Capstone Project: Used Cars Price Prediction | Regression (B)

Name: Juan José Astudillo

Executive Summary

This project proposes to carry out a regression model that allows predicting the price of used vehicles in India due to the great demand in the automotive market. Used vehicle data is available, collecting information such as kilometers driven, engine power, location, type of transmission, type of engine, owner number, among others. To obtain the model, different types of supervised learning techniques are approached: linear regression, Ridge and Lasso regularization, decision trees and Random Forest are explored, and through a comparison of performance metrics (R squared and Root Mean Square Error) the model chosen to be adopt is linear regression, given its high R squared in the test data, a low RMS value in both test and training data, and the simplicity of the model that does not require large computational calculations.

Problem Summary

There is a huge demand for used vehicles in the Indian market today. The sale of vehicles has decreased in addition to the manufacturing capacity is not capable of meeting the existing demand, a demand that has been increasing over time and currently exceeds the demand for new vehicles . Under this paradigm, technology companies can support the vehicle sales market in such a way as to maximize their profits.

In recent years there was a turning point, in which the demand for used vehicles exceeded the demand for new vehicles. The slowdown in the sale of new vehicles would be migrating to the purchase of used vehicles, largely due to a paradigm shift on the part of vehicle owners. has large uncertainties in both price and supply because their relationship to various factors can play a role in a car's value, so setting a selling price can be a real challenge.

Taking this into account, the price scheme of these used cars becomes important to position yourself correctly in the market.

Solution design

The use of machine learning helps our processes be more consistent and reliable, so machine learning techniques will be used as a work methodology. Since there is currently measurement of variables related to price, it is possible to classify the problem in the category of supervised learning. To obtain the price prediction model, 5 stages will be used, which are:

1. Collection of Data
2. Data Wrangling

3. Model Building
4. Model Evaluation
5. Model Deployment

In **collection of data** it is possible to supply data for the requirements of the problem from different sources, and could be in any format. CSV, XML.JSON, etc. In our case we are using an internal source in format .csv for collect data.

In **data wrangling** and data processing the main objective of this stage and focus its related to exploratory data analysis. In this stage we start understanding the given dataqset and helping up the given data set. With this we will better understand the features and the relationships between them. Then we extract the essensial variables and remove the non-essensial variables. After that we will handling missing values or human errors and identify outlier. The goal is maximize insight of dataset.

For **model building** is required to separate the data in both: training and test data. The training data is used to make sure the machine recognizes patterns of the data, the test data is used to see how well the machine can predict new answers based on its training. The train-test split procedure is used to estimate the ML performance of algorithms when they are used to make predictions on unseen data. With the training data we will train the model and fit/tune the models. In this project the split data is 30%-70% to test/train data. The model used are categorized as Regression mode, thus is the technique used for predicting continuous values. Are evaluated linear regression, Lasso/Ridge regularization, decision tree and random forest models.

Once the models are created, its possible to get a **model evaluation**, where the evaluations used to identify their performance are the coefficient of determination R squared and the Root Mean Squared Error.

Finally the **deployment** of a machine learning model require the integration of the finalized model into a production environment and get results to make business decisions. The Figure 1 presents a flow work of the mentioned stages.

Figure 1. Stages required for a machine learning model. [1]

Analysis and Key Insights

The collected data is separated into numerical variables and categorical variables, with a total of 9 numerical variables and 5 categorical variables initially. Of the numerical variables, 4 variables are identified that present biased distributions (left/right) which correspond to kilometers traveled, engine, power and price. A logarithmic transform is applied to redistribute the information and not cause problems in future learning of the model. There is a categorical variable with little representativeness (Name), for which it requires a processing that consists of a separation of characters in order to identify patterns, which is why split of data in Brand and Model was selected.

Additionally, a large number of missing values were identified, variables such as Mileage, Engine and Seats had a small number of missing values, however there are variables such as Power, New Price, Price that have a large number of missing values. Those data were identified and replaced by the median of the variable under study. This process allowed leaving a dataset ready to be trained.

To start training models, it is necessary to make some small readjustments to the dataset that allow the model to enter into greater precision. Categorical variables are transformed into multiple category variables, each category representing a feature. A total of 25 features are worked on to train the model. Additionally, the data is scaled in order to standardize its importance in the model. Finally, as already mentioned, the dataset is distributed in train/test data in a percentage of 70/30% respectively.

Four types of algorithms for the price prediction of the problem were tested and additionally they were evaluated with two performance metrics: R squared and RMSE. In the first instance, a linear regression algorithm was used, from which 6 of 25 features were identified that do not have an impact on the model results due to their high p-value. Of the remaining 19 values, 12 correspond to categorical variables and 7 numerical variables.The second algorithm used corresponds to Ridge/Lasso Regularization. Ridge Regularization behaves similarly to linear regression. When adjusting its Alpha value, there is no significant improvement with respect to linear regression. Lasso Regularization presents a low performance, so it is not a good learning mechanism for this case. The third algorithm used corresponds to the decision tree. When testing a decision tree without limiting its results, overfitting occurs. This occurs since the tree ends up creating a leaf for each piece of data. However, by adjusting the hyperparameters, the overfit can be reduced and models with fewer nodes can be obtained. For example, in Figure 2 you can see a decision tree with a length of 6 nodes. By incorporating the hyperparameters, it was possible to reduce the important variables from 25 to 13. Finally a Random Forest algorithm is proved. The algorithm was tested with and without hyperparameters. The results present an overfit in the training data, but reduced by the cross-validation of the process. There is no performance improvement from tuning the model hyperparameters.

