Predicting Used Cars Prices in the Egyptian Market: A Machine Learning Approach

2025 22nd International Learning and Technology Conference (L&T) | 979-8-3315-3221-5/25/$31.00 ©2025 IEEE | DOI: 10.1109/LT64002.2025.10941392

Mohamed Taher Gamal Eldeen
Information Technology and Computer Science,
Nile University
Giza, Egypt
mo.taher@nu.edu.eg

Sahar Fawzi
Information Technology and Computer Science,
Nile University
Giza, Egypt
sfawzi@nu.edu.eg

***Abstract*— Valuing used cars in the Egyptian market poses a significant challenge due to the scarcity of new vehicles and high inflation. Consequently, used cars dominated the Automotive market, causing price fluctuations and creating a complex pricing environment. Automating the valuation process presents a valuable opportunity to save time and minimize pricing errors. This study explores the use of machine learning models to predict used car prices using data scrapped from popular online listings websites. It compares the performance of top-performing algorithms, according to previous research, including XGBoost, Random Forest Regression, and Bagging Regressor when trained on price-range segmented data versus the full dataset to determine the optimal model. Results indicated that applying price-range segmentation reduced the Mean Absolute Error (MAE) by 23% compared to training without segmentation. XGBoost outperformed other models, achieving an MAE of 47,478, compared to 48,829 for both Random Forest Regression and Bagging Regressor.**

Keywords*— *Machine learning algorithm, RFR, Bagging regressor, XG Boost

I. Introduction

The demand for used cars has surged significantly worldwide over the past three years. Recent statistical research highlights this global trend, showing a notable rise in consumer interest in purchasing used vehicles. Manheim [1] introduced a methodology for calculating the Used Vehicle Value Index (UVVI), which was derived from analyzing an extensive database of over 5 million used vehicle transactions annually. This index provides valuable insights into market trends and pricing dynamics in the used car industry. UVVI vehicle index detected an unprecedented increase in used cars values as they increased by 6.81% in April 2021 compared to the March 2021as well as a 52.2% increase in value compared to April 2020. Stimulus payments and tax refunds caused this due to Covid-19 ramifications, and the decrease in car production globally. However, the index began to decrease till January 2024 by showing 33.8% which indicates a continuing shortage in global market. In addition to global crisis, the automotive Egyptian market was affected by the local economic crisis and the depreciation of the Egyptian pound which caused a noticeable scarcity of new cars and used cars became the available choice for the customer, therefore this led to a dramatic price increase for used cars [2]. Accurate car price prediction involves expert knowledge because price usually depends on many distinct features and factors, the most critical ones are generally the brand and model, year of manufacturing, horsepower, and mileage [3]. It is a fact that customers find it challenging to determine the appropriate price when buying or selling a used car. The primary reference for pricing is often online listings on used car advertisement websites. However, those seeking accurate pricing frequently encounter various issues, such as outdated listings, varying specifications, and uncertainty about the car's actual condition, making the search process difficult and unreliable. Furthermore, there is a major problem when searching for related car listings which is the variety of prices on the different listings. These issues made used cars valuation process very challenging [4]. This research aims to develop a reliable model for predicting used car prices by experimenting with various machine learning algorithms. The goal is to identify the most effective methodology for accurately valuing used cars, addressing the challenges in the current pricing process.

The research paper will be as follows. Section II will discuss the related work to the problem. Section III will discuss the methodology: data structure, exploratory data analysis, cleaning and transforming the data to be ready for training models, section IV will discuss models development, optimization and evaluation. Section V will be the conclusion and future work.

II. Related work

Abdulla Al Shared in his thesis[5] experimented with *RFR*, Linear regression, and Bagging regressor on the *UAE* used cars dataset scrapped from BuyAnyCar website and evaluated the results using MAE, RMSE MSE, and R2 score. He applied data preprocessing techniques on the dataset by removing duplicates, imputing missing data, and applying feature transformations to clean the data before modeling. The results showed that RFR showed the best R2 score performance with 90%, followed by Bagging regressor with 88% then Linear regressor with 85%.

