Cascade Model for Price and Time of Car Sales Prediction

Volodymyr Shymanskyi and Vasyl Liaskovets

Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine

Abstract

This article discusses the application of machine learning models to predict the price and time of car sales. The purpose of the work is to study and adapt various machine learning models for accurate and timely forecasting of car sales parameters in the secondary market. This paper delves into the predictive capabilities of various machine learning models, notably XGBoost and Random Forest, in estimating motor vehicle sale prices and times. Our analysis revealed that XGBoost outperformed other models in terms of accuracy, offering valuable insights for stakeholders in the automotive industry. The study underscores the efficacy of machine learning in transforming data-driven decision-making, marking a significant stride in the predictive analysis of vehicle sales. A database that includes information on more than 5386116 vehicles, including car characteristics was collected. Based on this 10 machine learning models were analyzed and the best model showed an accuracy of r2=0.94 in price prediction. The probability of using regression models to determine the time of car sales was also investigated. The resulting model has been integrated into a system that allows car owners or dealers to quickly get a forecast of the possible sale price of their vehicle. The study also includes a comparison of different forecasting approaches, an analysis of the importance of certain factors for pricing, as well as a study of market trends and demand for such forecasting systems. The practical value of the work lies in the possibility of optimizing the car sales process, reducing risks for sellers, and ensuring better information for buyers.

Keywords¹

Machine learning, cascade model, price forecasting, car market, time of sale forecasting.

1. Introduction

The modern world is full of data that is constantly being accumulated and analyzed to improve various aspects of our lives. One of the key industries that utilizes the power of data analysis is the automotive industry. In particular, predicting the price and time of sale of automotive vehicles is an important task that can contribute to more efficient inventory management, optimized pricing, and increased customer satisfaction.

The relevance of the study is driven by the growing need to automate and optimize processes in the automotive industry, as well as the large amount of available data on car sales. The use of machine learning models in this area can bring significant benefits, including improved forecast accuracy, speed of processing large amounts of data, and the ability to adapt to changes in the market.

Today, there is a wide range of tools and libraries for working with machine learning, such as TensorFlow, PyTorch, Scikit-learn, etc. These tools provide opportunities to create, train, and evaluate various machine learning models.

However, along with the opportunities, there are also challenges. These include a large amount of unstructured and incomplete data, the need to process large amounts of information, and the need to develop accurate and reliable models that can adapt to changes in data and market conditions.

D 0000-0002-7100-3263 (V. Shymanskyi); 0009-0001-0548-525X (V. Liaskovets)



© 2023 Copyright for this paper by its authors.
 Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
 CEUR Workshop Proceedings (CEUR-WS.org)

ProfIT AI 2023: 3rd International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2023), November 20–22, 2023, Waterloo, Canada

Svshymanskiy@gmail.com (V. Shymanskyi); (vasyl.liaskovets.mknssh.2022@lpnu.ua (V. Liaskovets)

The purpose of this paper is to study various machine learning models and determine their effectiveness for the task of predicting the price and time of sale of motor vehicles. The main tasks include selecting the optimal data processing methods, developing and comparing different models, and evaluating their accuracy and efficiency.

The results of this work can be useful for car dealers, automotive market analysts, automotive software developers, and researchers who study machine learning methods and their application in real-world settings.

Objectives of the study:

- Data collection;
- Investigate the parameters that affect the accuracy of the model;
- Find the best model for price estimation in terms of training time/prediction accuracy;
- Based on the price prediction model, to study the impact on the time of sale of motor vehicles.

The auto.ria.com portal has a car price calculator tool. Using this tool, based on the selected car characteristics, the user can see ads that have been published for a specific period - six months, a year, or several months. This service shows the minimum, maximum, and average price based on the submitted ads. However, this approach may not be ideal, especially if a certain characteristic is quite specific, such as a specific mileage for a certain car model.

The lack of a sufficient number of similar ads or limiting the values of the parameters for comparison can lead to an incorrect estimate. Therefore, the main goal of this paper is to use machine learning technologies to optimize the process of car valuation.

Machine learning makes it possible to determine the relationship between various parameters and the cost of a car. Regardless of the limitations of existing solutions, the new model will allow users to obtain more accurate cost forecasts based on the specified parameters. Thus, the purpose of this paper is to create a software solution for car valuation using machine learning technologies that provide users with the ability to obtain an accurate estimate based on specified characteristics.

2. Analysis of related publications

This paper conducts an analytical review of scientific sources by the standardized PRISMA methodology, which contains general recommendations for reviewing scientific and basic features of meta-analysis.

