PROJECT REPORT DATA SCIENCE FALL 2023

Price Prediction and Worth Determination of Used Cars in Pakistan

Abdullah Arif and Tahmooras Khan

Department of Computer Science, National University of Computer and Emerging Sciences, Lahore

December 3, 2023

Abstract

The rising costs of cars and the unpredictability of the Pakistani market pose difficulties for individual car owners. Our research endeavors to devise a comprehensive resolution to these problems through the utilization of a machine learning model supported by our data-driven understanding. By utilizing historical data and previous trends, the research attempts to forecast used car values, offering a useful resource to people navigating this intricate market environment. The initiative offers precise information for well-informed decision making by addressing the typical market problems of underselling, overselling, and other fraudulent actions. The main goals of this research are to design a hybrid system and demonstrate the price density of more sensible options.

The methodology used includes important phases such as feature engineering, data pretreatment, algorithm selection, and tweaks. The information itself serves as a good representation of the intricate and varied characteristics of the Pakistani car resale market. We tested and implemented a number of algorithms, including XGBoost Regressor, Huber Regression, Gradient Boosting Regressor, Decision Tree, and Linear Regression. XGBoost Regressor was shown to be the most accurate through an iterative process at handling the complexities present in the dataset, reading patterns, and adapting to variations in the data.

The result indicates the model's exceptional ability to forecast resale values. The minimal discrepancy between the actual and expected prices highlights the model's readability when the metrics are examined. Because of its capacity to recognize patterns, the model was an accurate and successful approach to predict automobile costs.

Keywords: Analysis, Research, Machine Learning, Random Forest, XGBoost, Decision Tree, Linear Regression are some of the terms used in this paper.

1 Introduction

Car costs in Pakistan are on the rise, and both purchasers and unsuspecting sellers face difficulties in the intricate resale market. In addition, clients face difficult hurdles from the prevalent market faults of overselling, overestimating, and fake marketplaces.

Our goal is to use a model that makes use of our car's data understanding without providing a one-stop shop. With the use of past data, this prototype aims to interpret car worth trends in order to determine a vehicle's potential resale value. Using a dataset that includes a variety of car parameters pertinent to the Pakistani automotive industry, this work tackles the problem of creating an accurate prediction model..

The introduction of the internet has led to a notable increase in the used-car industry in Pakistan in recent years. Buyers, sellers, and individual investors can all profit from understanding and predicting the changes in motorcar prices. In order to build a strong predictive architecture, this model makes use of a variety of datasets that are indicative of the used automobile industry in Pakistan. The worth of the automobile is predicted by a complex study of a number of variables, such as the car's attributes, location, kind of car, and past pricing patterns.

Car price forecasting requires a distinct technique due to Pakistan's unique socioeconomic and geographic features. By attempting to close the current gap and creating a thorough automobile price prediction model based on the Pakistani internet marketplace, this study tackles the difficulties.

2 Methodology

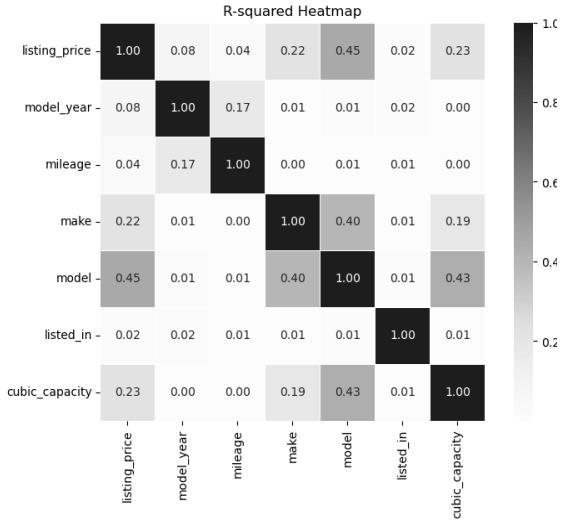
Existing literature underlines the necessity of using machine learning approaches to anticipate used automobile prices. The purpose of this literature study is to investigate present approaches and improvements in car pricing prediction models.

Traditional regression models were used in certain automobile price prediction techniques. S. Lessmann's (2011) research revealed a correlation between automobile pricing and key predictor factors such as car manufacturer, model, and year of release [1].

