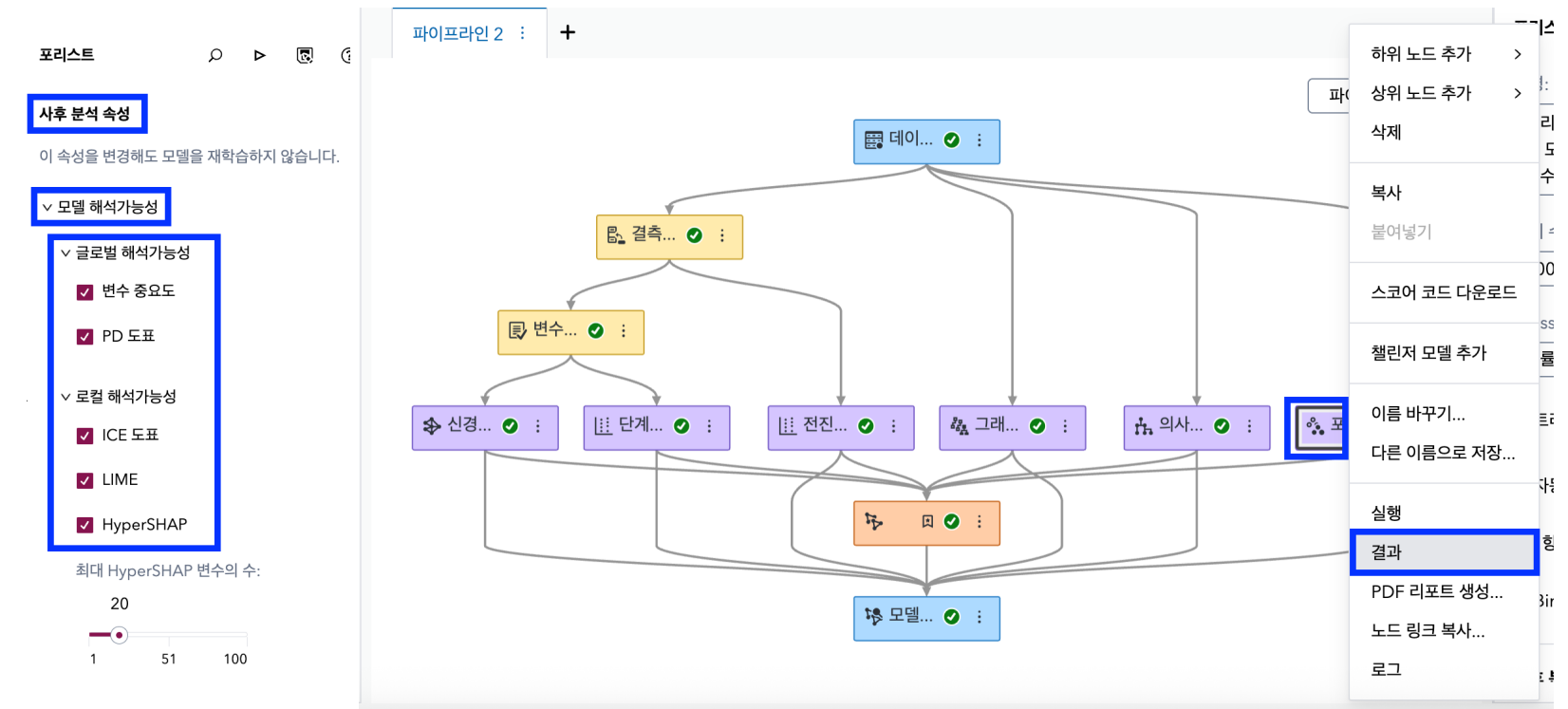
## 설명가능한 모델

이번에는 **모델 예측 결과를 해석**하는 단계입니다. 예시로 파이프라인에서 '포리스트' 노드를 선택하고 오른쪽 옵션 창에서 **사후 분석 속성 > 모델 해석가능성**으로 이동합니다.

글로벌 해석가능성에서는 **변수중요도와 PD 도표**를 선택하고, 로컬 해석가능성 옵션에서는 **ICE 도표, LIME, HyperSHAP**을 선택합니다.

그래프의 해석을 보기위해 다시 파이프라인을 실행한 후 완료시 **포리스트 노드 우클릭 > 결과**에 들어갑니다.

① **아이콘**에서 그래프 해석에 대한 안내를 볼 수 있으며, 글로벌 해석관점 (PD 도표, PD 및 ICE 중첩 도표)에서는 **입력 변수별로 모델이 어떻게 반응**하는지 확인할 수 있고, 로컬 해석관점 (LIME 설명, HyperSHAP)에서는 **개별 예측 사례에 대한 설명**을 확인할 수 있습니다.



