Neshat Jalali Heravi
October 2021

Building a Price Predictive Model for Cars: A Regression Story!

The recent pandemic has shaken many industries including automakers. Due to limited production and trading in the past year and a half, there has been a shortage of parts, including semiconductors, which caused the price spikes and fluctuation in vehicles[1]! The purpose of this project is to build a predictive model for cars mainly for Pacific Northwest. The outcome of this predictive model is to help car buyers by providing some price insights based on the recent cars trading data from Seattle and Portland areas.

Design

The goal of this project is to build and select the most accurate predictive model for car prices based on their features. With any predictive model there is going to be a margin of error. Thus, the values for evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are reported with each model! In addition, the R^2 score for each regression model is stated. Finally, the model with best metric values is chosen as our most accurate predictive model.

Data

The data is scraped from cars.com website using Python libraries and packages including Beautiful Soup library. Since the target area is Pacific Northwest, the extracted data belongs to Seattle, Portland, and areas within their 30 miles radius.

At the starting point, the dataset included 10,015 observations and 15 feature columns. After the initial cleaning, adding an interaction feature and eliminating the irrelevant columns during feature engineering the DataFrame had 8,819 rows and 17 columns including 9 categorical features. Mind you that the column counts increased post preprocessing and feature dummification of the categorical features!

Algorithms

After collecting the data in csv format, initial cleaning was performed. Although some features were independent and ready to use as is, some features required more thorough data cleaning. For example, features such as Built Year, Make and Model were extracted from the Name column.

Initial EDA was performed to have a better understanding of the data which led to eliminating the outliers in some of the features. The data set was split into train\validation\test (60-20-20) sets for training and testing each model. A simple linear regression model was built using the numerical features as the base model! The base model helped with understanding the need for any interaction features and eliminating the less important features. A log transformation was performed on the target values to make it more suitable for linear regression models. Moreover, the categorical features were transformed to become suitable for modeling.

Neshat Jalali Heravi
October 2021

Models such as Linear regression, Lasso and Ridge regularization methods, polynomial regression, gradient boosted regressor, and extreme gradient boosting (XGBoost) were used during this project. In each step the models were fine-tuned based on their evaluation metrics. Here is a summary of the metrics for each model:

- Base Linear Regression (numeric features only):
  - R^2 Training set: 0.35
  - R^2 Test set: 0.37
  - MAE: 8,940.20
  - RMSE: 11,545.34

- Linear Regression (post log transformation of target):
  - R^2 Training set: 0.45
  - R^2 Validation set: 0.47
  - R^2 Test set: 0.45

- Polynomial Regression:
  - R^2 Training set: 0.55
  - R^2 Validation set: 0.55
  - R^2 Test set: 0.54

- Lasso Regression:
  - R^2 Training set: 0.87
  - R^2 Validation set: 0.85

- Ridge Regression:
  - R^2 Training set: 0.88
  - R^2 Validation set: 0.85

- Gradient Boosted Regressor:
  - R^2 Test set: 0.80

- XGBoost Regressor:
  - R^2 Training set: 0.95
  - R^2 Validation set: 0.95
  - R^2 Test set: 0.90
  - MAE: 3,323.50
  - RMSE: 4,682.70

Tools

- Python libraries including BeautifulSoup, Pandas, and Numpy for web scaping, parsing and data manipulation.
- Statsmodels and Scikit-learn for modeling, hyperparameter tuning and evaluation metrics.
- Matplotlib, Seaborn and Yellowbrick for data visualizations.

Communications

You can find the presentation slides and details on the analysis, and the notebooks on my GitHub. In addition, the Tableau dashboard includes some interactive visualization of the results.

References

[1] “The high prices of used cars may finally be dropping: Sonic Automotive president” CNBC, Aug 1st, 2021.