부동산 가격 예측 모델 개발

팀 이름: 아파트

조　원: 이유리, 우병준, 이정인

Table of Contents

1. Project Overview

프로젝트 개요

매매가 예측

주제 선정 배경

서울시 아파트 매매가는 부동산 시장의 주요 지표. 따라서 시민과 투자자들에게 중요한 의사결정 정보를 제공

프로젝트 개요

서울시 아파트 매매가 예측 머신러닝 모델 개발

주요 데이터
아파트 매매 실거래가 데이터 등

데이터 전처리 및 분석
feature 선정
예측 모델 구성

활용 장비 및 개발 환경

언어
Python

라이브러리
Pandas, NumPy, Scikit-learn, XgBoost 등

분석 도구
Jupyter Notebook, VS Code

프로젝트 구조

데이터 수집

데이터 전처리 및 탐색적 분석

머신러닝 모델 구축 및 예측

2. Team Composition and Roles

팀 구성 및 역할

	이유리	우병준	이정인
수집 데이터	• 아파트 실거래가 • **아파트 브랜드** • 구/행정동/법정동 목록 및 경계 파일 • 버스정류장, 지하철역 • **정치 데이터** • 국토교통부 • **부동산 사이트** • 공공데이터 포털 및 브이 공간 월드 • 공공데이터 포털 • **재임 기간**	• 소비자 물가지수 • **환율, 금리** • 행정동별 인구밀도 • **행정동별 상가 소득 및 소비** • 범죄율(폐기) • 통계청 • **한국은행** • 공공데이터 포털 • **공공데이터 포털**	• 학교, 병원, 공원 위치 데이터 • 날씨(폐기) • 공공데이터 포털 • 기상청
전처리	• 그 외 모든 전처리	• 2019년, 2020년 아파트 도로명 주소 위도 경도 변환 • 행정동별 상가 소득 및 소비	• 학교 도로명 주소 위도 경도 변환
모델링	• XgBoost • Ridge, Lasso • Decision Tree	• Random Forest • 최고 성능 모델의 하이퍼파라미터 튜닝	
기타	• 발표자료 제작 (전반부) • 문서 작업	• 발표자료 제작 (후반부) • 문서 작업	

3. Project Execution Process and Methodology

수행 절차 및 방법

기획

프로젝트 주제 선정

- 부동산 시장 분석
  및 프로젝트 주제 선정
- 필요 데이터
  및 목표 정의

수행

데이터 수집

- 서울시 실거래가 데
  이터 수집
- 주변 시설
- 교통
- **정치**
- **환율, 금리**
- **소득-소비**
- **브랜드화**

데이터 전처리

- 아파트 기본 정보
  주소 위도경도 변환 등
- 주변 시설 및 교통
  아파트 기준
  최단거리 계산
  일정 반경 내 점수화
- 동별 직전 월의 거래 건수
- **환율, 금리, 정치**
  결측치 처리
- **인구밀도 및 경제 수준**

예측 모델링

- Ridge, Lasso
- Decision Tree
- Random Forest
- XgBoost

- 최고 성능 모델 (Gradient
  Boosting 알고리즘 구현)
  XGBoost 하이퍼파라미터 튜닝

4. Data Preprocessing

데이터 전처리

2019년 75,097건

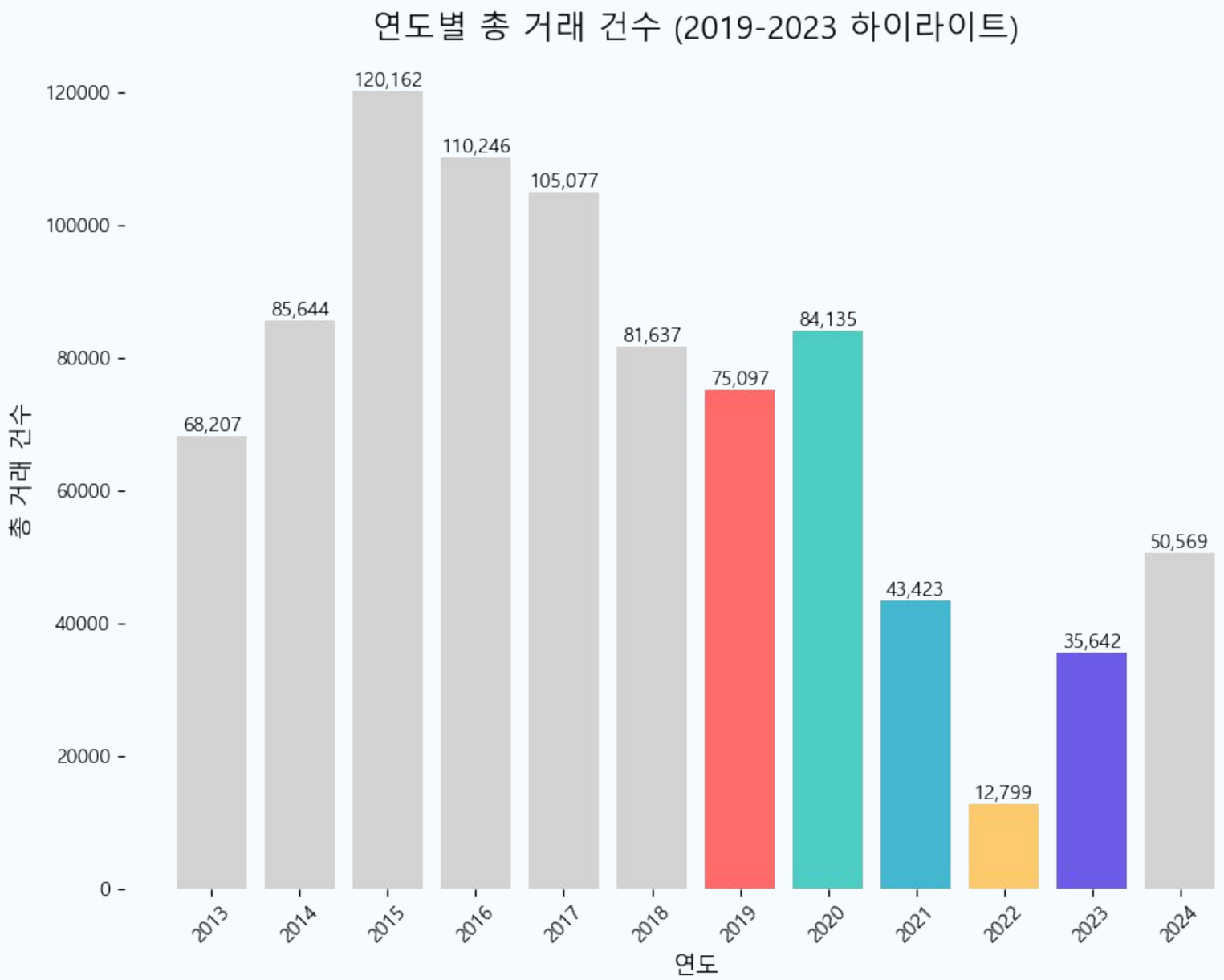
2020년 84,135건

2021년 43,423건

2022년 12,799건

2023년 35,642건

5개년 총합 251,096건



4. Data Preprocessing

데이터 전처리

불필요한 컬럼을 지운 서울시 아파트 실거래가 원데이터 (도로명 주소는 가림)

NO	시군구	단지명	전용면적(㎡)	계약년월	계약일	거래금액(만원)	층	건축년도
1	서울특별시 성동구 마장동	현대	134.790	201912	31	88,000	17	1998

