Sakarya University Journal of Computer and Information Sciences
http://saucis.sakarya.edu.tr/

DOI: 10.35377/saucis...1309103 RESEARCH ARTICLE

Price Prediction Using Web Scraping and Machine Learning Algorithms in the Used Car Market

Seda Yılmaz [1], İhsan Hakan Selvi [1]

[1] Information Systems Engineering Department, Institute of Natural Sciences, Sakarya University, Sakarya, Türkiye.

Corresponding author:
Seda YILMAZ, Information Systems Engineering Department, Institute of Natural Sciences, Sakarya University, Sakarya, Türkiye
E-mail address:
seda.yilmaz11@ogr.sakarya.edu.tr

Submitted: 02 June 2023
Revision Requested: 17 August 2023
Last Revision Received: 25 August 2023
Accepted: 28 August 2023
Published Online: 30 August 2023

Citation: Yılmaz S. and Selvi İH.. (2023). Price Prediction Using Web Scraping and Machine Learning Algorithms in the Used Car Market. *Sakarya University Journal of Computer and Information Sciences.* 6 (2) https://doi.org/10.35377/saucis...1309103

ABSTRACT

The development of technology increases data traffic and data size day by day. Therefore, it has become very important to collect and interpret data. This study, it is aimed to analyze the car sales data collected using web scraping techniques by using machine learning algorithms and to create a price estimation model. The data needed for analysis was collected using Selenium and BeautifulSoup and prepared for analysis by applying various data preprocessing steps. Lasso regression and PCA analysis were used for feature selection and size reduction, and the GridSearchCV method was used for hyperparameter tuning. The results were evaluated with machine learning algorithms.

Random Forest, K-Nearest Neighbor, Gradient Boost, AdaBoost, Support Vector and XGBoost regression algorithms were used in the analysis. The obtained analysis results were evaluated together with Mean Square Error (MSE), Root Mean Square Error (RMSE) and Coefficient of Determination (R-square). When the results for data set 1 were examined, the model that gave the best results was XGBoost Regression with 0.973 R^2, 0.026 MSE and 0.161 RMSE values. When the results for data set 2 were examined, the model that gave the best results was K-Nearest Neighbor Regression with 0.978 R^2, 0.021 MSE and 0.145 RMSE values.

Keywords: Web scrapping, machine learning, price prediction

1. Introduction

With the developing technology, the data size and traffic are increasing daily. Therefore, data analysis is very important in many different sectors and functions today. When data analysis is done correctly, it provides significant benefits in almost every field. Many websites contain large amounts of data. Obtaining this information with statistical summaries is important for individuals and businesses [1]. It enables businesses to achieve better results in decision-making. Accordingly, it provides important benefits such as competitive advantage, customer satisfaction, operational efficiency, and better management of risks.

However, data that needs to be collected and interpreted correctly only confuses. Therefore, data collection is as important as analysis. Most of the time, data cannot be copied directly. So, skills are needed to take as little time and effort as possible into analyzing data. The information is placed so that it is not easy to download; even if it is easy to download, it will be difficult to process. Web scraping is included in data science at this point [2]. Of course, data can also be collected manually from web pages. However, this process takes much time, and complex site structures may not allow information to be copied from sites so easily. For these reasons, the web scraping method has emerged [3].

With web scraping, collecting large amounts of data is automated, and the data is made more useful to the user. But in most cases, web scraping is a complex process. Websites come in different shapes and formats. This is why web scrapers vary in

functionality and features. Many websites now provide users with an API (Application Programming Interface) to access structured data stores that can be downloaded and used. There is no point in scraping if the website provides access to the API. One of the most important current uses of web scraping is for businesses to track the pricing activities of their competitors. Pricing can be created across an entire site in relatively short timescales and with minimal manual effort [4].

Accurate car price estimation requires expert knowledge because price often depends on distinguishing features and factors. To accurately estimate the price of a car, it has been shown that the knowledge of experts in this field is required, as well as many dimensions that must differ from the technical characteristics of the car, such as horsepower, engine capacity or transmission [5].

Asghar et al. [6] suggest using different ways to estimate sales prices based on the value of the tools used in marketing. The applications they recommend support both the buyer and the seller in buying and selling second-hand vehicles, can predict the best prices for themselves, and make a good personal and commercial determination. The proposed modeling performances reflect that their work includes effective and efficient methods and strategies.

Pandey et al. [7], in their study, say that machine learning algorithms are good solutions that can be used to solve the problems encountered in such areas. According to Pandey et al., machine learning algorithms and techniques can offer a good solution to such problems as the issue of used car price estimation.

