주차수요 예측 AI 경진대회

Presentation

1등에 주차

- statkwon
- duck
- 포효하는햄찌

0. CONTENTS

PPT 목차 설명

1. 데이터 설명

대회 설명 및 분석 목표, 데이터 소개

2. 데이터 전처리

NA 처리, 단지단위 데이터 생성

3. EDA

모형방향 설정 및 파생변수 생성

4. 외부변수추가

공공데이터 및 크롤링 데이터 활용 변수 생성

5. 예측 모형 선정

선형 모형 적합 및 성능 비교

6. 결과 및 성능

DACON 성적 (MAE 및 등수)

1. 데이터 설명

1등에 주차

Statkwon | duck | 포효하는햄찌

Introduce the Contest & Data

대회 배경설명 및 활용 목적

LH에서 제공한 데이터를 토대로, 유형별 **임대주택** 설계 시 단지 내 적정 P **주차 수요**를 예측

기존의 경우, 법정주차대수 및 장래주차수요 중 큰 값에 따라 주차대수 계산하는 방식을 사용.
이때, 장래주차수요는 인력조사로 진행, **오차가 발생할 우려**가 존재함.

➡ 새로운 모델을 통해 **주차 수요를 예측**하고, 이를 통해 임대주택 건축 시 **적정 주차면수를 계산**해내는 것이 목표

Introduce the Contest & Data

제공 데이터에 대한 기본적인 소개

데이터 제공

대회기간 중 수집된 임대주택 데이터

LH한국토지주택공사에서 공급하고 있는 임대주택들

age_gender_info.csv

지역내 성별 및 연령별 인구 구성비율

각 지역의 성별 및 인구를 10대 이하부터 100대까지,

구성 비율로 나타낸 데이터

Train.csv

총 423 개 단지로 구성된 train data

주로 상가에서 임대료, 임대보증금에서 NA값이 존재

지하철역, 버스정류장 수에서도 NA값 일부 존재

Test Data

총 150개 단지로 구성된 test data

자격유형, 임대료, 임대보증금에서 NA값이 일부 존재

지하철역, 버스정류장 수에서도 NA값 일부 존재

🎯 Target variable : 등록차량수 | 💡 Dependent variables : 임대료, 전용면적, 자격유형, 공급유형 등 12개 변수

제공데이터 변수설명

'train.csv', 'test.csv'로 제공된 데이터의 기본 변수들 설명

단지코드 (object)
각 단지의 고유 코드
단지별로 데이터 통합 후 삭제 예정

총세대수 (int)
각 단지의 총 세대수 (임대상가, 공공분양도 포함)

임대건물구분 (object)
아파트 | 상가로 구분
대부분의 상가는 한 세대당 row 하나씩

지역 (object)
Train데이터의 경우 인천광역시를 제외한
5대 광역시, 서울특별시, 제주, 세종 및 8개도,
Test의 경우 train과 동일하나 서울특별시 제외

공급유형 (object)
자격유형 (object)
자격유형의 경우, 비식별화 되어있음

전용면적
단지내 각 아파트 유형별 전용면적

전용면적별세대수
각 전용면적 type에 해당하는 세대수

공가수
해당 단지 내에서 비어있는 세대 수

임대보증금
각 전용면적에 해당하는 세대 별 임대보증금
상가 데이터에 대해 NA 다수 존재

임대료
각 전용면적에 해당하는 세대 별 월 임대료
상가 데이터에 대해 NA 다수 존재

도보10분거리내 지하철역 수

도보 10분거리내 버스정류장 수
지하철의 경우 대부분 0

단지내 주차면수
LH에서 임대주택 건설/매입 시 기록/예측한
단지내 주차면수

Target var.
등록차량수

왜 단지단위의 데이터를 사용했을까?

세대단위가 아닌 단지단위로 데이터를 합쳐서 모델링

유형별 주택단위 데이터

대회에서 제공된
train, test data에는
각 단지내 세대 유형별
로 row가 존재

타겟 변수인 '등록차량수'가
단지 단위로 주어져 있음

해당 변수를 주택 단위로 나누는
과정에서 추가적인 오차가 발생할
가능성 존재

→ 각 단지 데이터로 통합해 사용

단지단위 데이터

최종 데이터로는
각 단지 별 데이터로
통합 후 모델링

유형별 주택 단위의 데이터를 **단지 단위**로 통합하여 사용

주어진 데이터를 그대로 사용해 세대별 주차수요를 예측 후 통합하는 방식의 경우,
오차가 발생할 우려가 존재함.

