Predicting Used Car Prices with Machine Learning

By Matthew Malone

Problem Statement: Paying the true value for a used automobile is incredibly challenging in a market where prices greatly range. Can we create a pricing model that can assist individual buyers and sellers establish the fair market value of used cars on online platforms.

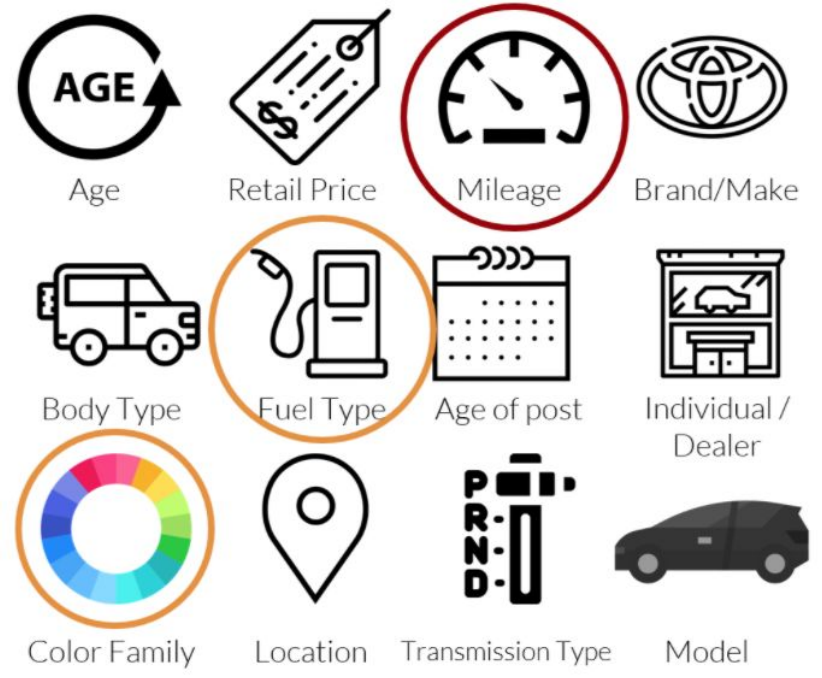
Price Estimates For a 2000 Chevrolet Four Door Lumina From the Same Month

