

# Keras와 DNN을 이용한 자동차 MSRP 예측 시스템

강지원\*, 윤현빈\*, 이상현\*, 최현호<sup>o</sup>, 문유진\*

\*한국외국어대학교 경영정보학과,

<sup>o</sup>한국외국어대학교 GBT학부

e-mail: {donumm64, hyonbinyun}@gmail.com\*, diem@kakao.com\*,  
777655256@naver.com<sup>o</sup>, yjmoon@hufs.ac.kr\*

## MSRP Prediction System Utilizing KERAS and DNN

Jiwon Kang\*, Hyonbin Yun\*, Sanghyun Lee\*, Hyunho Choi<sup>o</sup>, Yoo-Jin Moon\*

\*Dept. of Management Information Systems, Hankuk University of Foreign Studies,

<sup>o</sup>Division of Global Business and Technology, Hankuk University of Foreign Studies

### ● 요약 ●

본 논문에서는 Python 3의 Keras 모듈을 이용하여 특정 자동차에 대한 최적의 판매자권장소비자가격 (MSRP)을 예측하는 시스템을 제안한다. 이 시스템은 2004년에 미국에서 시판된 428종류의 자동차에 대한 정보를 제조사, 차종, 생산지, 엔진 크기, 실린더 수, 시내 주행 시 연비, 고속도로로 주행 시 연비, 마력, 차체 무게, 차체 길이의 독립변수를 사용하여 자체적으로 딥러닝한 회귀모델을 통해 특정 지표가 주어질 차량에 대해 종속변수인 판매자권장소비자가격을 예측한다. Optimizer를 adam으로, 학습률을 0.005로 설정한 경우의 검증 MAE 값이 3842.98로 가장 낮게 산출되었고, 해당 모델의 결과는 예측값과 실제값의 오차율이  $\pm 15\%$  정도 내외로 예측된 표본의 비율이 약 80.14%로 측정되었다. 위 모델은 향후 신차 가격 결정 및 중고차 시장에서 구매, 판매 결정을 돕는 등 특정 시장 내에서 다양한 자동차의 가치를 판단하기에 유용할 것으로 전망된다.

**키워드:** Keras, 판매자권장소비자가격(MSRP), 딥러닝(Deep Learning), 회귀모델(Regression Model)

## I. Introduction

일반적으로 자동차의 가격은 제조사가 임의로 설정하는 것이라고 생각한다. 그러나 자동차를 제조하는데 들어간 기술적 비용 등이 커지면 판매로 이익을 보기 위해 가격이 오를 것이라는 판단이 들었고, 여기서 자동차에 들어간 기술적 요소나 차종과 같은 변수에 따라 가격을 인공지능 모델을 통해 회귀적으로 분석할 수 있을 것이라고 예측해서 이번 연구를 시작했다. 이번 연구의 의의는 자동차에 대한 특정 정보가 주어졌을 때 판매자권장소비자가격 (MSRP)을 예측하고, 회사들은 잘 팔리는 자동차들을 학습시켜서 이 회귀 모델을 통해 합리적인 자동차 가격 책정을 위한 가이드라인을 제시하고자 함에 있다 [1].

## II. Preliminaries

### 1. Public Open Source Dataset

해당 프로젝트에서는 2004년에 미국에서 시판된 428 종류의 자동차에 대한 정보를 제조사, 모델명, 생산지 등 다양한 스펙을 표로

정리한 Cars.sas7 파일을 데이터 집합으로 사용했다. 해당 데이터 집합에서 종속변수로는 MSRP를 사용했으며, 독립변수로는 Make, Type, Origin, Engine Size, Cylinders, MPG City, MPG Highway, Horsepower, Weight, Length 등을 사용했다.

### 2. Regression Model

#### 2.1 전처리

해당 프로젝트는 Python 3.8.5 버전의 구글 Colaboratory 환경에서 실행되었으며 Tensorflow 2.3.0 버전의 Keras, Pandas, Numpy, 그리고 Matplotlib 모듈을 사용하였다. 모델 학습에 바로 사용이 불가능한 명목변수 Make, Type, Origin 들은 단순한 숫자를 부여하지 않고 각 항목에 해당하는 종속변수 MSRP의 평균값을 부여함으로써 보통의 사람이 인식하는 가치를 유형화시켜보고자 시도했다. 이를 토대로 Numpy 모듈을 이용하여 행렬로 변환시킨 후, 학습용 7할, 검증용 3할의 비율로 데이터를 분할하여 전처리를 완료하였다.