In the article [1], the authors used a linear regression model to predict car prices. They analyzed the results and concluded that the LR method is a suitable model for price estimation since characteristics such as brand, mileage, year, and transmission type of used cars directly affect their price. We observed that the success of the model increased as we removed illogical values in the dataset.

In a study [2], the authors applied the Random Forest Regressor Model to forecast car prices in Turkey. In this study, the authors conducted experiments on five regression models (Ridge, Decision Tree, Random Forest, Gradient Boosting, and XGBoost) to find the right model according to the data and predict the sale price of used cars with good accuracy. The model with the highest accuracy is a random forest. Results of the random forest model evaluation. After testing with five models for predicting car prices, the lowest MAE and RMSE error rate, then the random forest is the best.

In a study [3] the authors used three algorithms: linear regression, lasso regression, and ridge regression. The data was divided into two parts for training and testing using an SVM (Support Vector Machine) classifier, i.e. 75% of the data was used for training, and 25% of the data was used for testing machine learning. The accuracy of the three machine learning models was tested and compared with each other. The final result was predicted according to the algorithm that provides greater accuracy. The main drawback of this project was the smaller number of records that were used. In future work, they expect to collect more information and use further advanced

methods such as random forest, ANN (artificial neural network), and CNN (convolutional neural network) with a better computer user interface.

In [4] the authors investigated how the XGBoost+LightGBM Framework affects the results of the basic models for predicting used car prices.

In [5], the authors tried to solve the problem of price estimation in the used car retail market in China by developing machine learning functions (XGBoost, CatBoost, LightGBM) and an artificial neural network model based on collected data on 30,000 Chinese used car transactions. A direct feature selection algorithm was used to solve the optimal combination of features; a grid search algorithm was used to optimize the hyperparameters of each model.

In [6] the authors apply text mining techniques to create predictors for regressing car prices based on unstructured data, and textual descriptions in car ads. In this paper, the authors also investigate the effectiveness of reducing the dimensionality of the dictionary by applying stemming, lemmatization, or neither.

In [7] analyzes the factors that affect the price of used cars from three aspects: used car parameters, vehicle condition factors, and transaction factors, and establishes a used car price estimation system that includes 12 characteristic variables. Using web crawler technology to obtain used car transaction data, three prediction models, BPNN, GRABPNN, and PSO-GRA-BPNN, are developed to conduct comparative validation and analysis of the results.

In [8] the authors tested three machine learning algorithms: regression, decision tree, and random forest, and compared the results.

In [9] the paper proposes a model for predicting used car prices based on a neural network algorithm that is combined with mean coding to handle data features more carefully and avoid high-dimensional features. Experimental results show that this method has good accuracy in predicting used car prices. In this article, BP, decision tree regression, and SVM were used to model the same dataset. The same test data was used to validate the models, and different evaluation metrics were applied. The results showed that the improved BP neural network model proposed in this article has the best performance. In practical applications, customers can use the model proposed in this paper to estimate the sale value of cars in combination with their characteristics to help in decision-making.

In [10], the authors tested the algorithms of linear regression, ensemble of trees (regression), random forest (regression), Gradient Boosted Tree (regression), and simple regression tree performed on the same data set and analyzed it using the KNIME analytics tool. The result showed that the Gradient Boosted Tree (Regression) model has the highest R2 value regardless of the percentage of distribution (train and test set), as well as the highest linearity that can be observed from the scatter plot.

The objective of the study [11] is to use the linear regression (LR) algorithm to predict the price of a car and compare the accuracy with the support vector machine (SVM) classification algorithm.

In [12] proposes a model for predicting the price of a car by considering one important forecast variable and taking into account all important variables simultaneously. The important variables found by checking the correlation are used to create forecasting models. First, we tried with an individual linear regression model and calculated the average predicted price. The next MLR model considers all attributes simultaneously to find the predicted price. A single polynomial model is also prepared and finally, they are evaluated for fit.

In the paper [13], a model was obtained that can be used to predict the sale prices of used BMW cars and can be used to estimate and predict the price of used cars of other brands. The model is also adaptable to other methods through iterations and improvements. Several parameters were used for cross-validation to facilitate training and forecasting and to select the most optimal model. From the basic models of SVR, KNN, RF, and DT, it was seen that RF and KNN are the two basic models that can provide the highest accuracy.

In the article [14], three different methods were used to predict the price of used cars, respectively. Among them, the random forest model had the highest goodness-of-fit and the model evaluation result has the best performance compared to the other two models. In any case, all the proposed models can be used to predict the price of a used car.

In article [15] authors introduce a novel method for handling small datasets using a GRNN, demonstrating its superior accuracy compared to other methods. However, it is noted that the method incurs significant time delays compared to the basic GRNN.

