Used Cars Price Prediction

Final Presentation

Viktor Dimitrov / MIT ADSP May'21

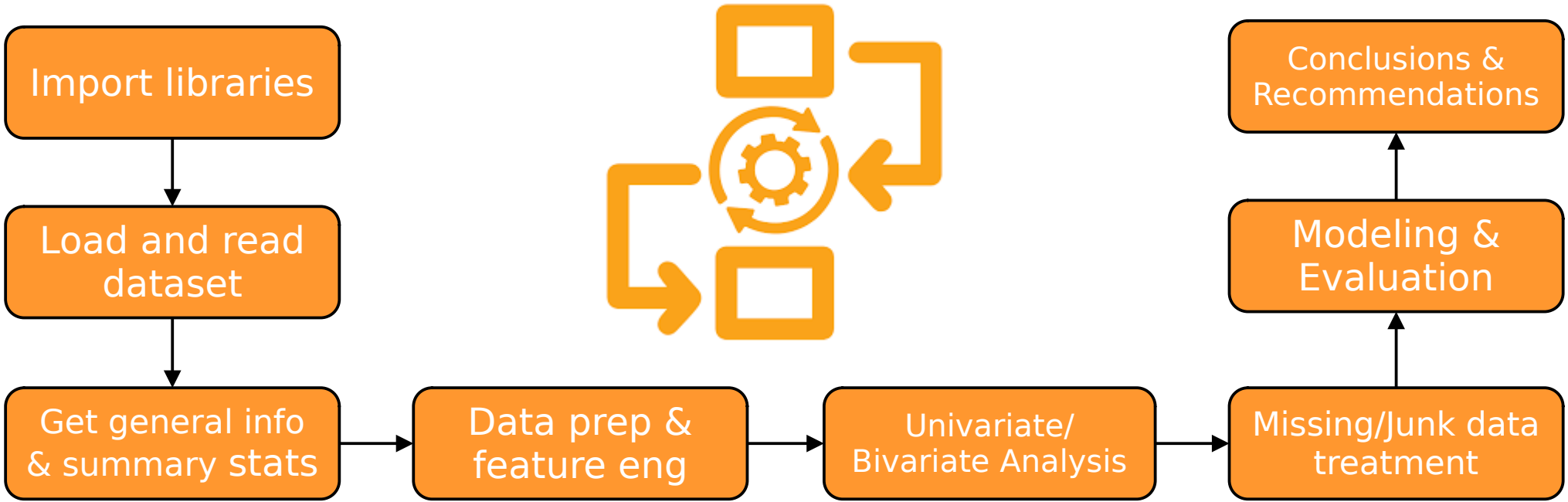
Executive Summary

Problem	The problem on hand is to predict the price of used cars based on the features provided in the used cars dataset. We need to identify the most important features as well as their correlations in the process. In order to achieve that, we need to use techniques for data pre-processing and build a regression model that predicts the prices for us.
Objective	Come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing.
Key questions	✔ What factors influence the actual worth of a car the most? ✔ How to come up with the right price for a used car?

Interesting facts

7,253	Total used cars
8	The average age of the cars | The % of NULL values
Maruti Swift	The most popular model (418 units)
0.03	The % of electric cars (2 units)
6,500,000	Kilometers “driven” by the biggest outlier
582	The BHP difference between the weakest and the most powerful car
33	Unique brands
5	Average number of seats
120	The mean price (in INR 100,000) of a Lamborghini
82	The % of cars having one owner
Bangalore	The city with most expensive cars

Solution workflow



Models performance evaluation & comparison

The performance of 5 regression models on the train and test data have been evaluated and compared, based on 5 key metrics (R-squared, MSE, RMSE, MAE, and MAPE):

Model	R-squared train	R-squared test	MSE train	MSE test	RMSE train	RMSE test	MAE train	MAE test	MAPE train	MAPE test
Linear	0.93	0.89	7.1	14.68	2.66	3.83	1.18	1.45	12.38	13.26
Ridge	0.93	0.89	7.1	14.67	2.66	3.83	1.18	1.45	12.38	13.26
Lasso	0.73	0.76	27.66	30.91	5.26	5.56	2.66	2.55	26.83	28.07
ElasticNet	0.71	0.71	30.16	37	5.49	6.08	2.73	2.73	27.2	27.71
Decision Tree	1	0.53	0	59.7	0.03	7.73	0	3.66	0.03	36.72
Random Forest	0.98	0.78	1.99	28.39	1.41	5.33	0.51	2.58	4.92	27.36

- ✔ Both **Linear Regression** and **Random Forest** performed very well. However, **Linear Regression** has the edge due to a better generalized performance on train and test data.
- ✔ With Alpha at 0.01 (best value) **Ridge Regression** is almost completely matching **Linear Regression**, with one exception being the slightly better MSE result on test data.
- ✔ The **Decision Tree** is clearly over-fitting the train data and under-fitting the test data.
- ✔ **Lasso** and **ElasticNet** have a very similar performance on train and test data. They seem to generalize quite well.

Business Insights and Recommendations

✔ Cars4U could benefit from investing in the following types of vehicles in this particular order:

- ✔ **New**/er
- ✔ **Powerful** (250+BHP)
- ✔ **Top-tier** brands like: "Mercedes", "Land Rover", "Mini", "BMW", "Jaguar"
- ✔ **Diesel/Electric**
- ✔ **Automatic**
- ✔ **First owner**

✔ Some southern markets like **Coimbatore** and **Bangalore** tend to have higher prices. It might be a good strategy to extend operations there. Markets like **Kolkata** and **Jaipur** on the other hand are quite risky and investments in these areas should be carefully re-considered.

✔ The **cost** aspect of business should be analyzed as well once we have the corresponding data.

✔ Going forward, the current model could be **improved further**, and **additional models** could be developed for different locations or car types as applicable.

Thank you!