인덱스용
NO컬럼 값 정제

NO
2019_00001

위도 경도 추출을 위한
주소 정제

NO	address
2019_00001	서울특별시 성동구 살 [redacted] 0

API 활용
위도 경도 변환

NO	latitude	longitude
2019_00001	37.56[redacted]	127.04[redacted]

Missing Value는 주소를 뒤에서 부터 잘라,

모든 주소 위도 경도 변환

(학교 주소도 같은 방식으로 변환)

4. Data Preprocessing

데이터 전처리

상위 1위 ~10위

고급 (4점)

상위 11위~28위

중급 (3점)

순위 밖 브랜드

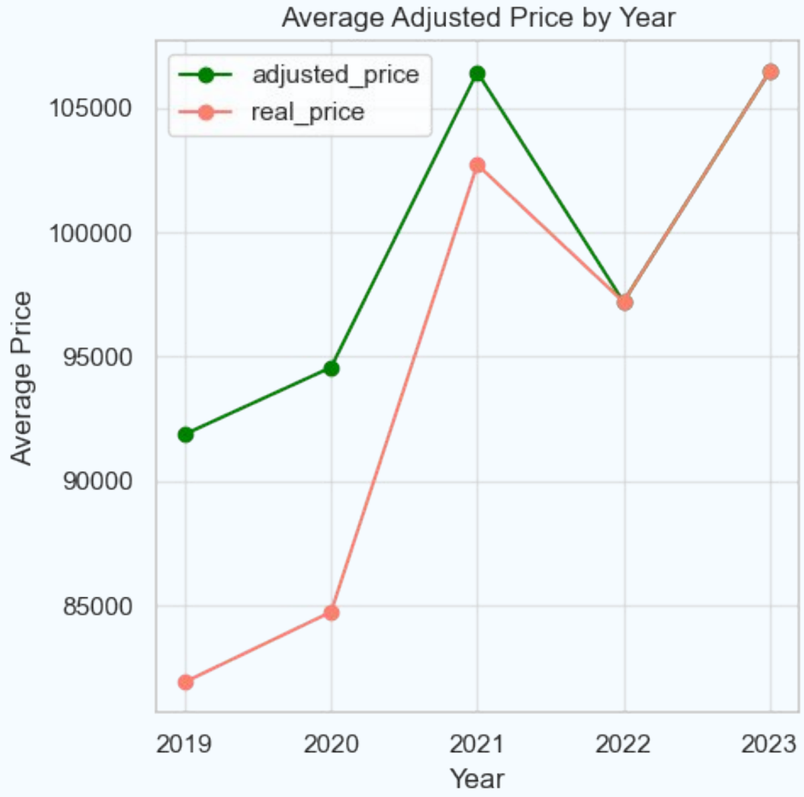
하급 (2점)

그 외

브랜드 아님 (1점)

4. Data Preprocessing

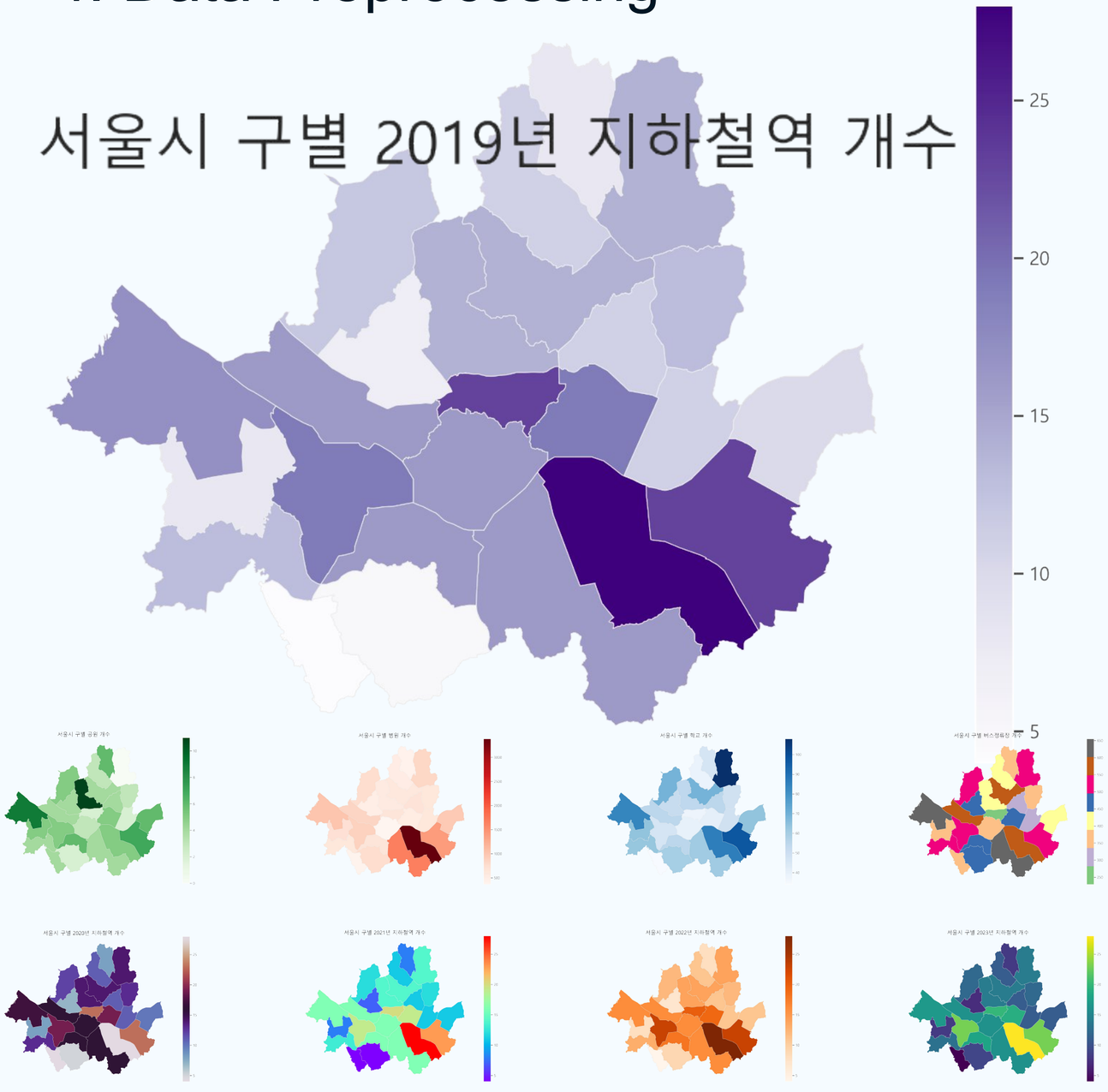
데이터 전처리



소비자 총물가지수 활용 → **2023 화폐가치로 전환**

4. Data Preprocessing

데이터 전처리



아파트 기준

버스 정류장, 학교, 병원, 공원
연도별 지하철역

거리 계산

점수 계산

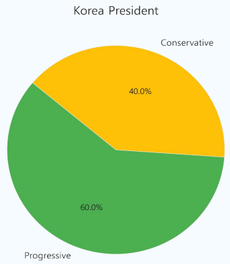
0~500m 10점 (도보 약 5분)
500m~1km 5점 (도보 약 10분)
1km~1.5km 1점 (도보 약 15분 이상)

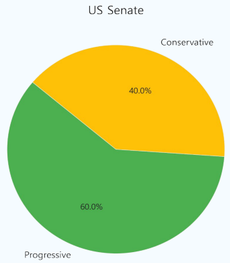
구/행정동 매핑

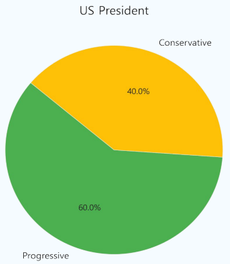
동별 직전 1/3/6개월 거래 건수 집계
행정동별 인구밀도 데이터 적용
행정동별 상가기준 소득-소비 데이터 적용

