



Cars4U: Predicting Used Car Prices

Analysis by Joe Balog

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

Context: Find foothold in the pre-owned car market as a tech-company selling cars and insights.

Problem: accurate pricing scheme and prediction power.

Implications: It will never advise someone to buy a used car that doesn't meet Cars4U customers' needs.

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations

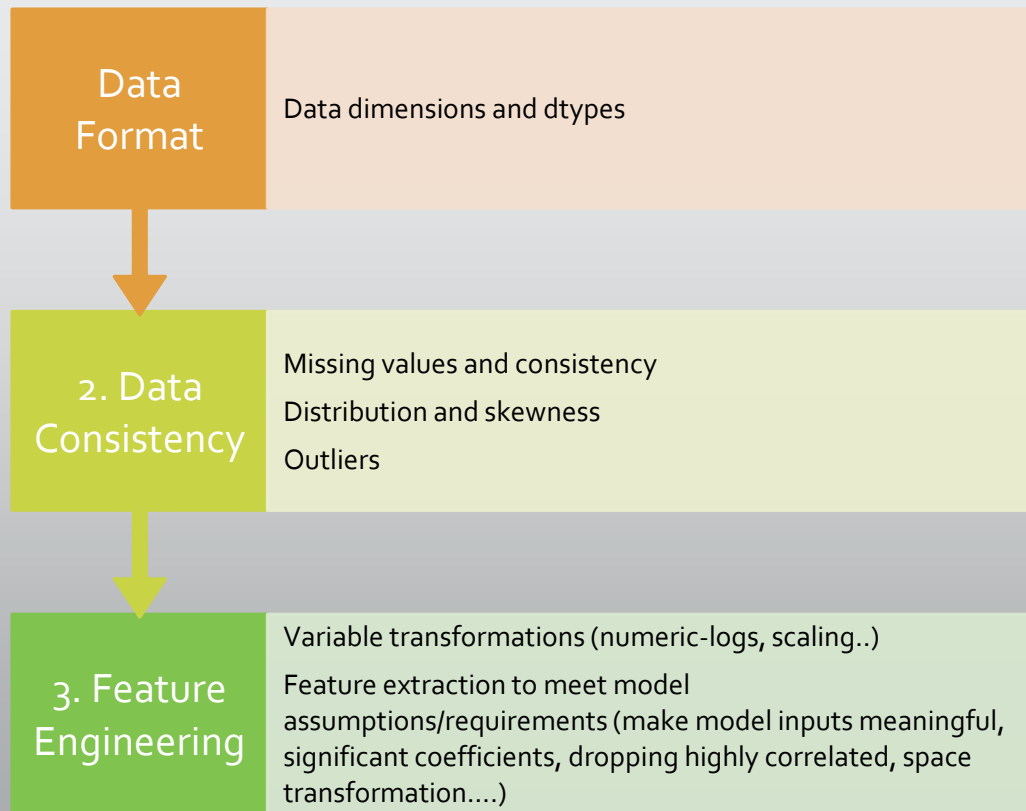
Data Overview: Description of data

A messy .csv file that has fictious data about used cars from places in India.

A look at the factors that effect people's decisions, like engine features and efficiency