2. 데이터 전처리

1등에 주차

Statkwon | duck | 포효하는햄찌

Data Preprocessing

오류단지 데이터 제거 | NA imputation | 단지단위 변수로 변환

오류데이터 제거

데이콘 공지사항에 등록된 **오류데이터**는 행에서 **제거**

Train data: 'C2085', 'C1397', 'C2431', 'C1649', 'C1036', 'C1095', 'C2051', 'C1218', 'C1894', 'C2483', 'C1502', 'C1988'

Test data: 'C2675', 'C2335', 'C1357' 단지 제거

NA 처리

NA는 0으로 대체하거나, 동일 단지코드 내의 값으로 대체

'도보 10분거리 내 지하철역 수(환승노선 수 반영)', '임대료' 등의 결측치는 0으로 대체

Test data의 경우, 동일 단지내에서 값을 찾을 수 있는 경우 자격유형 및 임대료 결측치를 동일 단지내 데이터 값으로 대체

단지단위 변수생성

변수 별 각기 다른 전략을 이용하여 **단지 단위로 변수를 합침**

임대료, 임대보증금, 전용면적: 전용면적별 세대수에 따른 **가중평균**

공급유형: 유형별 더미변수화

Data Preprocessing

오류단지 데이터 제거 | NA imputation | 단지단위 변수로 변환

✓ 자격유형이 공급유형에 포함된다고 판단하여

→ 자격유형 변수 제거

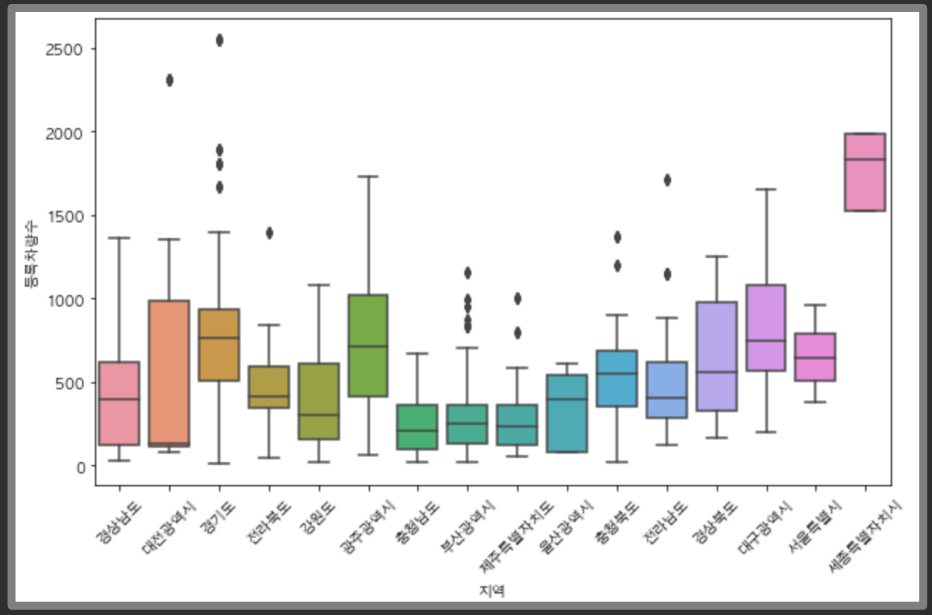
자격유형 (J, K, L, M, N, O)이 공급유형 행복주택으로 묶일 수 있음

자격유형	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
공급유형															
공공분양	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
공공임대(10년)	175	0	0	0	0	0	0	0	0	0	0	0	0	0	0
공공임대(50년)	31	0	0	0	0	0	0	0	0	0	0	0	0	0	0
공공임대(5년)	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
공공임대(분납)	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
국민임대	1508	21	0	0	34	0	9	155	0	0	0	0	0	0	0
영구임대	2	0	95	0	3	3	0	0	49	0	0	0	0	0	0
임대상가	0	0	0	562	0	0	0	0	0	0	0	0	0	0	0
장기전세	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
행복주택	0	0	0	0	0	0	0	0	0	103	33	33	2	30	1

