THE APPLICATION OF MACHINE LEARNING IN USED CAR PRICE PREDICTION AND RECOMMENDATION

Vukašin Vasiljević[1], Zoran Kalinić[2]

[1] Faculty of Economics,
The University of Kragujevac, Liceja Knezevine Srbije 3, 34000 Kragujevac
e-mail: vukasin.vasiljevic@ef.kg.ac.rs
[2] Faculty of Economics,
The University of Kragujevac, Liceja Knezevine Srbije 3, 34000 Kragujevac
e-mail: zkalinic@kg.ac.rs

Abstract:

Used car price prediction is still a topic of high interest due to various factors such as the global pandemic that lasted for two years and the unprecedented numbers of cars being purchased and sold. The motive for this research was to enhance and simplify user experience when purchasing used cars via e-commerce platforms. A primary objective of the project is to predict used car prices by using car specifications. Estimation of prices is possible with machine learning algorithms for regression, such as Extreme Gradient Boosting, Random Forest, Extra Trees, etc. In addition to machine learning algorithms, Sequential neural network was used to re-evaluate the model. After testing the regression algorithms, Extra Trees Regressor had the highest score. Similar results were obtained from Sequential neural network. Additionally, recommender system was developed to assist users when choosing used cars. This way, users are able to compare similar cars. Finally, this application was deployed to a cloud-based framework called Streamlit, which is part of the Python programming language library.

Key words: car price prediction, machine learning algorithms, recommender systems, artificial intelligence systems

1. Introduction

With an ever so demanding car market, on a global scale, companies have trouble keeping up with supplying customer demands for purchasing new cars. Due to a chip shortage that was caused by the global pandemic, car companies are lagging behind the deliveries of new cars. Nowadays, customers are frequently turning to an alternative solution – purchase of used cars instead of new ones. In addition, because of affordability and economic conditions, many will prefer pre-owned cars instead of new ones that cost, in many cases, a lot more. Having said this, accurately predicting used car prices requires domain knowledge due to the nature of their dependence on a variety of factors and features. Used car prices are not constant and they are changing more frequently than expected. Because of this, both buyers and sellers are in need of a system that could accurately predict used car prices and automatically update itself as the prices change on the market.

This project focuses only on the used car market in the Republic of Serbia. More specifically, the data has been collected from Serbia's largest online car market website called *Polovni automobili* [1]. Considering this, the scope of this project is limited to the national used cars market.

2. Methodology

This project incorporates some of the fundamental processes of a data science pipeline that converts raw data into actionable and representative answers. The following processes were followed to ensure automated and efficient data flow from the source to destination:

- Data collecting
- Data engineering
- Data modeling
- Model deployment

The Figure 1 illustrates the project pipeline.

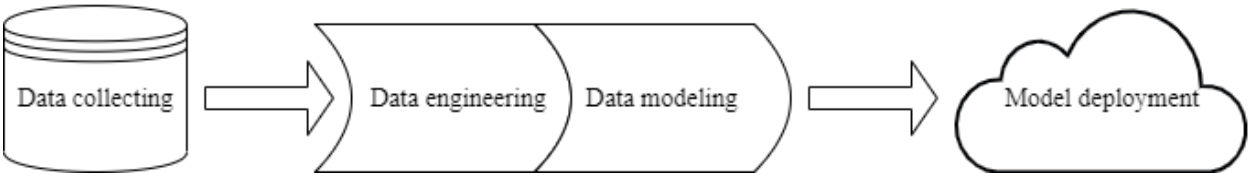

Fig. 1. Project pipeline

2.1 Data collecting

The dataset was collected by using the data mining technique called *web scraping* [2]. With this method, the researchers were able to collect data directly from a website in a raw form. Moreover, car specification is available on each car ad. Iterating through car ads allowed the following parameters to be downloaded: car brand, model, year, mileage, car type, fuel type, engine displacement (capacity), engine power, engine emission type, engine drive, shift, air conditioning, color of a vehicle, interior material and finally, the car price. The data was stored in a csv file later to be used to format the data into a more understandable form. This dataset consists of roughly 40,000 rows of data. The next step in this data flow process is to clean up the data.

2.2 Data engineering

Data engineering process is one of the most important steps in a data flow pipeline. Transforming raw data into meaningful information is an essential part of every data-driven project. This step's objective is to examine the data which signifies understanding of every feature, identify errors, missing values and corrupt data. After that, the goal is to clean the data by replacing, and/or filling missing values or errors. Using exploratory data analysis, the researchers were able to understand and visualize patterns and values in data. Furthermore, data augmentation was implemented to add more data to this dataset [3]. During the data engineering step, more insight into the data could be gained and conclusions drawn based on the aforementioned insights.

Almost every feature in the dataset had to be altered in some way. Most of the categorical features needed string formatting into standardized values. The numerical feature was also standardized and cleaned. The goal of this process is to encode data in a way that the algorithms can parse and interpret it more easily. After cleaning the data and data augmentation process, the final dataset consists of around 26,000 rows of cleaned data.