Enis Gegic[6] experimented with Artificial neural networks, and *RFR* on predicting used cars prices in Bosnia and Herzegovina used cars prices scrapped dataset. His research was focused on converting the regression problem to a classification problem to predict the price range for the cars, this was applied by pinning the data to price ranges (cheap, moderate, and expensive) and then evaluating the results using accuracy score on each class, the results showed that SVM scored the best accuracy score with 86%.

B Hemendiran [7] in his research experimented with *Random Forest Regressor, Extra Tree Regressor, Bagging Regressor, Decision Tree, and the XG Boost* algorithms to

979-8-3315-3221-5/25/$31.00 ©2025 IEEE

predict the price of used cars in India using historical information gathered from daily news articles, magazines, and from various standard websites. *Random forest* showed the best results.

Iqbal Singh Saini [8] in his paper compared five regression techniques of machine learning (linear regression, decision tree, random forest, k nearest neighbor, and extreme gradient boost) using a dataset of car resale prices in India. The experiment began with cleaning and transforming the data then training and evaluating each model by splitting the data into 80% train and 20% test. Models are evaluated using accuracy score, the results showed Random Forest, XG Boost, and Decision tree scored the best accuracy with 90.67%, 90.35%, and 85.7% respectively, Linear regression scored 66.05% and KNN scored 49.35%.

Sameerchand Pudaruth [3] in his research discussed car price prediction of used cars using *naive Bayes algorithm*, the prediction showed considerably less accuracy.

Daniel Aprillio Budiono [9] experimented KNN to predict used cars prices in Indonesia. The experiment is done on 504 used cars data collected through web scraping. The expereriment is evaluated using R2 score, showing 98.8% and error rate of 8.3%.

TABLE I: references results comparison

Reference	Model	R2 score	Aquired Data
Abdulla Al Shared [5]	*RFR*	99%	UAE scarped used cars listings from popular websiets
	Linear regression	88%	
	Bagging regressor	85%	
Enis Gegic[6]	SVM	86%	Bosnia and Herzegovina used cars prices scrapped dataset
	ANN	73%	
Iqbal Singh Saini [8]	Random Forest	90.67%	Car resale prices in India
	XG Boost	90.35%	
	Decision Tree	85.7%	
	Linear regression	66.05%	
	KNN	49.35%	
Daniel Aprillio Budiono [9]	KNN	98.8%	Indonesea used cars data collected through web scraping

As the related work shown above in table 1, authors proposed various prediction models for the pricing problem, the results showed a good chance for the tree-based models (Decision tree,RFR , XG Boost) to result in better results. However simple models like Naïve Bayes, Linear regression, and KNN didn't perform well.

For evaluation, the related research evaluated the problem in 2 perspectives: classification and regression. Classification evaluations were mainly using accuracy and regression metrics were mainly using R2 score.

III. Methodology

This section is focused on data exploration, preparation and modeling. This will be done using various data exploration and visualization techniques in addition to applying data cleaning and manipulation techniques then preparing the data for modeling by applying transformations and features engineering and finally experimenting and evaluating candidate machine learning algorithms on data.

A. Data preparation

The experiment is using scrapped data collected from the popular used cars listings websites in Egypt from the period January 2022 to February 2024, raw data were 895k rows.

TABLE II: dataset columns details

column	Datatype	Summary
make	categorical	Car manufacturer,40 unique values
model	categorical	car brand, 385 unique values
transmission type	categorical	2 unique values (manual-automatic)
year	numerical	From 2010 to 2024
kilometers	numerical	From 0 to 300K
price	numerical	From 0 to 10B

B. Checking duplicates

EDA showed that people post their ads on multiple sites to increase their chances for the car to be sold. The first step was dropping the duplicates which ads posted 2 or more times at the same time, 500k rows were detected as duplicates as the same car has been posted on more than 1 site. After removing duplicates, the data points became 390k rows.

C. Checking unwanted rare makes and models

This stage experimented with the 2 main categorical variables (make and model) to calculate the count of unique values to detect and remove rare models and makes.