Ensemble models combine forecasts from different base models, which can lead to more accurate and reliable results. This is particularly useful in situations where individual models may be vulnerable to noise or exhibit certain weaknesses [16].

By analyzing scientific articles and their number through the Scholar and Scopus search engines, we can analyze trends and conclude the relevance of the topic. We analyzed the issues that the authors faced in their research and the methods used in this or that paper, and based on this, we determined the purpose and objectives of the research.

3. Methodology

3.1. General Scheme of the Models

A cascade model for predicting the price and time of sale of a car based on various parameters and information about it is proposed. The proposed workflow of this model is shown in Fig. 1. This algorithm consists of the following steps:

- 1. Step 1: Data collection and preparation. A dataset is being formed, which includes various parameters of cars for sale. These parameters can include characteristics such as car make, year of manufacture, mileage, engine type, transmission type, body type, and many others.
- 2. Step 2: Train the model to predict the selling price. The prepared dataset is used to train a machine learning model to predict the selling price of a car. This model can use regression algorithms that analyze the relationship between car parameters and its selling price. After successful training, the model can predict the selling price for new cars.
- 3. Step 3: Expansion of the dataset. After the sales price prediction model is trained, an expanded dataset is created that contains information about the predicted sales price for each vehicle.
- 4. Step 4: Train the model to predict sales time. The final step involves training a machine learning model to predict when the car will sell. This model uses both the input from the previous model (predicted sales price) and other car parameters to predict how long it will take to sell the car on the market.



Figure 1: The workflow of the proposed cascade model

Overall, this cascading forecasting model provides a better understanding of what factors affect the time to sell a car and helps owners and sellers make informed decisions about selling cars in the market. Cascade models allow you to take into account new features to improve the accuracy of the model [17]. Parallelization algorithms make it possible to speed up the process of model implementation [18].

3.2. Dataset formation

The auto.ria.ua service was used to fill the dataset. auto.ria.ua is one of the largest automotive portals in Ukraine, which provides an opportunity to buy and sell vehicles. It not only helps potential buyers find cars but also provides a valuable analytical tool for studying the car market

The service is one of the most visited automotive websites in Ukraine. Its popularity is due to its wide selection of cars, intuitive interface, and reliable information. Additionally, many car dealerships and showrooms also use this portal to advertise their products.

After a detailed study of the data structure on the website and the capabilities provided by the API, it became clear that information about each individual car is stored in the form of a JSON file. This file has a unique identifier that is reflected in its name. This opens up new opportunities for data collection.

In particular, having the identifier of a particular car, you can access its data by going to a specific URL that leads to the corresponding JSON file. Taking into account this feature of the data structure, it is possible to develop a data collection methodology based on the recursive traversal of car identifiers.

For example, if you know the ID of the last car, you can start collecting data in reverse order, and thus get information about all the cars that were posted before it. This method will automate the data collection process and provide comprehensive information about cars that have been listed for sale in the past on AUTO.RIA.

It's worth noting that this method of storing data in static JSON files, where each car is associated with its unique identifier, has many advantages:

- 1. Saving resources: There is no need to use server resources to generate dynamic content every time a user makes a request. Instead, the user simply downloads a pre-prepared static file.
- 2. Loading speed: Static files tend to load faster than dynamically generated pages, which improves the user experience.
- 3. Scalability: Static files can be easily cached and distributed through content delivery networks (CDNs), which ensures high availability and speed of data access even with a large number of users.
- 4. Security: Since data is transferred as static files, the risk of potential injections or other attacks on the server is reduced.
- 5. Ease of management: Static files are easier to work with in terms of data storage, backup, and migration.

To summarize, this data structure not only simplifies the process of collecting data for analytical purposes but also provides several benefits for website owners and users.

Looking through the JSON file at the link, you can identify several main blocks that can be immediately discarded and not saved to reduce the dimensionality:

- 1. vinSvg svg file with the VIN code, it is not needed because we have a textual representation.
- 2. modelId, markId, modelNameEng internal identifiers for the service, and English names.
- 3. prices an object with prices in different currencies, we don't need it because we have the same values in the top scopes.
 - UAH, EUR because we will focus on dollars.
- 4. codedVin, optionStyles, levelData, infotechReport, secureKey, dontComment, sendComments other fields that do not contain useful information.

We also have the stateName, and regionName fields in stateData, you can get them and delete the parent object

- 1. photos a field with photos, we just convert it to have information about the number of photos.
- 2. dealer information about the dealer, you can delete the information about the dealer, and then just use the feature whether there are ads from the dealer or not.
- 3. color converts a color object to a color name.

In the context of the work, some graphs were created to analyze the price category. These graphs show the price distribution of the cars entering the dataset.