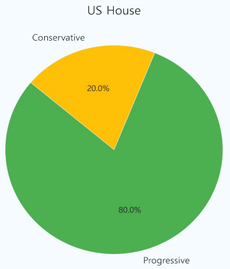
4. Data Preprocessing

데이터 전처리

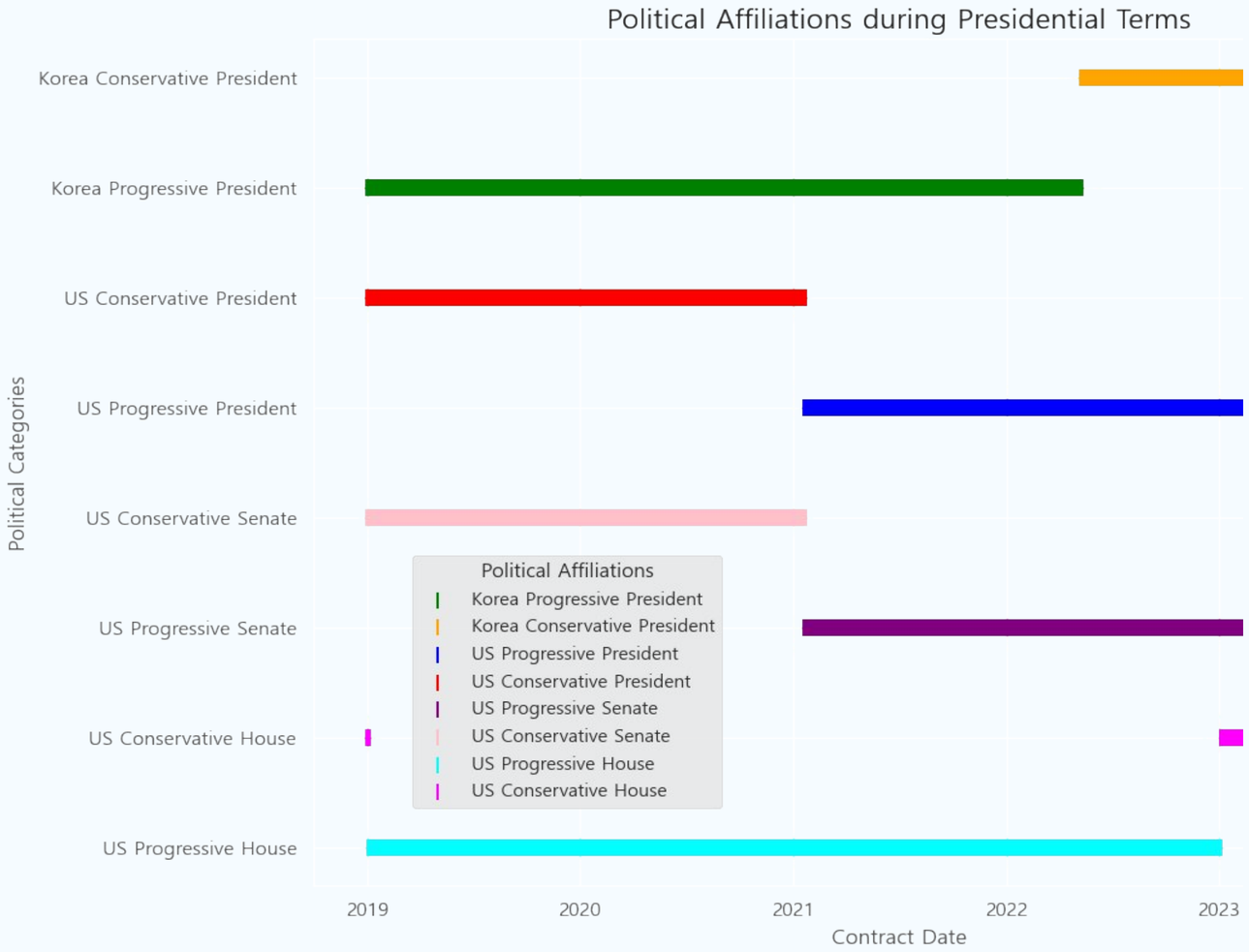


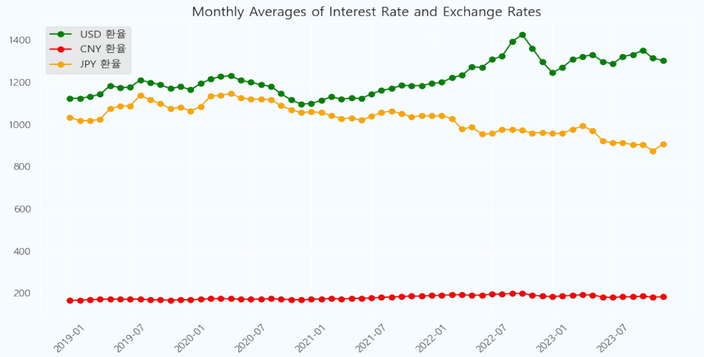






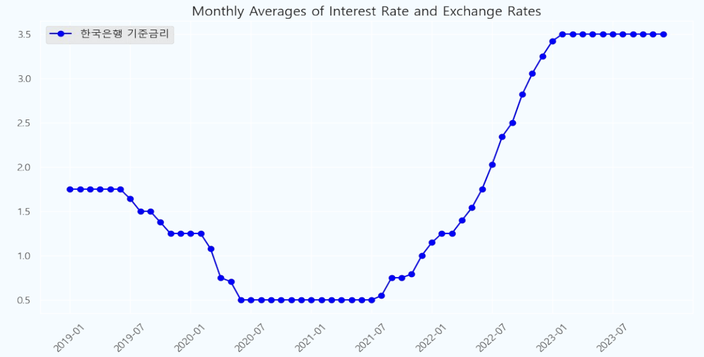
한•미 정치





환율

금리

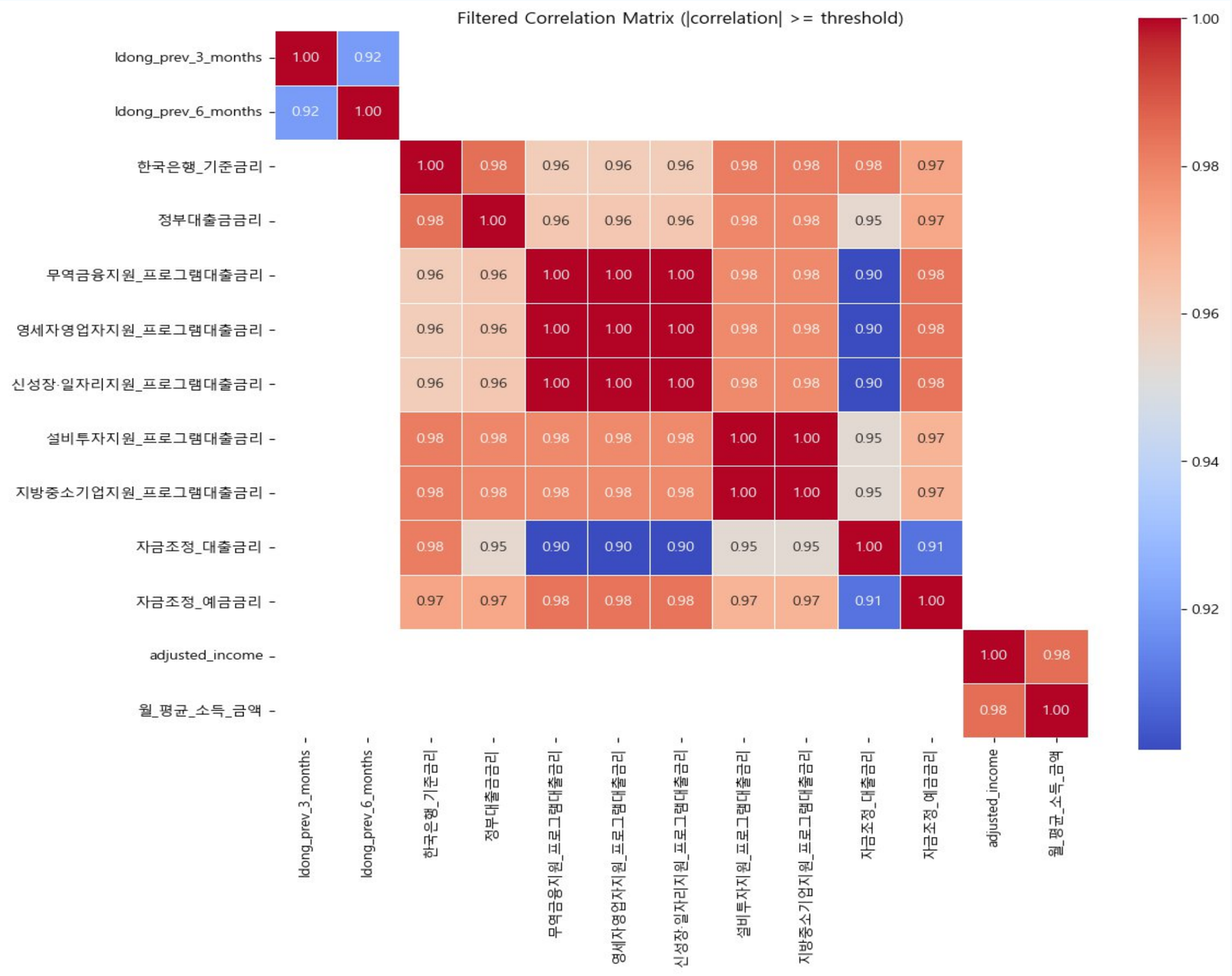


5. Project Outcomes Modeling

수행결과 - 모델링

Correlation Analysis

