Capstone Project: Regression

Used Cars Price Prediction

Carlos Medina - MIT-PE ADSP May '22 A

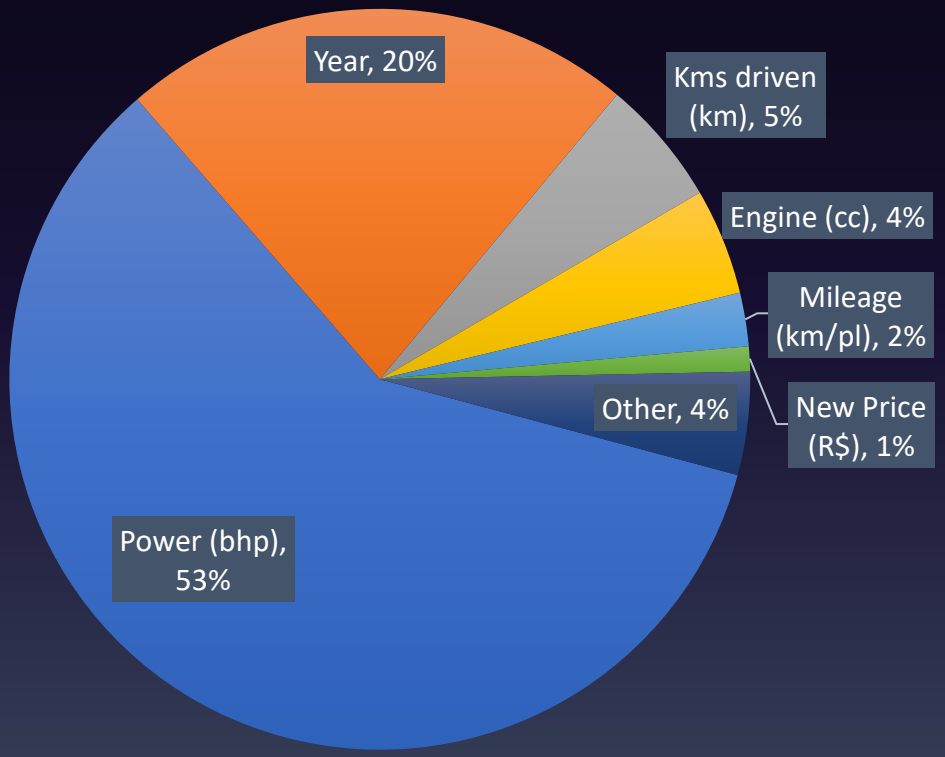
I. Executive Summary

Find a price scheme that allows to **easily predict** what should be the price of a used vehicle

Establish an objective criterion based on a **data-driven** approach

Identify the **critical features** and discover the **relationships** between them and opportunities to incorporate new ones

Key Takeaways



Power (bhp) + **Year of manufacture** both concentrate ~**70%** of the predictive power

Mileage, the **Engine** capacity (cc) and the **kilometers** traveled follows in importance

Iterations with different sample datasets can confirm this behavior.

Key Takeaways

Power (bhp) is **strongly** correlated with engine (cc) and price. An increase in power value implies an increase in the others two

Price and Year present a **slightly** direct relationship

Mileage is **inversely correlated** with the Engine and the Power

Key Takeaways

- Adding additional variables if possible, such as **category** (hatchback, SUV, sedan) can be useful

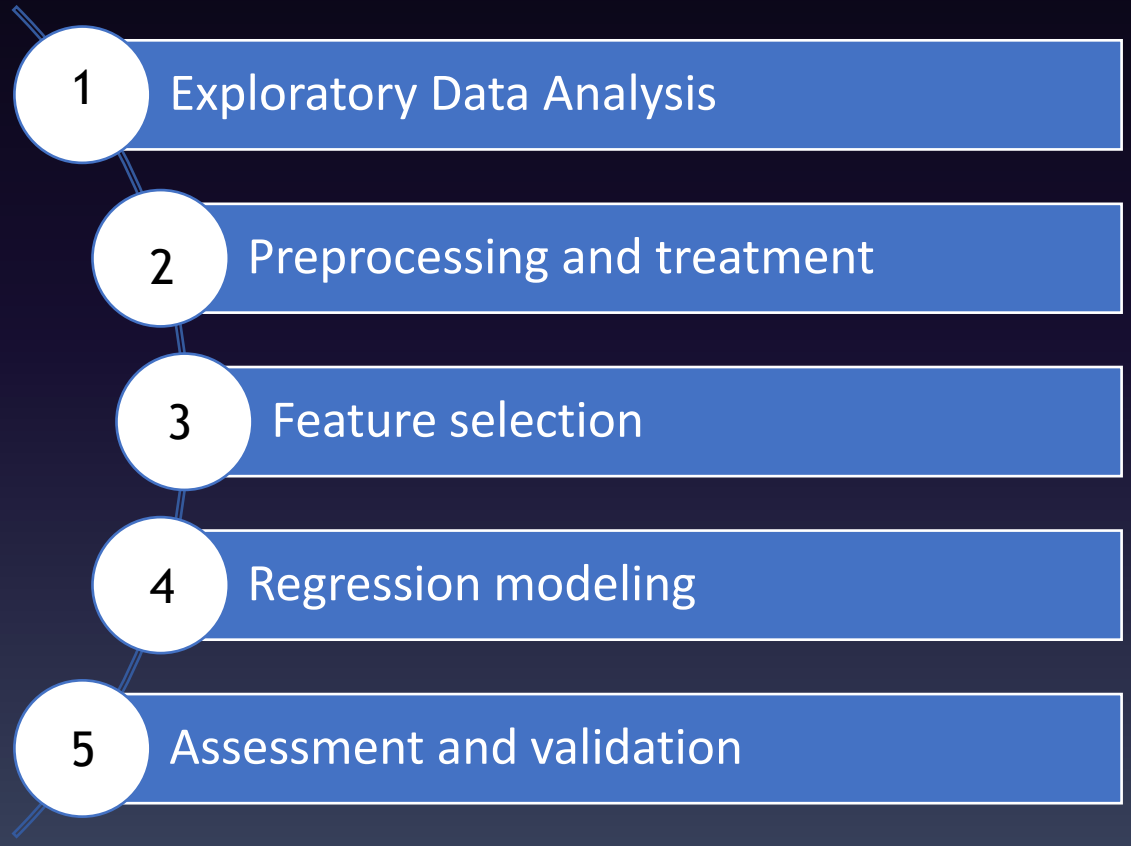
- Based on the data provided, some assumptions can be confirmed or discarded depending on the **weighted importance** of the available variables