Chen et al. [8] have put forward a comprehensive study by drawing attention to big data and usage areas in their studies. They emphasize in their writings that as the size of the data used increases, its importance also increases. Their work demonstrates how multiple linear regression and random forest techniques work with different complexities in the modeling part. As a result, it is argued in the study that the only purpose is to realize the most suitable model while dealing with a used vehicle price estimation.

Identifying needs is important in web scraping. First, which data will be collected from which website should be determined. Different tools and libraries are available for web scraping. Libraries like BeautifulSoup and Scrapy in Python are popular and frequently used methods for web scraping. By doing research, the most suitable method for web scraping can be determined.

Appropriate methods should be determined for the analysis of the obtained data set. These methods vary depending on the dataset's characteristics and the intended output. Machine learning, statistical analysis, data mining or other analytical methods can be used.

2. Materials and Methods

This study aimed to collect data from the website using web scraping techniques and to analyze the collected data with machine learning algorithms and make price estimations. The data used in the study were collected with the help of codes written in Python programming language from a website where used vehicles are sold and saved in a database.

Python is one of the most common programming languages in the data domain. It is preferred because it is easy to understand. Also, this programming language is flexible and scalable. Due to the many advantages of using this programming language, it has also found its way to writing codes and programs for web scraping. Web scraping is a relatively modern trend for developers, but the Python programming language is well suited for performing web scraping tasks. Python has a huge library of tools for developing algorithms to collect data using web scraping.

PyCharm is an IDE for Python with the toolset for effective Python programming. IDE (Integrated development environment) is an environment that helps users write code faster and more efficiently. It allows editing, compiling, or interpreting text, debugging and more. The IDE allows the development speed to be increased. Jupyter Notebook, on the other hand, is a server-client application where codes can be run, notes can be taken, and desired edits can be made in notepad format on a web browser. Thanks to its block logic, it is highly preferred because it allows observing the outputs by making changes in the desired parts of the code.

Selenium and BeautifulSoup were used in this study's web scraping phase for data collection. Selenium library is generally used to automate web applications for testing purposes. However, it is for more than this. It also allows data collection via the browser. The BeautifulSoup library is a library created for manipulating HTML files. It allows us to obtain the desired parts by parsing the HTML codes in the relevant source. Numpy, Pandas, Scikit-learn, and Mathplotlib libraries were mainly used in the preprocessing and analysis stages of the data.

After the data obtained by the web scraping method were converted into the appropriate format, they were written on the database in the form of tables. During these processes, pymysql and sqlalchemy libraries were used. With these libraries, the relevant table in the database can be accessed with Python and various select, insert, update, and delete operations can be performed.

The study was completed by following the steps below.