✓ 지역별 등록차량수에 차이가 적다고 생각하여

→ 지역 변수 제거

지역별 데이터 개수 차이를 고려하여 비교



그 외에도 타겟 변수에 유의미한 영향을 미치지 않는다고 생각되는 **임대건물구분, 전용면적별세대수** 변수 **제거**

Data Preprocessing

오류단지 데이터 제거 | NA imputation | 단지단위 변수로 변환

단지단위 변수 생성 example

같은 단지(‘C2515’) 내 변수 값들은 전용면적 및 세대 유형별로 각각의 row를 가짐

	단지코드	공급유형	전용면적	전용면적별세대수	자격유형	임대보증금	임대료
0	C2515	국민임대	33.48	276	A	9216000	82940
1	C2515	국민임대	39.60	60	A	12672000	107130
2	C2515	국민임대	39.60	20	A	12672000	107130
3	C2515	국민임대	46.90	38	A	18433000	149760
4	C2515	국민임대	46.90	19	A	18433000	149760
5	C2515	국민임대	51.97	106	A	23042000	190090
6	C2515	국민임대	51.97	26	A	23042000	190090

전용면적, 임대료, 임대보증금은 **‘전용면적별세대수’로 가중평균**

	단지코드	공급유형	전용면적	자격유형	임대보증금	임대료
0	C2515	국민임대	39.0	A	14035965.0	119432.0

3. EDA

1등에 주차

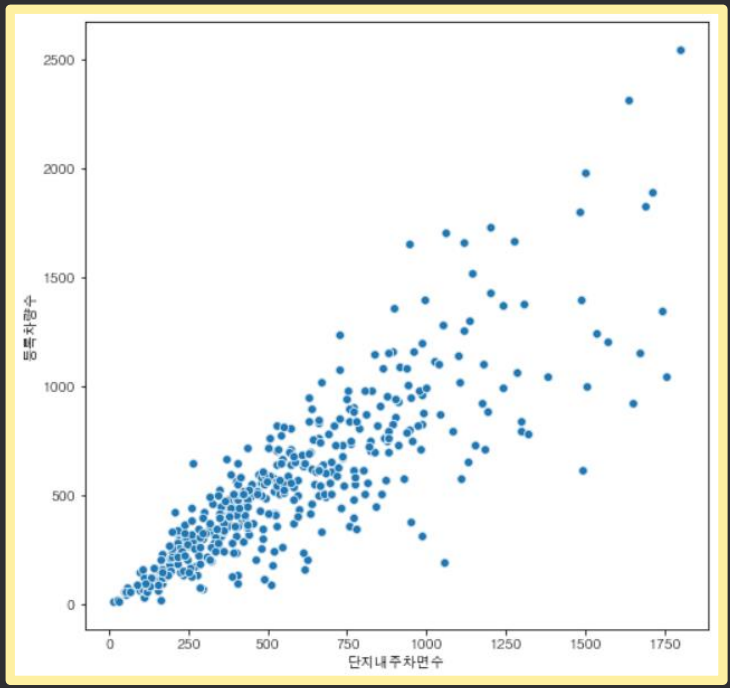
Statkwon | duck | 포효하는햄찌

Exploratory Data Analysis

변수에 대한 EDA 진행, 모델 방향 설정

Scatterplot

X: 단지내주차면수 | y: 등록차량수



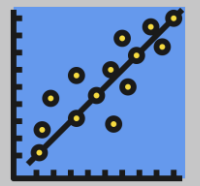
단지내주차면수와 등록차량수 간 **선형관계**가 보임

어떤 변수를 사용할 것인가

현재 있는 변수는 전용면적, 임대료 … etc
크롤링 데이터로부터 외부변수 추가

모형 선택: 선형모델

Linear Regression, Ridge, Lasso, MARS …
여러 모델 적합 후, 평균적으로 가장 좋은 CV score 내는 모형 선택



고민 : 등분산성?

