머신러닝 회귀 모델을 통한

중고차 판매가격 예측 API 서비스

Codestates
[AIB] 9기 안나

Contents

중고차 판매 가격 예측 API 서비스

주제 선정 이유

- 다양한 정보를 활용하여 각각 개인에게 **맞춤 서비스**를 제공하는 것이 중요한 시대
- **머신 러닝**을 활용하여 간단하면서도 필요한 정보를 제공 할 수 있도록 기획함
- **고관여 제품** (가격이 높고, 소비자들의 의사결정과정이나 정보처리과정이 복잡한 제품)을 선정하여 머신 러닝의 활용 가치를 높임

데이터 특성

- 중고차 데이터는 웹에서 수집하여 사용하므로 **허위매물, 잘못된 정보 등을 검증**하는 것이 중요
- 한국 데이터는 **교차 검증** 할 수 있는 자료가 유료인 경우가 많아 편의를 위해 **미국**사례로 선정
- 미국 정부자료 (자동차 이력 제공시스템, 미국 교통부 연방차량안전국 등)를 받아 제공하는 **API로 교차 검증**하여 사용

데이터 수집

kaggle

Craigslist(중고 매물 커뮤니티)
미국 중고차 매물
약 51만 건 스크레이핑 자료
→ VIN (자동차 등록번호) 추출

VinAudit

VIN 번호 교차 검증
API 자료 스크레이핑
→ 실제 매물 **약 7만 건** 확인
자동차 정보 및 가격정보 수정

DB 적재
자료 추출

PostgreSQL

ElephantSQL
클라우드 DB 사용

데이터 분석

scikit learn python

ML 모델링
(LinearRegression)
회귀 분석

Metabase

docker

EDA 및 시각화
대시보드 구성
((배포))