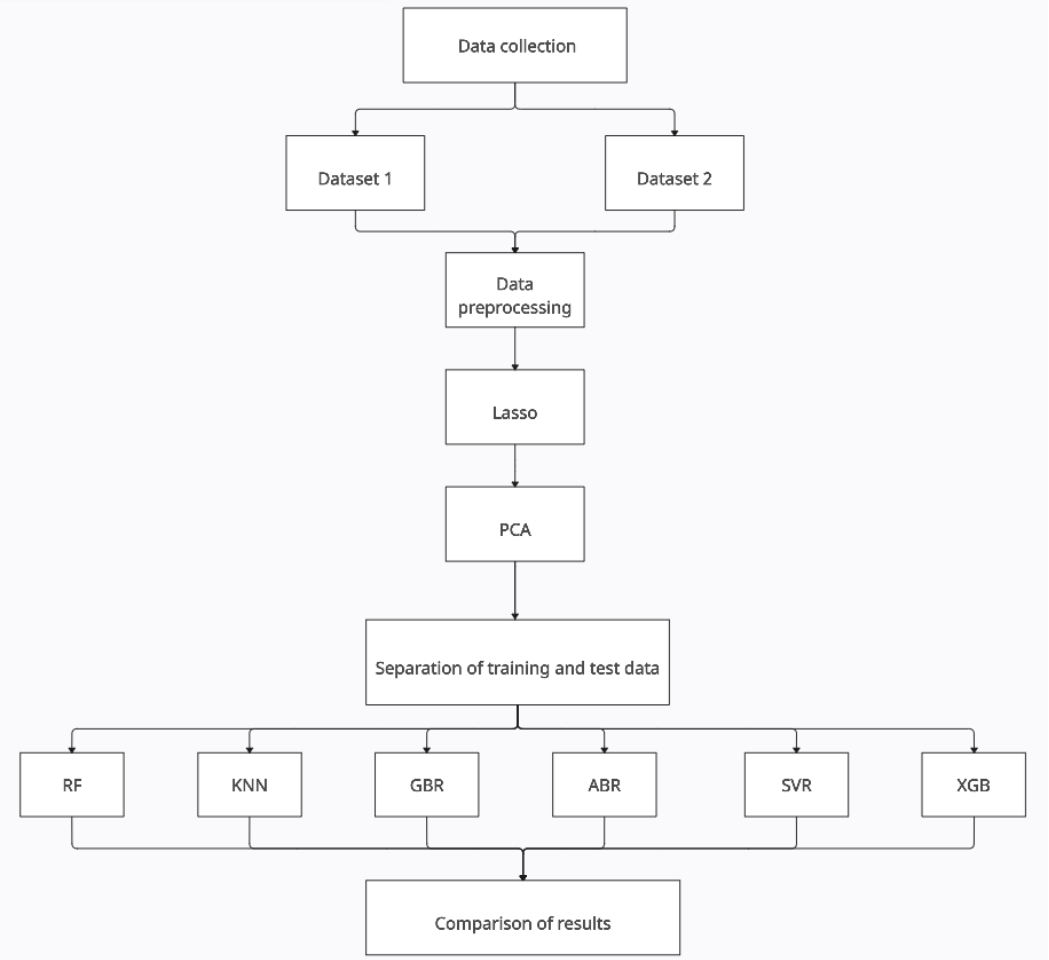

Figure 1 Prediction Method

2.1 Data collection

First, it was planned to collect follow-up data for car price estimation from a used car sales website: price, brand, model, series, type, from information, fuel, gear, engine power, and engine capacity. In addition, the car damage status, an important criterion in the used car market, was also wanted to be analyzed, and damage information was obtained from the template showing the car damage status on the site. The representative image is shown in Figure 2.

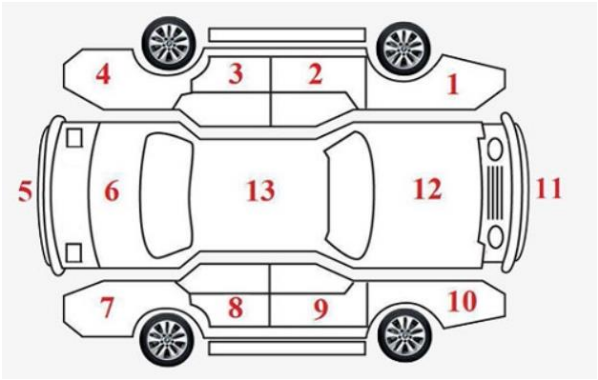

Figure 2 Representative image [9]

This information was added to the dataset in the following order: right rear fender, rear hood, left rear fender, right rear door, right front door, roof, left rear door, left front door, right front fender, engine hood, left front fender, front bumper, rear bumper. While these fields are being filled, 0 values are assigned if the relevant part of the car is original, 1 if it is painted, 2 if it is changed, and 3 if it is not specified.

In the web scraping process, the URLs of the requested products were first retrieved from the target website via Selenium. The source codes of the pages whose URLs were taken were collected, and the desired parts were parsed using BeautifulSoup.

The data collected with Numpy and Pandas were manipulated, converted into the appropriate format, and tabulated. The collected data was written to the database and used Pymysql and Sqlalchmy for database connection.

Within the scope of the study, two data sets with different data numbers were obtained to make comparisons during the analysis. Both datasets consist of 23 features. While collecting the data, the dates between 10.02.2023 and 31.03.2023 were taken as a basis. Dataset 1 contains 5557 rows of data, while Dataset 2 contains 11688.

2.2 Data preprocessing

The collected data were analyzed, and initially, some null fields in the columns reflecting the damage status were filled with the default (unspecified=3) value. Then, categorical data were converted into numerical data using the encoder method to be ready for analysis.

The price variable was set as the target variable. Lasso regression was used to determine which of the remaining 22 features would be used in the analysis. Thanks to Lasso, the coefficient of unimportant variables was reduced to zero.

After applying Lasso, 11 features were selected in the Dataset 1 result, and 9 were selected in the Dataset 2 result.

Dataset features before applying Lasso: brand, model, series, type, from who information, fuel, gear, engine power, engine capacity, right rear fender, rear hood, left rear fender, right rear door, right front door, roof, left rear door, left front door, right front fender, engine hood, left front fender, front bumper, and rear bumper.

After applying Lasso, dataset 1 features brand, model, series, from who information, fuel, gear, engine power, engine capacity, rear hood, roof, and front bumper.