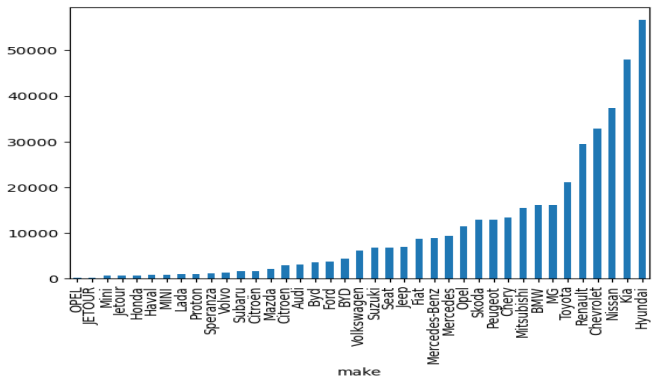

Fig. 1: unique makes counts.

- Figure 1 shows the counts of each make in the dataset we can see that the most common makes are those with counts > 10K otherwise are rare makes. First filtration step we will remove makes with counts < 10K.

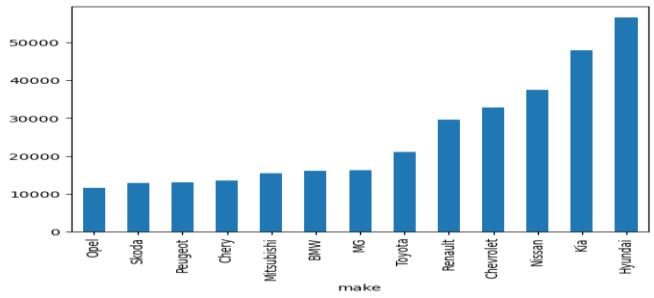

Fig. 2: top unique makes count after removing rare makes.

- Next step we will plot the bar plot for the counts of unique models.

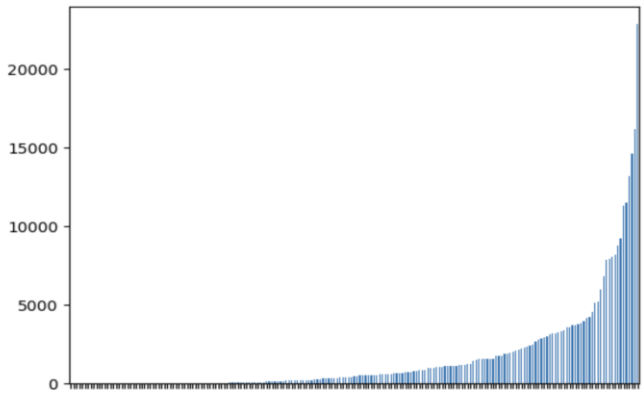

Fig. 3 distribution of unique models counts.

- Figure 3 clearly shows that most of the data is distributed in models with counts 3K or more.

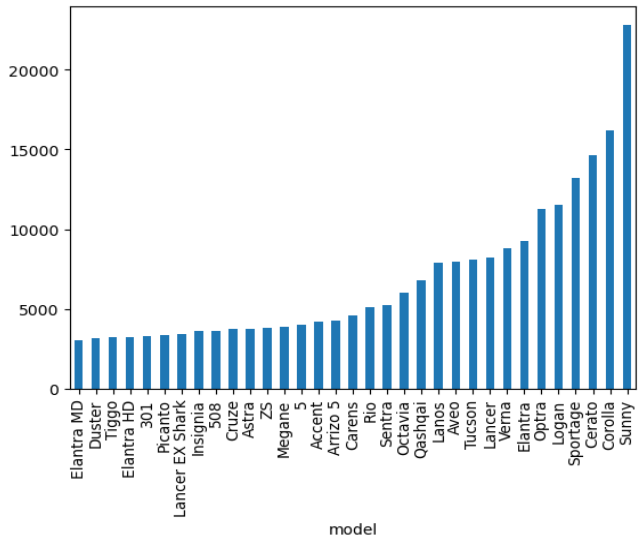

Fig. 4: Models with count >3K.

- After filtering rare makes and models we got 12 unique makes, 33 unique models, and 225 rows.

D. Checking outliers in Kilometers

This section will explore Kilometers distribution to remove outliers.