Kelley Blue Book: $7,875

NADA: $6,875

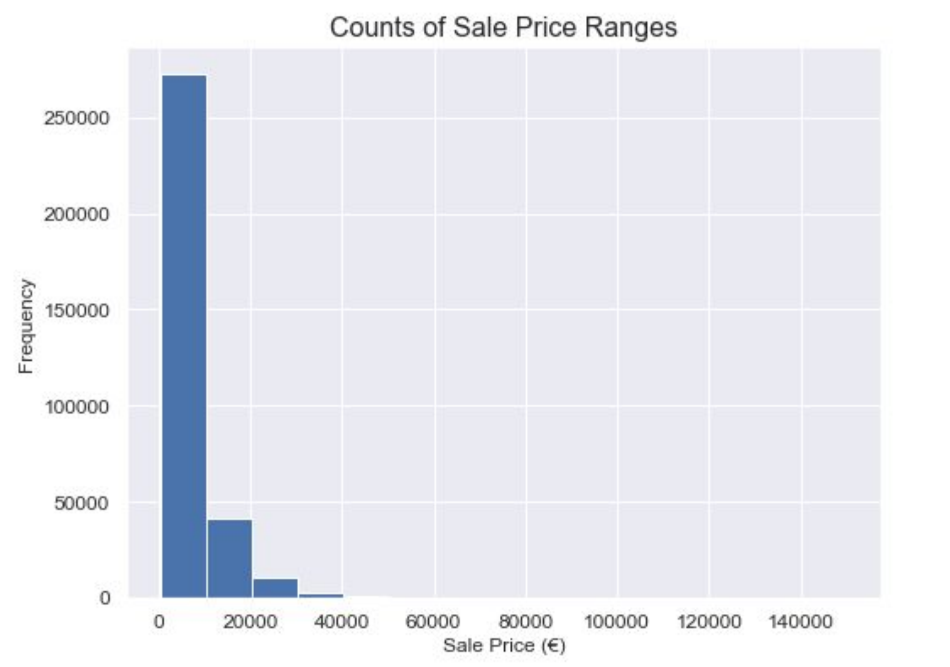
Black Book: $5,650 to $8,850

Data: Over 370,000 Used Car Listings Were Scraped From Ebay-Kleinanzeign, An Ebay Subsidiary Based In Germany.



Exploratory Data Analysis

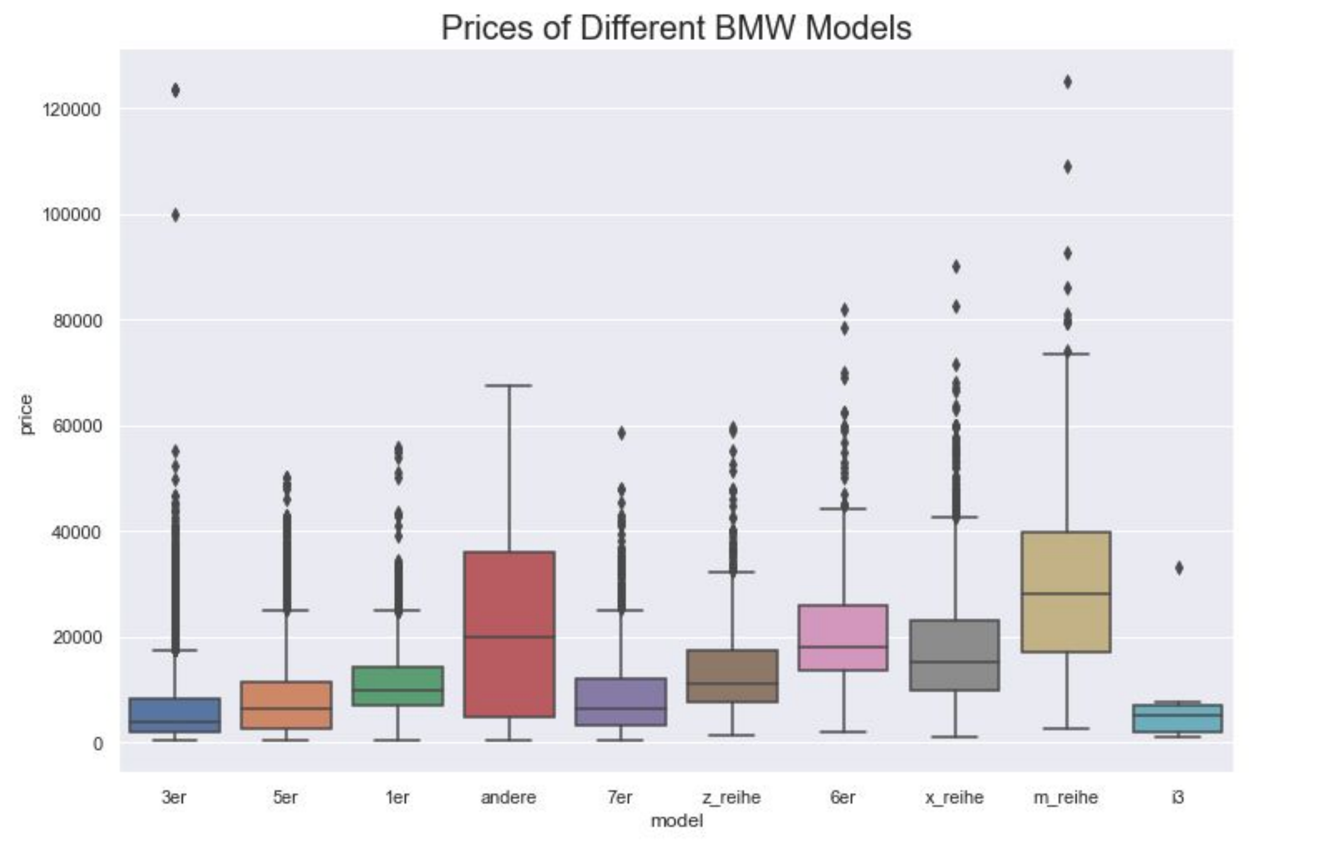
Target Variable: Price (Euros)