Dataset 2 features after applying Lasso: brand, model, type, from who information, fuel, gear, engine power, roof, and rear bumper.

Lasso regression is a technique used in feature selection and model parameter estimation. Unlike traditional regression analysis, it encourages model parameters to approach zero. Thanks to this feature, it can reduce overfitting problems and make the model simpler and interpretable.

Afterward, PCA analysis for size reduction was applied, and the variables included were determined. As a result of the PCA analysis, the number of features was reduced to 7 for both data sets.

Principal Component Analysis (PCA) is a statistical technique for dimension reduction in multivariate datasets. PCA is used to analyze the relationships between the variables in the data set and to create new variables that best explain these relationships.

Table 1 Dataset sizes after Lasso and PCA

	Lasso		PCA	
	Number of Samples	Number of Features	Number of Samples	Number of Features
Dataset 1	5557	11	5557	7
Dataset 2	11688	9	11688	7

2.3 Separation of training and test data

Training and test data were separated for use in machine learning algorithms. K-Fold cross-validation was used for this process. Cross-validation is a healthier method than the train-test split approach because it allows both testing and training to be performed on all data; for example, when the data is separated as 20% test and 80% training with a train-test split, some deviations and errors may occur because it is not known how the data is distributed here.

In the K-Fold cross-validation method, the dataset is divided into k different subsets. It is ensured that each piece is used as both training and test data. In this way, the errors that may occur due to its distribution are minimized. In this study, analyses were carried out by taking the k value of 10.

2.4 Machine learning algorithms

In the study, 5 different machine learning algorithms were used. Dataset 1, with a low amount of data and Dataset 2, with a higher amount of data, were run with each of these algorithms. GridSearchCV was used to determine the best parameters in the algorithms.

Random forest is one of the best-performing and widely used machine learning algorithms. Random forest regression generates multiple decision trees and combines the outputs of all these decision trees to arrive at a single result. It was first

introduced in 1984 [11]. Decision nodes and leaf nodes form decision trees. With the test function, decision nodes evaluate the samples and forward them to the relevant branches according to the results. RFR is a bagging community-based model that combines bagging benefits. Thus, it improves forecasting performance [12].

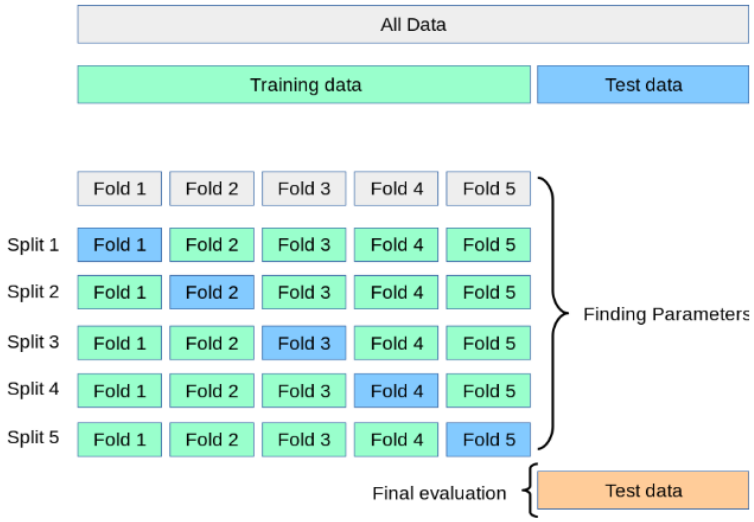

Figure 3 K-Fold working principle [10]

The K-Nearest Neighbors (KNN) algorithm is a machine learning algorithm for classification and regression problems. It is one of the supervised learning methods. It is trained with labeled data and used to classify or predict new samples. The KNN algorithm finds the k nearest neighbors to the sample and estimates by taking the neighbors' average. Other similarity criteria, usually Euclidean, can also be used to identify nearest neighbors [13].

Gradient Boosting Regression (GBR) is a machine learning technique that combines the predictions of multiple weak decision trees to create a strong prediction model. It is a powerful algorithm used for both regression and classification problems. In the gradient boosting method, weak decision trees are trained sequentially, and each next decision tree is trained to correct the errors of the previous decision trees [14]. The algorithm starts with an initial estimate, usually chosen as the mean value of the target variable for regression problems. It builds up a series of models in succession that minimizes the errors of the previous models.