- With the gained insight we can refine further analysis by dividing our offer through a **segmentation** for more specific prediction based on specific classes of vehicles

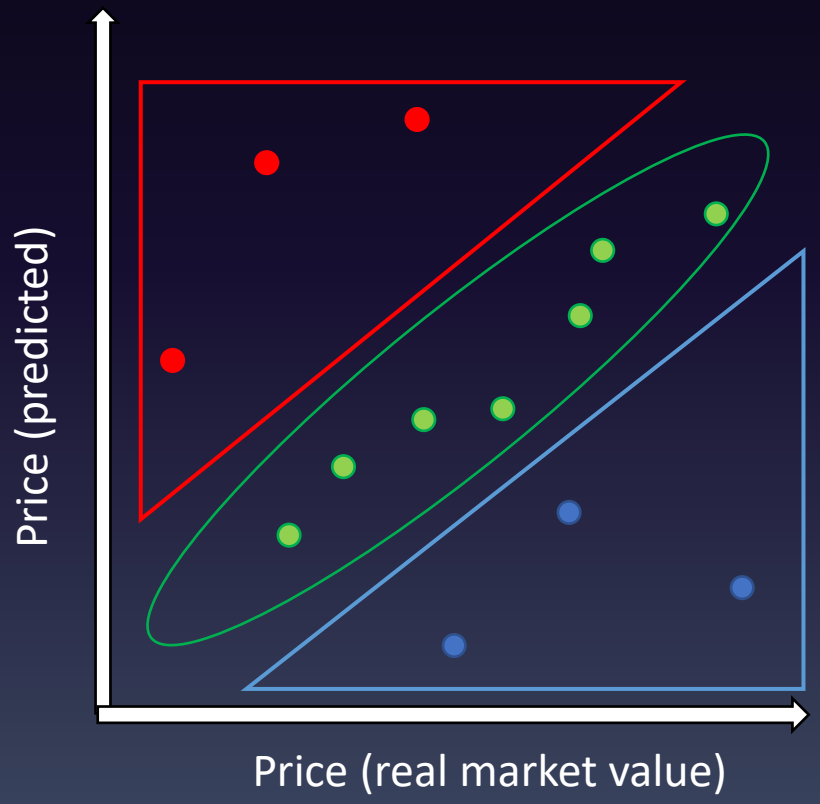
II. Problem and Solution Summary

The problem consists in find a **pricing model** that help the business to devising strategies for differential pricing

We implement a **Regression Model** which consists in a solution that using a set of variables from the data provided can find an equation able to effectively predict the price



Proposed solution



The graph shows conceptually that our price prediction calculation can fall into 3 zones clearly marked in red, blue and green.

Red zone: a poor calculation made subjectively lead to a predicted price higher than the real

Blue zone: an underestimation leads to sell below the standard market value

Both scenarios lead us to losses due to unrealized sales (low conversion ratio) or loss in profit margin (on inventory costs, advertising)

Choice of the proposed solution

Model	Train r2	Test r2
Linear Regression	0.88	0.60
Ridge Regularization	0.84	0.72
Lasso Regularization	-1.32	-0.11
Decision Tree	1.00	0.58
Decision Tree (depth = 7)	0.73	0.67
Decision Tree (tuned)	0.72	0.66
Random Forest	0.94	0.69
Random Forest (tuned)	0.82	0.66
KNN Regressor	0.70	0.70