The Fig. 2.a graph shows the prices of cars in the original dataset without any filtering. It gives you a general idea of the price range and price distribution between cars.

However, as you can see in the Fig. 2.a graph, there are some outliers in the upper price range. These outliers can affect the accuracy of machine learning models, so it was decided to filter the data.

The graphs present car prices after filtering to remove high-price outliers. The Fig. 2.b reflects car prices that are below 200000 USD, while the Fig. 2.c focuses on cars with a price below 50,000 USD.

Applying such filtering helps to get a more detailed view of the price distribution in the main range and can help to improve the efficiency of machine learning models used to predict the price of motor vehicles. Based on these graphs, the resulting dataset was filtered and cars with a price of less than \$50,000 were taken for the study.



Figure 2: Car price distribution in the dataset

Fig. 3 shows the distribution of cars by their mileage. Most of the data is concentrated in the range of mileage from 0 to about 500. The peak of the distribution is observed in the range from 150 to 350, where the largest number of cars have this mileage.

After 500 miles, there is a significant drop in the number of vehicles, with a few small outliers, indicating fewer high-mileage vehicles.

However, there is another peak around 800, which may indicate a certain group of cars with significantly higher mileage than the bulk.

In general, the graph shows that most cars have low to medium mileage, but there are also cars with high mileage.

Fig. 4 shows the plot matrix for car data, which displays their cost (USD), year of manufacture (year), mileage (raceInt), and condition (damage: True or False).

After analyzing the obtained distribution, the following conclusions can be drawn: most cars cost up to 10,000 USD; the prevalence of prices for cars that have been damaged (True) is lower than for cars that have not been damaged (False); most of the cars were manufactured after 1980; there is a tendency for older cars to have higher mileage; most cars have less than 200 kilometers; there is a connection between mileage and value: cars with lower mileage tend to be more expensive.

By colored points, you can see that cars that have been damaged (True) tend to have a lower value, an older year of manufacture, and higher mileage than cars that have not been damaged

(False). The condition of the car (damaged or undamaged) affects its value, mileage, and year of manufacture.



Figure 3: The distribution of cars by their mileage



Figure 4: The combined distribution of cars by their price, year, mileage, and condition

There is an expected relationship between the value, mileage, and year of manufacture of a car. Cars with lower mileage and a newer year of manufacture tend to be more expensive, while

cars that have been damaged tend to have lower values, an older year of manufacture, and higher mileage.

3.3. Data Preprocessing

The resulting data usually requires extensive preparation and cleaning. This step includes trimming or removing incorrect or missing values, as well as dealing with outliers and anomalies in the data. It may also be necessary to transform some parameters to improve the quality of the model.

To increase the effectiveness of the model, new features can be created based on existing ones. These new features can help the model better distinguish between cars and improve the accuracy of predictions. To ensure stability and optimal operation of the models, data scaling is performed. This means bringing the features to the same scale to avoid problems with different units of measurement and different feature weights.

Such actions make it possible to build a high-quality dataset that can be used to train machine learning models to predict car sales times. This data preparation is essential to achieve the best results in car sales forecasting and analysis.

It was also decided to create a new feature, engine size, for cars with internal combustion engines, from the fuelName column, with values shown in Fig. 5. It was decided to get the volume value and put it in the 'volume_num' column. Also, columns with volume_num=0 for internal combustion engines were removed.

fuelName		fuel	volue_num
Газ пропан-бутан / Бензин, 2.5 л.	→	Газ пропан-бутан / Бензин	2.5
Газ пропан-бутан / Бензин, 1.6 л.	→	Газ пропан-бутан / Бензин	1.6
Бензин, 2 л.	\rightarrow	Бензин	2.0
Дизель, 2 л.	\rightarrow	Дизель	2.0
Дизель, 1.5 л.	→	Дизель	1.5
Бензин, 1.6 л.	\rightarrow	Бензин	1.6
Бензин, 3.6 л.	\rightarrow	Бензин	3.6
Бензин, 1.8 л.	\rightarrow	Бензин	1.8
Газ пропан-бутан / Бензин, 2 л.	\rightarrow	Газ пропан-бутан / Бензин	2.0
Бензин, 2.4 л.	\rightarrow	Бензин	2.4

Figure 5: Creating attributes fuel and volue_num based on attribute fuelName

We also have columns markName, modelName. It was decided to create a field mId - which is a car model that consists of a name and a model (Fig. 6). To translate it into categories later. It was decided not to use the model and brand separately, because, for some brands, model names can have the same value.

markName	modelName		mld
УАЗ	2206	\rightarrow	УАЗ 2206
Volkswagen	Golf	→	Volkswagen Golf
Nissan	X-Trail	\rightarrow	Nissan X-Trail
Chrysler	Sebring	→	Chrysler Sebring
Renault	Megane	\rightarrow	Renault Megane
BMW	Х3	\rightarrow	BMW X3
Skoda	Octavia	\rightarrow	Skoda Octavia
Daewoo	Matiz	\rightarrow	Daewoo Matiz
Audi	80	→	Audi 80

Figure 6: Concatenation of car make and model names

To estimate the time of sale, four fields display the time values addDate, updateDate, expiresDate, soldDate.