Adaboost regression is an ensemble learning algorithm used in machine learning [15]. AB (Adaptive Boosting) aims to create a strong learner by bringing together weak learners. It is usually implemented using a set of decision trees. It advances by reinforcing relatively weak models that made erroneous predictions in previous steps. While it reduces the weights of the decision tree points with correct predictions, it increases the weights of those with incorrect predictions. This creates a combination of weak models whose weights are adjusted at each step and aims to obtain a strong model with the best predictive ability [16].

Support Vector Regression (SVR) is a version of support vector machines (SVM) adapted to regression problems. Support Vector Regression divides data points with a hyperplane and makes predictions using support vectors on this hyperplane [17]. It aims to minimize the regression error by classifying the data points into hyperplanes as much as possible [18]. Support Vector Regression can work effectively on low-dimensional or high-dimensional datasets. It is especially preferred because it resists outliers and shows good generalization ability in various data distributions.

XGBoost is a gradient-boosting-based learning algorithm. Under the Gradient Boosting framework, it develops extensible parallel classification and regression trees (CARTs) that can quickly and accurately solve data science problems. This approach combines weak learners (usually decision trees) and creates a strong prediction model [19]. XGBoost's high performance, strong predictive ability, scalability, resistance to outliers, feature importance scores, automatic intersection processing, and flexibility stand out.

3. Conclusion and Discussion

Analyzes were carried out using the materials and methods in Chapter 2. As a result of the analyses made with machine learning algorithms, the following outputs were obtained. The table includes R-square scores, MSE, and RMSE values.

Table 2 Results of analysis with Dataset 1 and Dataset 2

		RFR	KNN	GBR	ABR	SVR	XGB
Dataset 1	R^2	0.889	0.854	0.970	0.676	0.811	**0.973**
	MSE	0.110	0.145	0.029	0.323	0.188	**0.026**
	RMSE	0.333	0.382	0.172	0.568	0.434	**0.161**
Dataset 2	R^2	0.960	**0.978**	0.967	0.658	0.820	0.973
	MSE	0.039	**0.021**	0.032	0.341	0.179	0.026
	RMSE	0.197	**0.145**	0.179	0.584	0.423	0.164

When the results were examined, it was seen that increasing the amount of data for RF, KNN, and SVR models affected the results positively; the scores increased, and the errors decreased. For other models, it is seen that the performance decreases when the amount of data increases. The best results are shown in bold in Table 2. A learning curve graph was used to show the performance in the analysis results. A learning curve is a graph that shows how the performance of a model changes with the number of training samples. As the number of training samples increases, the model's performance on training and validation sets is examined.

According to Table 2 and graphics, it was determined that the Adaboost regression model gave worse results in both data sets.

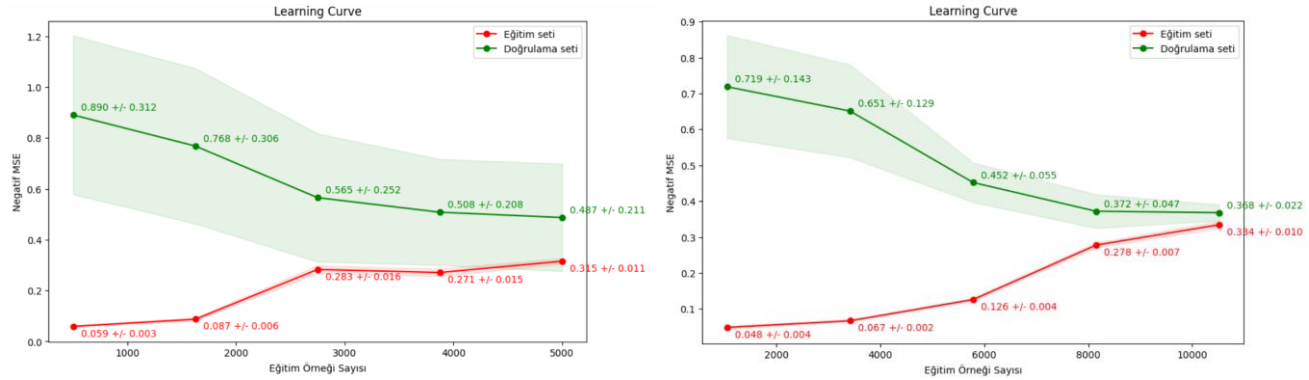

Figure 4 ABR Learning Curve graph (Dataset 1 and Dataset 2)

In the graphs, it is seen that the errors in the training set and test set are high. The high standard deviation in the test set errors indicates that the model's predictions are quite different from each other, and the reliability of the predictions is low. This indicates that the model may perform worse on some data points, and its predictions must be evaluated more carefully. As a result, error values and high standard deviation, according to the graph, show that the model's generalization ability could be stronger, and there may be an overfitting problem. Increasing the data size does not improve the results.