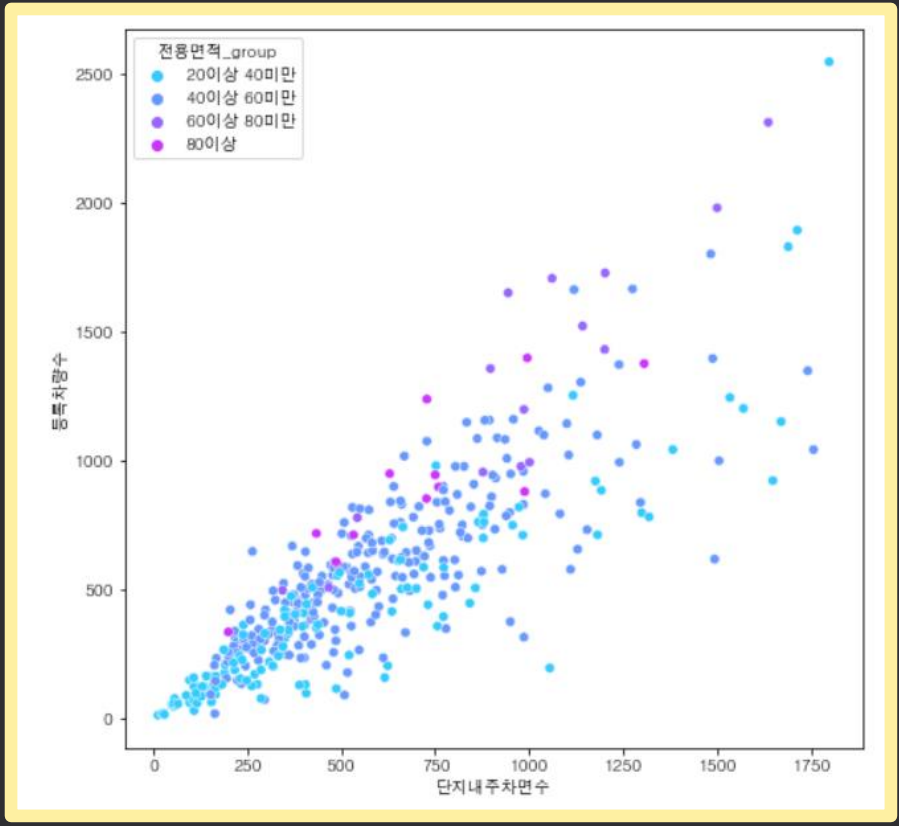
선형회귀 사용 시, 등분산성에 대한 고민 필요
Weighted Least Square, Quantile Regression 시도