API 서비스

Flask

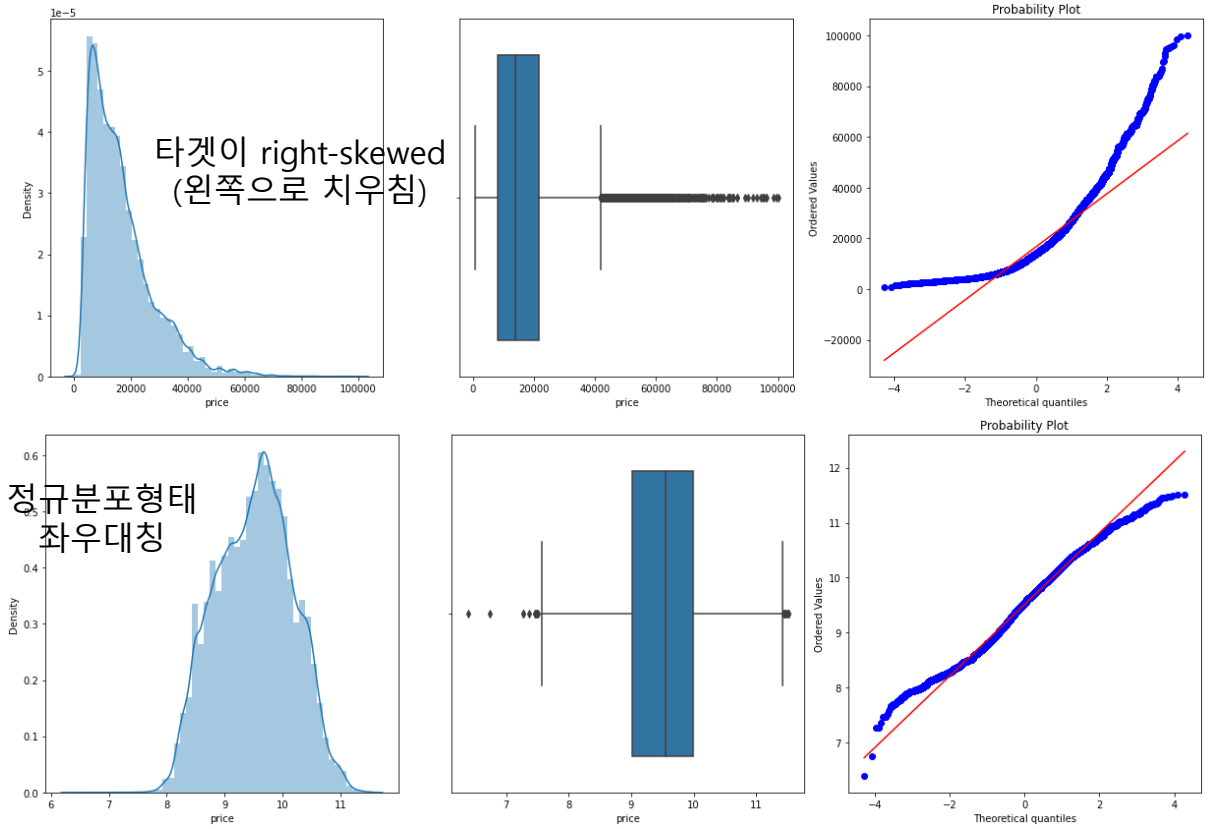
HEROKU

API 서비스 구현
((배포))

분석 Target : **가격**

- **회귀분석**을 통해 변수들 사이에서 나타나는 경향성을 설명
- 회귀 분석은 타겟 변수가 **정규분포**에서 좋은 성능을 보이며, 비대칭 형태인지 확인 필요