Two additional columns were created:

- 1. hours soldDate addDate, in hours.
- 2. hours2 soldDate updateDate, in hours.

The data was collected in portions. It was decided to use downloading portions of ads with the ability to store them on the server (datalake). For example, for a quick analysis, we do not need to have a large amount of data, and to analyze the fields, an initial set of one portion is enough. The portion in this study was 50000. There are some advantages to downloading in batches:

- 1. Loading control: if the initial amount of bottom is not enough, you can request the next portion.
- 2. Caching: one portion is stored only once, and when you request it again, the result is taken from the cache.
- 3. Scheduling: you can schedule the download of a specific portion at a time when theoretically fewer people use the service, so as not to load the portal.

Data acquisition was divided into several stages

- 1. Retrieving a portion of raw data and storing unstructured data on a temporary server.
- 2. Process fields, delete unnecessary ones, and create new characteristics.
- 3. Adding the portion of data to the main list that will be used in the study.

The first and second points of data storage were done in nodejs because it works very well with parallel tasks and is quite fast. The third step was done in Python and Jupyter Notebook because they are very powerful tools for data analysis and data viewing.

At the final stage, a data frame of dimension (4410940, 33) was obtained.

4. Results and Discussion

4.1. The Price Forecasting Task

The purpose of the study is to investigate machine learning models that would allow predicting the cost and duration of car sales. To achieve this goal, it was decided to focus on regression models.

Regression models are known for their ability to work efficiently in tasks where you need to predict numerical values based on given input data. Since our goal is to predict parameters such as price and time of sale, which are quantitative characteristics, regression models are an ideal choice. They can take into account the relationships between various car characteristics and their price or time to sale, providing accurate and reasonable forecasts.

To evaluate machine learning models, we wrote a function that allowed us to compare which model is best suited for a given task based on the backward models and the given evaluation functions.

We selected 10 models for analysis:

- 1. Linear Regression: This is one of the most basic regression models that establishes a linear relationship between the dependent and independent variable(s). Using linear regression can help determine the impact of certain car characteristics on its price.
- 2. Linear SVR (Support Vector Regression): This model uses the principles of support vectors but for regression tasks. It tries to find the hyperplane that best approximates the real values.
- 3. MLPRegressor: Multilayer perceptron is a neural network that can overcome nonlinearities in data. It is suitable for complex datasets with multiple characteristics.
- 4. Stochastic Gradient Descent: This model uses an iterative optimization method to minimize the cost function. It is often used in large-scale problems.
- 5. Decision Tree Regressor: A decision tree splits data into subsets based on certain criteria, allowing the model to make predictions based on decisions made at each level of the tree.

- 6. Random Forest: This ensemble model uses multiple decision trees to work with different subsets of data and then averages their results.
- 7. XGBoost: is an optimized gradient boosting library that is effective for analytical tasks, including regression.
- 8. Ridge Regressor: This is a form of linear regression that includes L2 regularization. It can help in avoiding overlearning.
- 9. Bagging Regressor: This model uses the bagging technique to build multiple submodels based on data samples and then averages their results.
- 10. Extra Trees Regressor: This model is similar to the random forest, but uses the entire dataset to build each tree, providing additional randomness.

We also added the following metrics to evaluate performance:

Relative error: A measure of how well a model predicts outcomes compared to actual data, normalized to the mean of the target variable.

Maximum absolute error: Shows the largest deviation between the predicted and actual values.

$$\mathbf{MAE} = \max\left(|y_1 - \widetilde{y}_1|, |y_2 - \widetilde{y}_2|, \dots |y_n - \widetilde{y}_n|\right)$$
(1)

MSE (Mean Square Error): Shows the average of the squared differences between the predicted and actual values.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$
⁽²⁾

RMSE (Root Mean Square Error): This is the standard deviation of the difference between the predicted and observed values. The smaller the RMSE, the better the model predicts.

$$\mathbf{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}$$
(3)

MAPE (Mean absolute percentage error): Shows the relative size of the error as a percentage. This helps to understand how large the model error is relative to the actual values.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \tilde{y}_i}{\tilde{y}_i} \right|$$
(4)

MedAE (Median Absolute Error): This metric shows the median value of the absolute differences between predicted and observed values. It is less sensitive to outliers than MAE.

$$MedAE = median(|y_1 - \tilde{y}_1|, |y_2 - \tilde{y}_2|, ... |y_n - \tilde{y}_n|)$$
(5)

At the initial stage, three tests were made with different amounts of input data: 50000, 100000, and 250000, this was done in order to understand how the models behave with different amounts of input data and to choose the best possible model. It was also necessary to estimate the time and dimensionality of the resulting model.