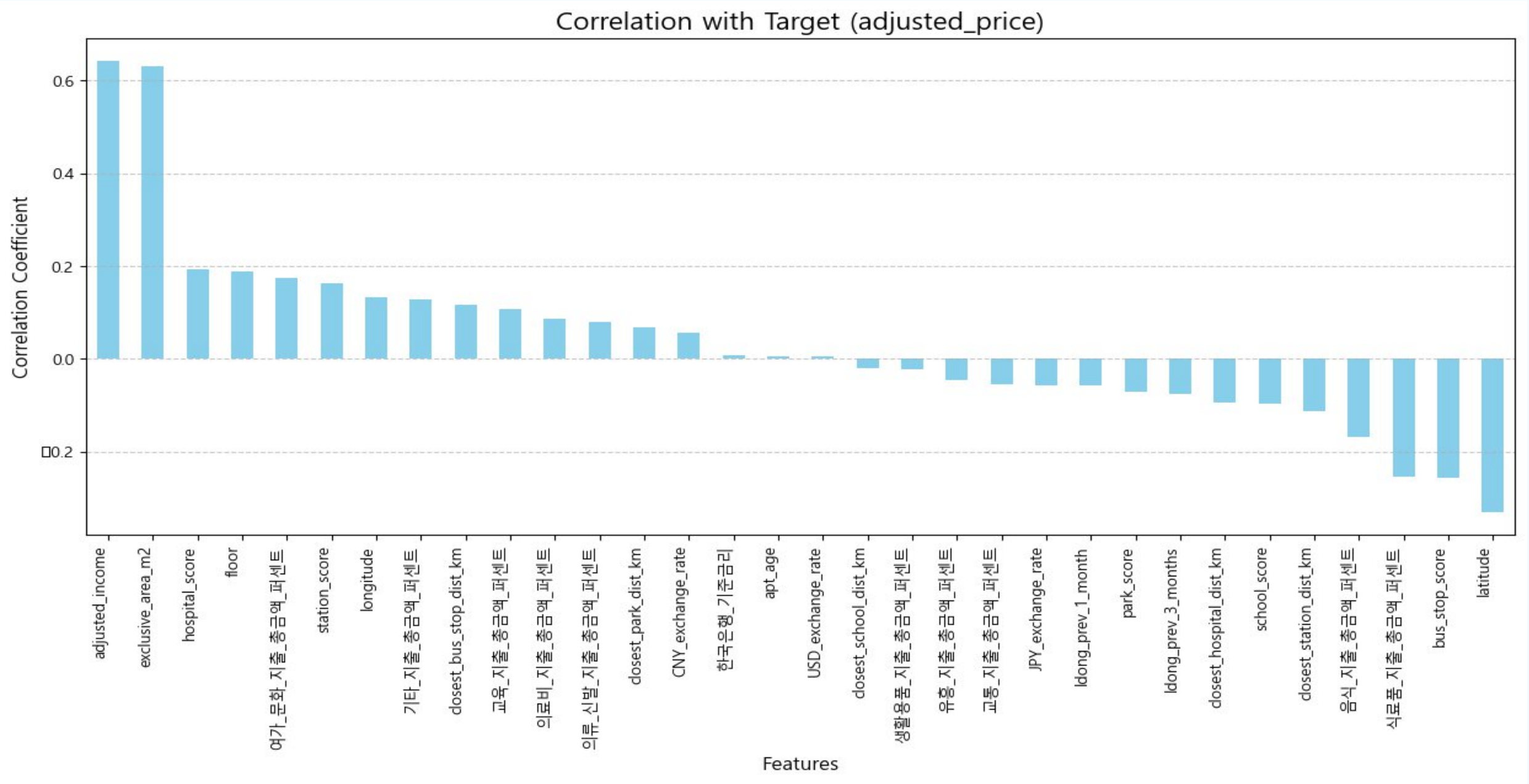
상관 관계 분석



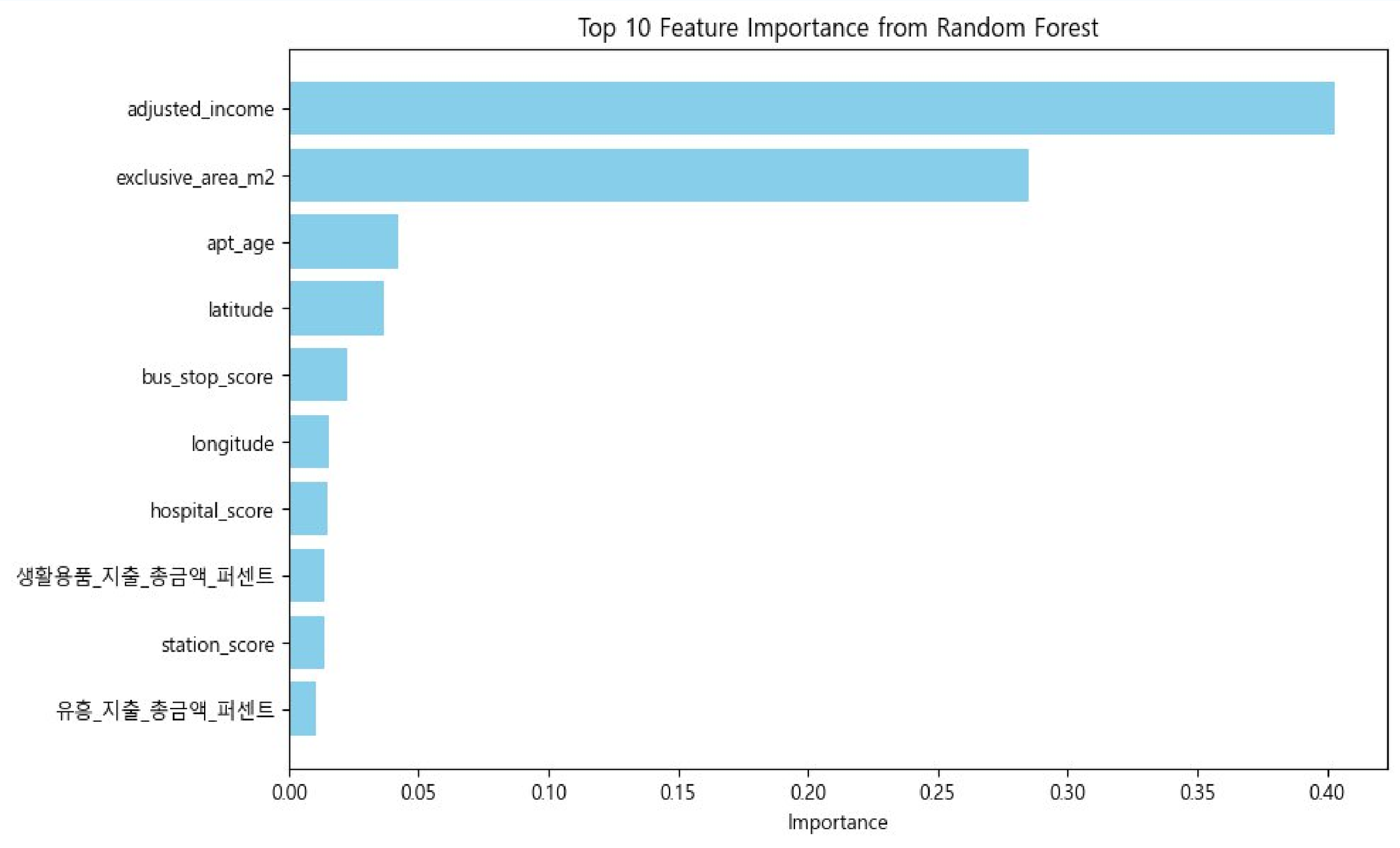
특성 간 상관관계 분석을 통한
다중공선성 탐지

Correlation Analysis

상관 관계 분석

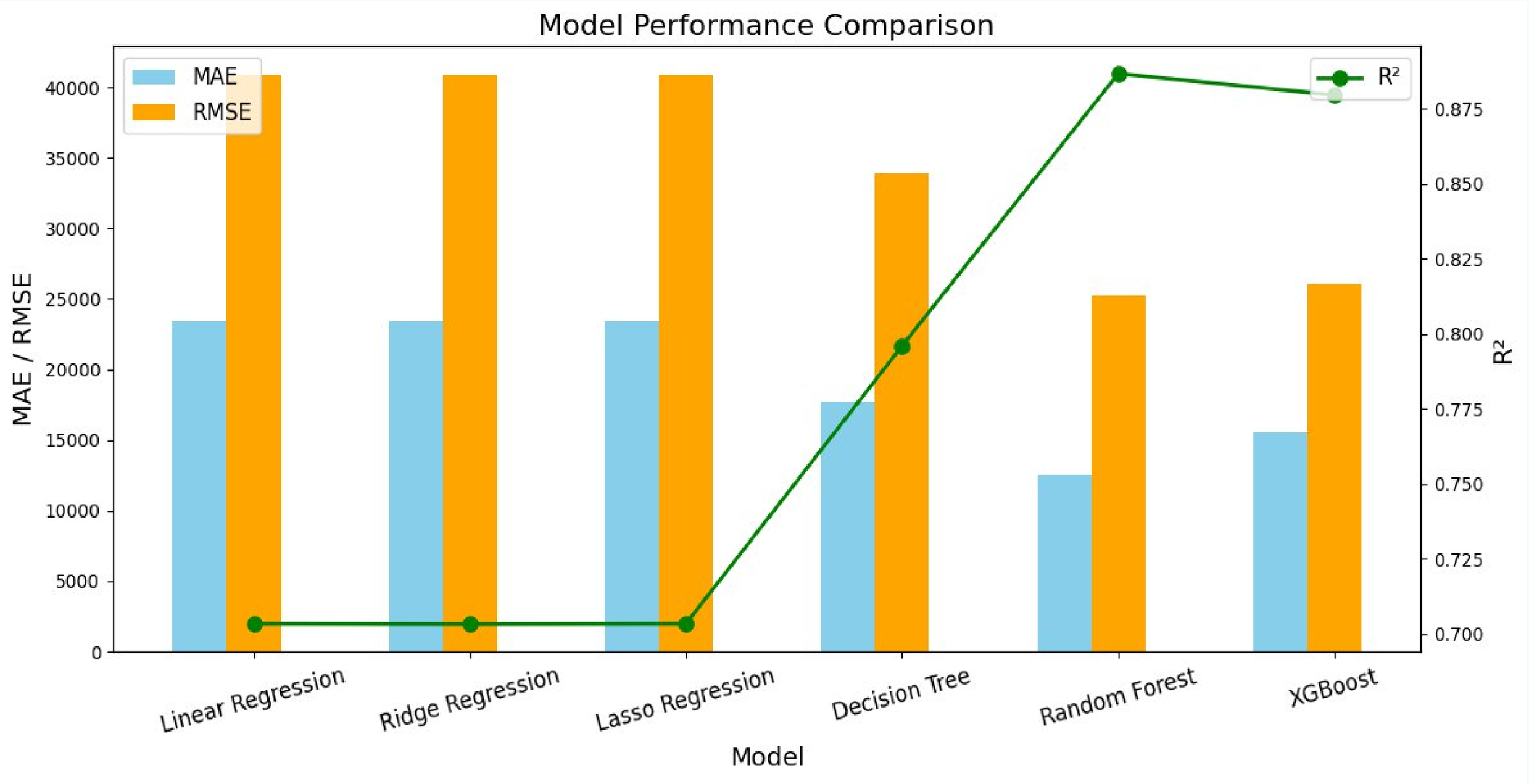


Feature Importance



Model Selection

모델 선택 과정



RandomForestRegressor

랜덤 포레스트

Error Rate

0.13

- **R^2**
  훈련 세트 R^2 : 0.984