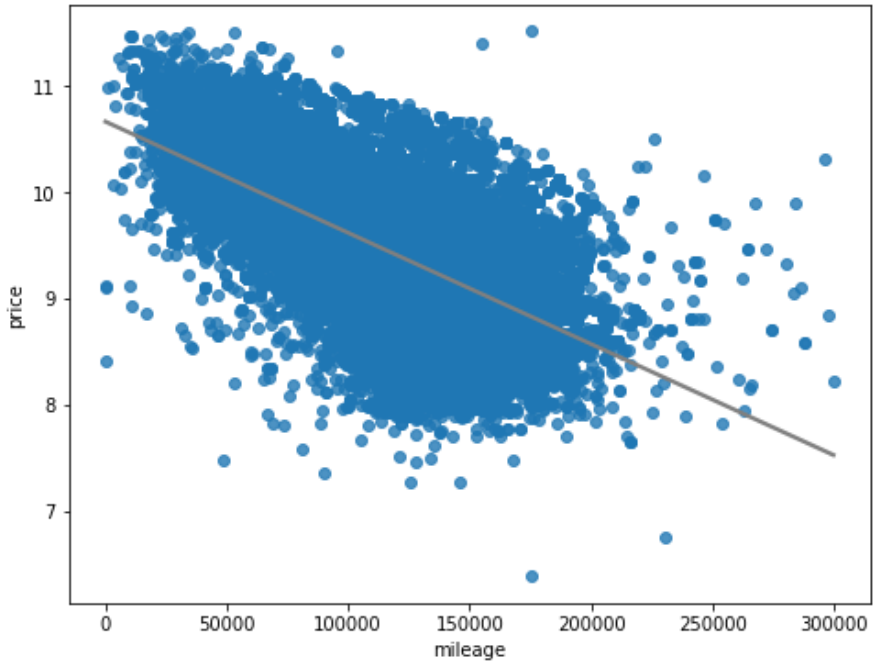
로그변환 사용
비대칭 분포형태를
정규분포형태로 변환



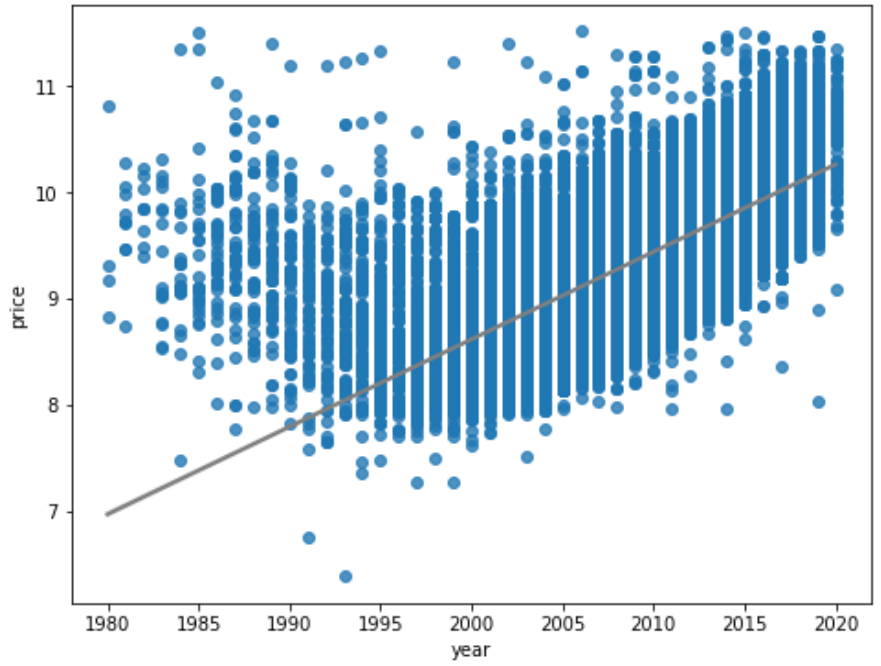
변수 분석

- 분석하고자 하는 **가격**에 영향을 줄 수 있는 독립변수가 다양함
- **독립변수** : 연식, 주행거리, 엔진 기통수, 연료타입, 구동방식, 차량 색상 등
- 수치형 변수인 **운행거리(마일), 연식**과의 상관관계를 우선 파악

운행거리가 증가하면 가격이 낮다 (음의 상관관계)

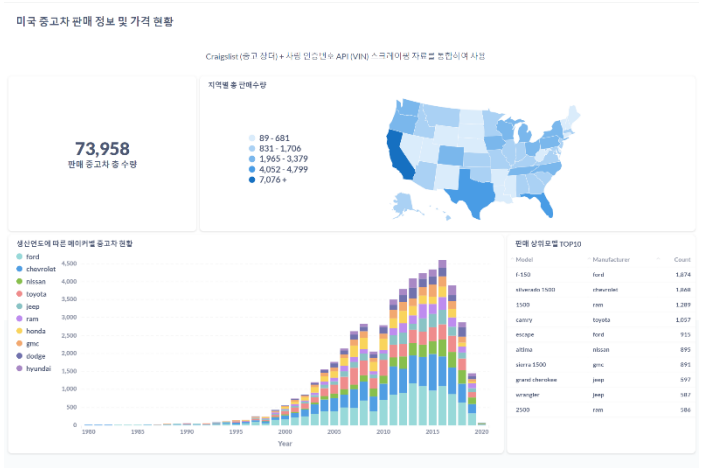


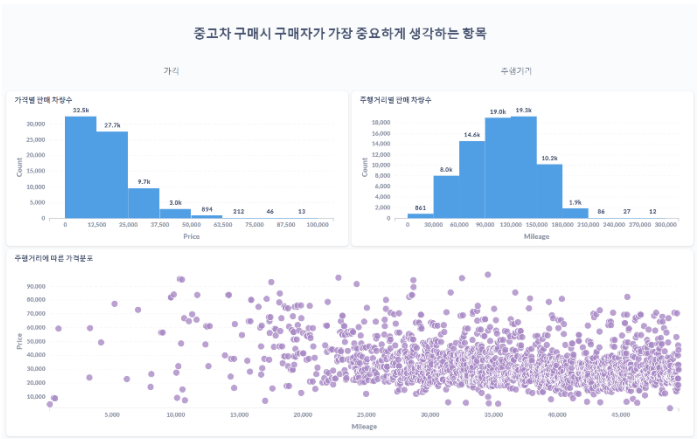
출시된 연도가 최근이면 가격이 높다 (양의 상관관계)

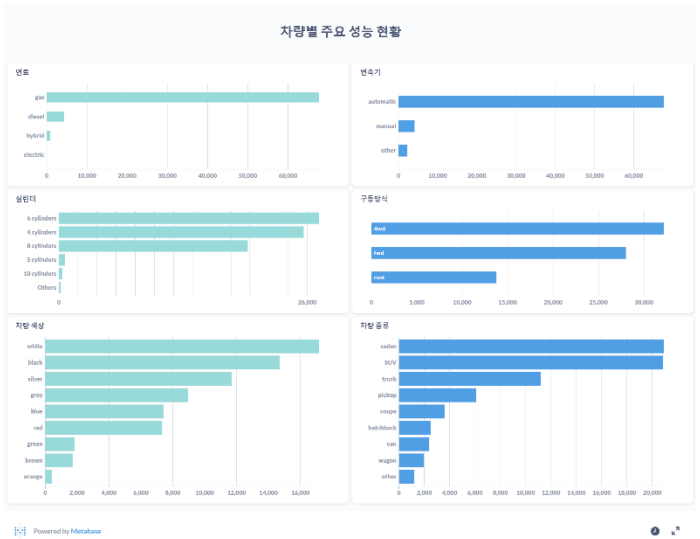


대시 보드 구성

- 다양한 변수들의 관계는 대시보드를 통해 쉽게 파악 가능
- **Metabase**를 사용하였고, **온라인 배포** 하여 주소로 접속할 수 있음
- 지역 별 총 판매 수량, 판매 상위 모델 및 판매 현황 등을 한눈에 확인 할 수 있음