The historical information on the ownership history and the year of purchase

Data Overview: Pre- Processing Steps

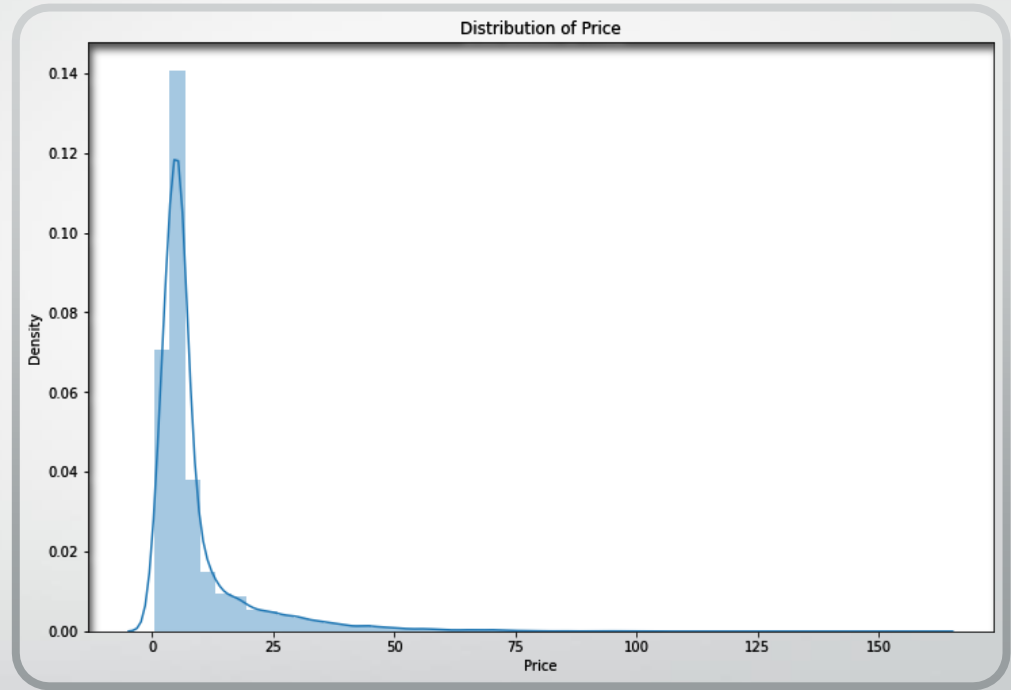


Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations

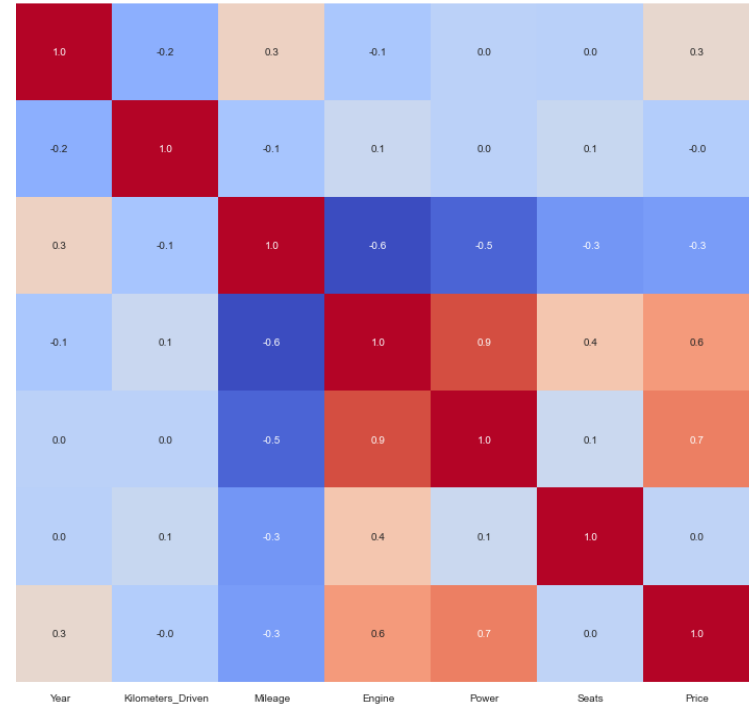
EDA: Price

- Skewed towards cheaper cars.
- Not a normal distribution.
- Outliers are Lambourghini and Ambassador brands.