When the results of the other models are examined, it is seen that the results of the GBR and XGB models based on Gradient Boosting for Dataset 1 containing 5557 rows of data are better than the remaining models. When the two models are compared, it has been determined that the XGBoost model gives better results.

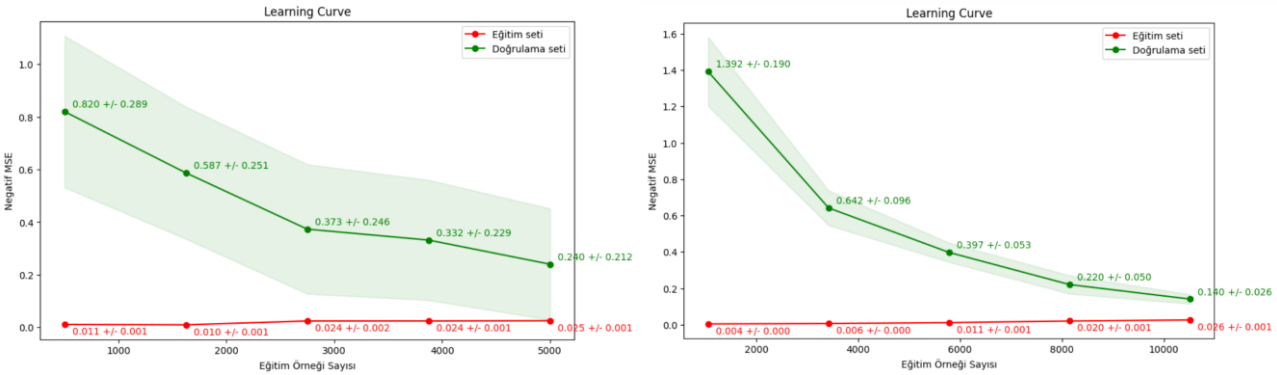

Figure 5 XGB Learning Curve graph (Dataset 1 and Dataset 2)

The training set has a low initial error value and standard deviation. This indicates that the model has learned the training data well and has a low error. However, the validation set initially appears to have a higher error value and standard deviation. As the data size increases, the error value in the validation set and the standard deviation decreases. A decrease in the error

value indicates that the model's predictions are more accurate and better fitted to the data. In summary, while the model performs well on the training set, it initially performs poorly on the validation set. However, as the data size increases, the model's generalization ability improves, and the error in the validation set decreases.

When the results for Dataset 2, which contains 11688 rows of data, are examined, it is seen that the KNN and XGB models have the best results. When the two models are compared, the results of KNN were found to be better than XGB.

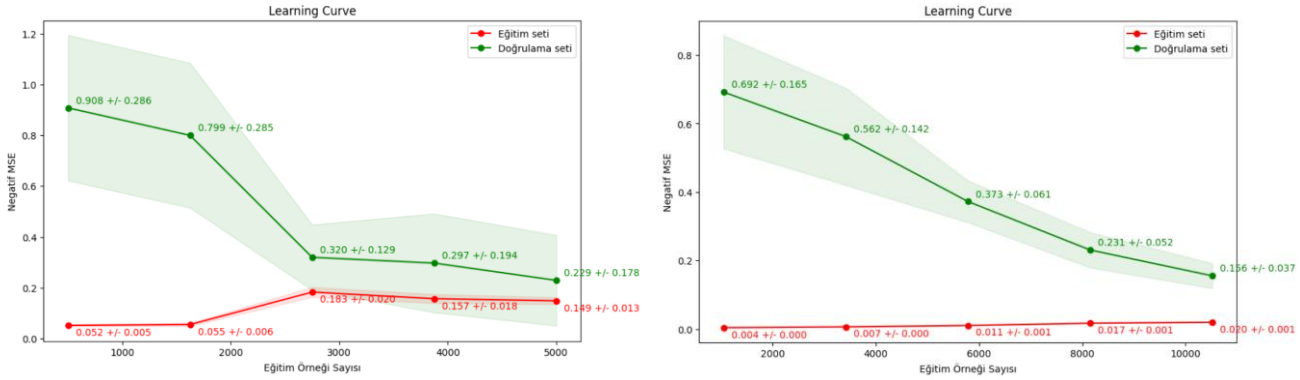

Figure 6 KNN Learning Curve graph (Dataset 1 and Dataset 2)

The error values obtained in the training set are quite low, indicating that the model performs well on the training data. The standard deviation values are also quite low, indicating that the model has low sensitivity to the training data and that the predictions are consistent. The error values obtained in the test set are higher than those in the training set. The standard deviation value is higher, indicating that the model's performance on the test data is more variable and the predictions have a wider error range. As a result, it is seen that the overall performance of the model improves with the increase in the amount of data. The error and standard deviation decrease on the validation set indicates that the model generalizes better with more data.