➡ 단지내주차면수와 등록차량수 간 유의미한 선형 관계가 있다고 판단, 여러 변수 추가한 뒤 **선형모델** 사용

파생변수 생성

myhome data crawling | 외부 변수에 대한 EDA | 외부 변수 추가

전용면적 크기별로 분포가 달라지는지 확인
X: 단지내주차면수 | y: 등록차량수



전용면적이 작은 소형세대의 경우
세대 내 거주하는 인원이 적을 것이며,
차량 보유 인구 또한 적을 것이라 추측

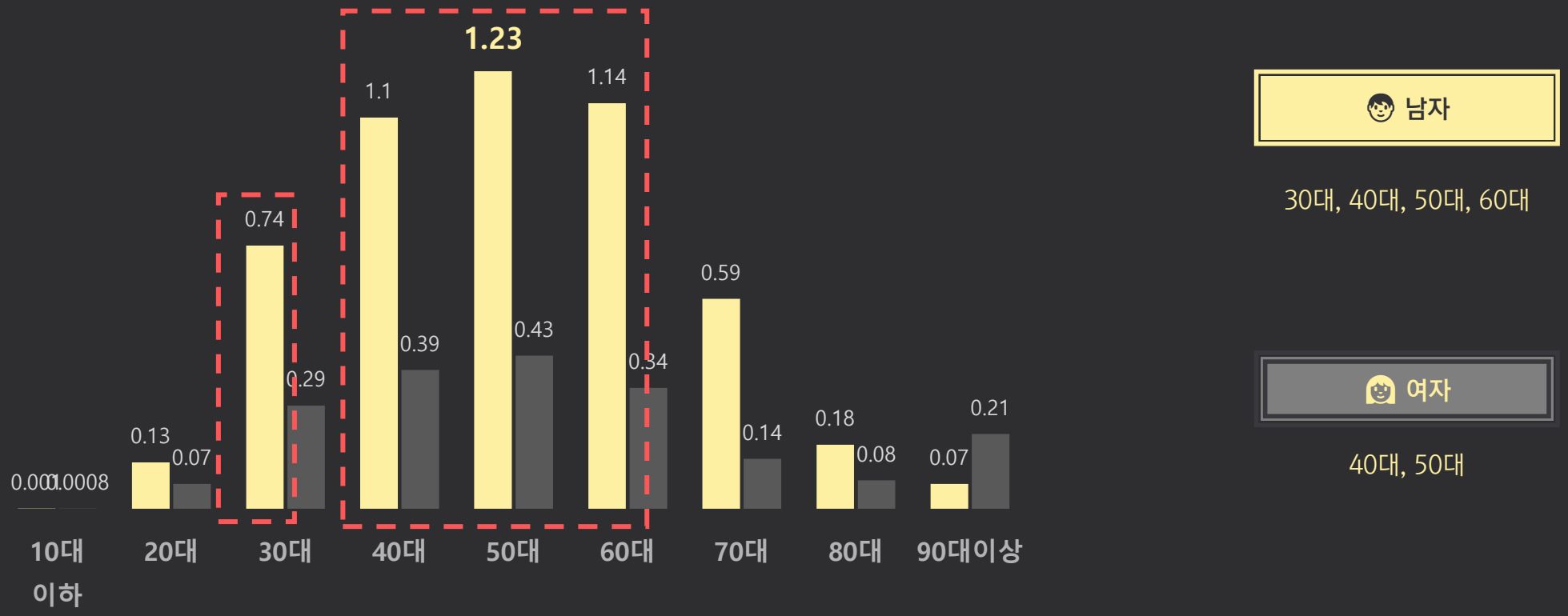
전용면적이 큰 세대의 경우
소형 세대에 비해 세대내 거주 인원이 많을 것
자녀 등 가족구성원 문제로
차량보유수가 비교적 많을 것이라 추측