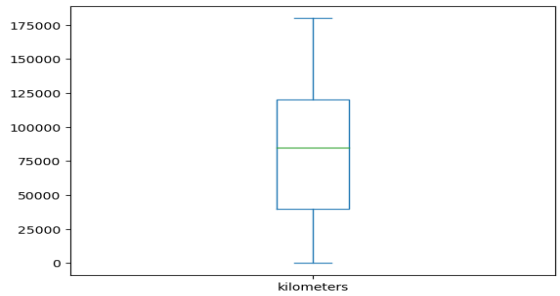

Fig. 5: Distribution of kilometers.

Figure 5 shows the distributions for kilometers, it shows that the data is well distributed in an acceptable range between 0 and 175K kilometers.

E. Checking outliers in price

Next step we will plot the distribution of price using a boxplot. Figure 6-Plot A shows the original distribution of the price feature, which shows extreme outliers, first step prices of more than 2 million will be removed from the dataset to get better visualization for the price.

Figure 6-Plot B shows the price distribution after removing extreme prices of more than 2 million, this shows a better distribution for the data. However, data needs to be filtered again using IQR filter to remove the remaining outliers.

Figure 6-Plot C shows the results after applying IQR filter.

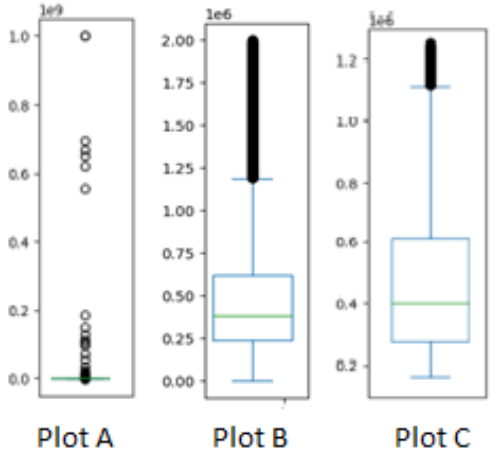

Fig. 6: Price distribution

F. Transforming data for modeling

This section will discuss suggested features engineering methods and transformations that will be applied to the data to prepare it for the model.

For year feature transformation, a simple normalization technique will be applied to convert a large range (2010-2024) to a smaller range, a new feature 'years passed' will be calculated by the equation "years passed = 2024 – year", this will change the years passed to range from (0-24).

For the Kilometers feature, the data is distributed between 20K-140K. For a large range of numerical features, it's recommended to normalize the data to mitigate the influence of varying scales [10]. In this case, the min-max normalizer is applied to the kilometers feature.

For the timestamp 'created at' feature, Wei-Han Lee showed a technique for transforming timestamps to a numeric feature that can be used for data transformation using a sliding window [11], in this case, the time window will be each month without intersection. Therefore, a new feature will be created using the timestamp feature by calculating the months passed from the beginning of the time series.

For the 'transmission type' feature, there are 2 cases for transmission (Automatic and manual) a binary encoding will be applied to map manual to 0 and automatic to 1.

For categorical features (make and model) one hot encoding will be applied to the features.

Enis Gegic[6] in his study introduced an interesting technique to transform the data to 3 categories (cheap , moderate and expensive) price. The process was pinning the data to 3 pins, in this case Plot C in Figure 6 will be used as reference for data pinning. Pins will be: category_1 'cheap cars' price< the 400K (data median), moderate prices category_2 will be between 400K and 600K (between 50th and 75th quantile), expensive cars category_3 will be cars > 600K (price > 75th quantile). Table 3 shows data distribution after pinning.

TABLE III price pins distribution

Price category	count
Category_1 'cheap' (price<400K)	72337
Category_2 'moderate' (400K>price<600K)	33845
Category_3 'expensive' (price>600K)	38219

G. Model Development

The research will experiment with RFR, XG Boost, and Bagging regressor on the data. First part of the experiment is to find the best fit for each model using grid search.

- Random forest regressor [12] is a machine learning algorithm that is constructed by using a collection of decision trees based on the training data. Instead of taking the target value from a single tree, the Random Forest algorithm makes a prediction using the bagging technique which uses the average prediction of a collection of trees. The decision trees themselves are constructed by fitting randomly drawn sample groups of rows and columns in the training data. This method is called bagging, and results in a reduction of bias as each tree is built on different parts of the input at random. The method of averaging the predictions of decision trees reduces the overfitting that can occur when using single decision trees.

