CodeStates Project 3

Data Pipeline

AIB_05 최중훈

목차

1. 서비스 소개

1. 서비스 소개

중고차 가격 예측 서비스

- 중고차 사이트에 공개된 정보를 바탕으로 중고차 가격을 예측할 수 있는 머신러닝 모델을 개발

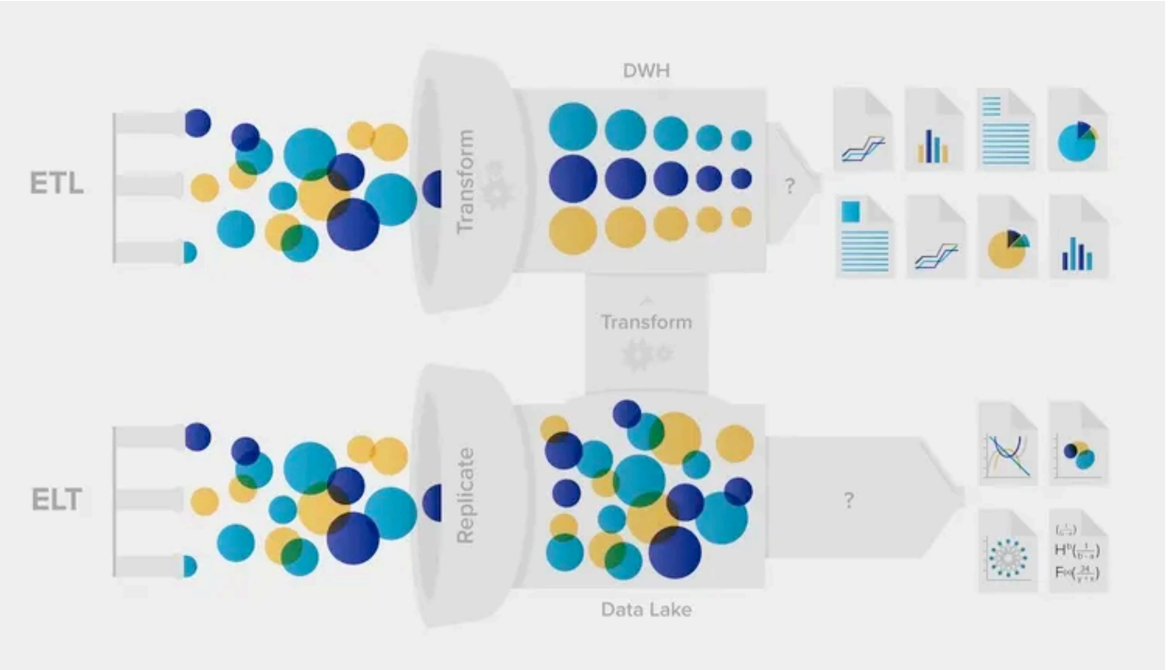
- 개발된 모델을 쉽게 사용할 수 있도록 웹 페이지에서 차량 정보를 입력하고, 결과를 받을수 있도록 배포

2. ETL

2. ETL

Extraction, Transformation, Load

- ETL이란 Extraction, Transformation, Load 의 약자



2. ETL

스크래핑 데이터 선정

- 중고차 사이트 스크래핑
- 제조사명, 모델명, 등록 연월, 주행거리, 연료 유형, 거래 지역, 차량 가격 정보를 수집
- 시간관계상 경차 차종만 데이터 수집

2. ETL

스크래핑 방법

- 동적 웹 스크래핑을 위해 selenium 사용
- html 소스보기로 확인한 결과 원하는 데이터가 없어 동적 페이지로 판단
- BeautifulSoup으로 html 소스 파싱

