Used Car Price Prediction using Machine Learning: A Case Study

Mustapha Hankar
LAROSERI Lab. Department of computer science
Faculty of sciences, Chouaib Doukkali University
El Jadida, Morocco
hankar.mustapha@gmail.com

Marouane Birjali
LAROSERI Lab. Department of computer science
Faculty of sciences, Chouaib Doukkali University
El Jadida, Morocco
birjali.marouane@gmail.com

Abderrahim Beni-Hssane
LAROSERI Lab. Department of computer science
Faculty of sciences, Chouaib Doukkali University
El Jadida, Morocco
abenihssane@yahoo.fr

2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC) | 978-1-6654-8724-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIVC54825.2022.9800719

***Abstract*— In many business fields that are related to statistics and machine learning (ML), multiple linear regression (MLR) models are often used to estimate and fit a linear relationship between a continuous response variable and other explanatory variables. In our case study, we applied several regression techniques based on supervised machine learning to predict the resale price of used cars given many factors such as mileage, fuel type, fiscal power, mark, model, and the production year of the car. In all tested models, gradient boosting regressor showed a high R-squared score and low root mean square error.**

Keywords— regression analysis, prediction, estimation, Avito, machine learning, log transformation, used car price, regression assumptions.

I. INTRODUCTION

The market of new car sales has been growing in recent years very fast in Morocco due to the big investments allocated by the government for car industries planted in the north of the kingdom. However, many customers choose to buy a used car with an affordable price instead of engaging in bank loans to purchase a new car. Several ecommerce web platforms offer their third-party services between sellers and buyers in the market of used cars such as **Moteur**[1] and **Avito**[2]. For many buyers and sellers, it is very helpful to have a prior knowledge about the reasonable pricing value of a car before making any decision of buying or selling it. In order to help customers to buy a used car to estimate the pricing value of the used car they are willing to buy or sell, we build a machine learning model to predict the car price based on many features such as mileage, year of production, fuel type, fiscal power, car mark, and its model.

In this paper, we applied several supervised machine learning algorithms to predict used car prices based on the car features cited above. The dataset used to train and test the regression models is collected from an online ecommerce website called **Avito**. In all tested models, gradient boosting regressor (GBR) showed a high R-squared score and a low root mean squared error (RMSE).

The remainder of this paper is organized as follows. In the second section, we introduce other related work to our case study. The third section is dedicated to describe data collection and preprocessing tasks. In section 4, exploratory data analysis and feature selection are performed. In section 5, we suggest many machine algorithms to train and evaluate on data. The obtained results are presented and discussed in section 6. Finally, our work is concluded in section 7.

II. RELATED WORK

Regression analysis methods are widely applied in many fields like business intelligence, environmental modeling, and financial forecasting. Peerun et al. [1] applied artificial neural networks to predict second-hand car prices. The dataset provided for this study contains 200 cars from different sources that were gathered and fed to four different machine learning algorithms. In [2], Listiani applied support vector machines (SVM) to estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression.

Wu et al. [3] used a neuro-fuzzy knowledge-based system to predict the price of used cars. Only three factors used: the brand of the car, the year in which it was manufactured, and the engine style. Pudaruth et al. [4] used supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions of all trained models (multiple linear regression analysis, k-nearest neighbors, and naïve Bayes and decision trees, etc.) are evaluated and compared in order to find those which provide the best performances.

Liang Han et al. [5] suggested a regression model to predict the prices of online second-hand items. The system predicts the price of an item based on its uploaded images and text descriptions with other statistical features collected from shopping platforms. N. Kanwal and J. Sadaqat [6] used a multiple linear regression model to predict vehicle car price. They used a feature selection method to find the most relevant variables. The used dataset contains only selected variables to form the linear regression model. The result was impressive with a R^2 score of 98%.

III. DATA COLLECTION AND PREPROCESING

The dataset provided for this study was collected from **Avito,** a local online ecommerce website**.** To collect data, we coded a python script to crawl all offered-to-sell used car links in the website using BeatifulSoup library. After then, we loop over the retrieved links to extract the most important

[1] https://www.moteur.ma/
[2] https://www.avito.ma/

978-1-6654-8724-5/22/$31.00 ©2022 IEEE

features of used cars in the platform such as mileage, fuel type, production year, mark, model, and fiscal power. The raw collected data is then cleaned from null values and outliers. Some missing values in columns like mileage and fiscal power are imputed by their mean to avoid information loss. The final preprocessed dataset contains 8000 values about car features.

IV. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a crucial step to solve every data science problem. In this phase, we intend to explore the main characteristics about data features using statistical methods such as mean, standard deviation, distribution, etc., and visualization charts like bar charts, histograms, scatter plots, etc. For example, the pair plot in Fig. 2 shows that mileage, production year, and fiscal power are linearly related to the prices of the car.

