Ai를 활용한 중고차 가격 예측 및 추천 사이트

정어진, 류건우, 연건모, 이성욱*
16106 경기도 의왕시 철도박물관로 157 한국교통대학교 데이터사이언스학과

Used car price prediction and recommendation site using AI

Eojin Jeong, Geonu Ryu, Kunmo Yeon, Songwook Lee*
Major in Data science, Korea National University of Transportation, Uiwang, 16106, Korea

ABSTRACT

Used car sites, which are mainly used by people, play an important role for both buyers and sellers of vehicles, Buyers often find it difficult to understand complex information, such as performance records and vehicle conditions, making rational decisions.

In particular, users who lack experience in purchasing used cars have difficulty determining the reliability of the information provided.

To solve this problem, we have developed 'Chachaza', a service that predicts used car prices and recommends suitable vehicles.

Keywords: Web Crawling, Machine Learning, Openai

Ⅰ. 서론

시중의 중고차 구매 사이트는 차량 구매자와 판매자 모두에게 중요한 역할을 하지만, 구매자는 성능 기록부나 차량 상태와 같은 복잡한 정보를 이해하기 어려워 합리적인 결정을 내리기 힘든 경우가 많다. 특히, 중고차 구매 경험이 부족한 사용자들은 제공된 정보를 가지고 적절한 가격인지 판단하는 데 어려움을 겪는다.

이러한 문제를 해결하기 위해 우리는 중고차의 적정 가격을 예측하고 적합한 차량을 추천해 주는 서비스인 '차찾자'를 개발하게 되었다.

'차찾자'는 엔카 API와 웹 크롤링 기술을 활용하여 차량 데이터를 수집하고, 이를 바탕으로 인공지능 알고리즘을 통해 가격 예측 및 구매 추천을 제공한다. 이를 통해 사용자들이 보다 쉽고 정확하게 원하는 중고차를 구매할 수 있도록 돕는 것을 목표로 하고 있다.

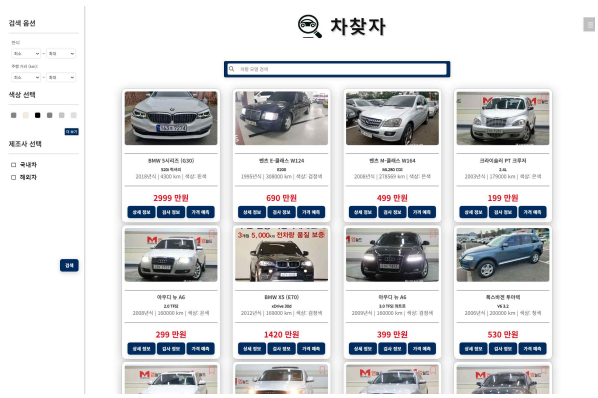

Fig. 1. Screen of "Chachaza" site

* leesw@ut.ac.kr, Tel: 031-460-0585

Ⅱ. 시스템 설계

1. 전체 구조 개요

‘차찾자’ 는 크게 Web Crawling을 이용한 데이터 수집, SQLite를 이용한 데이터 관리, Random Forest 모델 학습 및 가격 예측, OpenAi를 활용한 사용자 인터페이스로 구성된다.

데이터 수집 모듈로는 엔카 API와 성능기록부를 Selenium, BeautifulSoup를 활용하여 차량의 기본 정보 23가지를 수집하였다. 수집한 정보로는 차량의 모델명, 가격, 이동거리, 연식, 연료 타입, 배출가스, 사고여부 등이 있다.

수집된 데이터는 SQLite 데이터베이스에 저장 및 정리하며, 학습에 적합한 형태로 전처리 한 후, 주행거리, 연식, 최초등록일, 배출가스, 일산화탄소 데이터에 노이즈를 주는 방식으로 데이터를 증강시켜 데이터의 수를 약 5배 증가시켰다.

머신러닝 모델 중 ‘Random forest’ 모델을 사용하여 가격을 예측하였고, 기존의 데이터와 예측한 가격을 OpenAi를 사용해 사용자가 읽기 쉽도록 요약하고, 웹 기반의 검색 시스템을 마련해, 사용자가 쉽고 직관적으로 차량 정보를 탐색하고, 비교할 수 있게 설계하였다.