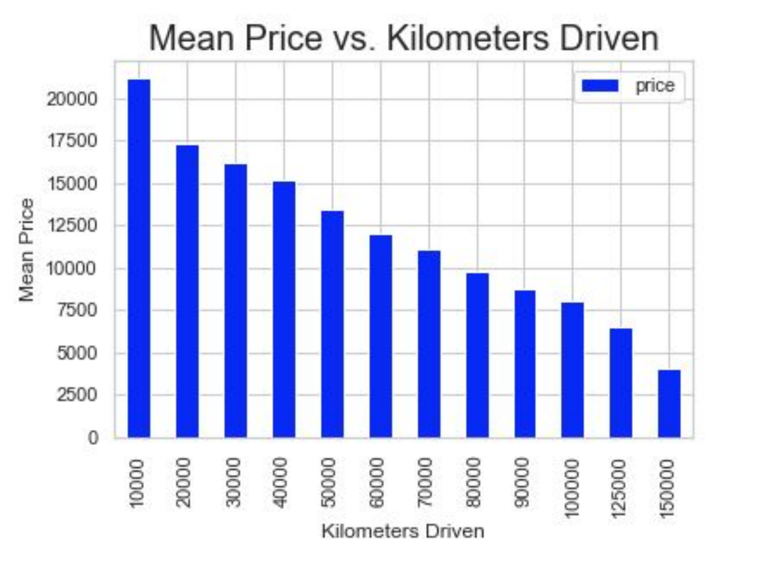
-Average car price: 4,878.12 Euros

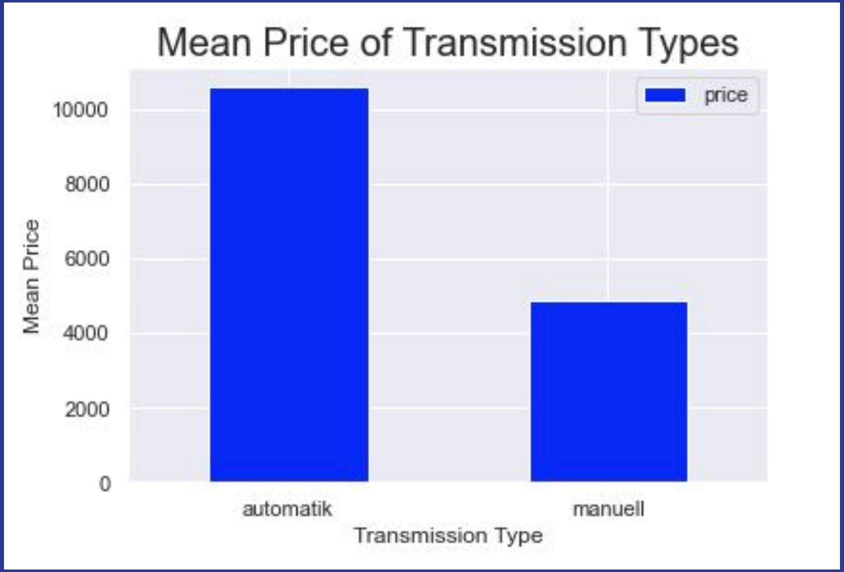
-Cars priced at below 500 Euros will be removed from data as cars at this price point and lower are considered “junk status”