2.3 Data modeling

After cleaning and encoding the data, the next step is to model it. This process includes testing the data with different machine learning algorithms and evaluate given results to determine the model that gives the best price prediction. The researchers in this project used the following metrics to evaluate the model performance:

- **Root mean square error (RMSE):** RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

- **Mean absolute error (MAE):** This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset.

- **R^2 score:** R^2 score represents total variance explained by model. This score varies between 0 and 1, where 1 represents that variables are perfectly correlated, i.e. with no variance at all. A low value would show a low level of correlation between variables.

Algorithms used in this research are following:

1. Random forest regressor
2. Extra trees regressor
3. Extreme gradient boosting regressor

A train-test split of 70/30 with a 10-fold randomized search was used in all of the experiments. The results for each model on test data, after parameter tuning, are shown in Table 1.

Model	RMSE	MAE	R^2
Random forest	329	210	0.98
Extra trees	285	184	0.98
Extreme gradient boosting	325	219	0.98

Table. 1. Model results

From these results, it can be seen that the best performing model is the Extra trees regressor with R^2 score of 98%, MAE of 184 and MSE of 285 and it can be concluded that the errors of Extra trees algorithm are lower in comparison with Random forest and Extreme gradient boosting regression models. Hence, Extra trees algorithm outperformed.

Along with the price prediction, the researchers developed a recommender system to suggest similar cars to customers. This recommender system is formed on the rule-based method [4]. Figure 2 shows predicted price in function of actual price on a test set.

2.4 Model deployment

Following the development of the data model that predicts used car prices and recommends similar cars, the researchers deployed it to an online open-source platform Streamlit [5]. This platform incorporates both front and back-end development via easy-to-use Python library. The appearance of the webpage is presented in Figure 3.

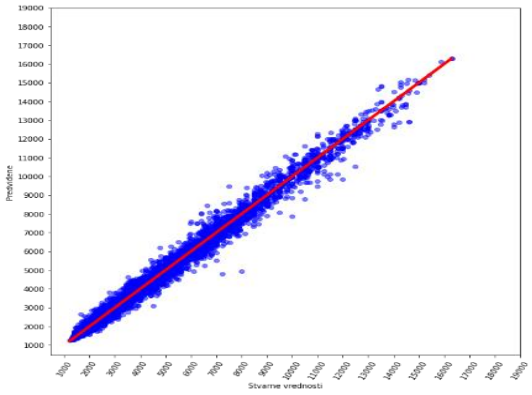

Fig. 2. Actual vs. Predicted price

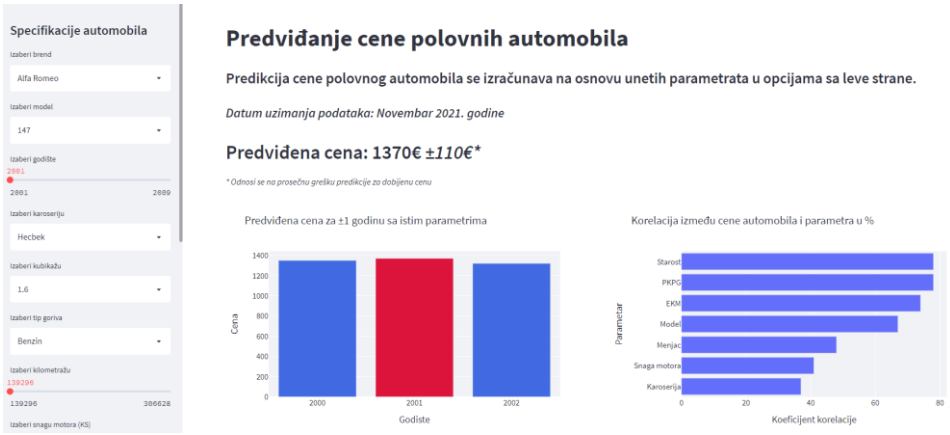

Fig. 3. Webpage appearance

3. Conclusion

Using the data mining technique in combination with applying an artificial intelligence system, this project proposed a framework for prediction of the used car prices in the Republic of Serbia. An efficient machine learning model was built that could predict the price and recommend similar cars with high accuracy and in real time. This methodology ensures a scalable framework that can be automatically updated with new data.

References

[1] Polovni automobili (used cars website), https://www.polovniautomobili.com/
[2] Ramageri BM, Data mining techniques and applications. Indian Journal of Computer Science and Engineering 1(3):301-305, 2010.
[3] Shorten C, Khoshgoftaar TM, A survey on Image Data Augmentation for Deep Learning. Journal of Big Data 6:60, 2019 https://doi.org/10.1186/s40537-019-0197-0
[4] Apolloni B, Bassis S, Mesiti M, Valtolina S, Epifania F., Francesco. (2020). A rule based recommender system, Advances in Neural Networks, 87-96, 2020 http://dx.doi.org/10.1007/978-3-319-33747-0_9
[5] Streamlit, https://docs.streamlit.io/