2. ETL

변환, DB 저장

- 데이터베이스에 저장하기 적합한 형태로 변환 후 저장
- 스크래핑한 데이터를 관계형 데이터베이스인 PostgreSQL에 저장

3. 모델링

3. 모델링

CatBoost

- 수집한 데이터중 명목형 변수의 비율이 높아 명목형 변수를 처리하는데 특화된 CatBoost를 사용

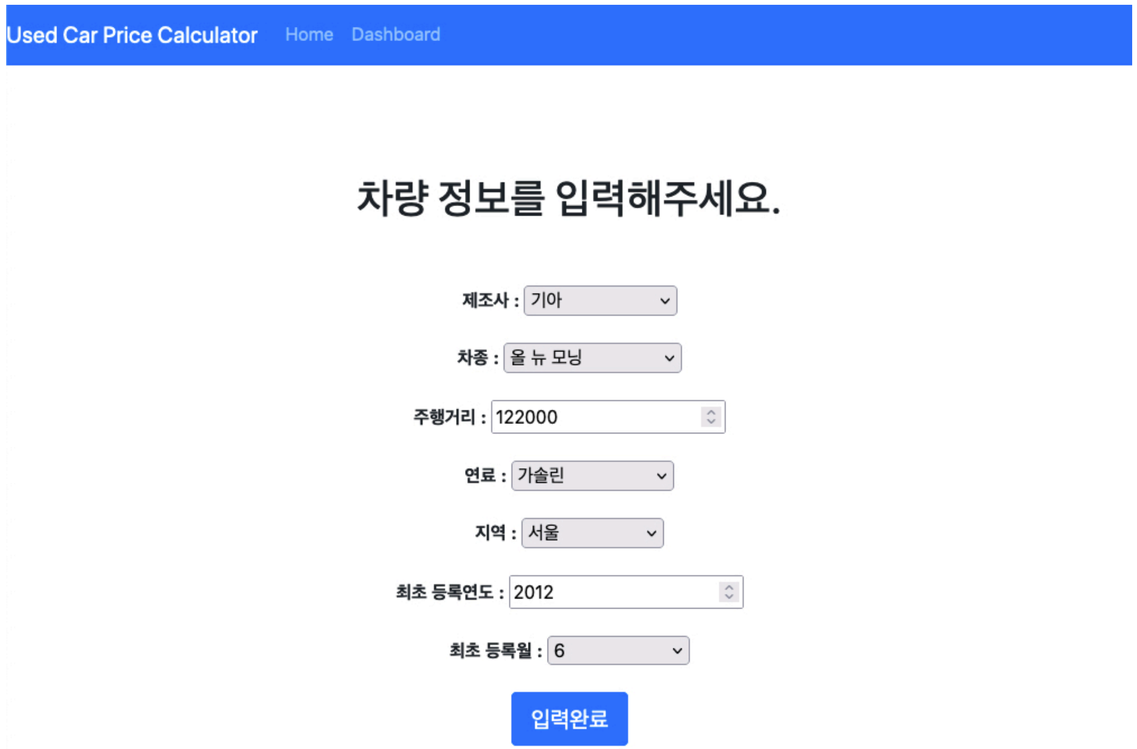
- 모델링 결과 R^2값이 0.9로 성능이 우수

4. 웹 어플리케이션 배포

4. 웹 어플리케이션 배포

Flask, Heroku

- 웹페이지는 Flask로 구성하고, Heroku로 배포