Exploratory Data Analysis



Exploratory Data Analysis





Modeling

Model Evaluation Metric: Root Mean Squared Error (RMSE)

Baseline Model: 4580.45

Models:
Linear Regression
Ridge
Lasso
Decision Tree
Random Forest
Extra Trees
Gradient Boost

Model Selection

Random Forest

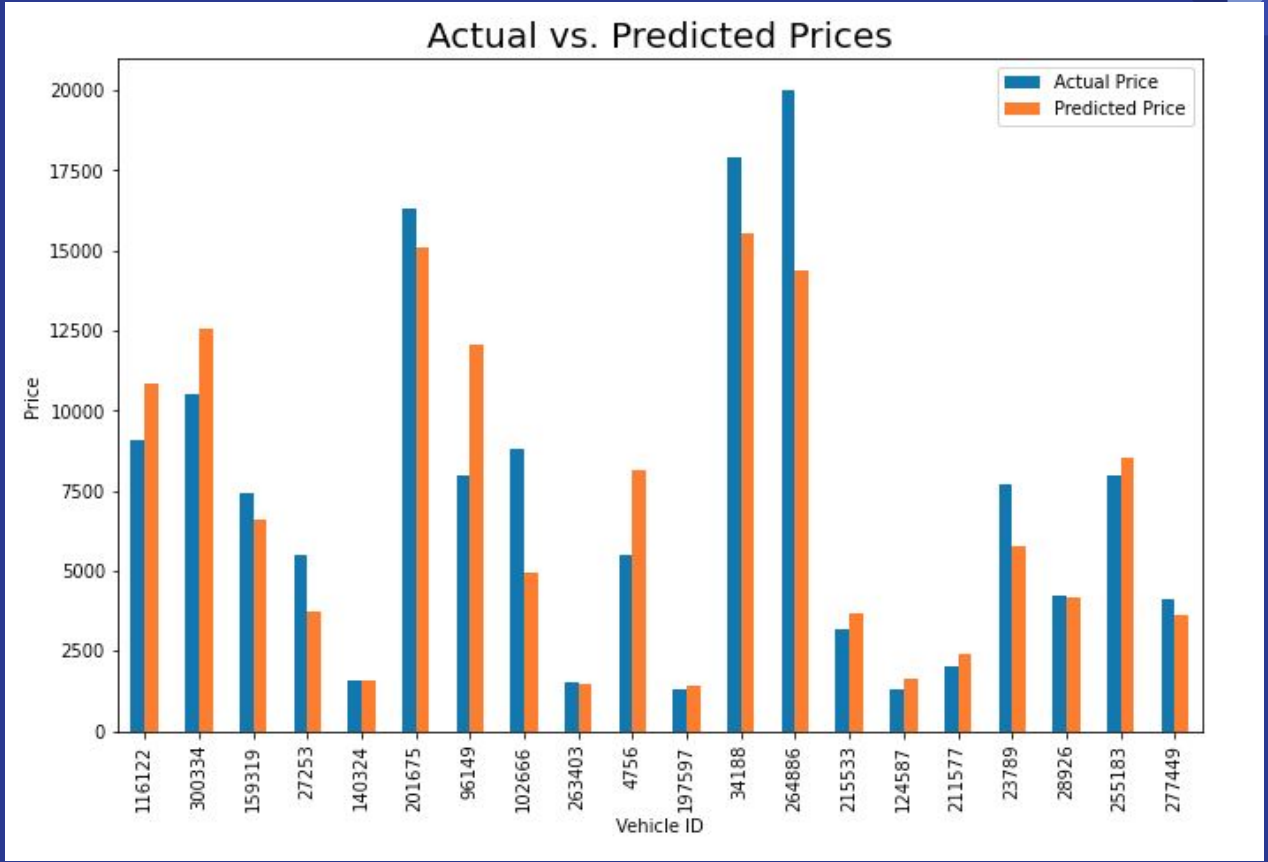
RMSE Training Score: 1,274.47
RMSE Testing Score: 1,710.21