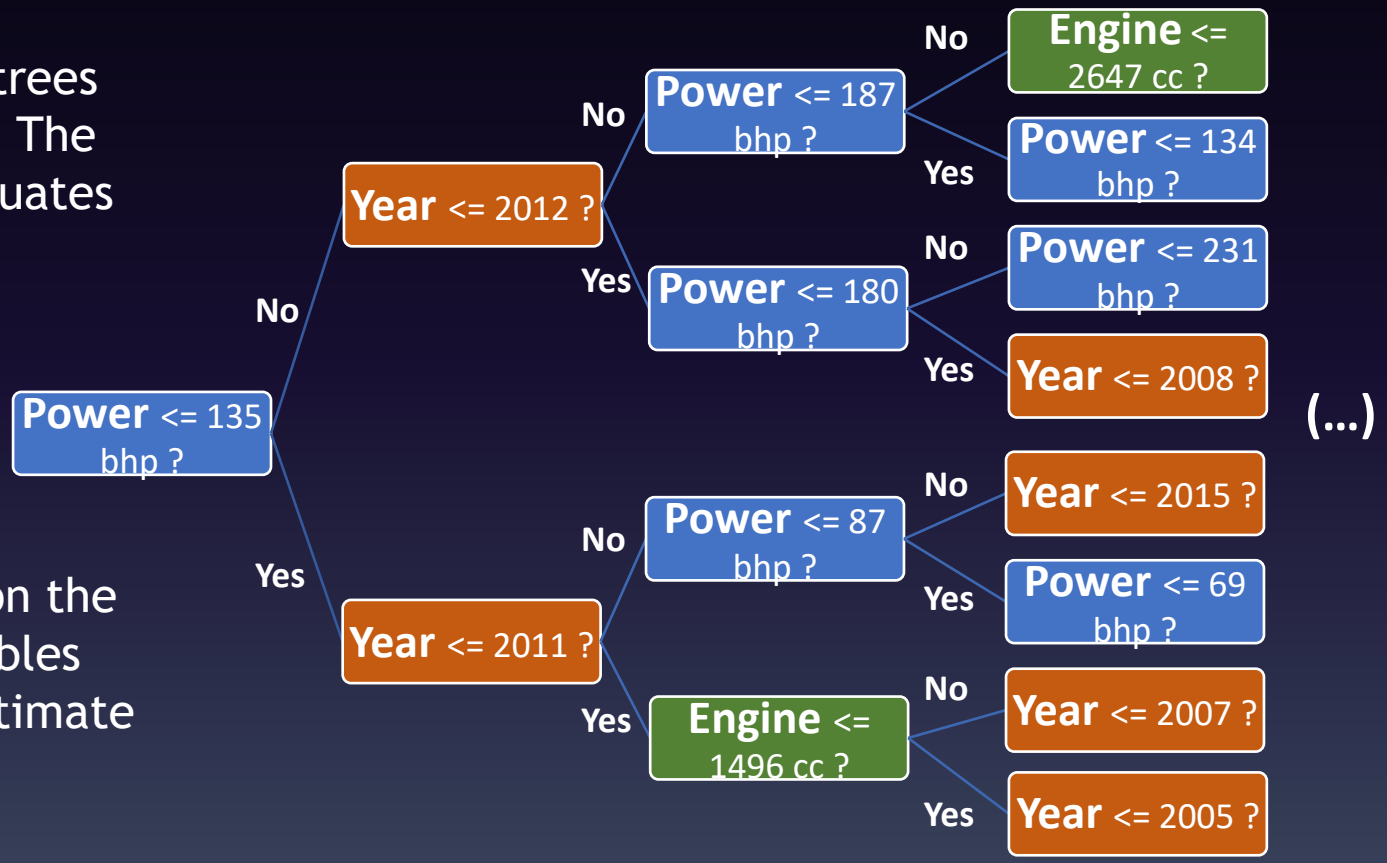
After applying different **regression methods**, the **performance** results show an adequate balance between the test and training data for the Random Forest model

This has greater interpretability of the characteristics that make up the price prediction. It follows a scheme very similar to a **flowchart** to explain to stakeholders

Proposed Random Forest Model

The first 4 levels of one of the decision trees that make up the Random Forest model. The decision flow starts with Power and evaluates the value related to the vehicle

Then performs a new evaluation based on the Year and so on through the critical variables that contribute the most to the price estimate obtaining the predicted price after



III. Recommendations for Implementation

For Sales: it is recommended to perform a **segmentation** of the offer. From compact vehicles of low power through segments up to high performance vehicles

For Finance / Acquisition: Knowing in advance the maximum price acceptable to sell in the market, we can maximize the profit margin in the process of buying the vehicles from their original owners. For them it is recommended to improve the collection of information on prices (**17% of the prices was missing**)

For Marketing: Advertising strategies can be **targeted by segments**, in order to control budgets on channels, this can be supported with the financial area for further investment in advertising

Risks and Considerations

- **Missing values / Outliers:** We have detected about 17% of price variable, and problems may arise when discarding (**loss of information**) or when imputing (**add noise to model**)

- **Maintain the performance:** Market prices, inflation, fluctuations in demand, exchange rates, which can quickly degrade the predictive power of the model.

- **Additional features:** Can enrich the pricing model, as well as discarding some that have proven not to be necessary.

Questions