  테스트 세트 R^2 : 0.871

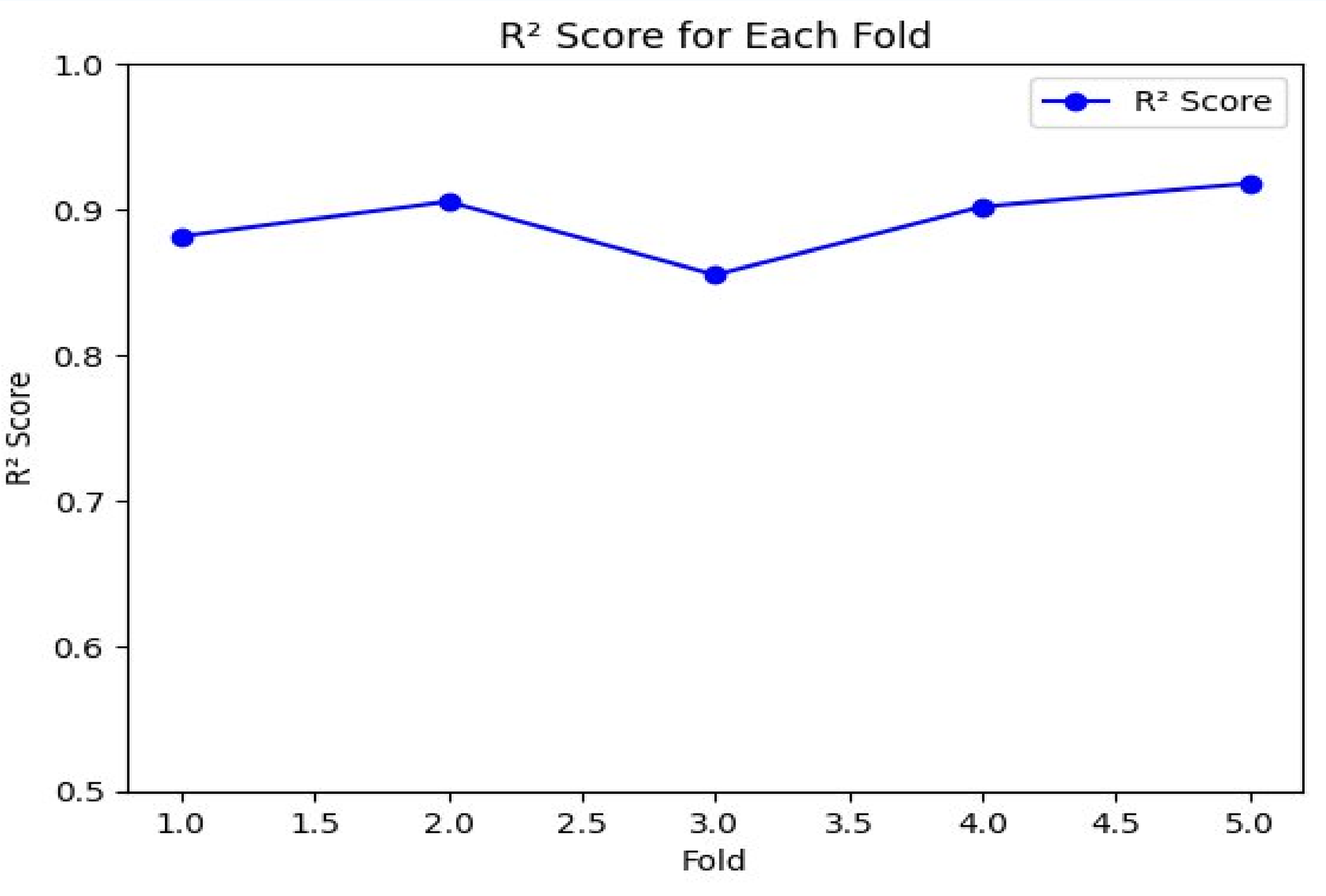
- **MAE**
  12731.622
- 평균 절대 오차

- **MSE**
  724951722.668
- 평균 제곱 오차

- **RMSE**
  26924.928
- 평균 제곱근 오차

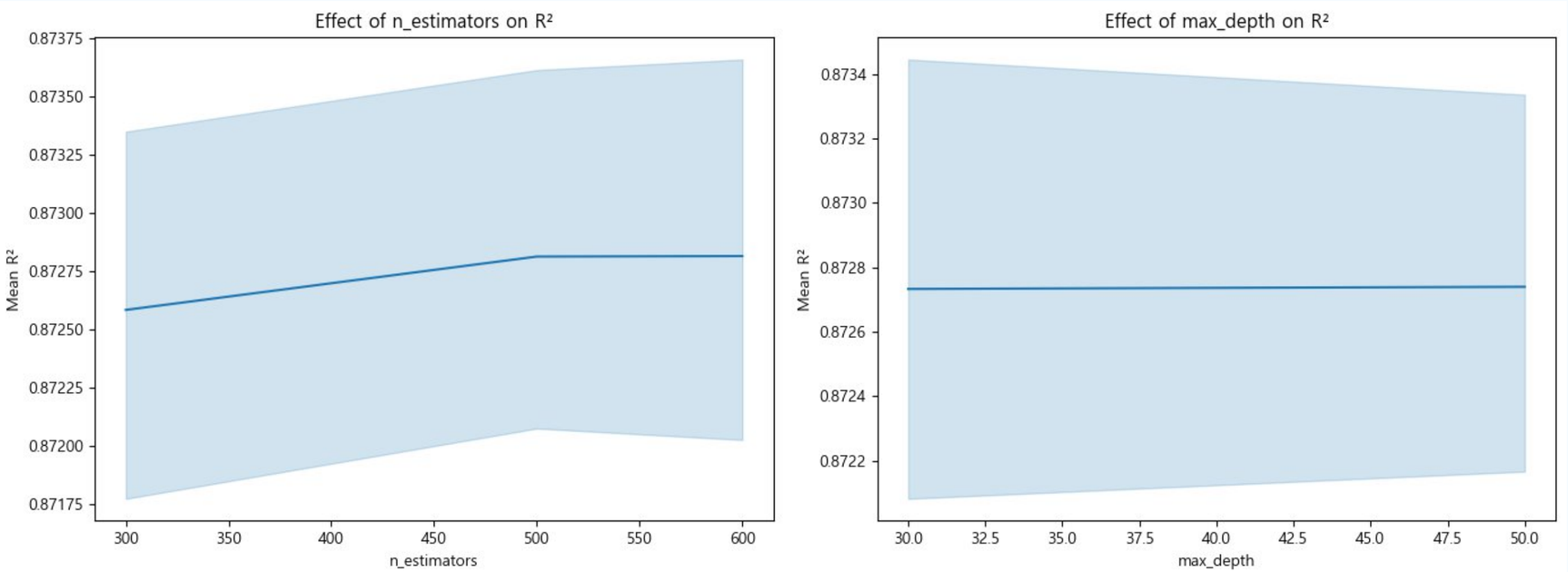
Cross Validation

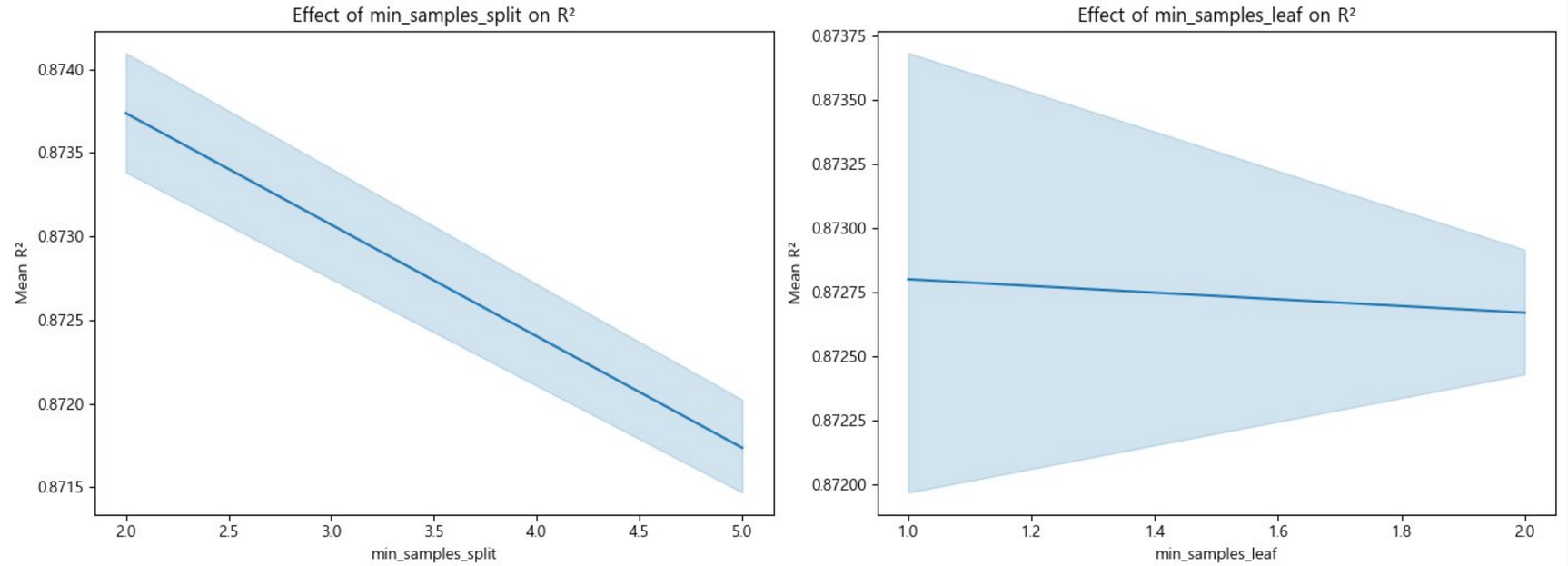
교차 검증



Hyperparameter tuning

하이퍼 파라미터 조정





R^2 = 0.875