Gradient Boost

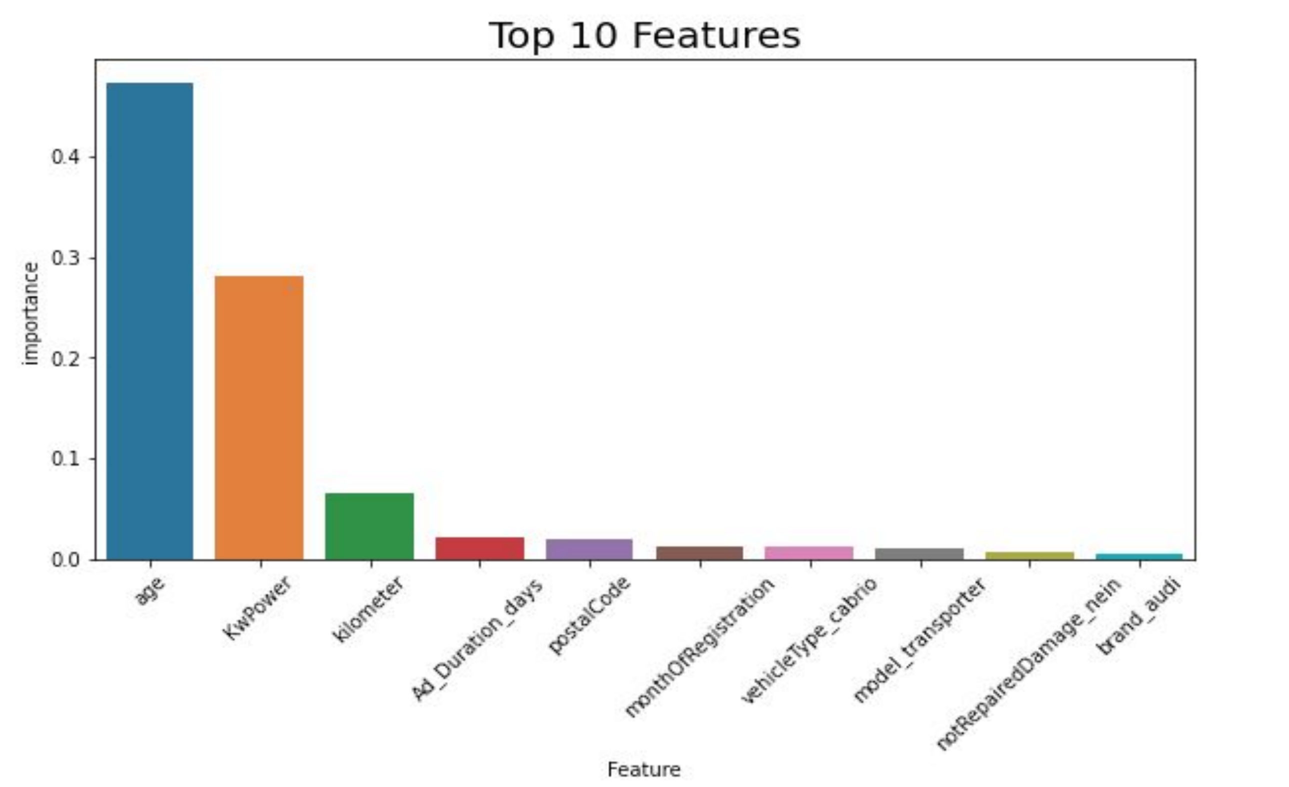
RMSE Training Score: 1,871.87
RMSE Testing Score: 1,882.51

Gradient Boost Was Selected Because It Achieved The Second Best Scores Of All Models And Had Relatively Low Variance. There Was Cause For Concern With The Random Forest Model Because It Had Such High Variance. We Would Be Concerned About Introducing New Data To The Random Forest Model