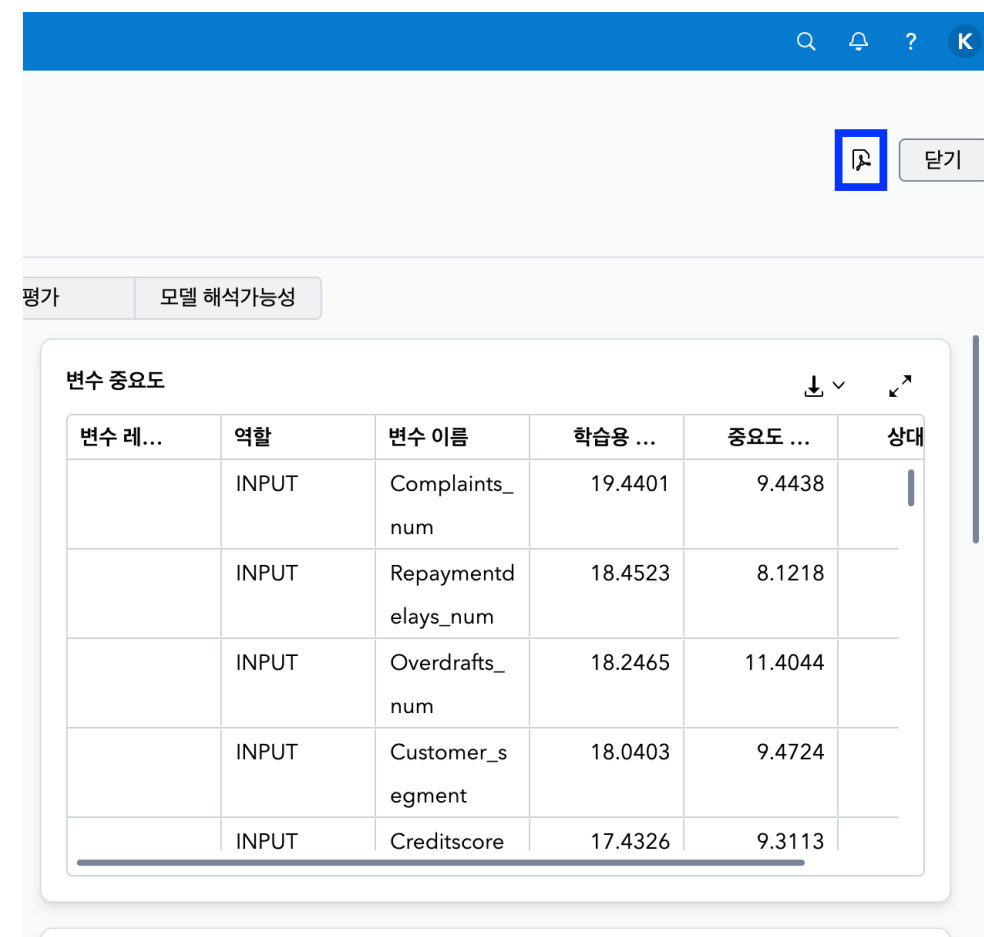
# 8. 모델 리포트

# 8. 모델 리포트

## 모델 문서화

지금까지 수행한 모든 모델링 작업은 반드시 **문서화**되어야 합니다. 그 이유는 프로젝트 차원 뿐만이 아닌, 규제 대응을 위해 개별 모델 단위에서도 필수적으로 요구되는 경우가 있기 때문입니다.

모델 리포트를 저장하기 위해 예시로 **포리스트** 모델을 선택한 뒤 **결과** 창으로 이동합니다. 화면 우측 상단의 **PDF 아이콘**을 클릭하고 **내보내기**를 선택하면 **해당 모델의 정보 및 시각화 결과를 포함한 PDF 리포트를 자동으로 생성**합니다.



평가 모델 해석가능성

변수 중요도

| 변수 레... | 역할    | 변수 이름               | 학습용 ... | 중요도 ... | 상대 |
|---------|-------|---------------------|---------|---------|----|
|         | INPUT | Complaints_num      | 19.4401 | 9.4438  |    |
|         | INPUT | Repaymentdelays_num | 18.4523 | 8.1218  |    |
|         | INPUT | Overdrafts_num      | 18.2465 | 11.4044 |    |
|         | INPUT | Customer_segment    | 18.0403 | 9.4724  |    |
|         | INPUT | Creditscore         | 17.4326 | 9.3113  |    |