- Extreme gradient boosting (XGBoost) [13]is a machine learning algorithm that is constructed by using an enhanced gradient boosting machine using the tree ensemble boosting process. The decision trees are constructed by fitting randomly drawn sample groups of rows and columns in the training data. This process ends in the sum of the outputs from all the trees.

- Bagging Regressor [14] is a "bootstrap"' ensemble method that creates individuals for its ensemble by training each regression model on a random redistribution of the training set. Each regression model's training dataset is generated by randomly drawing, with replacement, N examples - where N is the size of the original set; many of the original examples may be repeated in the resulting training set while others may be left out. After the construction of several regression models, averaging the predictions of each regression model performs the final prediction.

H. Model Training and optimization

Training process will use grid search cross validation to find the best hyper-parameters for each model, data is divided for 80/20 train test split for each price category (cheap, moderate and expensive) and an alternative approach will training on all data without splitting. The grid search will run and evaluate each candidate model on 5 folds cross validation to find the best hyperparameters combination. Table 4 shows candidate hyperparameters for each model, best parameters after grid search and the number of fits for each experiment and table 5 ,6 shows grid search results.

TABLE IV shows hyper-parameters selection for each price category

ML Model	Hyper parameters	Best parameters (cheap prices)	Best parameters (Moderate)	Best parameters (Expensive)
RFR	n_estimators: [20,50,100] min_samples_ split:[0.5,1.0,2.0]	n_estimators:100 Bootstrap:True min_samples_ split:8	n_estimators: 100 Bootstrap:True min_samples_ split:8	n_estimators:100 Bootstrap:True min_samples_split:8
Bagging regressor	n_estimators: [20,50,100] max_samples: [0.5,1.0,2.0] bootstrap: [True,False]	n_estimators:100 max_samples:0.5 bootstrap:True	n_estimators: 100 max_samples: 0.5 bootstrap: True	n_estimators:100 max_samples:0.5 bootstrap:True
XG Boost	learning_rate: [0.05,0.1,0.15] max_depth:[5,6,8] gamma: [0.0,0.1,0.2] min_child_weight: [3,5,7]	Learning_rate:0.1 Max_depth:8 Min_child_weight:7	Learning_rate: 0.15 Max_depth:6 Min_child_weight:1	Learning_rate:0.15 Max_depth:8 Min_child_weight:5

TABLE V shows hyper-parameters selection without price pinning

ML Model	Hyper parameters	#fits	Best parameters
RFR	n_estimators:[20,50,100] min_samples_split:[0.5,1.0,2.0] bootstrap:[True,False]	90	n_estimators:100 Bootstrap:True min_samples_split:8
Bagging regressor	n_estimators:[20,50,100] max_samples:[0.5,1.0,2.0] bootstrap:[True,False]	90	n_estimators:100 max_samples:0.5 bootstrap:True
XGBoost	learning_rate:[0.05,0.1,0.15] max_depth:[5,6,8] gamma:[0.0,0.1,0.2] min_child_weight:[3,5,7]	135	Learning_weight:0.15 Max_depth:8 Min_child_weight:5

IV. Model Evaluation

Model evaluation is done on the best estimator for each model and the results will be evaluated using mean absolute error and R2 score.

MAE is the measure of errors between paired observations expressing the same phenomenon. MAE is calculated as the mean of the absolute difference between prediction and ground truth which is ad posted price.

$$MAE = \frac{\sum_{n=1}^{N} \mid \hat{r}_n - r_n}{N}$$

R2 score [16] is a measure that provides information about the goodness of fit of the regression model, it is a statistical measure that tells how well the plotted regression line fits the actual data. The coefficient of determination is defined as the sum of squares due to the regression divided by the sum of total squares. Usually, R2 is interpreted as representing the percentage of variation in the dependent variable explained by variation in the independent variables.

First part is calculating the sum of squares due to regression (SSR)

$$SSR = \sum_{i=1}^{n} (yi - \hat{yi})^2$$

Second part is calculating the sum of total squares (SST)

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

R2 score is calculated as the 1 – (SSR/SST)

$$\text{R-Squared } = 1 - \left(\frac{SSR}{SST}\right)$$

Table 7 shows the results for each model on 3 price categories, it can clearly shows that XGBoost outperforms other models on all price categories.

