Ali Hushyar

11.1 What Drives the Price of a Car?

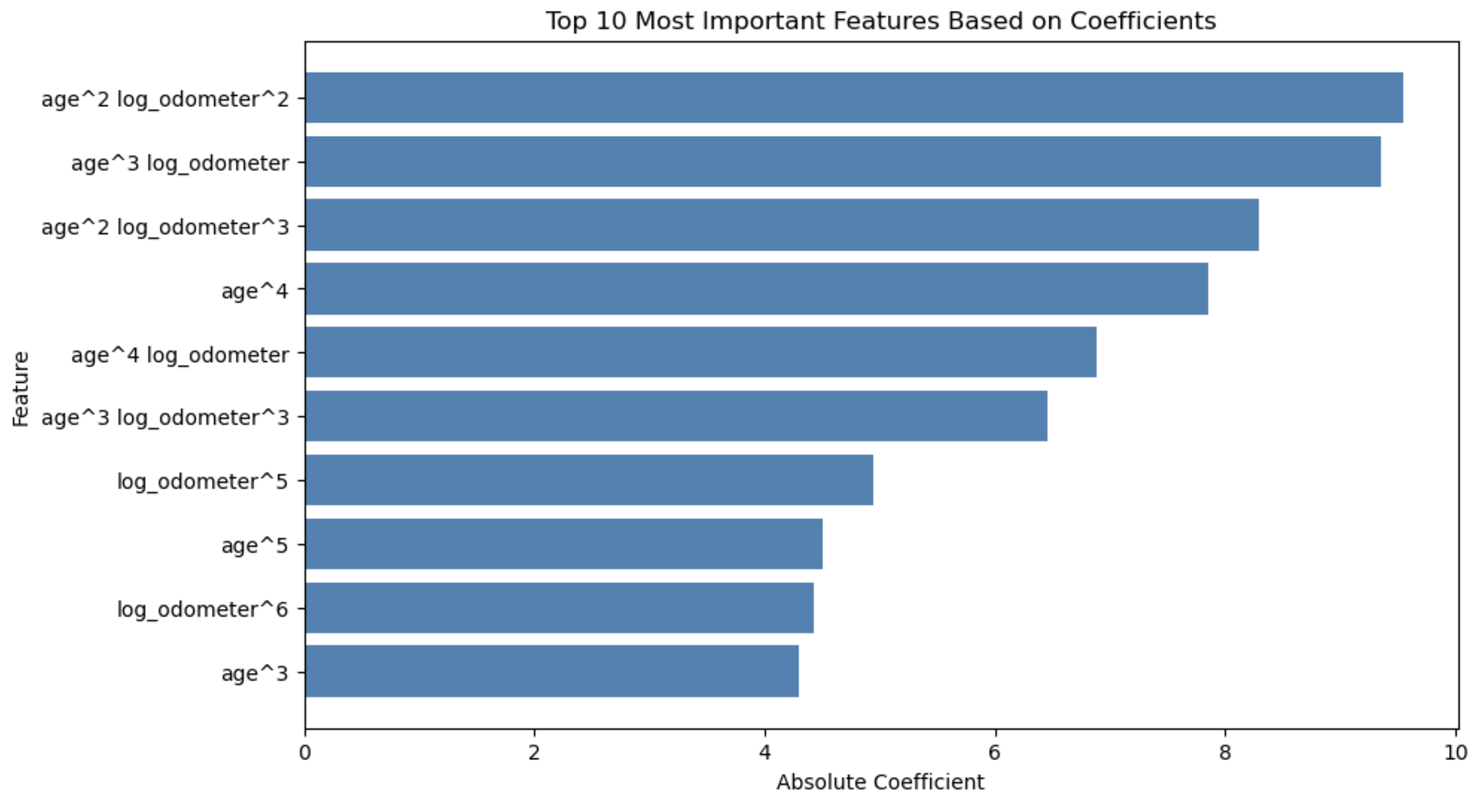
The problem statement here is to understand what features from the given dataset are the best predictors for determining the selling price of a used car. The dataset contains over 400K records with varying data quality.

The first step was to understand the nature of the data and filter out records that would not contribute to a working model. The approach here is to first focus on records with a positive price target. I then removed columns that I felt would not contribute to the model given my domain knowledge of the topic. Features like VIN, paint color, state and region were deemed to be weak predictors and hence removed. All other features were left for modelling purposes. Out of the original 400K dataset, I ended up with approximately 140K records for the linear regression.

I noticed that the numerical features include age and odometer and these did not show any linear relationship with the price target. I log transformed these columns in order to see a linear correlation. All other attributes were categorical with the “condition” feature being ordinal in nature. Other categorical attributes were one hot encoded.

I then performed two GridSearchCV linear regressions. The first was standard linear regression using Lasso for feature selection. The second was a Ridge regression model. For numerical features, both models employed polynomial features and then applied standard scaler to them. I found both of these yielded similar results using test MSE as the evaluation metric.

Using the Ridge model, I outputted the features and associated coefficients that contributed the most to the prediction model. Polynomial interaction features combining age and odometer had the most impact. See below:



Regarding categorical attributes, coefficients shown the following impact:

* **cyl\_category\_High:** (-1.170) Strong categorical impact.
* **fuel\_diesel:** (0.352) Notable impact from this categorical feature.