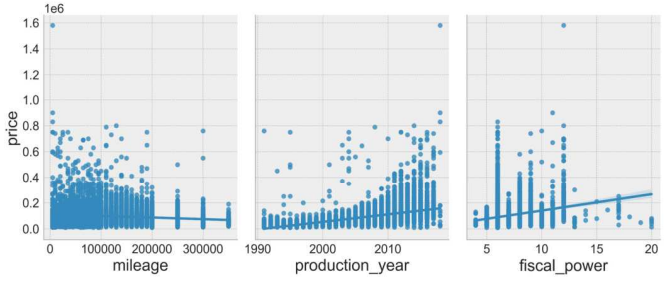

Fig. 1. Pair plot between numerical features and the price

The heatmap matrix in Fig. 3 shows a presence of correlations between the numerical features and the prices, and absence of collinearity among them. But we can notice that mileage and production year are slightly correlated with each other.

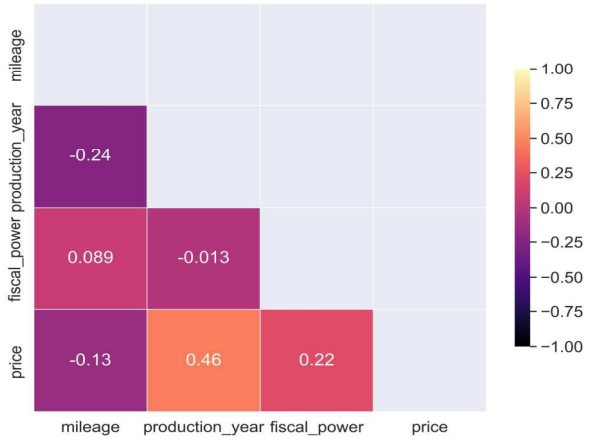

Fig. 2. Correlation matrix of numerical features

After we perform an EDA, a feature selection task is carried out using recursive feature elimination (RFE) method to only maintain the most relevant features to car prices. The selected features sorted by importance are: year of manufacture, mileage, mark, fuel type, fiscal power, and model. Features like city, type, and sector are removed because they are irrelevant to the car price. Label encoding of categorical features (fuel type, mark, model) and standard feature scaling are performed before training and testing machine learning models.

V. METHODOLOGY

In a simple linear regression problem, we investigate the relationship between one independent variable and another one dependent variable. Multiple linear regression models are used to estimate a linear (or nonlinear) relationship between multiple input variables and an output variable [7]. The general form of multiple linear regression is given by the equation (1) below:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \qquad (1)$$

where y is dependent variable and $\beta_0, \beta_1, \beta_3, \ldots, \beta_p$ are regression coefficients and $x_1, x_2, x_3, \ldots, x_p$ are independent variables and the error term ε.

To compare a baseline multiple regression model with other regression models, we have trained four other regressors in the validation phase. Namely, K-nearest neighbors regressor (KNN), random forest regressor (RFR), gradient bosting regressor (GBR), and artificial neural network (ANN) based regressor. The models are trained on a portion of 80% of the whole dataset, and the remaining 20% is left for testing the models. After training, regression models are evaluated using two different metrics. The first one is the coefficient of determination score given in the equation (2) which measures the rate of variance in the target variable that is explained by the predictor variables [8]. The second is a loss function method given in the equation (3) measures the root mean squared errors between the observed values and the predicted values [9].

$$R^2 = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad (2)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}} \qquad (3)$$

where y_i is the observed value, $\hat{y}_i$ are predicted value, $\bar{y}$ is the mean value of all observations y_i, and N is the number of samples. The best model is to be selected in the validation phase, and the regressor is tested on unseen data to examine its performance before deployment. The overall process of our work is illustrated in Fig. 3

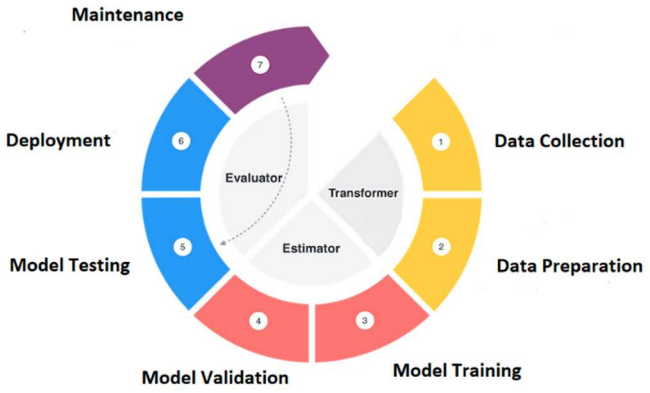

Fig. 3. Machine learning life cycle

The table I below shows all tested models and their tuned parameters. We mention that all parameters were selected using grid search optimization algorithm in order to fit the best regression models having the highest R-squared score and the lowest RMSE error.