PDF 내보내기

페이지 설정    개체 선택

페이지 설정

용지 크기: Letter

방향: 세로

가장자리

단위: 센티미터

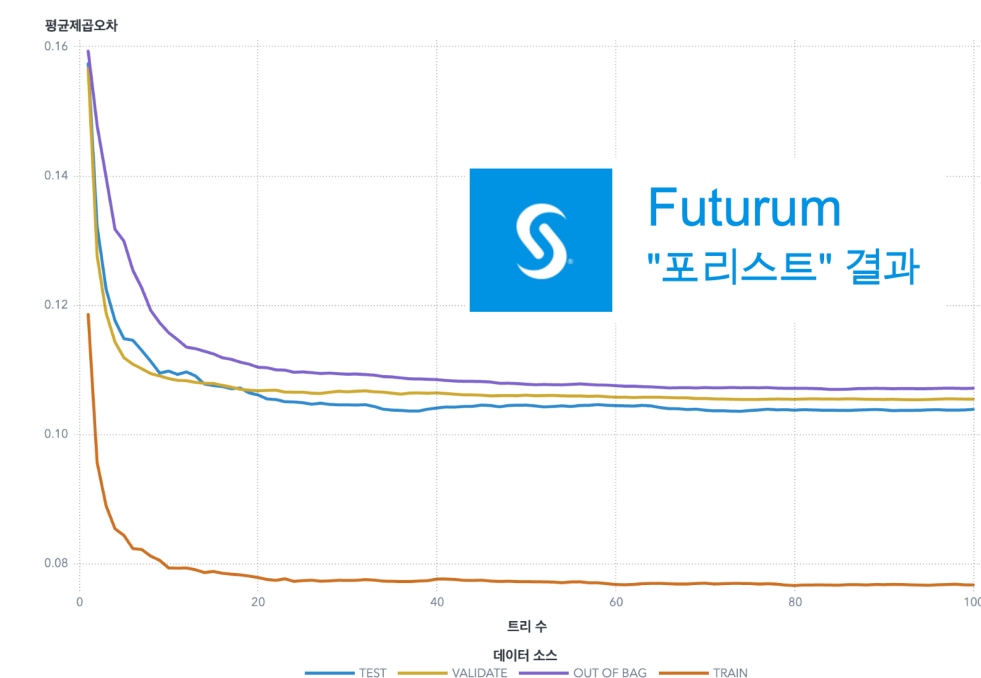
위: 2.54    오른쪽: 2.54

옵션

- 목차 포함
- 페이지 번호 표시
- 제목 페이지 포함
- 상세 정보 포함

내보내기    취소

평균제곱오차



이 도표는 포리스트의 트리 수가 증가함에 따라 평균제곱오차가 어떻게 변화하는지 보여줍니다. 학습용 오차는 일반적으로 트리 수가 증가함에 따라 감소하지만 VALIDATE 분할에 대한 오차는 모델이 얼마나 잘 일반화되는지에 대한 지표를 제공합니다. 이 모델의 경우 VALIDATE 분할에 대한 최소 오차는 0.105이며, 93개의 트리에 대해 발생합니다.

# 9. 파이프라인 비교 (Pipeline Comparison)

# 9. 파이프라인 비교 (Pipeline Comparison)

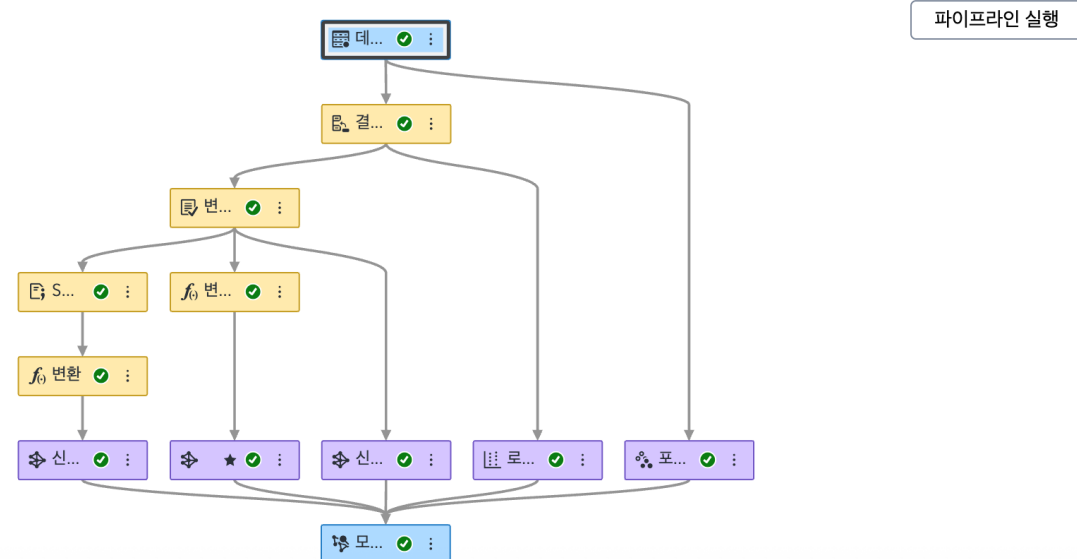
현재 저희가 생성한 파이프라인은 하나이며,  
**파이프라인 비교**하기 전에 **새로운 파이프라인**을 생성하겠습니다.