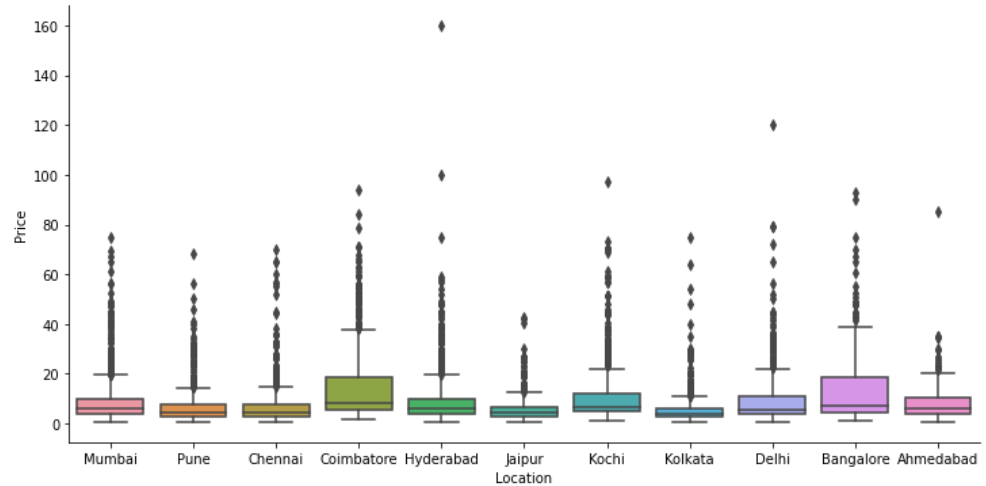
EDA: Impacting Price

- The relationship between the maximum power in an engine and price is positive.
- The average Break-Horse-Power is 112 which covers Brands like Mahindra & Mini Cooper.
- The 10 cheapest cars have have 76 - 227bhp.

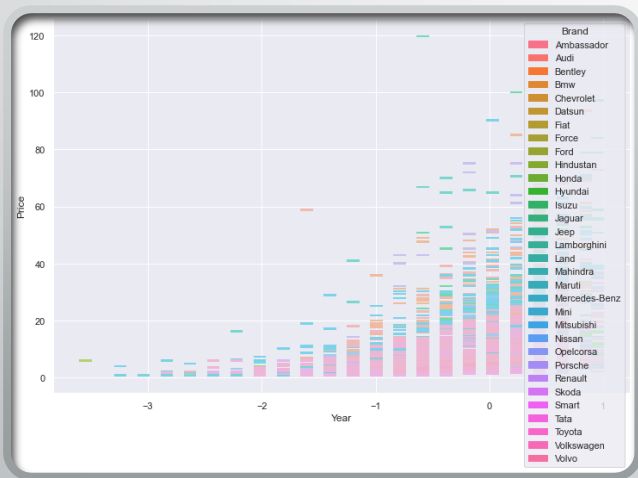


EDA: Impacting Price

- Coimbatore has the highest average price amongst all place in India
- The center of India's high-tech industry, Bangalore, has the second highest.
- Mumbai and Delhi, the largest cities, are low on the average price of used cars.



EDA: Impacting Price



- The spread of Price is not even across car Brands with outliers for those high-cost ones.
- You can see some clustering at lower prices in certain brands.
- There are some brands which are not well represented.

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations

Model Performance Summary: model and its parameters

- $\text{Price} = \beta_0 + \text{Year}_1x_1 + \text{Engine}_2x_2 + \text{Power}_3x_3 + \text{Kilometers_Driven}_4x_4 + \text{Mileage}_5x_5 + \text{Location}_n x_n + \dots + \text{Transmission}_n x_n + \text{Make}_n x_n + \text{Fuel_Type}_n x_n$
- The model includes both numerical data (Engine, Power, etc) and categorical data (Location, Transmission, etc).
- A 70/30% split between Training and Test sets to build a model that fits well between them and can be tested.
- Built one that has a parameter of intercept and constant coefficient that is very low.