MAE = 12632.60
RMSE = 26503.194

Error Rate = 0.129

Unsampled Data

랜덤 포레스트

- **R^2** 0.891

- **MAE** 12170.569
- 평균 절대 오차

- **MSE** 540423738.4162
- 평균 제곱 오차

- **RMSE** 23247.015
- 평균 제곱근 오차

Error Rate

0.128

XGBRegressor

XGBoost

오차율 **0.105**

- **R^2**
  훈련 세트 R^2 : 0.982
  테스트 세트 R^2 : 0.936

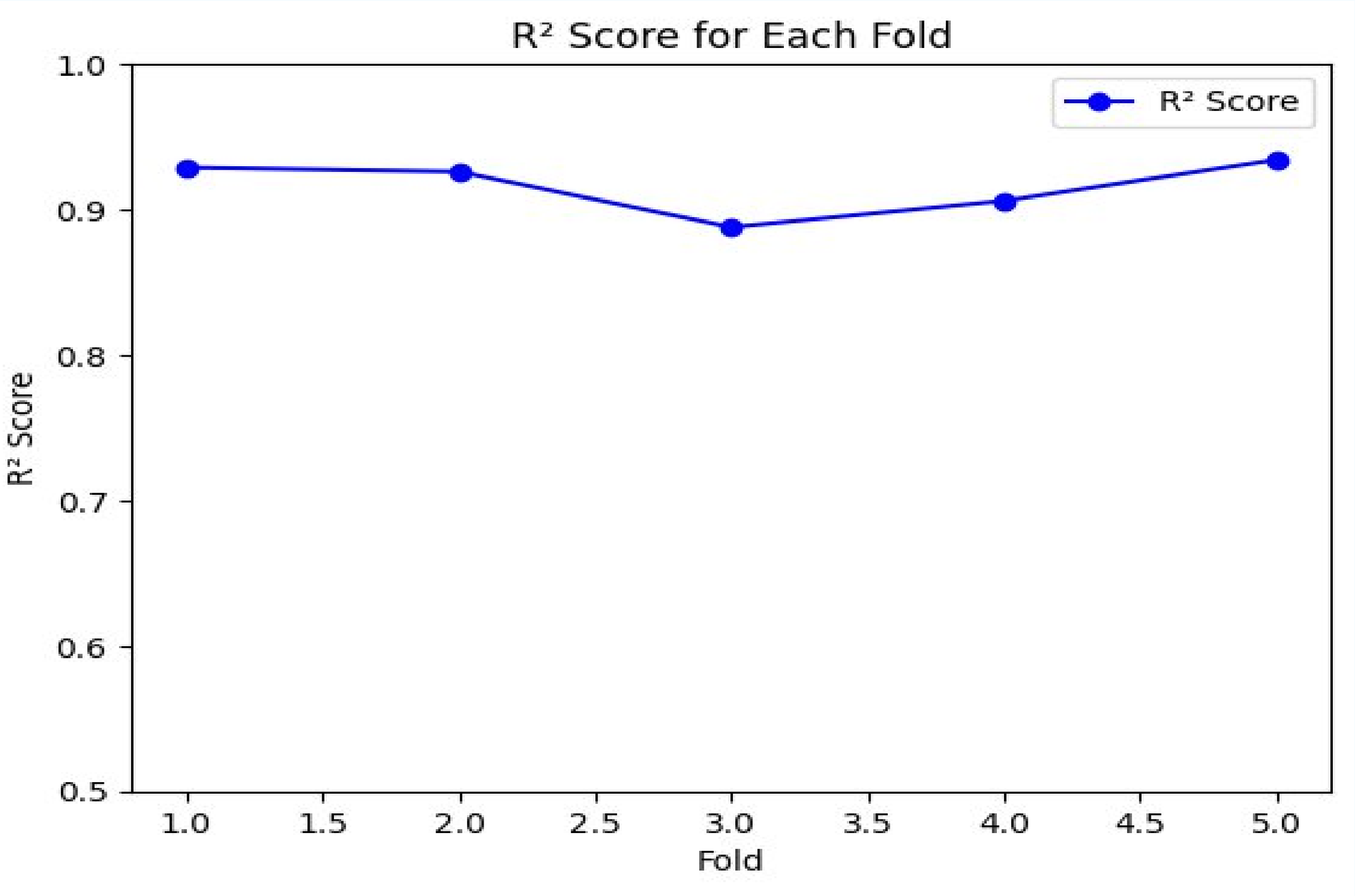
