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. The model requires 12 different predictors.

3. Model's metric: $R^2 = 0.82$, meaning that 82% of the variance is described by the model.

4. The features with the highest importance are: vehicle year, engine liters, mileage, and the vehicle maker.

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

2. The model requires 52 different predictors.

3. The model was evaluated on the base of different metrics; precision: 86%, recall: 86%, accuracy: 86%, and F1: 84%.

4. The predictors with the highest importance are: the vehicle price label, the presence of alloy wheels, and other car features.

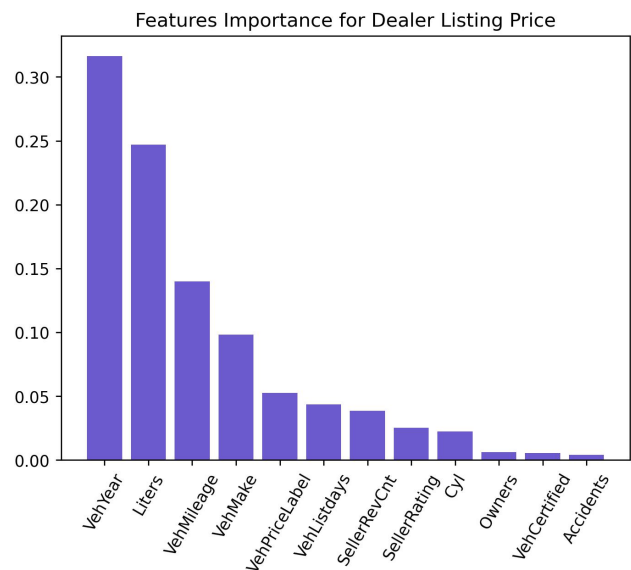V. Ganci, GitHub:
https://github.com/Zipter?tab=repositories

Figure 1: Top 12 features importance scores of the predictors selected for the dealer listing price model.
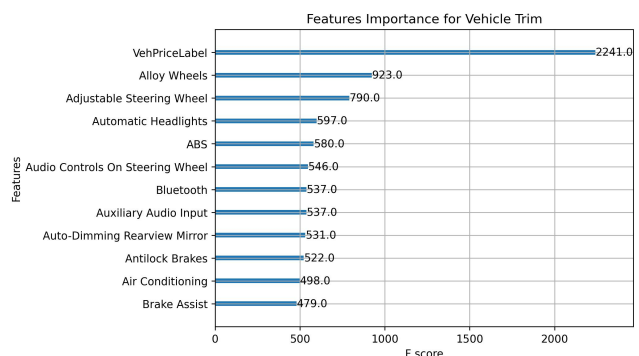

Figure 2: Top 12 features importance scores of the predictors selected for the vehicle trim model.

Next Steps

1. Apply hyperparameter tuning to each model via GridSearchCV to improve the models metrics.

2. Retrieve data of low represented trim classes.

3. Apply oversampling techniques to deal with the unbalanced trim classes.

4. Feature engineering of the geographical data with additional data, for example the average income per city, to check if the dealer listing prices depend on the sellers location.