이번에 생성할 파이프라인은 **파이프라인 자동 생성**을 통해 만들어지는 파이프라인이며,  
**SAS Viya**는 이처럼 추천 템플릿을 활용하는 방법 이외에도 **맞춤형 템플릿 생성 기능**을 제공합니다.  
자동생성을 위해 걸리는 **시간** 또한 설정할 수 있으며,  
시간은 **1분**으로 설정하고 **저장**을 눌러 **파이프라인 생성**을 시작합니다.

**파이프라인 생성이 완료**될 시 **파이프라인 실행**을 클릭합니다. 파이프라인 실행이 완료되면, **파이프라인 비교** 탭으로 이동합니다.

파이프라인 2    SAS 자동 생성 파이프라인    +

이 파이프라인은 원래 SAS 자동화에서 생성되었습니다. 프로세스가 지정된 시간 제한인 1분에 종료되었으며 그 시점에 최량의 파이프라인을 표시했습니다.



| 선택                                  | 챔피언 | 등록 | 이름      | 알고리즘 이름 | 파이프라인 이름          | KS(Youden) | 관측값 수 |
|-------------------------------------|-----|----|---------|---------|-------------------|------------|-------|
| <input checked="" type="checkbox"/> | ☑   | ☑  | 양상블     | 양상블     | 파이프라인 2           | 0.5230     | 1,000 |
| <input type="checkbox"/>            |     |    | 신경망 (1) | 신경망     | ⊖ SAS 자동 생성 파이프라인 | 0.5110     | 1,000 |