TABLE VI: shows the results for each regressor on each price category

Model	MAE cheap	R2 score cheap	MAE Moderate.	R2 score Moderate.	MAE expensive	R2 score Expensive
Bagging regressor	32408	0.79	36605	0.68	78278	0.75
XG Boost	**31409**	**0.82**	**35707**	**0.75**	**75319**	**0.8**
RFR	32307	0.81	36291	0.7	77891	0.78

Table 8 benchmark applying pinning vs not applying pinning on data by comparing avg MAE of each model on all price categories vs MAE when evaluating data without pinning. The result shows that XGBoost gave the best results on both experiments and pining helped to reduce MAE with about 23%.

TABLE VII: shows the avg MAE results

Model	MAE (on all data without pinning)	MAE (on mean of 3 models on 3 categories results)
Bagging regressor	64400	49097
XG Boost	61680	47478
RFR	64100	48829

V. Conclusion and Future Work

Accurately valuating used cars in Egypt is a challenging task for such a rapidly changing market as it relies on many subjective and economic factors. Machine learning algorithms showed promising results in such a task. The results showed that using machine learning algorithms got promising results in used cars valuation. XG Boost is slightly better than Bagging Regressor and RFR. Pinning price ranges showed a tangible enhancement to reduce MAE. Future works may include pinning the data based on kilometrage ranges or training a separate model on each make or model to reveal the implicit nature of the data.

References

[1] Manheim, "Used Vehicle Value Index, 2021." [Online]. Available: https://site.manheim.com/en/services/consulting/used-vehicle-value-index.html

[2] Huaxia, "Used car market flourishes in Egypt amid hiking prices of new cars," 2023. [Online]. Available: https://english.news.cn/20230918/86a8ff153e324af284f53b25509e2ea9/c.html

[3] R. Siva and M. Adimoolam, "Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm," ECS Transactions, vol. 107, pp. 12953-12964, 2022.

[4] A. S. Pillai, "A Deep Learning Approach for Used Car Price Prediction," J. Sci. Tech., vol. 3, no. 3, pp. 31–50, Jun. 2022.

[5] A. AlShared, "Used Cars Price Prediction and Valuation using Data Mining Techniques," 2021. [Online]. Available: https://repository.rit.edu/theses/11086/

[6] E. Gegic, B. Isakovic, D. Kečo, Z. Mašetić, and J. Kevric, "Car Price Prediction using Machine Learning Techniques," TEM Journal, vol. 8, pp. 113-118, 2019, doi: 10.18421/TEM81-16.

[7] B. Hemendiran and P. N. Renjith, "Predicting the Prices of the Used Cars using Machine Learning for Resale," IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023, pp. 1-5, doi:10.1109/SCEECS57921.2023.10063133.

[8] I. S. Saini and N. Kaur, "Comparison of Various Regression Techniques an International Journal of Computer and Information System (IJCIS), vol. 5, No. 1, 2024. doi: https://doi.org/10.29040/ijcis.v5i1.147

[9] M. Lones, "How to avoid machine learning pitfalls: a guide for academic researchers," 2021, arXiv:2108.02497. [Online]. Available: https://arxiv.org/abs/2108.02497.

[10] W.H. Lee, J. Ortiz, B. Ko, and R. Lee, "Time Series Segmentation through Automatic Feature Learning," 2018, arXiv:1801.05394. [Online]. Available: https://arxiv.org/abs/1801.05394.

[11] G. Biau, "Analysis of a Random Forests Model," Journal of Machine Learning Research, vol. 13, pp. 1063-1095, 2010.

[12] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[13] K. Samruddhi and R. Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," International Journal of Innovative Research in Applied Sciences and Engineering, vol. 4, pp. 629-632, 2020.

[14] D. Figueiredo, S. Júnior, and E. Rocha, "What is R² all about?," Leviathan-Cadernos de Pesquisa Política, vol. 3, pp. 60-68, 2011, doi: 10.11606/issn.2237-4485.lev.2011.13228.