https://usedcardashboards.herokuapp.com/public/dashboard/34de8990-8db3-4116-9301-5503336a8b71

머신러닝(ML) 모델링

- 여러 개의 독립변수를 가지고 가격을 예측하기 위한 회귀 모형 (**다중선형회귀 모델**) 사용
- 다양한 평가지표를 검증 후 모델이 **전체 현상을 얼마나 설명하는지 알려주는 R2 Score**로 최종 비교

기준 모델

- 가장 간단하면서도 직관적
- 최소한의 성능을 나타내는 기준
- **회귀**: 타겟의 **평균** 사용

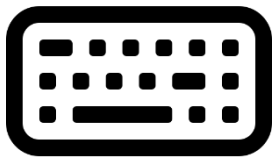
	Metric	Score
0	MSE	0.426039
1	MAE	0.538630
2	RMSE	0.652717
3	R2	0.000000

다중선형회귀

- 사용 특성 수에 따라 **R2 score**의 차이가 커서 전체 데이터 사용 (일부 특성 사용: 0.5, 전체사용: 0.7, 설명력의 차이가 크다)
- 이후 서비스 구현 시 유저 편의를 위해 필수데이터는 유저 입력, 다른 데이터는 기존 DB자료를 활용 예정
- 다른 평가 지표도 개선되었으며, R2 Score가 0.7로 모델의 설명력이 적당하다.

	Metric	Score
0	MSE	0.124160
1	MAE	0.265203
2	RMSE	0.352364
3	R2	0.708571

유저 정보 입력



자동차 모델명
운행 거리 (마일)
연식 (연도)

submit

API

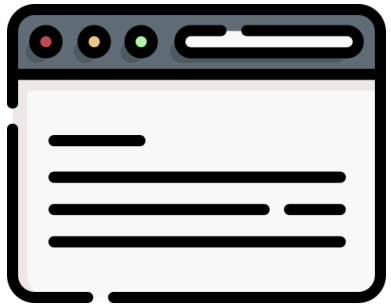
모델명과 매치되는
기타 정보를
중고차 매물 **DB에서**
추출해서 사용

제조사
실린더
구동방식
연료
변속기
차량 종류
....