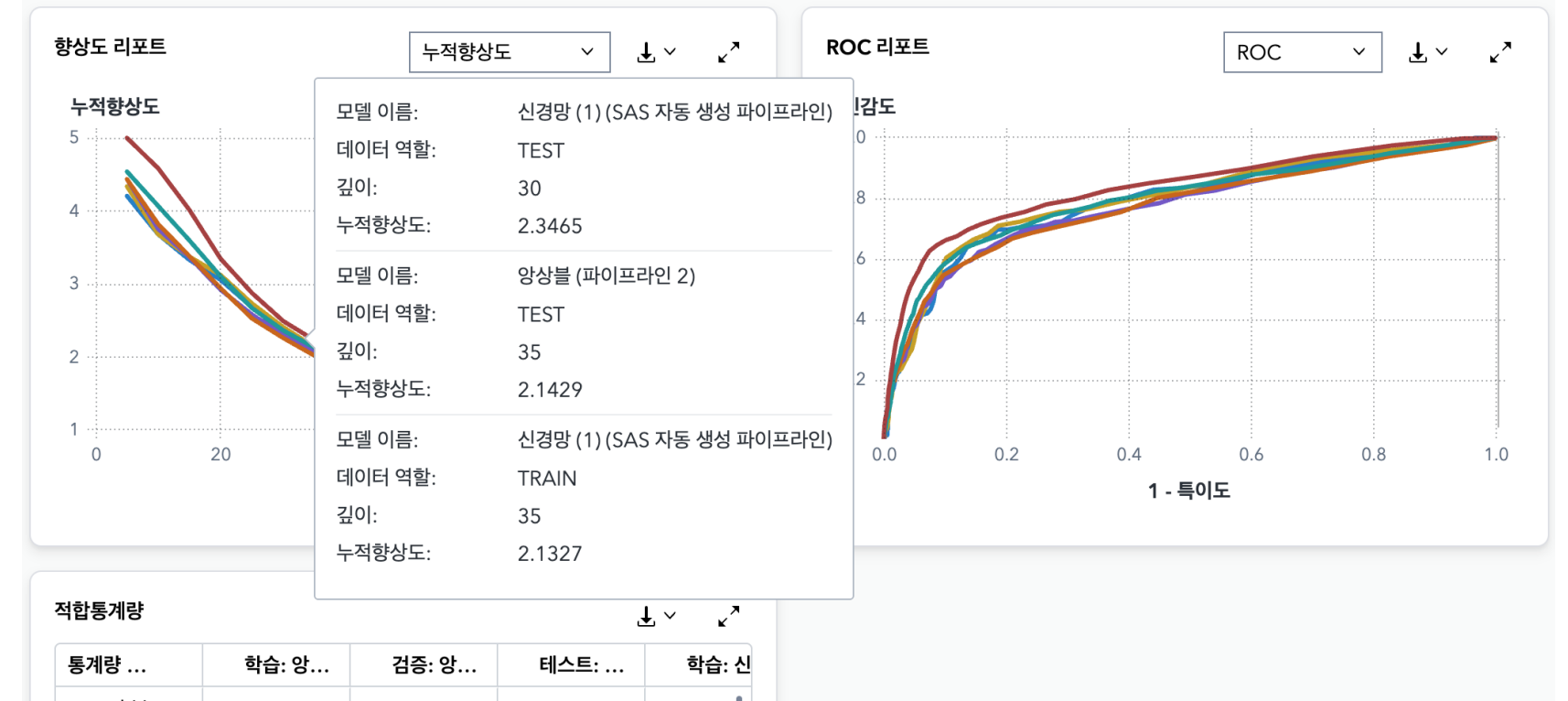
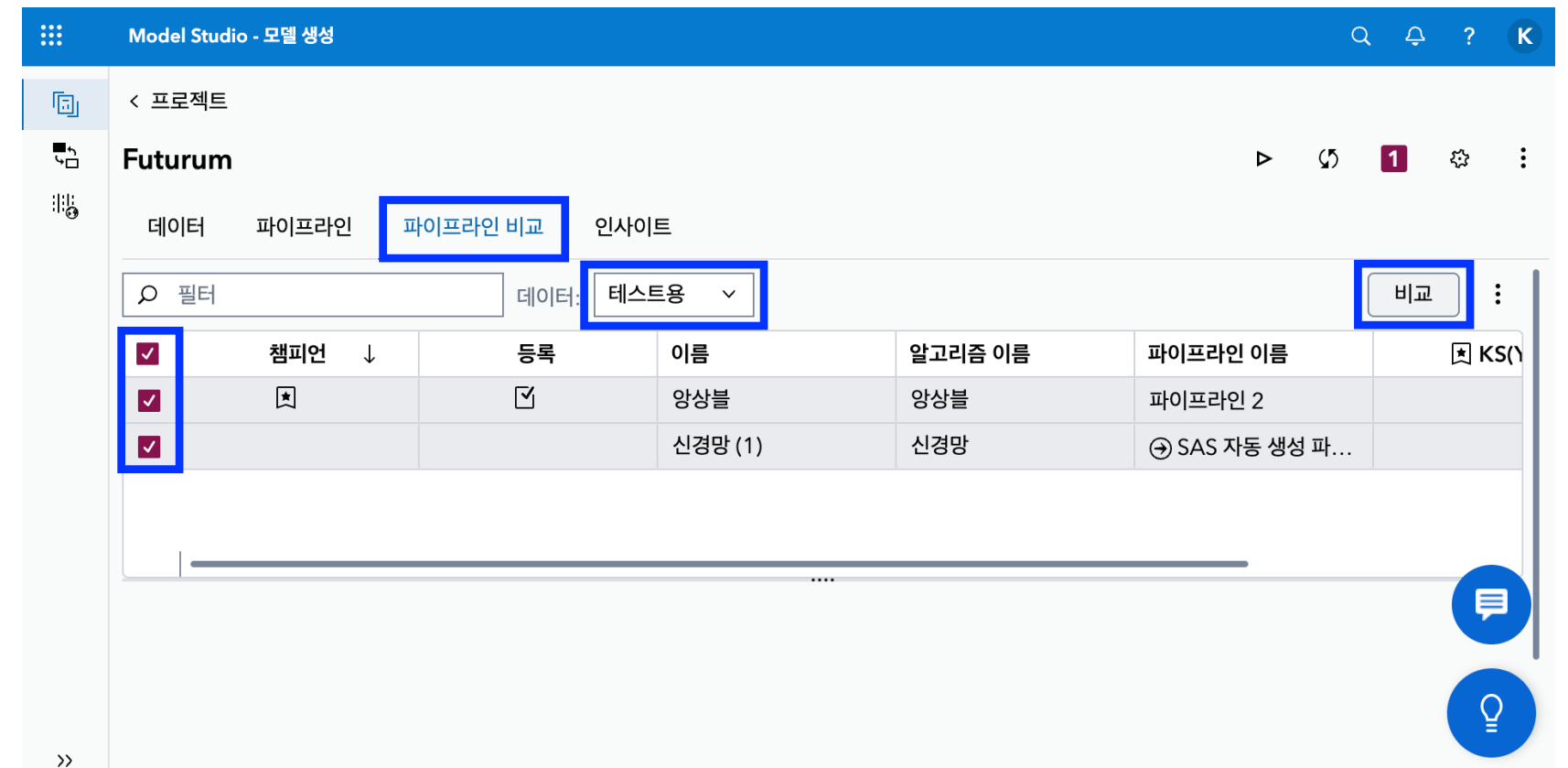
위 사진서 볼 수 있듯이 **새로운 모델(신경망 (1))**이 추가된 것을 확인하실 수 있으며, 기존 양상블 모델이 여전히 챔피언 모델의 자리를 유지하고 있음을 알 수 있습니다.