The learning curve graphics of other models are as follows. When the graphs are examined, it is seen that increasing the amount of data in RFR and SVR models improves the results. In the GBR model, on the other hand, it is seen that the performance has decreased slightly. However, for all models, it is seen that the performances in the validation set increase and the standard deviations decrease.

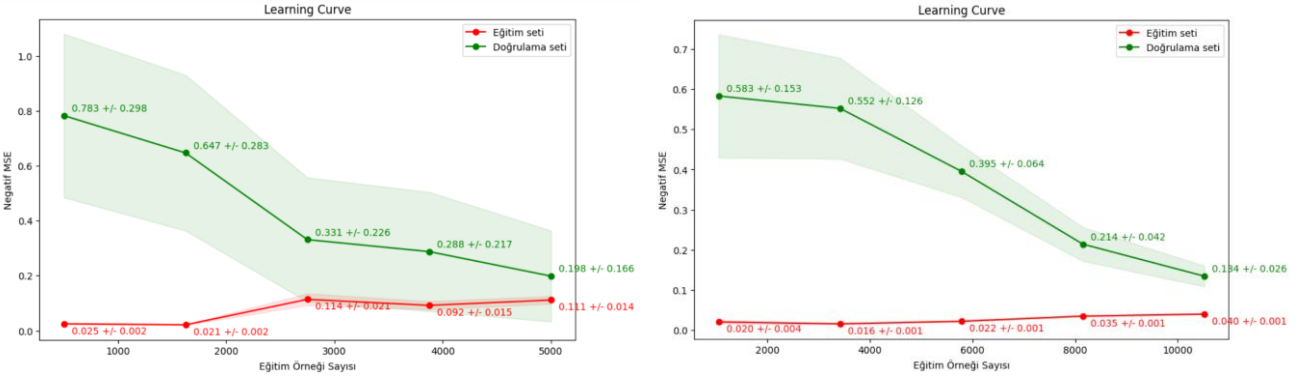

Figure 7 RFR Learning Curve graph (Dataset 1 and Dataset 2)

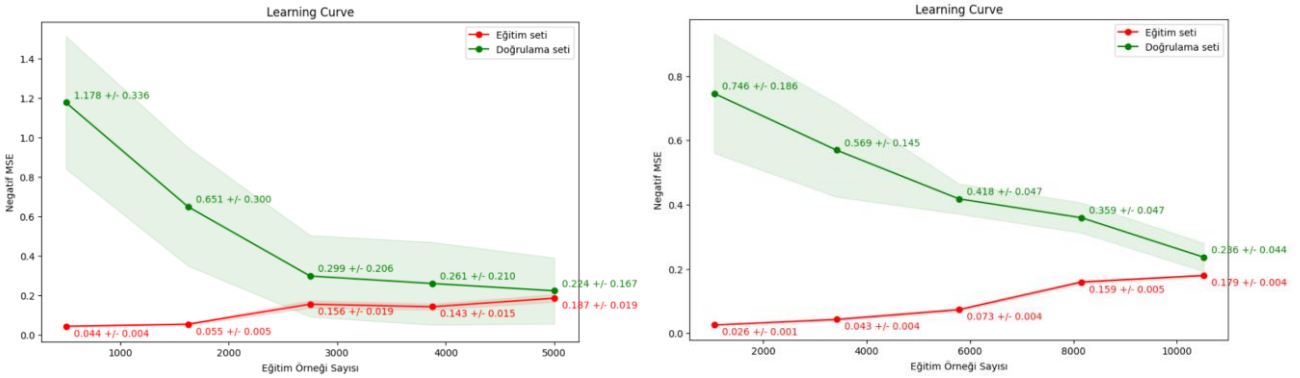

Figure 8 SVR Learning Curve graph (Dataset 1 and Dataset 2)