세대별 전용면적별로 데이터의 분포가 달라지는 양상을 보임
단지내 총 세대수에서 **소형세대(전용면적 40 이하)가 차지하는 비율**을 나타내는 변수, **'소형세대'** 추가

파생변수 생성

myhome data crawling | 외부 변수에 대한 EDA | 외부 변수 추가

'age_gender_info.csv'로부터 '**차량보유인구비율**' 변수 생성, 추가



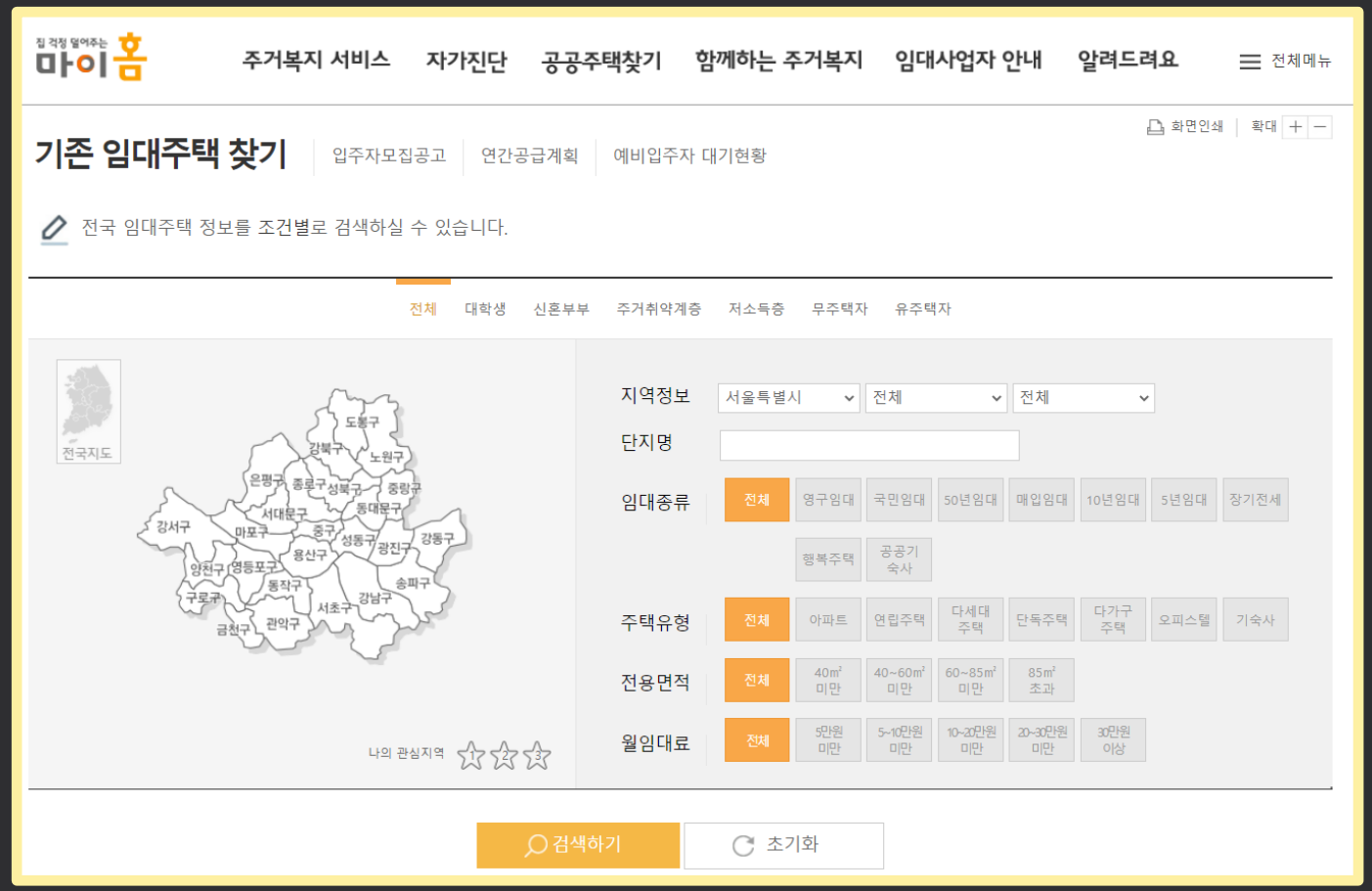
{30대(남), 40대(남), 50대(남), 60대(남), 40대(여), 50대(여)}를 차량보유인구로 정의
해당 파일로부터 지역별 '**차량보유인구비율**' 변수 생성 후 merge