Fig. 2. Example of summary information

2. 데이터 수집 설계

데이터 수집 과정은 차량 기본 정보와 성능 기록부 데이터 두가지로, 엔카에서 공개 되어 있는 엔카 API를 통해 차량 기본 정보인 차량 ID, 제조사, 모델명, 가격, 주행거리, 연식, 연료 유형, 사진, 옵션을 가져오고, 여기서 구한 차량 ID를 통해 차량의 성능기록부를 Selenuim을 통해 크롤링 한다.

이 과정에서 데이터의 일관성과 품질을 유지하기 위해 중복 제거와 누락된 데이터를 보강한다.

3. 데이터 관리 설계

수집된 데이터는 SQLite 데이터베이스에 저장되며, 학습 모델에 적합하도록 가공된다.

예를 들면, 최초등록일 데이터의 경우 연-월-일의 형식으로 데이터가 저장되어 있는데, 이를 일단위로 풀어 저장하고, 차량 제조사, 모델명, 모델 데이터를 모두 합쳐 하나의 데이터로 저장한 후, 합친 모델명을 정규화 하는 과정이 있다.

4. 모델 설계

중고차 가격 예측을 위해 Random Forest 알고리즘을 사용하였다.

이 모델은 ‘선형 회귀(Linear regression)’ 모델보다 복잡한 패턴의 모델에서 유용하고, ‘SVM(support vector machine)’ 모델보다 데이터 크기가 큰 규모에서 유용하기 때문에 이 모델을 사용하기로 결정하였다.

Random Forest 모델을 사용할 때, 트리 개수를 100개로 설정하고 최대 깊이를 62으로 제한하였으며, 데이터셋의 80%를 학습에, 20%를 검증에 사용하였다.

5. 사용자 인터페이스 설계

사용자 인터페이스는 직관적이고 간결한 디자인을 목표로 완성하였다. 주요 기능으로는 검색 옵션, 차량 목록 조회, 찜 목록 관리, 가격 예측 결과 확인 등이 있다.

사용자는 제조사, 모델, 연식, 주행거리, 색상과 같은 다양한 조건을 설정하여 원하는 차량을 검색할 수 있으며 OpenAi를 통해 요약한 결과를 보고 합리적인 구매 결정을 내릴 수 있다.

Ⅲ. 시스템 구현

1. 데이터 수집 모듈 구현

Python의 BeautifulSoup를 활용해 엔카 API를 호출하여 차량의 기본 정보를 수집하였고, 수집된 데이터에는 차량 ID, 제조사, 모델명, 가격, 주행거리, 연식, 사용연료, 추가옵션, 사진주소가 포함되었다.

이후 수집된 차량 ID를 기반으로 Selenuim을 통해 각 차량의 성능 기록부 페이지를 크롤링하여 세부 정보를 추출하였다.

수집된 데이터는 중복 데이터를 제거하고 누락된 값을 보완하여 SQLite 데이터베이스에 저장하였다.

2. 데이터 관리 및 증강

차량의 기본 정보와 성능 기록부 데이터를 저장하기 위해 SQLite를 사용하였으며, 구조화된 테이블을 설계하여 효율적인 데이터 관리를 구현하였다.

모델 학습에 필요한 데이터를 확보하기 위해 주행거리, 연식, 최초등록일, 배출가스, 일산화탄소 데이터에 노이즈를 추가하여 데이터를 증강하였다.

이를 통해 기존의 208,844개의 데이터를 1,253,064개로 5배 증강하였다.

3. 머신러닝 모델 구현

중고차 가격 예측을 위해 Random Forest 알고리즘을 사용하였다. 학습에는 모델명, 연식, 최초등록일, 변속기 종류, 사용연료, 주행거리, 배출가스, 일산화탄소, 사고이력, 변속기 상태, 실린더 누유 상태(부위별), 냉각수 누수 상태(부위별) 데이터가 사용되었고, 데이터 증강을 통해 확보된 대규모 데이터셋을 바탕으로 학습되었다.

모델의 성능을 평가하기 위해 평균 MSE(Mean Squared Error)을 활용하였으며, 증강 전 기존 데이터의 MSE는 약 485596 이었던 반면, 증강 이후의 MSE는 약 39049로 증강 후 오차율이 많이 줄어들었다는 사실을 알 수 있었다.

예측된 가격 데이터를 바탕으로 차량의 특징 및 성능 기록부 내용을 OpenAi의 GPT-4o 모델을 활용해 요약하여 사용자에게 직관적으로 설명하는 기능을 구현하였다.

프롬프트는 차량의 정보를 모두 입력한 다음, “차량의 장단점과 구매 추천 여부를 간단하게 평가해주세요” 와 같이 작성하였다.