TABLE I. FINE-TUNED PARAMETERS FOR REGRESSION MODELS

Model	Parameters
MLR	*{regression_coefficients:* {production_year_coeff = 0.395, mileage_coeff = -0.004, fuel_type_coeff = -0.213, mark_coeff = 0.059, model_coeff = -0.013, fiscal_power_coeff = 0.129}, *intercept* = 96608.59}
KNN	{k = 10, weights = 'uniform'}
RFR	{n_estimators = 100, criterion = 'mse', max_depth = 50}
GBR	{loss ='ls', max_depth = 6}
ANN	{neurons_per_layer = (32,16,8), activation_function ='relu', loss ='mse', optimizer='adam'}

VI. REGRESSION ANALYSIS ASSUMPTIONS

The effectiveness of such regression analysis results always relies on various conditions related to the nature of data features we feed to the model. Therefore, it is very essential to pre-check that many regression assumptions are not violated before making any decision about the usefulness of a regression model. Most common regression assumptions which are not robust to violation are [10, 11]:

i. No multicollinearity: predictor variables are not supposed to be highly correlated between each other.
ii. Linearity: there must exist a linear relationship between predictor variables and target variable.
iii. Normality: the residuals of the model follow a normal distribution.
iv. Homoscedasticity: this assumption assumes that errors are constant across the values of independent variables.

VII. RESULSTS AND DISCUSSION

To evaluate ML models, we used R^2 score to specify the proportion of variation in the car prices explained by predictor features (mentioned in section 1), and RMSE errors to obtain the residuals between the predicted price values (y-hat) by the model and the observed price values (y). The range of R^2 score values vary between 0 and 1, and higher values of R^2 score implies that the model is a best fit. Low value of RMSE error means that the model is highly predictive. In Table II, we present the evaluation results for all tested regressors.

TABLE II. REGRESSION MODELS EVALUATION RESULTS

Model	R^2 score	RMSE
MLR	0.57	63933.52
KNNR	0.70	51224.96
RFR	0.74	44939.79
GBR	0.80	44516.20
ANN	0.67	54957.98

From the results above, we notice that Gradient Boosting Regressor outperformed other tested models and approximately reached 0.80 of R^2 score. This means that the model explains 80% of the variation in car price values. Although, this is not a great R^2 score to evaluate the model as a good fit. We also observe that GBR model recorded a low value on minimizing RMSE errors. The results still remain with no significance if regression assumptions (mentioned in section 5) are not validated. Therefore, these assumptions must be checked before making any statistical interpretations or decisions about the usefulness of GBR model. The first assumption to check is the absence or lack of multicollinearity between independent variables. The correlation matrix in Fig .2 showed that there are no correlations among predictor variables expect a weak negative correlation between milage and production year (-0.13).

The second assumption verifies the existence of linearity between the response variable and the predictor variables. The scatter plots from Fig. 3 showed that mileage, fiscal power, and production year have a linear relationship with car prices.

In the third assumption, we check that the normality assumption is not violated for GBR model. The distribution plot from Fig. 4 shows that GBR residuals are normally distributed across the x-axis.

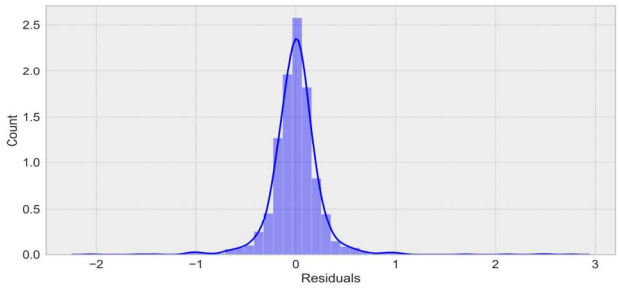

Fig. 4. GBR residuals distribution

The fourth assumption to check is homoscedasticity. This means that model residuals must have a constant variance for each value of the predictors. When homoscedasticity is violated (presence of heteroscedasticity), the results of the regression analysis become biased and hard to trust. Specifically, heteroscedasticity increases the variance of the regression coefficient estimates. In Fig. 5, we plot the residual values versus fitted values of GBR model, and we notice that error terms are constantly spread across the values of independent variables. As a result, error terms remain unchanged with the value of the response variable.