4. 외부 변수 추가

1등에 주차
Statkwon | duck | 포효하는햄찌

외부 변수 추가

마이홈 데이터 크롤링



마이홈포털

전국 임대주택에 대한
단지명, 주소, 위/경도, 준공일자 등
데이터 존재

지역, 총세대수 등을 기준으로 기존
데이터에 단지명을 Mapping

https://www.myhome.go.kr/hws/portal/main/getMgtMainPage.do

외부 변수 추가

차량보유인구비율

지역별 차량보유확률이 높은 연령대[30대(남), 40대(남), 50대(남), 60대(남), 40대(여), 50대(여)]에 대한 비율을 구함

소형세대

총세대수 대비 전용면적이 $40m^2$이하인 세대수의 비율을 구함

차량보유입주민수

아래의 두 공공데이터를 이용하여 세대별 총입주민수 데이터를 생성.
총입주민수 중 상대적으로 차량보유비율이 높은
30대(남), 40대(남), 50대(남), 60대(남), 40대(여), 50대(여) 의 입주민수를 구함.

- [한국토지주택공사_임대주택 단지별 연령대별 성별 정보] https://www.data.go.kr/data/15059813/fileData.do
- [myhome crawling data] https://www.myhome.go.kr/hws/portal/sch/selectRentalHouseInfoListView.do

외부 변수 추가

최종 사용 변수

기존변수(변형)
단지내주차면수, 전용면적, 총세대수, 공가수

추가변수
소형세대, 차량보유인구비율, 차량보유입주민수

공공데이터 활용 파생변수
행복주택, 영구임대, 임대상가, 공공임대(10년), 국민임대

5. 예측 모형 선정

1등에 주차
Statkwon | duck | 포효하는햄찌

Modeling

여러 선형모델 적합 후 score(MAE) 비교, 최종 모델 선택

MAE

Linear Regression

113.74

표준편차: 11.208430511575735

Ridge

113.20

표준편차: 12.676535165769227

LASSO

113.31

표준편차: 13.484610888884106

Polynomial Regression

$1.12 * 10^{11}$

표준편차: 311848184333.64246

Ridge with Polynomial Features

106.64

표준편차: 10.512571328786787

LASSO with Polynomial Features

105.54

표준편차: 10.896002954365605

- Linear Reg./Polynomial Reg./Ridge/Lasso 비교
- Score 기준은 MAE
- Standard scaler 이용해 scaling 후 모델 적합
- Ridge, Lasso의 경우 MAE 최소화하는 최적의 alpha값 이용해 모델 적합