# 9. 파이프라인 비교 (Pipeline Comparison)

이제 두 파이프라인을 통해 생성된 **두 모델을 비교**하겠습니다. **파이프라인 비교** 탭에 들어가면 테스트 데이터를 기준으로 각 파이프라인의 적합한 모델이 나열된 것을 확인하실 수 있습니다.

본 예시에서는 테스트 데이터 기준으로 두 모델을 비교했으며, 사용에 맞게 훈련 데이터와 검증 데이터를 기준으로 한 비교 또한 가능합니다.

파이프라인 비교 후 기본 설정인 정확도 기준으로 설정된 챔피언 기준 모델이 아닌 **다른 모델을 챔피언으로 선정하고 싶은 경우** 우클릭 > **'챔피언으로 설정'**을 통해 변경 가능합니다.



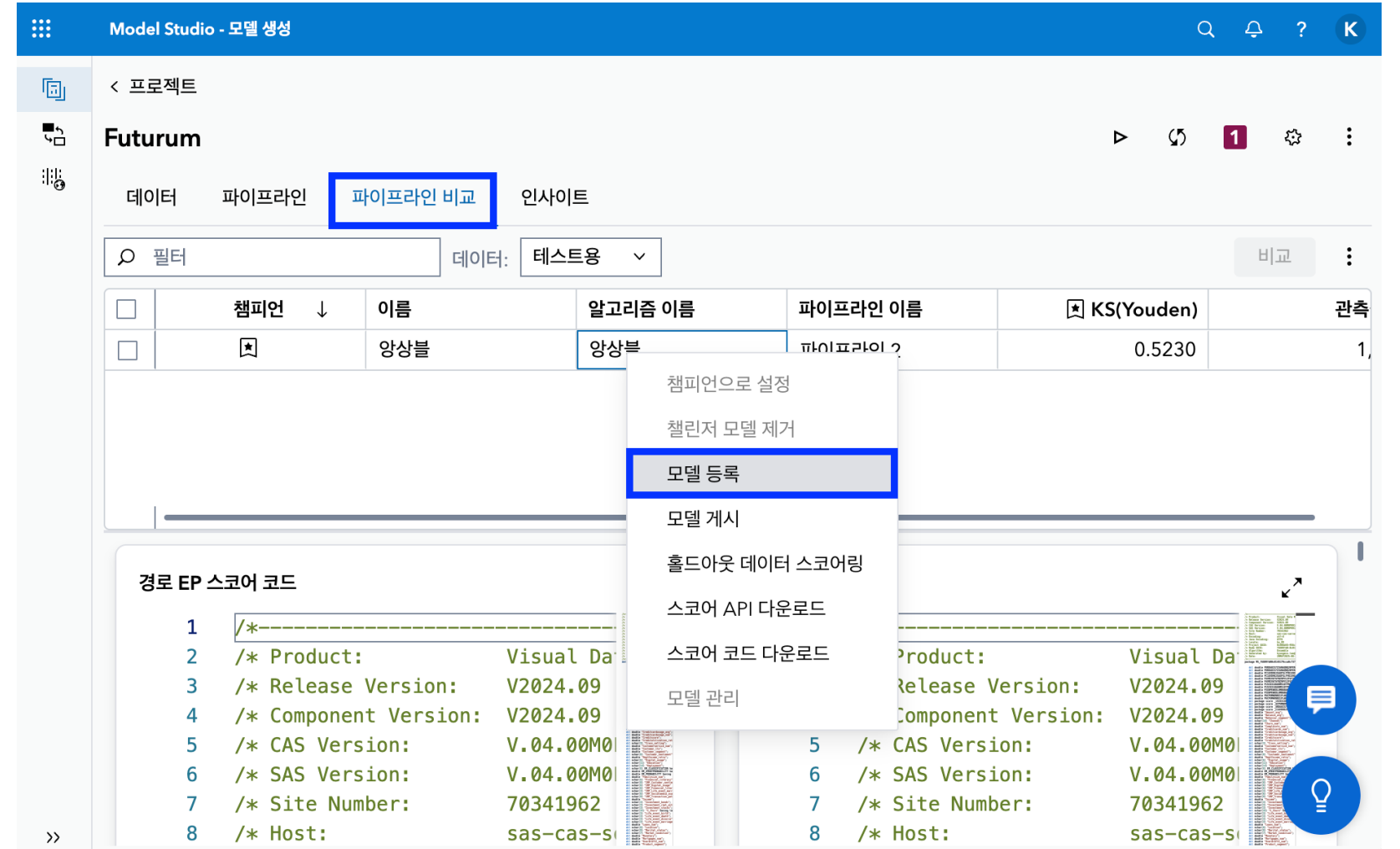
# 10. 모델 등록 (Model Registration)