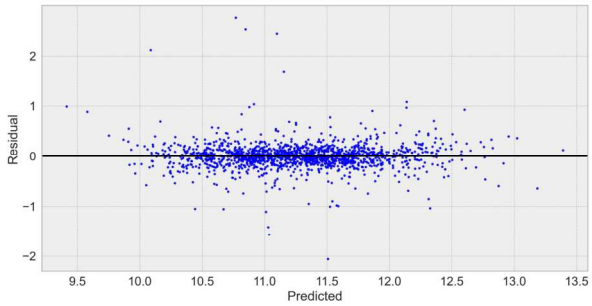

Fig. 5. Predicted prices vs residuals of GBR model

While the prices of the car are not normally distributed and highly right skewed, we applied a natural log transformation on the price variable before training the models. The result is an obvious increase of R^2 score and RMSE errors are decayed even more. We also mention that the residuals, after the log transformation, have become more normally distributed with a zero-mean value.

Finally, we plot the real prices versus the GBR predicted prices in Fig .6 to visualize the comparison between the real prices and the model predictions. The chart illustrates that GBR model has effectively learned a linear relationship between input variables and the target variable from the given dataset.

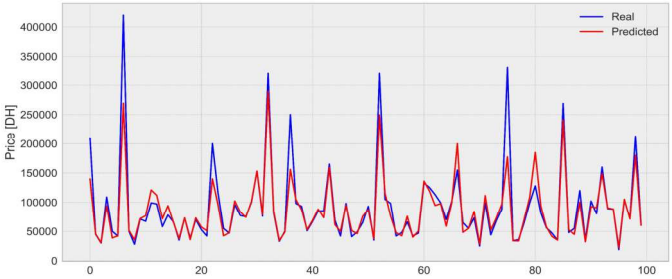

Fig. 6. Real vs predicted prices of GBR model

VIII. Conclusion

In this paper, we proposed a regression model to predict the resale value of used cars in Morocco to help buyers and sellers of used cars to pre-estimate its pricing value. The results showed that gradient boosting regressor outperformed all tested models with a highest R^2 score and a minimized root mean squared error. As a future work, we intend to increase the performance of the model by scaling the training data and adding more other variables to the feature set.

Acknowledgment

Authors of this paper express their sincere gratitude to the research team, and all colleagues who contributed to make this work done. We also thank our supervisor Pr. Beni-Hssane Abderrahim and the administration staff of the Faculty of Sciences in El Jadida for their corporation and help.

References

[1] S. Peerun, N. H. Chummun, and S. Pudaruth, “Predicting the Price of Second-hand Cars using Artificial Neural Networks,” The Second International Conference on Data Mining, Internet Computing, and Big Data, no. August, 2015, pp. 17–21.

[2] LISTIANI, M, “Support Vector Regression Analysis for Price Prediction in a Car Leasing Application, ” Thesis (MSc), Hamburg University of Technology, 2009.

[3] WU, J. D., HSU, C. C. AND CHEN, H. C, “An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference,” Expert Systems with Applications. Vol, 36, Issue 4, pp. 7809-7817, 2009.

[4] Pudaruth, Sameerchand, “Predicting the Price of Used Cars using Machine Learning Techniques,” International Journal of Information & Computation Technology. Vol. 4, no. 7, pp. 753-764, 2014.

[5] Liang Han, Zhaozheng Yin, Zhurong Xia, Minqian Tang, and Rong Jin, “Price Suggestion for Online Second-hand Items with Texts and Images,” The 28th ACM International Conference on Multimedia (MM '20)., New York, USA, october 2020, pp. 2784–2792.

[6] N. Kanwal and J. Sadaqat, “Vehicle Price Prediction System using Machine Learning Techniques,” International Journal of Computer Applications, vol. 167, no. 9, pp. 27–31, 2017.

[7] Yan, Xin, Linear Regression Analysis: Theory and Computing, World Scientific, 2009.

[8] Zhang, Dabao, “A Coefficient of Determination for Generalized Linear Models,” The American Statistician, vol. 74, no. 4, pp. 310-316, 2016.

[9] Hyndman, Rob J.; Koehler, Anne B "Another look at measures of forecast accuracy". International Journal of Forecasting, vol. 22, no. 4, pp. 679–688, 2006.

[10] Osborne, Jason W. and Waters, Elaine, "Four assumptions of multiple regression that researchers should always test," Practical Assessment, Research, and Evaluation, vol. 8, no. 2, 2002.

[11] Williams, Matt N., Grajales, Carlos Alberto Gómez, & Kurkiewicz, Dason, “Assumptions of Multiple Regression: Correcting Two Misconceptions,” Practical Assessment, Research & Evaluation, vol. 18, no. 11, 2013.