- Mean absolute percentage error (MAPE) measures the accuracy of predictions as a percentage.
- MAE score is calculated as the average of the absolute error values. Absolute or `abs()` is a mathematical function that simply makes a number positive.
- A constant model that always predicts the expected value (or mean) of y R^2 of 0.0.
- Adjusted- R^2 tells us R^2 changed based on number of variables in the model. The higher the better the model fit is.

Model Performance Summary: factors used for prediction

Model Performance Summary: key performance metrics

With our linear regression model we have been able to capture ~76% information in our data. R^2 is 0.79, i.e., the model variance of its errors is 79% less than the variance of the dependent variable. So overall, the model is satisfactory.

Metric	Training Set	Test Set
MAPE	inf	inf
MAE	0.31	0.31
R^{MSE}	0.39	0.39
Adjusted R^2	75%	75%

Linear Regression Model

The model indicates that the most significant predictors of price of used cars are -

- 'Car_category_Ultra_luxury'
- 'Car_category_Luxury_Cars'
- 'Car_category_Mid-Range'
- 'Fuel_Type_Diesel'
- 'Year'

The following features have a positive influence on Price:

- 'Year'
- 'Power'
- 'Fuel_Type_Diesel'

The following features have a negative influence on Price:

- 'Transmission_Manual'
- 'Seats'
- 'Owner_Type_Third'

$$\hat{y} = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon_i$$

Scope of Improvement

- Binning categorical variables into larger features (i.e., Location by region) could improve the model performance.
- The empty values were replaced with the median value from the feature and could have been improved by a mean replacement.
- A stronger transformation on the exponential Price feature or the semi-normalized predicative variables could have made a stronger model.

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA
- Model Performance Summary
- Business Insights and Recommendations


Business Insights:

- Automatic transmission used cars tend to be newer and have higher prices.
- Used cars with a Third owner tend to be used before 2015 and have lower prices than other owners levels.
- The mean price of petrol-fueled used cars is higher and tends to be in model years around 2010.

Business Insights and Recommendations

- Year and Power data navigate inference towards higher priced vehicles when looking to predict used car prices.
- Flag incoming data with any Electric vehicles as high-cost and Hindustan makes as lower-cost. Other fuel types have similar effects to each other and will tend to be lower cost as a whole. The other makes will be higher cost but all together have the same effect on Price, roughly speaking a -13 to -24 point decrease in Lakhs.
- Location has a negative effect on the price across any region or city in India, so don't bias your recommendations based on location.

- Give clients a predicative assessment of a vehicle they are interested in purchasing by looking at the factors using in the model (Power, Location, Fuel type, etc...) and add the real-world car values to the appropriate feature and run the model to come up with a car (you may have to drop other unneeded columns).
- For example, you can get a Power (bph) or of a used car and review the model for the predictive co-efficient value to determine the correlation to price. You can get a sense of what the asking price should be at the Power level.
- Get out there and test your predictive scheme on more used vehicles and add them to the data running into the model.



Model Implementation in the Real World

Potential Benefits of Implementing the Solution

- Giving clients who are purchasing used cars a sense of confidence in their purchasing decision.
- Giving price-point predictions that are more accurate than other firms (KellyBlueBook, Cars.com, etc) that can give you a competitive edge in the market.
- A starting point to do on-going development of improved models, whether Linear Regression or not, to give even more accurate predictions on used car prices, based on even new features.

Advice to grow the business



Try out the model on a small sample of clients and examine the results. See if the model lines up with reality. If it doesn't, go back and review what's wrong about the model and add any needed features from the real-world.



Without more data, the extend to which the model will be useful will run out in 5 years or so. It's important to build a data-pipeline that will give you cleaner, fresher, and accurate data on a regular basis.



To that effect, look at hiring a database administrator or data engineer to create a safe, secure, and reliable database system that will support on-going analysis and modeling.

Thank You!

- Joseph Balog
- 231-313-7225
- josephmbalog@gmail.com