Results of different machine	learning models	for price	prediction task

Table 1

Model	Time	Size	r2_train	r2_test	D_train	D_test	RMSE_ train	RMSE _test
XGB	11.517s	4.74	0.97	0.93	0.17	0.21	0.18	0.26
Bagging Regressor	12.260s	158.18	0.98	0.92	0.09	0.22	0.12	0.28
Random Forest	53.582s	11.47	0.9	0.89	0.27	0.28	0.31	0.33
Decision Tree Regressor	0.942s	0.12	0.89	0.87	0.29	0.3	0.34	0.36
Extra Trees Regressor	17.289s	10.83	0.86	0.86	0.33	0.33	0.37	0.38
MLPRegressor	35.219s	0.02	0.82	0.82	0.38	0.38	0.42	0.42
Linear Regresion	0.096s	0.0	0.6	0.61	0.58	0.57	0.63	0.62
Ridge Regresion	0.481s	0.0	0.6	0.61	0.58	0.57	0.63	0.62

Stochastic Gradient Descend	0.276s	0.0	0.6	0.61	0.58	0.57	0.63	0.63
Linear SVR	18.966s	0.0	0.56	0.57	0.54	0.53	0.66	0.65

Table 1 shows the results of different machine learning models. Here is a brief analysis of their effectiveness.

XGB (XGBoost) has the highest quality scores (r2_train = 0.98, r2_test = 0.93) among all models. This confirms the effectiveness of XGBoost in many machine learning tasks. Note that the RMSE and MSE for this model are also quite low compared to the others, which is positive. The Bagging Regressor and Random Forest also showed high quality on the training data (r2_train 0.98 and 0.92 respectively). However, the quality of the test data is slightly lower compared to XGB. Models such as Linear Regression, Ridge Regressor, Stochastic Gradient Descent, and Linear SVR have relatively low quality scores on both training and test data compared to other models. Decision Tree Regressor and Extra Trees Regressor showed average results. Given these results, XGBoost seems to be the best choice for this particular dataset.

Fig. 7 shows the relationship between the actual values (y_test) and the model predictions (y_pred). Ideally, all the points would lie along a straight line with a slope of 45 degrees, as this would mean that the model predictions are in agreement with the actual values.

Here are a few observations after analyzing the results in Fig. 7:

- 1. General trend: Most points are concentrated along the line, illustrating a certain degree of correlation between actual values and predictions.
- 2. Scatter of values: In the low range of y_test (up to about 10,000), the model's predictions seem to be quite accurate. However, as the y_test values increase, the spread of predictions also increases.
- 3. Large values: For very large values of y_test (around 40,000-50,000), the model seems to be less accurate and the predictions are more scattered.

Overall, the model seems to be quite effective at predicting low and medium y_test values but may be less accurate for high values. Further tuning or consideration of other aspects of the data may be required to improve prediction accuracy at these high levels.



y_test vs y_pred

Figure 7: Actual vs. predicted values scatter plot

4.2. The Sales Time Forecasting Task

With a model that can predict the market value of cars with high accuracy, we can assess how much the price of a particular car deviates from market standards, defining it as overpriced or underpriced. This aspect plays a critical role in determining the pace of vehicle sales. For example, a car whose price is significantly lower than the market price has a good chance of being sold in the blink of an eye, as it represents a bargain for the customer. On the contrary, overpriced cars are likely to stay on the market for a long period, scaring off potential buyers. Thus, our model serves as an invaluable tool in determining the optimal pricing policy, which undoubtedly contributes to the efficiency and speed of car sales.

By working with the actual and predicted prices of a car, we can calculate not only the absolute price difference but also determine by what percentage the actual price differs from the predicted price. This indicator can be calculated using the formula:

$$Deviation = \frac{Actual \, Value - \, Forecasted Value}{Forecasted Value} \tag{6}$$

This indicator helps you get a clear picture of how significant the price deviation is and can be a useful tool when making decisions about selling a car.

Impact on the time of sale:

- 1. Underestimated price: If the actual price is significantly lower than the predicted price, the percentage deviation will be negative. This situation usually leads to a quick sale, as buyers see it as a good deal.
- 2. Overpriced: The percentage deviation will be positive if the actual price is higher than the predicted price. Such cars are likely to stay on the market longer, as buyers may consider the price too high.