- **MAE**
  6735.414
- 평균 절대 오차

- **MSE**
  89093354.404
- 평균 제곱 오차

- **RMSE**
  9438.927
- 평균 제곱근 오차

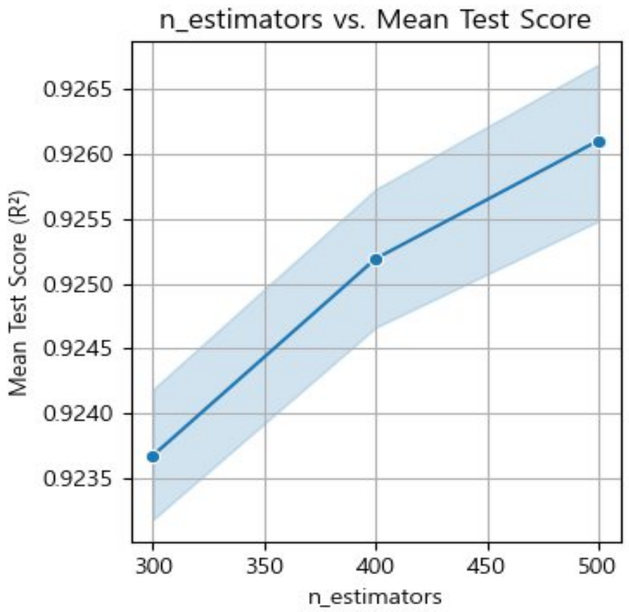
Cross Validation

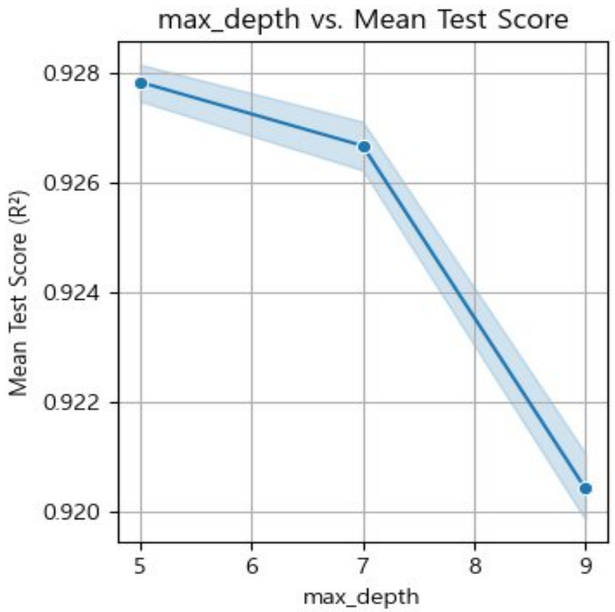
교차 검증

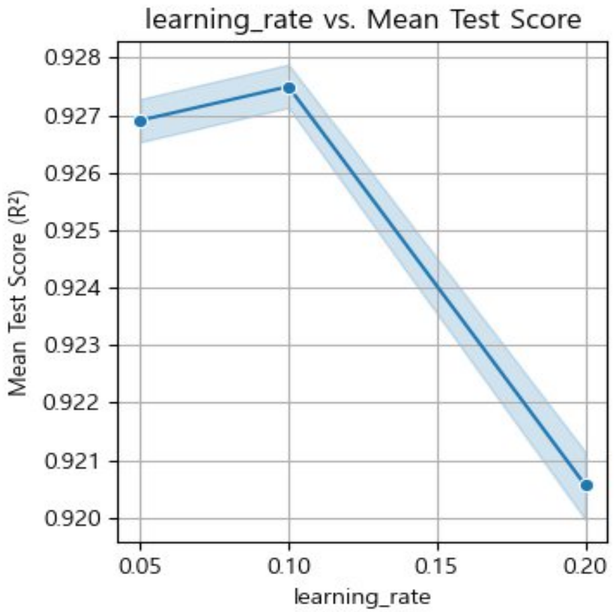


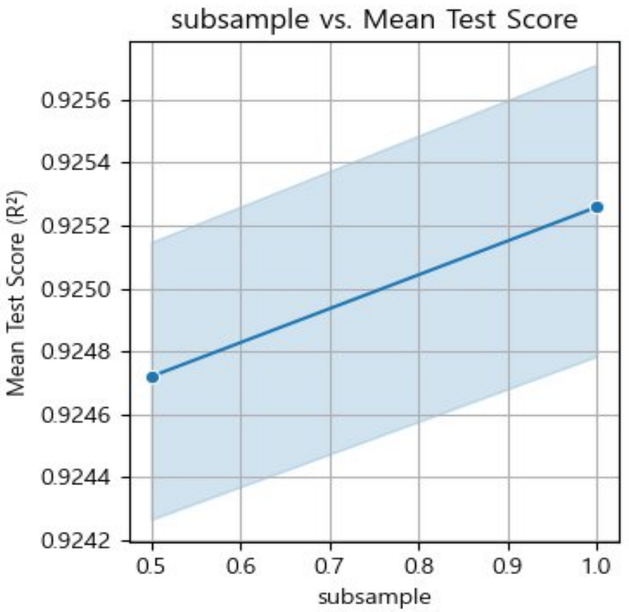
Hyperparameter tuning