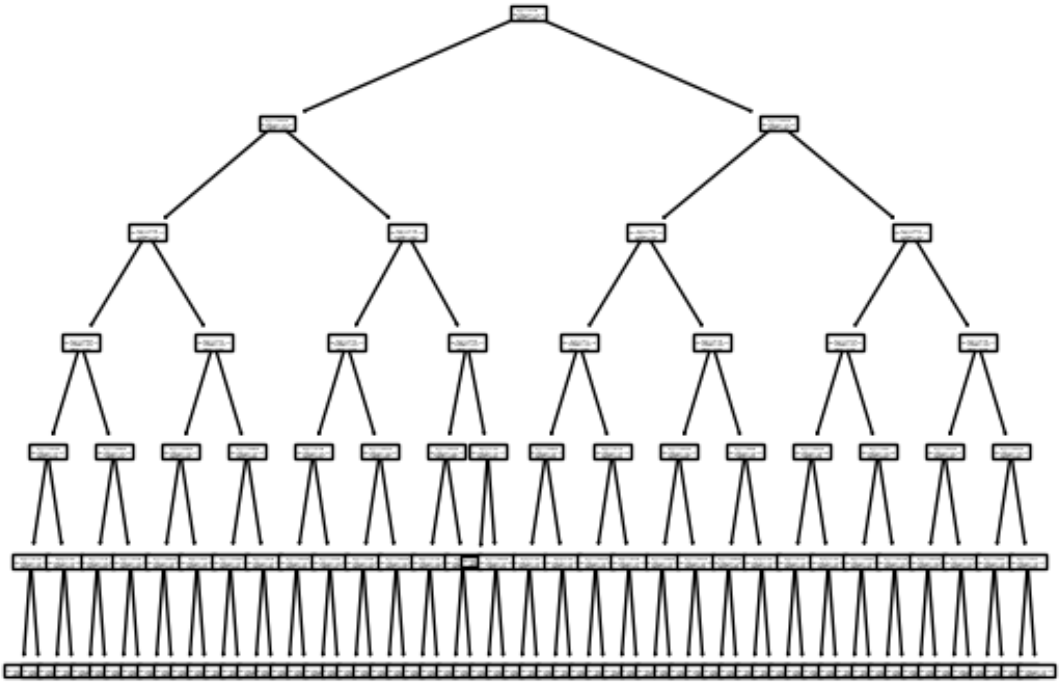

Figure 2. Decision tree with 5 stages.

Finally, the results obtained are presented in Table 1 and compared to select the model with the best performance. It can be seen that the Decision Tree and Random Forest models present an overfit in their training data, which decreases in the test data. In the opposite direction, the linear regression and Ridge Regularization present a greater bias in the training, but they are better adjusted for the test models. Most of the models with the exception of Lasso Regularization present a good R squared that would result in an optimal prediction. When comparing the RMSE, the model with the lowest RMSE corresponds to Random Forest, however, they correspond to values not far from the other models. Given the good performance obtained in the models, the **linear regression** model is chosen as the best method for price prediction, given the simplicity of its use and ease in interpreting its results.

	R squared		RMSE	
Model	Train	Test	Train	Test
Linear Regression	0.704	0.872	6.238	3.708
Decision Tree	1.000	0.865	0.009	3.810
Ridge	0.706	0.872	6.217	3.714
Lasso	0.102	0.095	10.875	9.860
Decision Tree Tuned	0.873	0.852	4.091	3.992
Random Forest	0.977	0.918	1.725	2.970
Random Forest Tunes	0.967	0.916	2.083	3.001

Table 1. Model evaluation of regression models.

Limitations and Recommendations for implementation

Some of the limitations of the model correspond to some assumptions that were not validated in the present work. It is necessary that there is no multicollinearity, this means that there is no correlation between the independent variables. Additionally, the model errors must present a normal distribution and there have to be homoskedasticity of the error terms.

Confirmation of these assumptions will allow the linear regression model to be used as the chosen model, however it is important to take into account that it will not predict correctly for those data that do not present a linear relationship, as could be the case of outlier. Precisely along the same lines, linear regression models are sensitive to outliers, so if there were no preprocessing of the data, our model would lose reliability. It will also be underadjusted compared to other models with more robust mechanisms, however, with a good feature engineer and EDA it was possible to obtain a competitive linear model in its performance.

Finalmente, para una correcta implementation para una empresa tecnológica, es necesario implementar el modelo en una nube, que permita la recopilación de manera continua, y con esto ir sacando métricas de desempeño, esto quiere decir, ir actualizando continuamente la base de datos con que se retroalimenta el modelo, y analizando como estos van afectando el Rsquared and RMSE.

Finally, for a correct implementation for a technology company, it is necessary to implement the model in a cloud, which allows for continuous collection data, and with this, to obtain performance metrics, this means, to continuously update the database with which the model is fed back, and analyzing how these are affecting the Rsquared and RMSE.