Car Price Prediction

...

September 22, 2019

ROHIT

UG Student at IITM,
Interested in using data
science techniques to make
informed business decisions

Agenda

- Business Understanding
- Data Exploration and Preparation
- Model Building
- Hyper-parameter Tuning and Model Evaluation
- Result / Outcomes

Business Understanding

Business Understanding

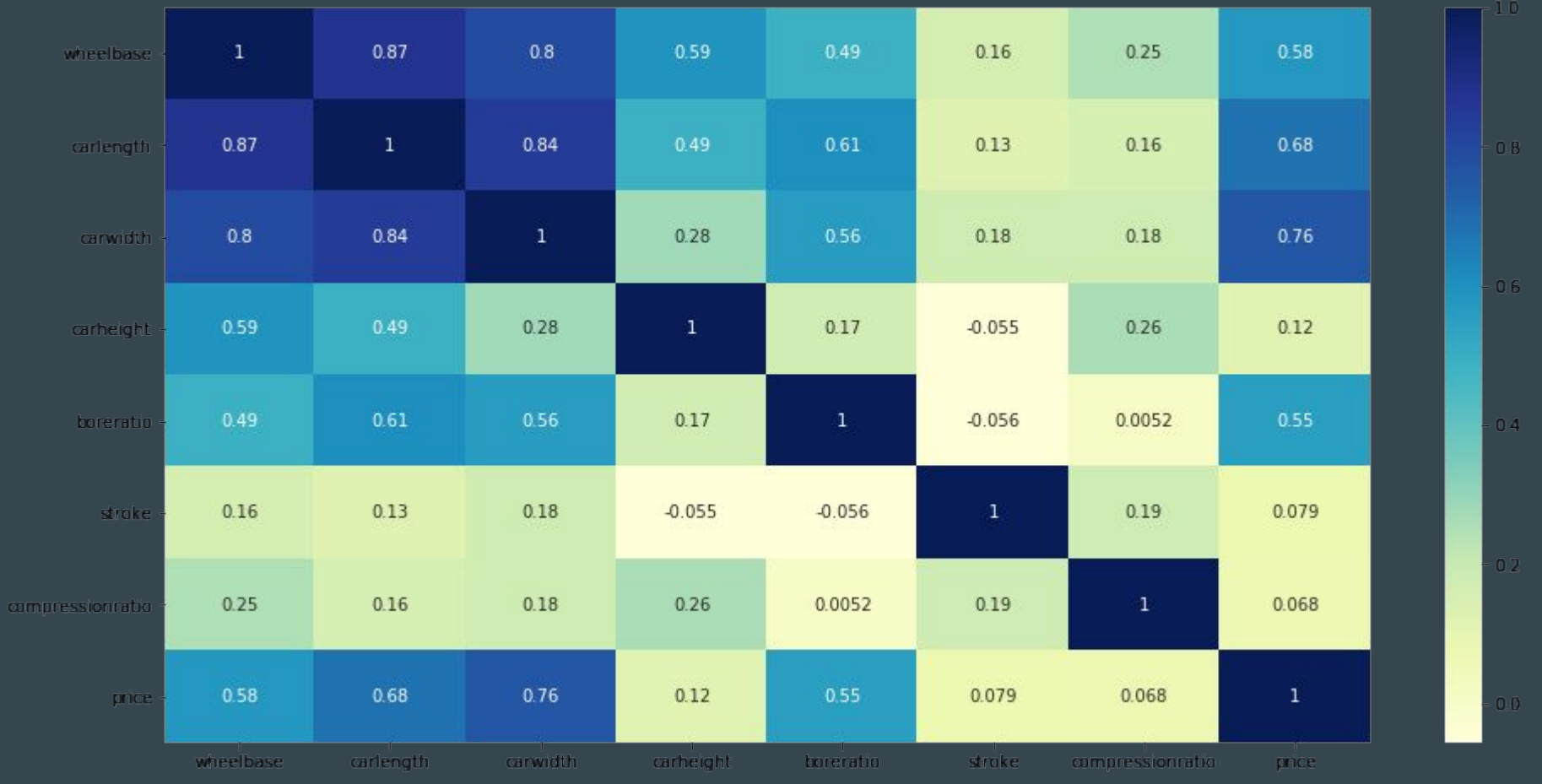
- Problem Statement: Predict Car Pricing by analyzing correlation among various parameters
- Problem Motivation: By devising such a prediction algorithm, the buyer can anticipate car prices based on car features and various other factors and make informed decisions while planning for purchasing a car
- Data was collected about various car brands
- What can this data tell us about car prices
- Can these data science techniques be applied to other areas?

Data Exploration and Preparation

Data Exploration and Preparation

- All coding done in Python 3.
- Extensive use of pandas, numpy, matplotlib, as well as seaborn and sklearn packages.
- Dataset contained 26 different features on 205 car brands and type
- Features were both categorical and numerical. Target variable was binary (“Yes” or “No”).
- Pandas package was imported and a dataframe was created.
- Categorical variables were looked at first. Visualizations were created using the seaborn package
- Heatmap using seaborn package was created to show us any particularly strong correlations between the independent variables and the target variable outcome..

Correlation Heatmap :



The heatmap shows some useful insights:

- Correlation of price with independent variables:

  Price is highly (positively) correlated with wheelbase, carlength, carwidth, curbweight, enginesize, horsepower (notice how all of these variables represent the size/weight/engine power of the car)

  Price is negatively correlated to citympg and highwaympg (-0.70 approximately). This suggest that cars having high mileage may fall in the 'economy' cars category, and are priced lower (think Maruti Alto/Swift type of cars, which are designed to be affordable by the middle class, who value mileage more than horsepower/size of car etc.)

- Correlation among independent variables:

  Many independent variables are highly correlated (look at the top-left part of matrix): wheelbase, carlength, curbweight, enginesize etc. are all measures of 'size/weight', and are positively correlated

Thus, while building the model, we'll have to pay attention to multicollinearity (especially linear models, such as linear and logistic regression, suffer more from multicollinearity).

Model Building

Model Building (1/2)

Linear Regression

- sklearn.linear_model.LinearRegression
- Its a classification model

RFE-Recursive feature elimination

- sklearn.feature_selection.RFE
- Simple to understand and effective

Model Building (2/2)

Statsmodels

- statsmodels.api

Results / Outcome

Best Model and Feature Importance

- This is a simple model , the final predictors seem to have high correlations.
- Below is the heatmap of multicollinearity

Thank You