4. 사용자 인터페이스 구현

HTML, CSS, JavaScript를 활용해 사용자에게 친화적인 웹페이지를 설계하였다.

사용자가 제조사, 모델명, 연식, 주행거리 등의 조건을 입력하면 해당 조건에 맞는 차량 목록을 표시하도록 검색 및 필터 기능을 구현하였다.

또한, 로컬 스토리지를 활용하여 사용자가 관심 있는 차량을 찜 목록에 추가하고 관리할 수 있도록 하였다.

예측된 가격과 차량의 세부정보를 화면에 표시하고, AI를 통해 요약된 내용을 제공하여 사용자가 쉽게 차량의 조건을 확인할 수 있도록 하였다.

Ⅳ. 결론 및 향후 과제

우리는 중고차에 대해 잘 모르는 사용자도 쉽게 중고차의 가격을 판단할 수 있는 중고차 가격 예측

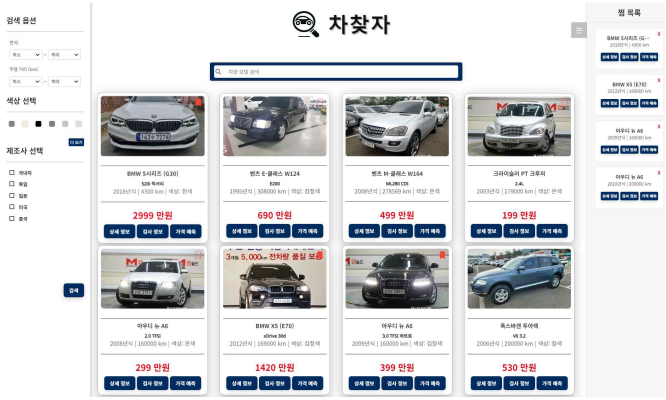

Fig. 3. Screen of “Chachaza” site with a wish list

및 추천 서비스 ‘차찾자’ 를 개발하였다.

‘차찾자’ 는 엔카 API와 WebCrowing 기술을 통해 중고차 데이터를 수집하고, Random Forest 알고리즘을 기반으로 차량 가격을 예측해 사용자에게 합리적인 구매 결정을 돕는다.

또한 OpenAi의 자연어 처리 모델을 활용하여 사용자 맞춤형 차량 설명 및 성능기록부 요약 기능을 제공함으로써 사용자 경험을 개선하였다.

이를 통해 중고차 시장의 정보 비대칭성을 해소하고, 사용자들이 보다 정확하고 효율적으로 차량을 선택할 수 있도록 돕는 효과를 확인할 수 있다.

향후 ‘차찾자’ 서비스를 더욱 발전시키기 위해서는 다음과 같은 개선점들이 필요할 것이라고 생각한다.

첫 번째는 데이터 실시간 수집이다. 현재 데이터 수집은 개발자가 직접 일주일에 한번씩 데이터를 업데이트 하는 정적인 방식으로 이루어져, API를 주기적으로 호출하는 스케줄링 기법 등을 사용한 실시간 데이터를 반영할 수 있는 자동화된 시스템 구축이 필요하다.

두 번째는 속도 개선이다. 사용자가 가격을 예측하고, 성능기록부를 요약하는 과정에 있어서 Web Crawling과 Random Forest모델, OpenAi를 통한 문장 생성이 한번에 이루어지기 때문에, 상당히 오랜 시간이 소모되는데, 이를 해결하기 위해, 예측된 가격을 데이터베이스에 추가로 미리 저장하는 방식이나, 다른 머신러닝 모델을 사용하는 방식등이 필요하다.

마지막으로 사용자 경험 강화이다. 웹사이트의 사용자 인터페이스(UI)를 개선하고, 개인화된 추천 기능을 더욱 정교화 하여 사용자 만족도를 높힐 계획이다.

이러한 과제를 해결함으로써 ‘차찾자’ 는 더 많은 사용자에게 신뢰성 있는 중고차 가격 예측 및 추천 서비스를 제공할 수 있을 것으로 기대된다.

참고문헌

1. Encar , http://www.encar.com/, 2024
2. Selenium, https://www.selenium.dev/, 2024
3. BeautifulSoup, https://www.crummy.com/software/BeautifulSoup/bs4/doc/, 2024
4. OpenAi, https://openai.com/, 2024
5. SQLite, https://www.sqlite.org/, 2024
6. IBM, https://www.ibm.com/topics/random-forest, 2024