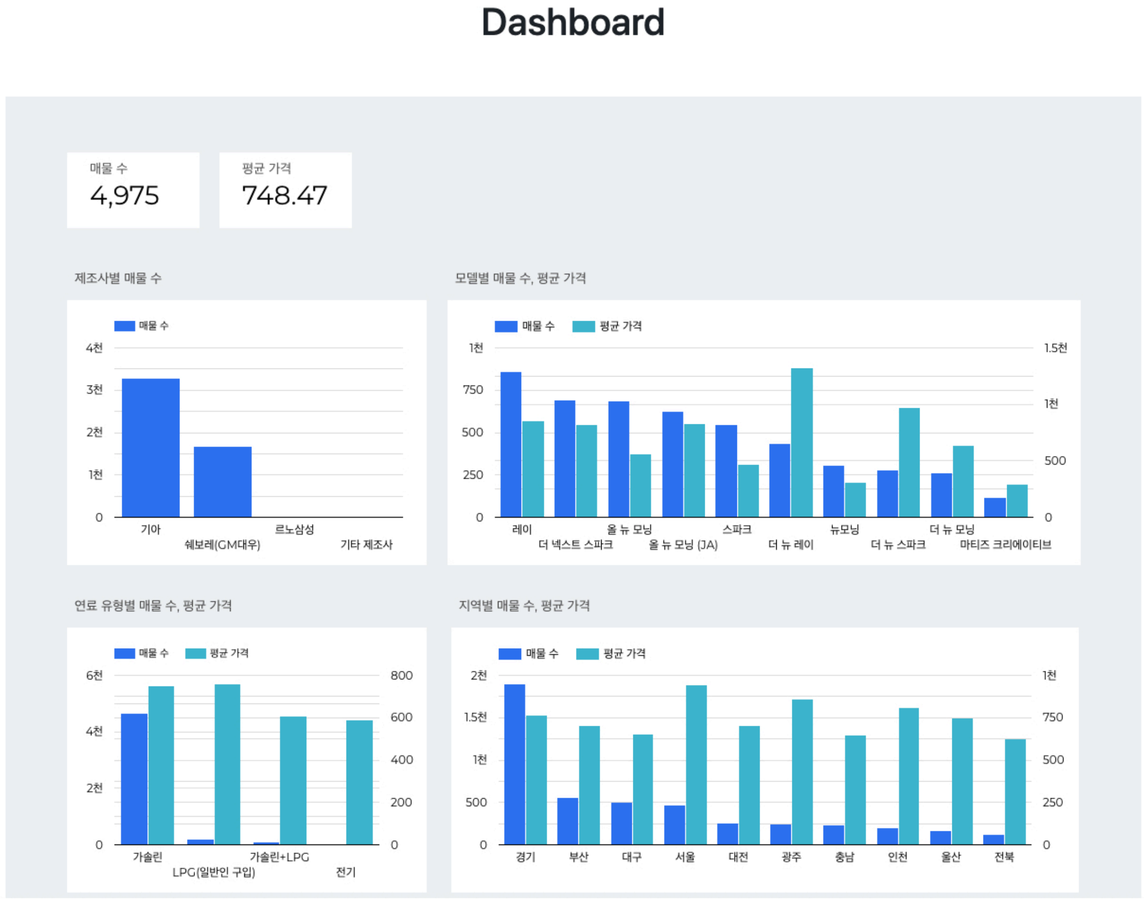


5. 대시보드

5. 대시보드

구글 데이터스튜디오

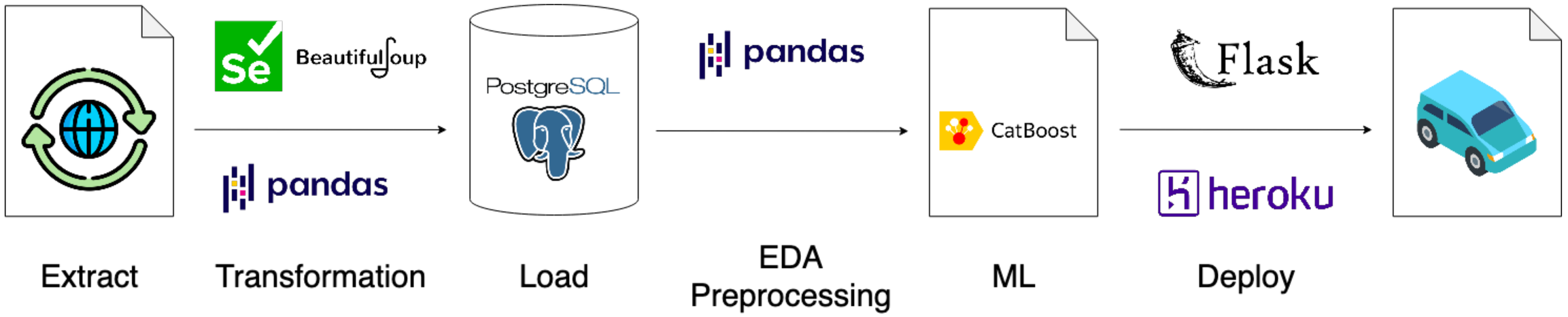
- 구글 데이터스튜디오를 이용하여 대시보드 제작
- 배포된 웹 어플리케이션에서도 대시보드를 열람할 수 있도록 페이지를 구성



6. 정리

6. 정리

데이터 파이프라인



6. 정리

국토교통부 자동차 종합정보 API서비스

- 차량 번호만으로 자동차 기본정보 32건, 제원 70건, 정비이력 16건의 정보를 열람
- 사용자는 더 간편하고 더 고도화된 모델로 중고차 가격확인 가능