Model Evaluation

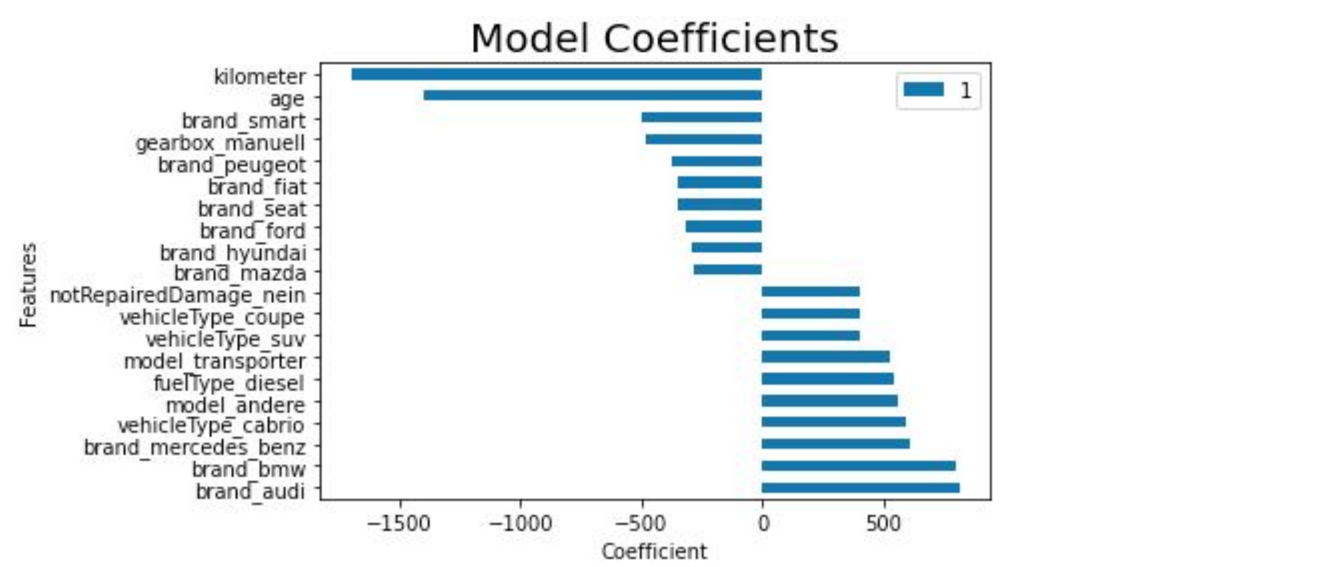


Model Evaluation



Model Evaluation

Interpretable Model: Ridge Regression



Conclusion

- Model Failed to Perform well enough were it could be deployed for real world use. Did not meet goal of a RMSE score within 10% of mean value.

- The best RMSE score 1882 is too far off from the actual price of a car. This would cause a consumer to overpay or a seller to under sell.

- Used Car Sales are incredibly nuanced. Negotiation often comes into play when purchasing a used car.

- How a car runs is an important feature of its resale price. There are a considerable amount of factors that determine how a car runs. This is a difficult feature to include in a model

Recommendations

- Create sub-groups of car models (tier 1, tier 2, etc.) based off of their mean price to reduce the number of features that added noise to the model.
- Incorporate car history reports with more comprehensive information about accidents & repairs (information can be obtained from web sites like Car Fax)
- Target specific brands to increase accuracy of predictions. Volkswagen made up 20%+ of our data, could create a Volkswagen price prediction model

Sources

-Consumer Reports, May, 2014, consumerreports.org/cro/2012/12/how-much-is-the-used-car-really-worth/

-Nerd Wallet, Jeanne Lee Oct, 2015, https://www.nerdwallet.com/blog/loans/dealers-set-car-prices/

-InCharge Debt Solutions, Sept, 2017, incharge.org/understanding-debt/auto/the-truth-about-used-car-prices/