# 10. 모델 등록 (Model Registration)

본 프로젝트에서는 **두개의 파이프라인**을 생성하였지만, 실무에서 **데이터 사이언티스트**는 여러 파이프라인을 통해 다양한 모델을 생성하는 것과 더불어 파이프라인 비교 탭에서 챔피언 모델을 선정할 수 있습니다.

이제 **중앙 모델 저장소에 등록**할 차례입니다. **챔피언 모델을 우클릭** > **모델 등록**을 합니다. 모델 등록이 완료된 시점인 **현재 이후의 작업들은 ModelOps 엔지니어를 통해** 진행될 것입니다.

**하나의 프로젝트에서 여러 개의 모델을 등록해 모델 간 비교**를 할 수 있으며, 시간에 따른 **성능 변화나 안정성을 비교 분석**할 수 있습니다. 또한, 필요시에 **데이터 사이언티스트는 챔피언 모델을 교체**하여 운영 환경에 적용할 수도 있습니다.



## 모델 등록

### 모델

|     |       |
|-----|-------|
| 이름: | 상태:   |
| 양상블 | ✓ 등록됨 |

# 11. 프로젝트 인사이트 보고서 – 문서화

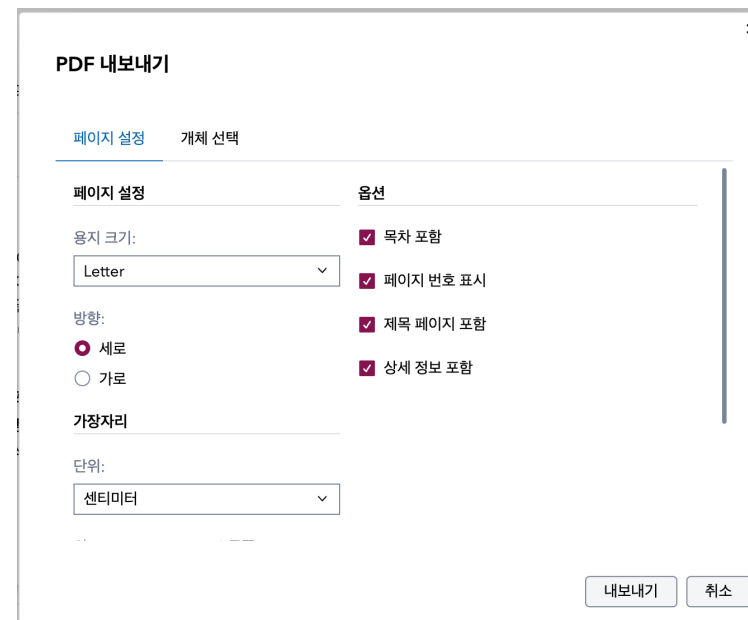
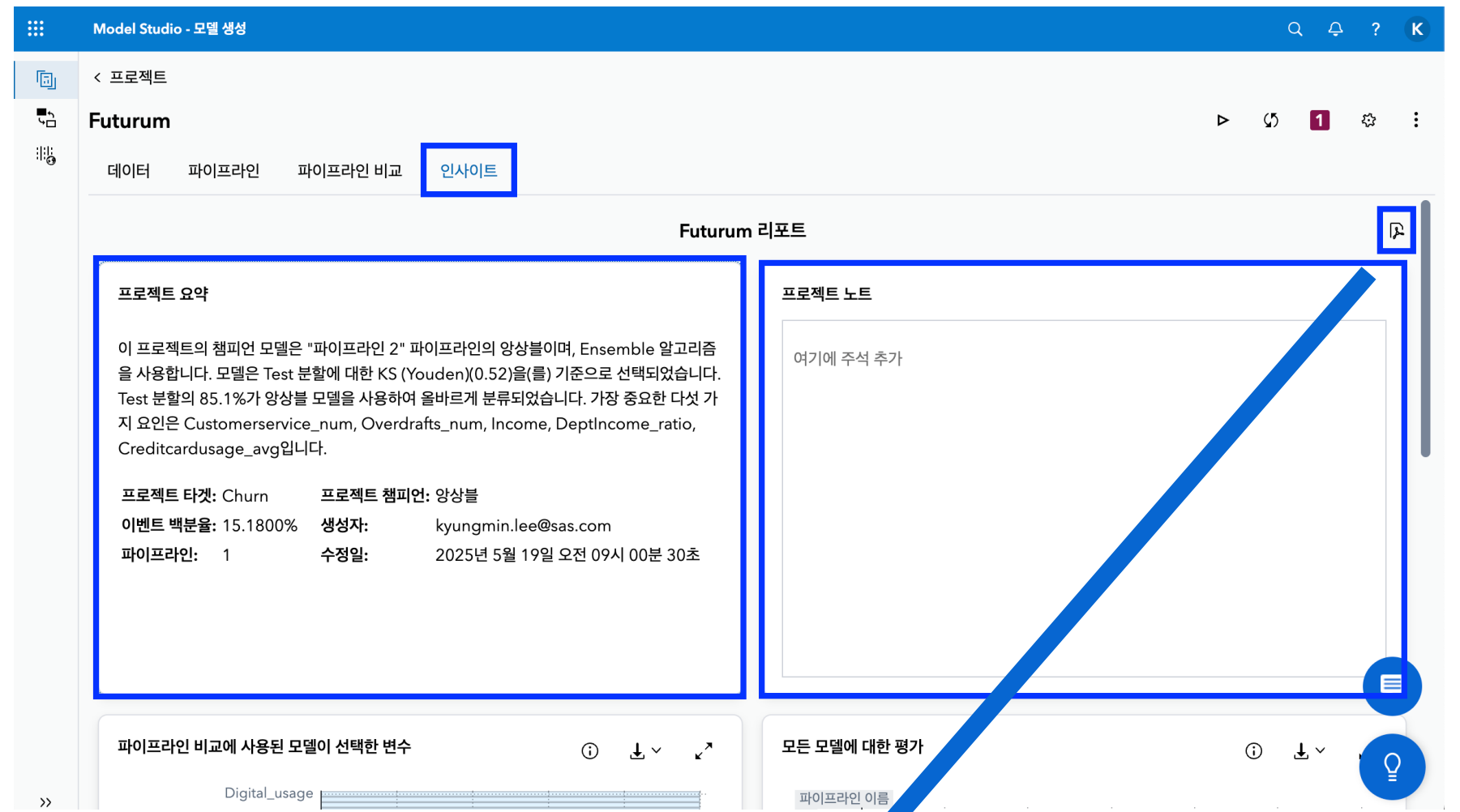
# 11. 프로젝트 인사이트 보고서 – 문서화

