Used Cars Price Prediction

Capstone project based on Machine Learnig / Reggresion

Juan José Astudillo

Topics

- Overview of the problem
- Approach for the solution
- Key findings & insights
- Key takeaways
- Recommendations & next steps.

Overview of the problem

Huge demand for used cars in the Indian Market today

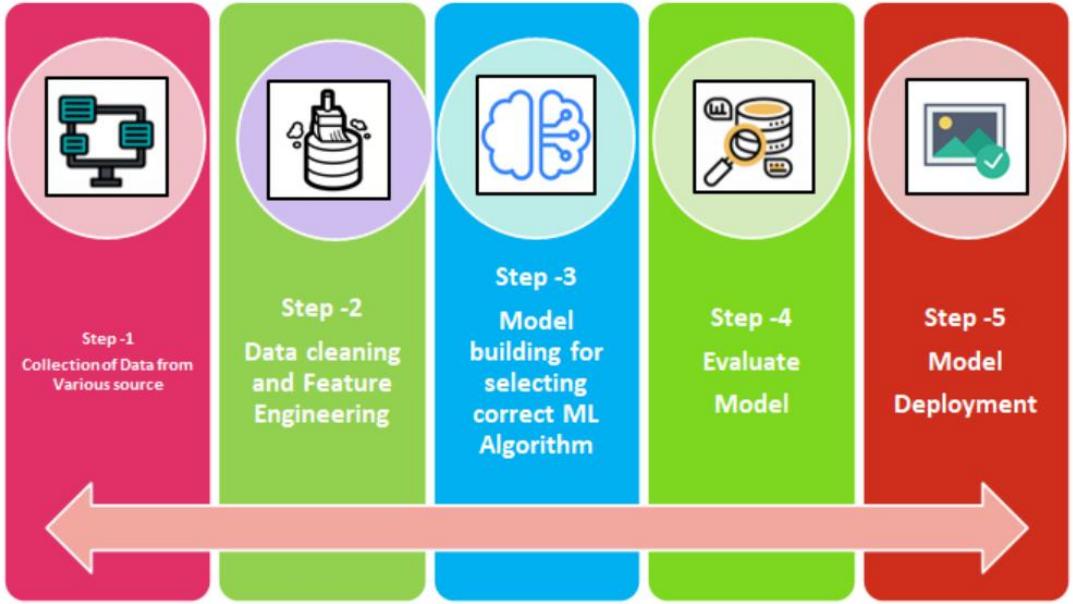
- The used car market has large uncertainties in both pricing and supply.

- Several factors, including mileage, brand, model, year, etc. can influence the actual worth of a car

Objective

Pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing.

Approach for the solution



Key findings & insights

Processing data:

- **Log transformation** to skewed distributed data (Kilometers_driven, engine, power and price)
- Split Name data into **Brand** and **Model**.
- **Replace missing data** of variables (Milage, Engine, Seats, Power, New Price, Price) by the median Brad.

Key findings & insights

Model building:

- Converts the categorical variable into dummy or indicator variables (**25 features**)
- Data distribuited in **70% train data** and **30% in test data**.
- **Probed algorithm**:
  - Linear regression
  - Ridge/Lasso regularization
  - Decision Tree
  - Random Forest

Key takeaways

Linear Regression:

- **6 of 25 features** were identified that do not have an impact on the model.
- Remaining 19 values, **12** are **categorical variables** and **7 numerical variables.**

Ridge/Lasso Regularization:

- In Ridge, adjusting alpha value, there is no significant improvement with respect to linear regression.

- Lasso Regularization presents a low performance.

Key takeaways

Decision Tree:

- **Overfit in training data**. The tree ends up creating a leaf for each piece of data.
- Adjusting the hyperparameters, the overfit can be reduced and DT with fewer nodes can be obtained (length of 6 nodes).
- **Reduce** the important variables from **25 to 13**.

Random Forest:

- Overfit in the training data, but **reduced** by the **cross-validation process**.
- No performance improvement from tuning the model hyperparameters

Key takeaways

- DT and RF present an overfit in their training data, which decreases in the test data.
- LR and Ridge Reg. present a high bias in the training, but they improve for the test data.
- There are no big differences in RMSE except for the Lasso Reg.

	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.

- 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.

Recommendations & next steps

- Assumptions of LR were not validated in the present work:
  - ✓ No multicollinearity
  - ✓ Normal distribution error
  - ✓ Homoskedascity of error terms.
- Its necessary to implement the model in a cloud:
  - ➢ Continuous collection data.
  - ➢ Continue performance metrics.
  - ➢ Automatic update of database for predict Price.