The price is in line with the market: If the actual price is very close to the predicted price, the percentage deviation will be small and the car will likely be sold within the average time to market.

Thus, by analyzing the percentage deviation of the actual price from the predicted price, you can get important insights into the time of car sales, as well as make an informed decision about the necessary adjustments in pricing.

There are four columns (addDate, updateDate, expireDate, soldDate) in your dataset that capture information about dates and times related to the car sales process. Each column plays a different role and has a different purpose. Understanding how these four columns interact with each other and what insights they can provide is key to analyzing data and improving the car sales process.

By redrawing these fields, we created two new features based on the time and dates in your dataset that can be useful for a car sales time prediction model. Here is their description:

- 1. hours: calculated as the difference between "soldDate" (date and time of the car sale) and "addDate" (date and time of the ad addition). This feature is intended to reflect the total time that the car has been on sale, from the moment it was added to the market to the moment it was sold.
- 2. hours2: calculated as the difference between "soldDate" and "updateDate" (date and time of the last update of the listing). This attribute can reflect the market reaction time to an ad update, which can be important for assessing how quickly buyers react to changes in the ad.

Both of these attributes are measured in hours, which provides a more detailed measurement of the time a vehicle has spent on the market and can help your model better understand sales dynamics.

Let's analyze and display the distribution for the two columns shown in Fig. 8.



Figure 8: Histogram for hours and hours2

Fig. 8 has two main distinct areas or peaks. One peak is centered around the value of 1 (10^0), while the second peak is near the value of 1000 (10^3). The red dashed line indicates the median value of the selling time for hours. This value is located between the two main peaks, but closer to the peak around 1000 (10^3). The time to sell mainly varies between about 0.01 hours (10^-2) and 10^4 hours.

The histogram for hours2 shows one very distinct peak around the value of 1 (10^{\circ}0). The red dotted line shows the median value of the sale time for hours2, which is very close to the main peak. The selling time is concentrated in the range of about 10^{\circ}-6 to 10^{\circ}4 hours.

Negative values for the time of sale are illogical and may indicate data errors. The following steps are usually taken in this case: data origin analysis; delete or replace; and filtering.

We removed negative values and zero (Fig. 9). Because it is difficult to imagine a scenario where the car was sold immediately after filing.



Figure 9: Hours histogram distributions with zero removes

Thus, based on the graphs presented, we can conclude that after replacing the negative values with the median in the hours and hours2 columns, there were no significant changes in the

distribution of the data. The distribution of values in the histograms remained virtually unchanged, and the median values in both graphs also did not change. Therefore, it can be argued that the manipulations performed on the data did not affect the overall distribution. Removing negative values leads to a higher concentration of data near the origin, indicating that negative values have had an impact on the distribution of the data. In both datasets, most events occur within a short period, and their number decreases over time.

You can also look at Fig. 10 of car sales additions, a general graph that will help you see trends and get more information about the dataset. The graph shows the dynamics of added and sold cars over a certain period.



Figure 10: Dynamics of added and sold cars over a certain period

Main observations:

• General trend: Both the "added" and "sold" lines show some seasonality, with higher peaks in certain periods of the year, as well as a slight decline in 2022.

• A big drop in 2022: There is a noticeable decline in both the number of items added and the number of items sold starting in February 2022. This coincides with the date of Russia's military aggression against Ukraine, indicating that this event had a major impact on the market. This event severely affected the typical distribution of added and sold vehicles, leading to a prolonged depression in the market that lasted for several months.

• Recovery: After the initial decline, we see a certain recovery in market activity, but the levels of items added and sold are not returning to their previous highs.

• 2023: Again, there is an increase in the number of items added, but the items sold do not show the same growth, which may indicate a certain gap between supply and demand.

As for the automotive market, Russian aggression has obviously affected the market, leading to a decrease in the number of announcements and sales. The unfavorable conditions created by military activities may cause uncertainty among consumers and sellers, which in turn may lead to a decline in market activity. In addition, possible logistics and supply disruptions could also affect the automotive market.

It is important to take this event into account when analyzing data, as it could distort the usual market trends and influence the decisions of both buyers and sellers.

Based on these results for the prediction of sale time, it is possible to draw conclusions and develop approaches that will provide tools and methods for analyzing the actual sale price of vehicles in the future.