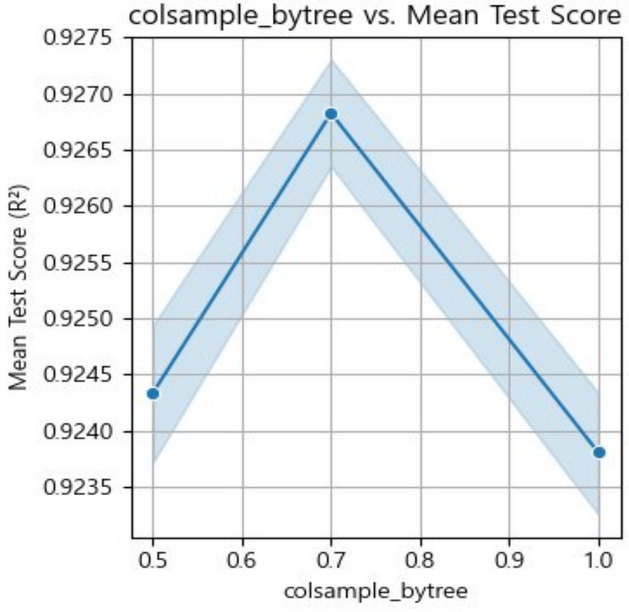
하이퍼 파라미터 조정

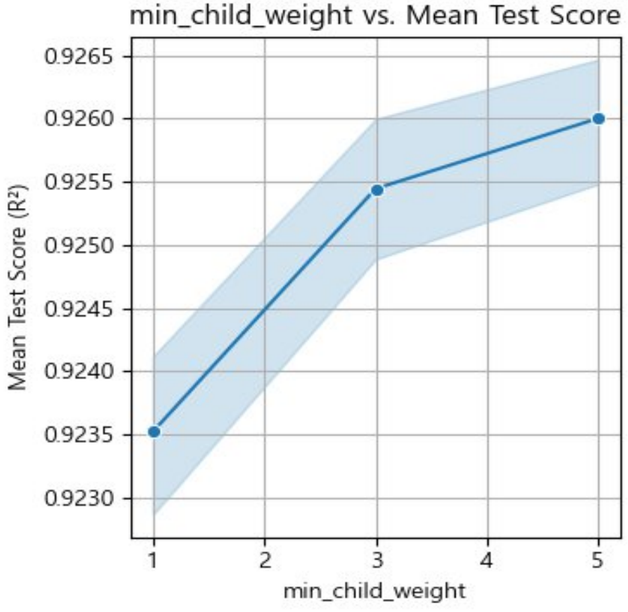


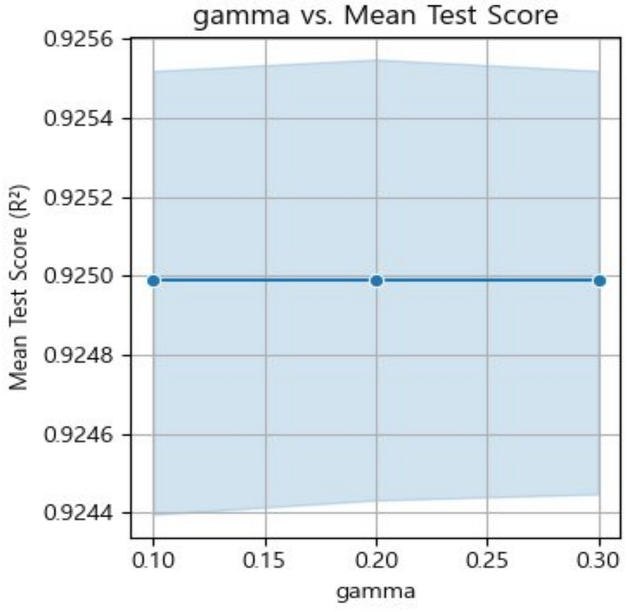












Unsampled Data

XGBoost

오차율 **0.096**

- **R^2** 0.936

- **MAE** 9506.353
- 평균 절대 오차

- **MSE** 316722287.767
- 평균 제곱 오차

- **RMSE** 17796.693
- 평균 제곱근 오차

Thank You

아파트