## 2.2 회귀 모델

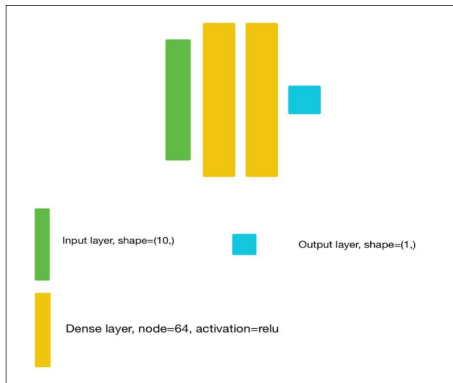


Fig. 1. Regression Model Visualization

이 연구는 Fig. 1의 DNN 회귀모델을 사용했으며, 학습에 사용되는 7개의 변수를 입력층으로, 64개의 노드를 가지며 활성화함수로 Relu 함수를 사용하는 2개 층을 은닉층으로, 최종적으로 자동차의 가격을 출력하기 위한 노드 1개로 이루어진 출력층으로 구성하였다. 최적화함수는 Adam을 사용했고, 이때 학습률은 0.005로, Epoch는 1000으로 설정했다. 손실함수는 MSE, 평가함수는 MAE로 선택했다. 학습을 진행하는 과정에는 K-fold 방식을 선택했다. 최종 MAE 값은 3842.98 정도로 나왔는데, 이는 평균적으로 오차가 3842.98 정도 발생한다는 것을 의미한다. 학습시킨 후 시험용 데이터로 시험을 해본 결과, 예측값을 기준으로 실제값과의 오차가 15% 이내인 데이터가 75~80% 정도로 도출되었다.

## III. The Proposed Scheme

### 1. Valid MAE Analysis

분석 시 인공지능 모델이 과적합하는 것을 미연에 방지하기 위해 MAE (평균절대오차) 손실 함수를 토대로 검증을 거쳐 MAE값이 다시 상승하기 시작하는 Epoch의 수를 확인하였다. 해당 시점을 확인하여 학습의 횟수가 해당 시점을 넘어가지 않도록 설정하여, MAE값이 약 2배 개선되었음을 확인할 수 있었다.

### 2. Result Analysis

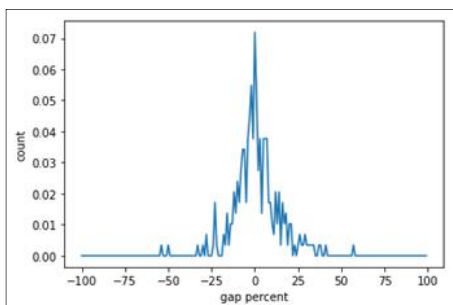


Fig. 2. Analysis by Percentage

데이터의 전처리와 ANOVA 결과에 따라 데이터 집합의 특정 부분을 제외한 후 모델 학습을 진행하였다. Optimizer를 rmsprop으로, 학습률을 0.05로 설정하여 학습한 결과, 검증 MAE 값이 4966.06으로 산출되었다. 또한, 해당 모델에서 학습률을 0.005로 변경하고 Optimizer를 adam으로 교체한 결과, 검증 MAE 값이 3842.98로 산출되었다. 따라서 검증 MAE 값이 가장 낮게 산출된 모델을 사용하였고, 이 경우 Fig. 2에서 볼 수 있듯이 예측값과 실제값의 오차율이  $\pm 15\%$  정도 내외로 예측된 표본의 비율이 약 80.14%로 측정되었다. 향후 해당 모델에 대하여 차량 제조사가 지향하는 브랜드 가치 및 실제 자동차의 기능적 가치에 대한 변수를 추가로 고려하게 된다면 모델의 성능이 보다 향상될 것으로 추정된다.

## IV. Conclusions

위 모델을 통해 임의의 자동차에 대한 제품가격을 입력했을 때 적절한 MSRP에 근접한 값을 도출할 수 있음을 확인했다. 향후 해당 모델이 사용될 수 있는 분야는 신차 가격 결정 및 중고차 시장에서 구매, 판매 결정을 돕는 등 특정 시장 내에서 다양한 자동차의 가치를 판단하기에 유용할 것으로 전망하는 바이다. 이외에도 자동차의 가치를 결정하는 다양한 변수를 사용하여 모델을 개선하기 위해서는 해당 분야에 대한 전문적인 지식이 요구된다.

## REFERENCES

- [1] A. Yezioro, B. Dong, and F. Leite, "An applied artificial intelligence approach towards assessing building performance simulation tools," *Energy and Buildings*, vol.40, no.4, pp. 612-620, 2018.