Predicting Vehicle Trim and Dealer Listing Price

Problem

Create one or more models capable of predicting vehicle trim and dealer listing prices of used cars given the other twenty-six labeled variables provided.

Solution

We are aiming at predicting a continuous target variable (dealer listing price) and a categorical one (vehicle trim). Since the predictors are labeled, we are going to use a supervised machine learning regression model for the first target, and a supervised machine learning classification model for the second.

Key insights

1) The random forest regression model created, provides a framework for predicting the dealer listing prices of used cars.

2) Model metric: R2 = 0.82, meaning that 82% of the variance is described by the model.

3) The xgboost classification model created, provides a framework for predicting the vehicle trim of used cars.

4) Metrics. Precision: 86%, Recall: 86%, Accuracy: 86%

5) Stronger predicting features:
- Production year
- Price label
- Engine liters
- Mileage
- Different vehicle features
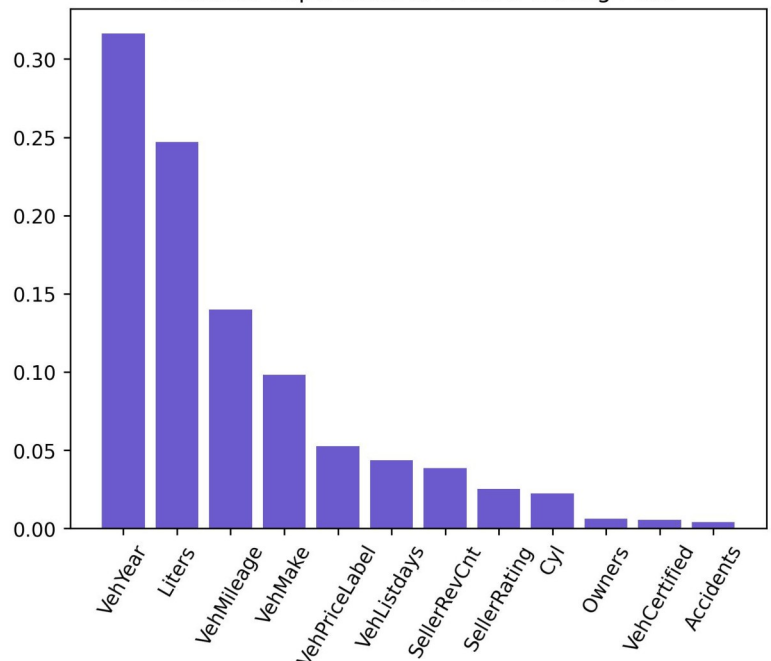
Barplot showing the top 12 features importance scores of the predictors selected for the dealer listing price model.
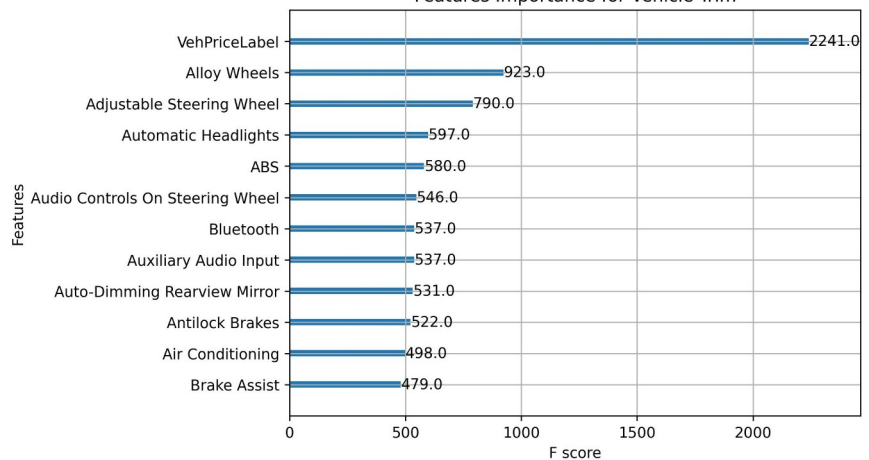
Barplot showing the top 12 features importance scores of the predictors selected for the vehicle trim model.

Next Steps:

1) Models hyperparameters tuning

2) Retrieve more data of low represented vehicle trim classes

3) Apply oversampling techniques to deal with the unbalanced vehicle trim classes