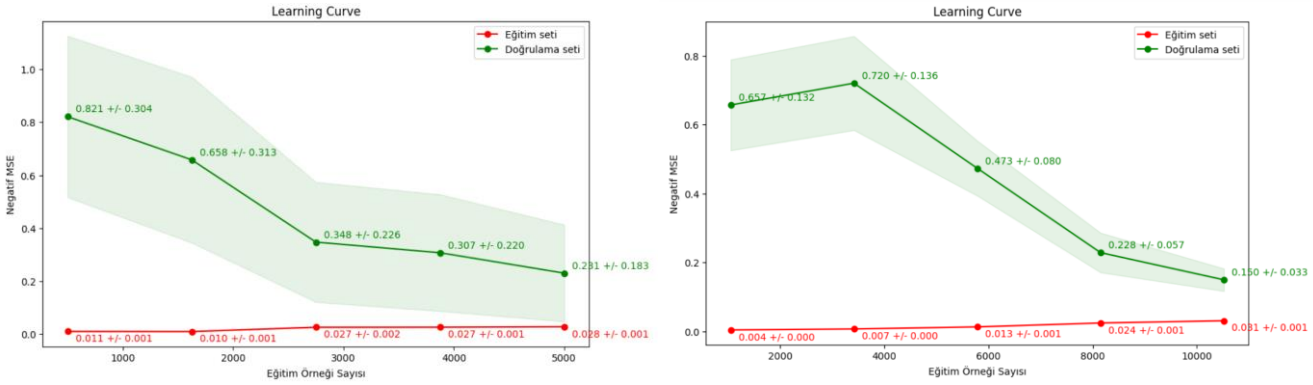

Figure 9 GBR Learning Curve graph (Dataset 1 and Dataset 2)

As a result of the study, the data collected from the website with web scraping methods were analyzed with machine learning algorithms, and the algorithms that gave the best results for different data sets in size for price estimation were determined. It has been seen that increasing the amount of data for some models has a positive effect on the results, while for some models, it has a negative effect.

Based on the study, it can be deduced that similar data from different sources can be collected by web scraping methods, the data set can be expanded, and performance predictions can be made by selecting the appropriate model after the analysis. If developed, this study can continuously extract data from websites by performing web scraping with various methods, and data can be collected, the collected data can be analyzed, and the forecast results can be shared.

References

[1] Milev, P., Conceptual Approach for Development of Web Scraping Application for Tracking Information. Economic Alternatives, 475-485, 2017.
[2] Khder, M., Web Scraping or Web Crawling: State of Art, Techniques, 73 Approaches and Application. International Journal of Advances in Soft Computing and its Applications, 2021.
[3] Banerjee, R., Website Scraping, Happiest Minds Technologies, 2014.
[4] Haddaway, N., The use of web-scraping software in searching for grey literature. Grey Journal, 11(3):186-190, 2015.
[5] Gegic, E.; Isakovic, B.; Keco, D.; Masetic, Z.; Kevric, J. Car price prediction using machine learning techniques. TEM J. 2019, 8, 113.
[6] Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M., Used Cars Price Prediction using Machine Learning with Optimal Features. Pakistan Journal of Engineering and Technology, 4(2), 113-119, 2021.
[7] Pandey, A., Rastogi, V., & Singh, S., Car's selling price prediction using random forest machine learning algorithm. In 5th International Conference on Next Generation Computing Technologies, 2020.
[8] Chen, K.-P., Liang, T.-P., Yin, S.-Y., Chang, T., Liu, Y.-C., & Yu, Y.-T., How serious is shill bidding in online auctions? evidence from eBay motors. work, 1–51, 2020.
[9] Yolcu360, Available: https://yolcu360.com/blog/oto-ekspertiz-raporunda-ne-yazar. [Accessed: 03-May-2023].
[10] Scikit-learn, Available: https://scikit-learn.org/stable/modules/cross_validation.html. [Accessed: 04-May-2023].
[11] Breiman, L., Random Forests. Machine Learning, 45(1), 5-32, 2001.
[12] Breiman, L., Bagging Predictors. Machine Learning, 24(2), 123-140, 1996.
[13] Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician, vol.46, s. 175–185, 1992.
[14] Friedman, J. H., Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232, 2001.
[15] Freund, Y. and Schapire, R. E. "Experiments with a new boosting algorithm", Icml, 96, 148-156, 1996.
[16] Schapire, R. E., Explaining adaboost. In Empirical Inference, pp. 37–52, Berlin Heidelberg., 2013.
[17] Vapnik V., The Nature of Statistical Learning Theory, 1995.
[18] Awad M. and Khanna R., Efficient Learning Machines, Apress, 2015.
[19] Chen, T., Guestrin, C., XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu, 785–794, 2016.

Conflict of Interest Notice
The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical Approval and Informed Consent
It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Availability of data and material
Not applicable

Plagiarism Statement
This article has been scanned by iThenticate ™.