5. Conclusions

The task of the paper and the steps to its fulfillment were set. The task was to investigate machine learning models for predicting the price and time of sale of motor vehicles. We analyzed and tested 10 models, of which the XGBoost library proved to be the best at price forecasting. Based on the results, it was possible to analyze and find the prices with the largest deviations and find false ads. There was also an attempt to predict the speed of car sales based on the estimated price and the current price. In the course of the study, it was found that a significant part of the ads did not correctly reflect the time of sale of the car, because many users, instead of removing the ad from the publication, marked it as sold, which distorted the data used to determine the time of sale.

In conclusion, while our study has provided significant insights into the application of machine learning models in predicting vehicle sale prices, it opens avenues for future research. Further exploration could involve the utilization of more diverse datasets, including a broader range of vehicle types and incorporating regional variations. Additionally, investigating the impact of emerging technologies in machine learning and data preprocessing methods could further refine predictive accuracy. Such future endeavors will not only enhance the model's applicability but also contribute to a deeper understanding of market dynamics in the automotive industry.

To conclude, our study not only demonstrates the power of machine learning in predicting vehicle sales outcomes but also sheds light on its potential applications in broader contexts. While the current findings are robust, future research could expand on these models to include more diverse datasets and explore their applicability in different market conditions. The study's limitations also pave the way for further investigation into the interplay of external factors like economic trends and consumer preferences. Ultimately, this research contributes to a more nuanced understanding of the automotive market, offering a valuable tool for decision-makers in this dynamic industry.

References

- [1] Muti S., Yildiz K. "Using Linear Regression For Used Car Price Prediction". International Journal of Computational and Experimental Science and Engineering. 2023.
- [2] Kriswantara B., Sadikin R. "Used Car Price Prediction with Random Forest Regressor Model". Journal of Information Systems. 2022.
- [3] Bharambe P., et al. "Used Car Price Prediction using Different Machine Learning Algorithms". International Journal for Research in Applied Science and Engineering Technology. 2022. Vol. 10, No. 4. pp. 773-778. https://doi.org/10.22214/ijraset.2022.41300
- [4] Cui B., et al. "Used Car Price Prediction Based on the Iterative Framework of XGBoost+LightGBM". Electronics. 2022. Vol. 11, No. 18. pp. 2932. https://doi.org/10.3390/electronics11182932
- [5] Huang J., et al. "Used Car Price Prediction Analysis Based on Machine Learning". ICAID 2022, Atlantis Press International BV, 2023. pp. 356-364. https://doi.org/10.2991/978-94-6463-010-7_37
- [6] Gonçalves R. M. G. "Text Mining Techniques for Car Price Prediction", 2021.
- [7] Liu E., et al. "Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network". 2022. https://doi.org/10.3390/su14158993
- [8] Gaikwad D. V., et al. "Resale Car Price Prediction". 2022. Vol. 21, No. 11, pp. 2205-2219
- [9] Kan D., et al. "Prediction of used car prices using back propagation neural network model based on mean encoding". SPIE, 2023.
- [10] R C. A., C D. A. B. "Performance Analysis of Regression Algorithms for Used Car Price Prediction: KNIME Analytics Platform". International Journal for Research in Applied Science and Engineering Technology. 2023. Vol. 11, No. 2. pp. 1324-1331. https://doi.org/10.22214/ijraset.2023.49180

- [11] Siva R., M A. "Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm". ECS Transactions. 2022. Vol. 107, No. 1. pp. 12953-12964. https://doi.org/10.1149/10701.12953ecst
- [12] Roy A., Ghosh D. "Learning Based Data Science Model for Car Price Prediction", 2022. https://doi.org/10.1007/978-981-99-4677-8_10
- [13] Kuo P.-H., et al. "A Multimodel Transfer-Learning-Based Car Price Prediction Model with an Automatic Fuzzy Logic Parameter Optimizer". Computer Systems Science and Engineering. 2023. Vol. 46, No. 2. pp. 1577-1596. https://doi.org/10.32604/csse.2023.036292
- [14] Chen J., et al. "Comparisons of different methods used for second-hand car price prediction". SPIE, 2022.
- [15] Izonin I., et al. "A GRNN-based Approach towards Prediction from Small Datasets in Medical Application". Procedia Computer Science, Vol. 184, 2021, pp. 242-249, https://doi.org/10.1016/j.procs.2021.03.033.
- [16] Natalya Shakhovska, et al. "A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system". Mathematical Biosciences and Engineering, 2022, 19(6): 6102-6123. doi: 10.3934/mbe.2022285
- [17] Semkovych, V.; Shymanskyi, V. "Combining OCR Methods to Improve Handwritten Text Recognition with Low System Technical Requirements". LNDECT, 2023, pp. 693–702. https://doi.org/ 10.1007/978-3-031-24475-9_56.
- [18] Lesia Mochurad. "Optimization of Regression Analysis by Conducting Parallel Calculations". COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine, pp. 982-996.