최종모델 선택

LASSO with Polynomial Features

6. 예측결과 및 성능

1등에 주차
Statkwon | duck | 포효하는햄찌

예측결과 및 성능

test data 로 예측한 결과

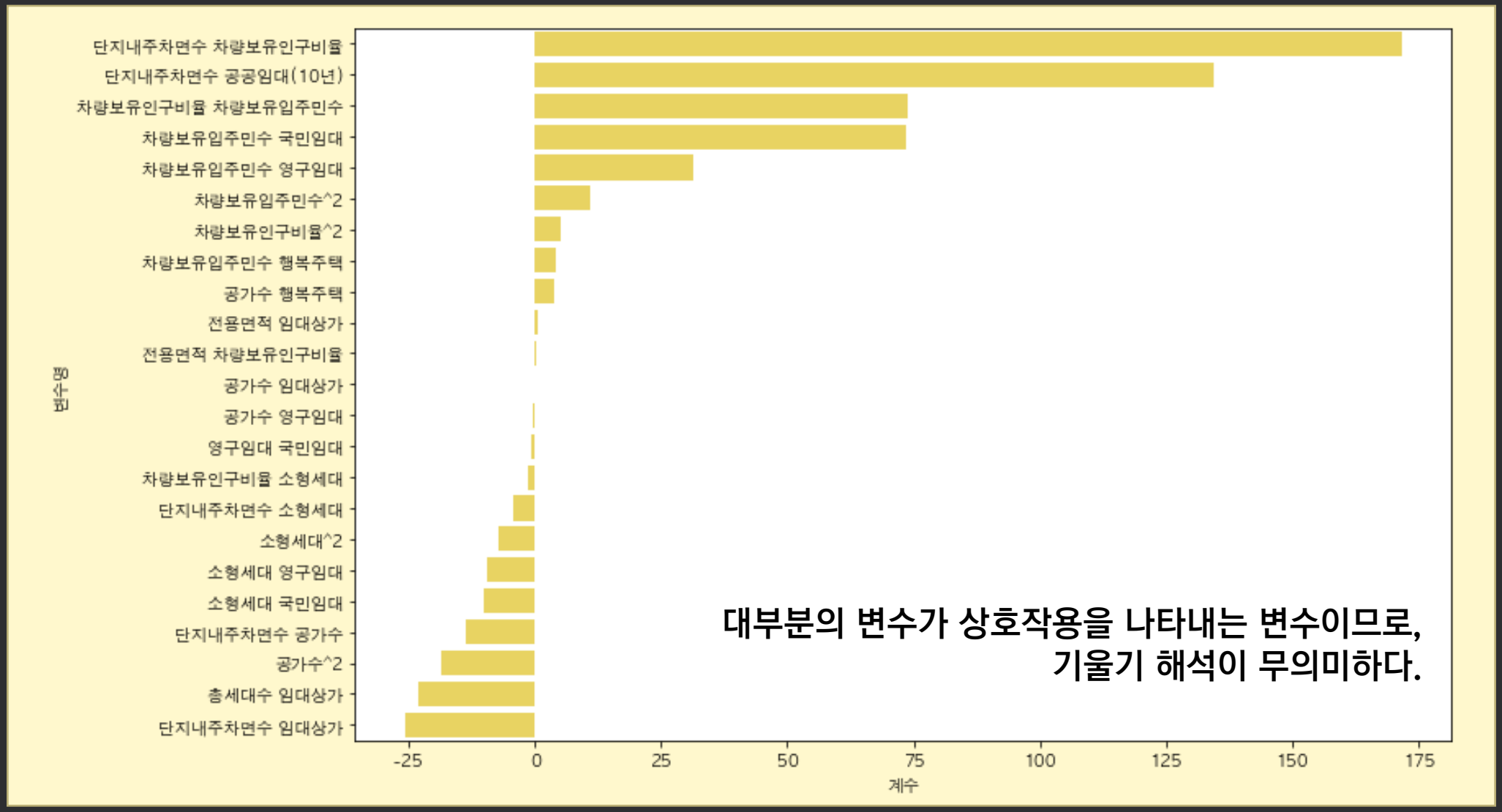
(CV) Estimated MAE: 105.54305623701052

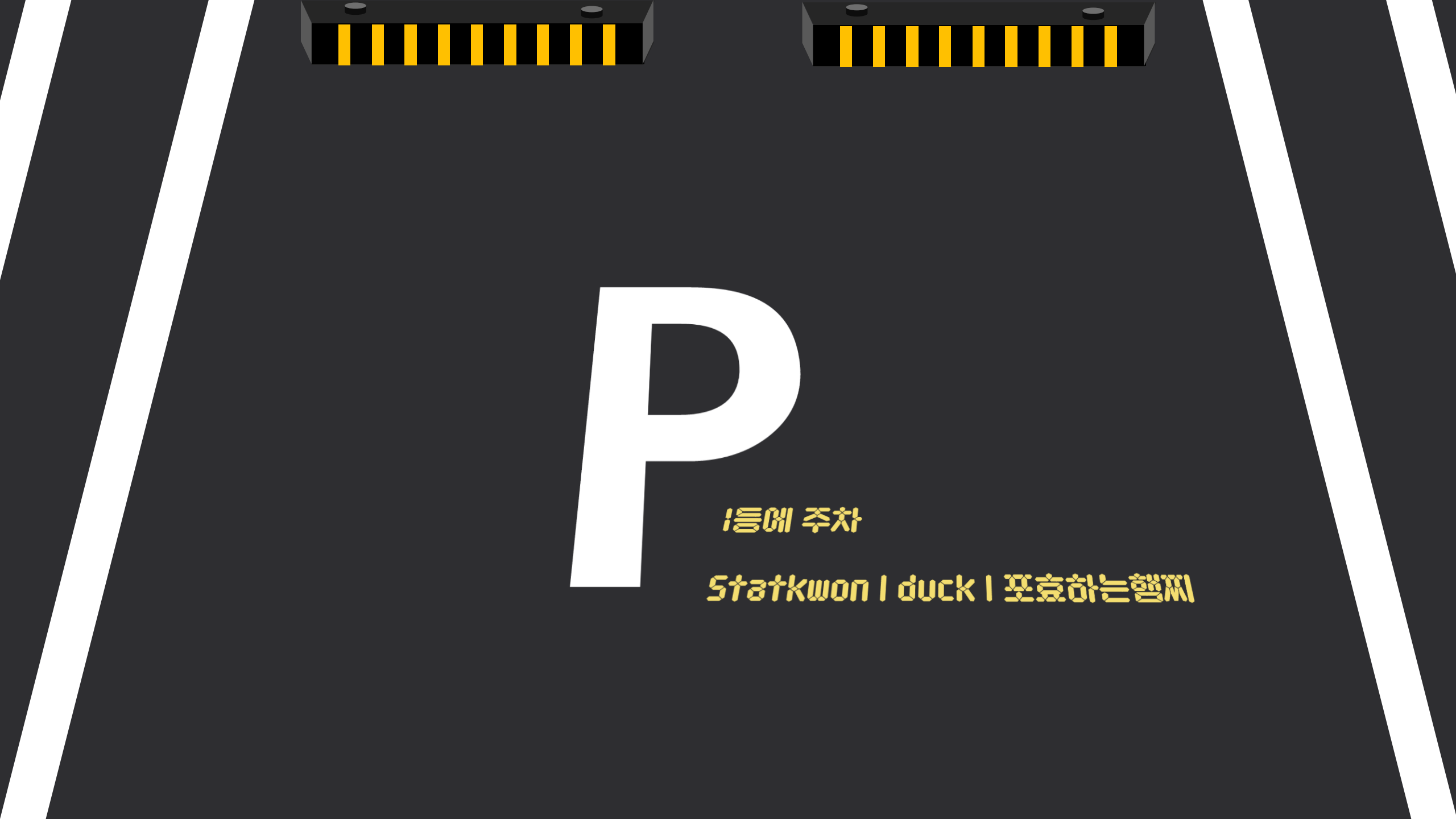
Dacon Public MAE	105.4968492188
Dacon Private MAE	105.5256734854

최종 결과: 상위 4%(16등/503팀)

최종 모델 해석

최종 모델에서 선택된 변수들, 해석




P
1등에 주차
Statkwon | duck | 꼬효하는햄찌