프로젝트의 마지막이자 가장 중요한 단계는 인사이트를 문서화하는 것입니다.

Model Studio에서는 이 과정을 클릭 한 번으로 간편하게 수행할 수 있습니다.

인사이트 탭으로 이동하면 **프로젝트의 전반적인 요약**과 챔피언 모델의 **주요 성능 지표**를 확인할 수 있으며, 화면 우측 상단의 PDF 아이콘을 클릭해 보고서 형태로 저장할 수도 있습니다.

또한, 프로젝트와 관련된 비즈니스 정보나 이해관계자와 공유할 내용을 **프로젝트 노트**에 직접 기록할 수 있어 실무에서도 매우 유용하게 활용됩니다.



# 12. 프로젝트 공유 - 읽기/쓰기 권한 설정

# 12. 프로젝트 공유 - 읽기/쓰기 권한 설정

Model Studio에서는 작업 내용이 자동 저장됩니다. 또한, 공유 및 협업으로 이동해 생성한 프로젝트를 찾아 우클릭 > 공유를 선택하면 다른 사용자와 프로젝트를 손쉽게 공유할 수 있습니다.

비즈니스 관계자나 경영진에게는 핵심 정보만 확인할 수 있도록 읽기 전용 모드로 공유할 수 있으며, 다른 데이터 사이언티스트와 협업이 필요한 경우에는 읽기 및 편집 권한을 설정하여 효율적인 공동 작업이 가능합니다.

분석 라이프사이클

정보 에셋 검색

데이터 관리

탐색 및 시각화

모델 생성

모델 관리

의사결정 생성

공유 및 협업

코드 및 플로우 개발

The screenshot displays the SAS Drive interface for sharing a project. On the left, a sidebar shows navigation options like '내 즐겨찾기', 'My Folder', and '공유'. The main area shows a folder named 'Futurum' with a context menu open, highlighting the '공유...' option. On the right, a '공유' (Share) dialog box is open, showing the folder name 'Futurum' and a dropdown menu set to '읽기 가능' (Read Only). Below this, it indicates '다음 사용자 및 그룹과 공유됨(0)'. At the bottom right of the dialog are '공유' and '취소' buttons.

# End of the Guide