Scientists began combining sophisticated machine learning model strategies to improve prediction accuracy with the introduction of machine learning (ML) and related research. Regression with XGBoost has shown to be a reliable method that works well with both continuous and discrete/categorical data. The effectiveness of XGBoost in capturing intricate non-linear correlations was demonstrated in the works of K. Gopi and K. Subarna (2023) [2].

There were several crucial phases in the process. Preprocessing of the data includes managing outliers, standardization, and missing values. To improve the efficiency of the model, feature selection approaches were employed in conjunction with feature engineering to extract pertinent information. According to Uçar, M. K. & Nour (2020) [4], the data set was divided into training and testing sets, of which we utilized 80% and 20%, respectively.

Algorithms like Linear Regression and XGBoost Regressor were used for training. We also examined a few more algorithms: Hubers Regression, Decision Tree Regression and Gradient Boosting.



3 Experiments

A number of different algorithms must be carefully considered when implementing machine learning to estimate auto prices, with XGBoost finally showing to be the most accurate when covering the most data. Because automobile pricing datasets are varied, XG Boost is well-suited to handle missing values. The model was able to recognize patterns in the data thanks to its regularization mechanism and ensemble learning technique. The framework operated robustly thanks to the auto pruning mechanism.

Because of its sensitivity to proximity and difficulty with mixed-type variables, Linear Regression was disqualified from consideration along with the other models. This also applied to Random Forest. Due to the difficulties in managing categorical variables like {transmission}, Random Forest was rejected, and HuberRegressor was deemed superfluous for the dataset that did not present any major problems. It took an average of 1.9 seconds to calculate 70,000 items every iteration, which caused problems with bias adjustment and fine-tuning as well as time complexity disruption.

The chosen XGBoost model, which was trained using the unique characteristics of vehicle price data, is a dependable and accurate solution for predicting the car's resale value. It is memory efficient due to cache optimization and out of memory computation. Regularization's outstanding feature protects the model from the curse of overfitting and keeps the model's variance under control. Because it is a decision tree at its foundation, auto pruning allows our model to be robust and successful.

```
R-squared: 0.9712
Mean Absolute Error (MAE): 276760.32
Root Mean Squared Error (RMSE): 644123.44
Mean Squared Error (MSE): 414895003152.55
```

4 Results & Discussion

After a fun process of fine-tuning and modifications, it shows impressive accuracy in automobile price prediction.

The smallest variation between expected and actual automobile pricing is shown by the "RMSE" and "MAE," which appeared as 6.4 e5 and 2.7 e5, respectively. These mistakes also highlight how exact, precise, and dependable the model is in making predictions. At 0.97, the coefficient of determination was determined. The efficacy of the model is demonstrated by the fact that it can account for a sizable percentage of the variation in the data. In our instance, the model—which is notably quite precise—explains 97% of the data.

The high correlation coefficient (R) of 0.98 signi1es a robust linear relationship between predicted value and actual price. Also considering the same relation, it is strong and positive. Overall, the results highlight the success of the implemented ML model in providing accurate and reliable prediction of used cars price.

5 Conclusion & Future Work

In a nutshell, by deploying a strong predictive model, this project successfully navigated the constraints of the dynamic and uncertain Pakistani automobile resale market, with XGBoost emerging as the ideal method.

The complete methodology ensures the model's strength by including data pretreatment and algorithm selection. Rejecting specific models due to their limitations reduces the dependability of the chosen algorithm.

Future work might improve accuracy and relevance by continuously refining the prediction model, including real-time data, and collaborating with stakeholders. Furthermore, many studies and projects are entirely open-source, allowing anybody to contribute and maintain them up to date. The project is also accessible to assist users in understanding market patterns and conducting more research or analysis on them.

6 References

[1] Voß, S. and Lessmann, S. (2011). Resale Price Prediction in the Used Car Market[1][1], Pg. 2-4.

[2] Gopi, K., Ramya Harika, K. B., Suvarna Jyothi, B., Suvarna, K., Venkateswarlu, B., & Chaitanya Kumar Reddy, G. (2023)

[3] Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old car price prediction with machine learning. International Research Journal of Modernization in Engineering Technology and Science, 3(3), 284-290.

[4] Uçar, M. K., Nour, M., Sindi, H., & Polat, K. (2020). The Effect of Training and Testing Process on Machine Learning.