Model.pkl

머신러닝 모델을
적용하여
가격을 예상

웹 페이지 출력



예상 가격은

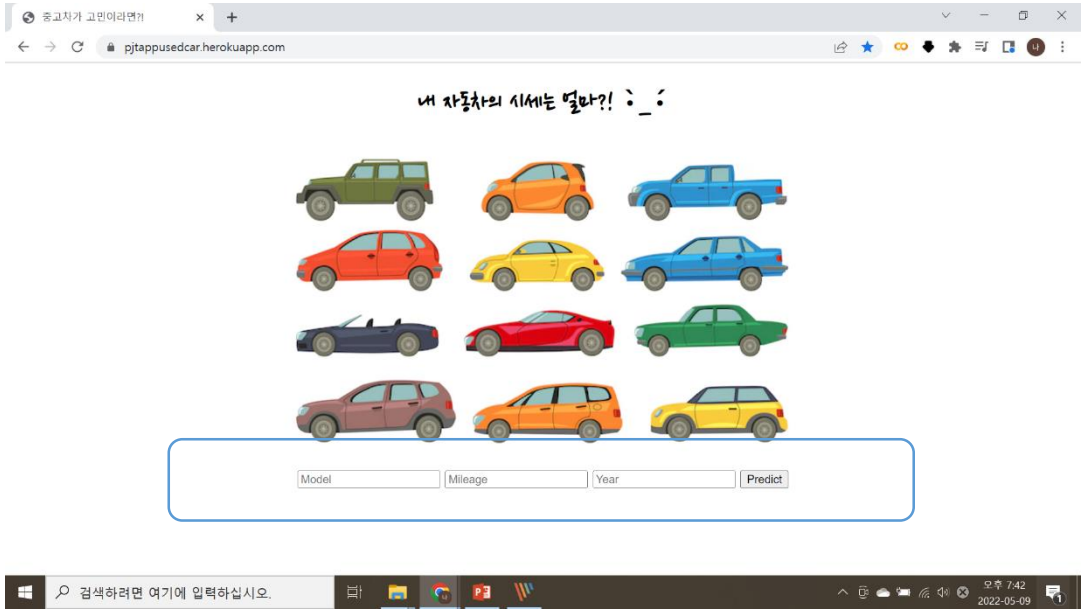
$ 100

대시 보드 및 앱 서비스 시연

중고차 판매 가격 예측 API 서비스

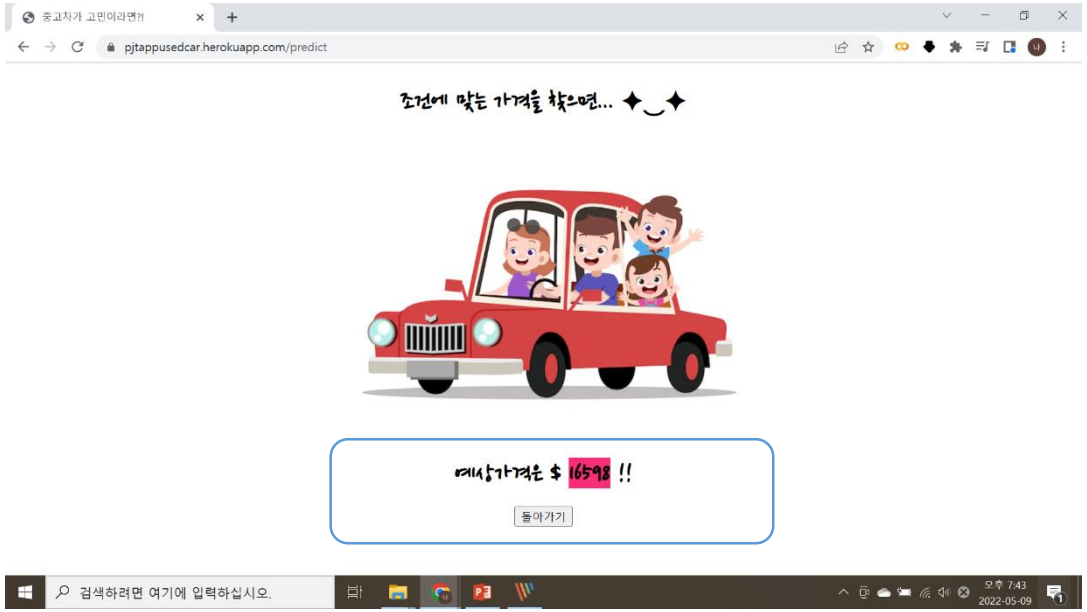
https://pjtappusedcar.herokuapp.com/

메인 페이지



자동차 모델명, 운행 거리 (마일), 연식 (연도)
필수 값 입력창

결과 페이지



예측 가격 안내

중고차 판매 가격 예측 API 서비스

느낀 점

- 다양한 tool들이 서로 어떻게 연관되어 있는지 생각하며 전체 데이터 파이프라인을 온전히 구축해 볼 수 있어서 보람 있었습니다.
- 또한 머신러닝 모델을 단순 분석이 아닌, 간단하게라도 실제 고객에게 제공할 수 있는 서비스로 직접 구현하여 시야가 넓어진 느낌이 들었습니다.

이슈 사항

- ML 모델 예측결과를 ①운행거리에 따른 판매가격 변화, ②연식에 따른 판매가격 변화의 그래프 2가지로 구현하고 싶었는데, 웹 페이지 구현이 어려워 간략히 결과 값만 출력하게 되어 아쉬웠습니다.
- ML모델(LinearRegression)을 활용한 서비스를 기획 했는데, 딥러닝을 활용하면 모델의 성능을 개선 할 수 있을 것 같습니다. 반응속도를 감안하여 결과를 비교해보면 더욱 좋을 것 같습니다.

Thank you for your attention!