ID5059 Knowledge Discovery and Data Mining. Assignment 1

Erna Kuginyte 220013309

Aims

The aims of this assignment were to clean the provided cars dataset, investigate it, select a small number of important variables, construct a few simple machine learning models to then predict the numerical attribute of listing prices, and perform an appropriate evaluation of the results of those models. My code ended up being quite extensive as I wanted to explore as much as I can of both the machine learning principles and the Python language itself.

Introduction

The multivariate data set of cars (66 variables) comprises information about both new and used cars, including variables such as cars' features, manufacturing date, and the target variable, price. The data wrangling process involved removing redundant variables, dropping variables with more than 25% of missing values, filling most of the remaining missing values with mean values, filling some missing variables according to another variable's values, and transforming character variables into categorical and numerical types; additional variables such as such as "savings_per_day" were also considered. The initially selected variables for the data set have at least 0.15 correlation with the price variable, a total of 12 explanatory variables.

As the categorization was completed during the data wrangling process, the pipeline that transforms the training and test data sets only handles numerical values. Since the original data sets were not randomly divided, it would be inappropriate to train the machine learning model on one data set and test it on another, therefore, the downloaded chunk of medium size data was split into training and testing sets. The sets are split to have equal representation across the price categories. The price categories are defined by quantiles (*see fig. 2*). Although variables are heavily skewed, this statistical analysis aims to predict cars listing prices, which is why the variables have not been transformed to log scale.

Model Fitting and Selection

Models that were fit: linear, regression with automated variable selection, ridge, lasso regression models, and decision tree, random forest models.

Linear, ridge, and lasso regression models require assumptions such as multivariate normality, a linear relationship between predictors and the prediction variable, minimal multicollinearity, no autocorrelation, and homoscedasticity. In this analysis, the normality assumption is not met, as evidenced by the histograms of the data in the exploratory analysis section of the code. Although transforming the data to an exponential log scale could potentially improve normality, this approach was not explored in this study. The issue of multicollinearity has been addressed. However, the assumption of autocorrelation has not yet been evaluated and should be considered in future analyses. For both Lasso and Ridge models grid search has been applied to find the optimum alpha values, a higher alpha value causes more shrinkage and a simpler model with smaller coefficients, whereas a lower alpha value causes less shrinkage and a more complex model with larger coefficients.

Decision Tree and Random Forest (fine-tuned by using the grid search approach) models have been fit. The evaluation shows poorer performance as the mean cross-validation score is great, but the standard deviation at least doubles the mean value; both RMSE and MAE are extremely high (*see fig. 2*) indicating poor fit.

Final Model

The top-performing Lasso regression model with regularisation parameter alpha = 0.873 includes variables back legroom, engine displacement, fuel tank volume, height, highway fuel economy, horsepower, length, mileage, previous owner count , sales person id, width (dropped 2 variables from the initial selection). It exhibits overfitting, as indicated by an R-Squared value nearing 1 and a Root Mean Squared Error (RMSE) higher than the Mean Absolute Error (MAE), as well as it is visually visible (*see fig. 3*). Despite the mean cross-validation score (CI [0.73957492, 0.87656736]) being lower than other models', it is still greater than the recommended threshold of 0.7. The Akaike Information Criterion (AIC) is substantially lower compared to other linear models. The Lasso model's residuals exhibit strong heteroscedasticity indicating poor goodness-of-fit (*see fig. 4*).

Clearly, more models should be considered as the current best model performance is still quite poor, this includes better selection of prediction variables and their scale. Another aim would be to train and test the large data sets.

Appendix

	Linear Regression	**Linear Regression Reduced**	**Ridge Regression regularisation alpha = 0.001**	***Lasso regularisation alpha = 0.873***	**Decision Tree**	**Random Forest**
R-squared	1.000	1.000	0.999	*0.999*	0.999	0.999
RMSE	3.227e-11	1.752e-11	0.001	*0.811*	368.518	368.518
MAE	2.417e-11	1.317e-11	6.902e-4	*0.557*	16.283	16.283
Mean CV	2.814e-11	2.266e-11	0.0017	*0.910*	2369.550	2121.080
CV standard deviation	4.872e-12	6.766e-12	0.001	*0.330*	4805.413	4770.098
AIC	-1032667.508	507314.599	-291209.376	*3818.323*	-	-

Figure 1. Model Performance

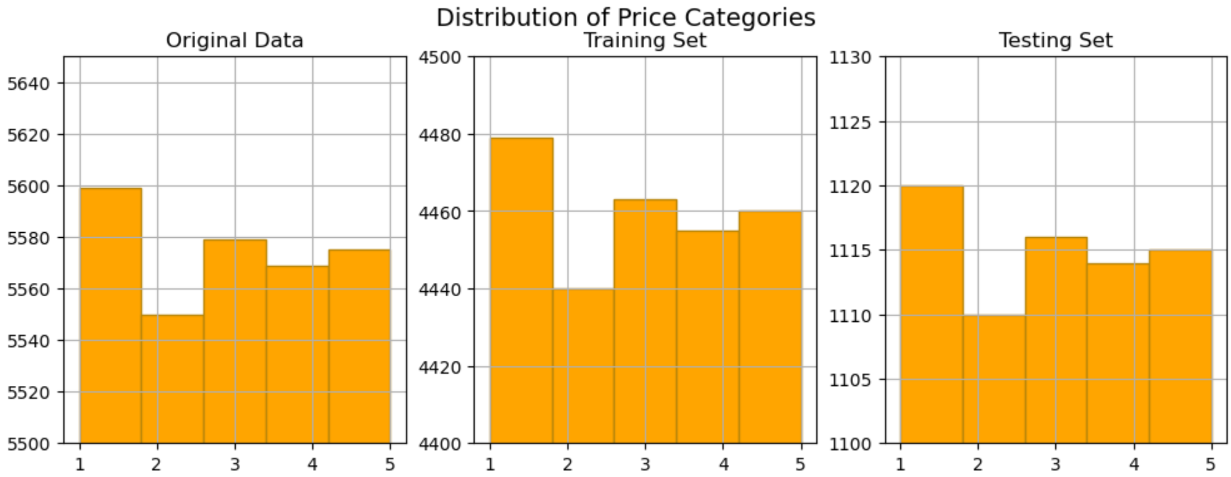

Figure 2. Distribution of Price Variable Across Data Sets

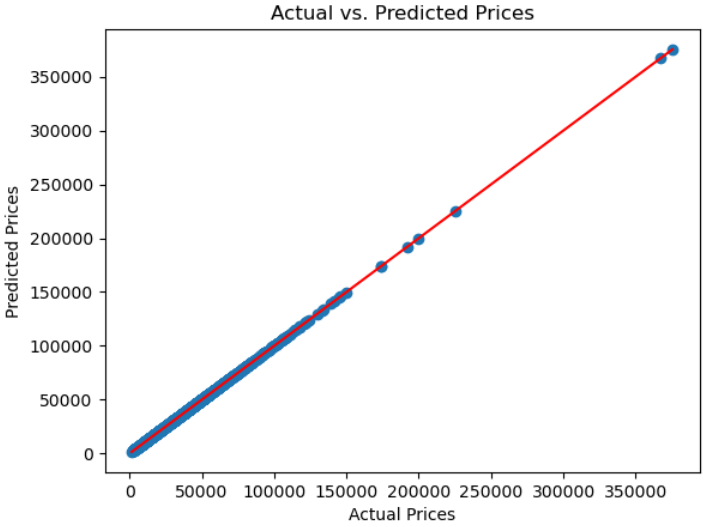

Figure 3. Observed Price Variable against Fitted

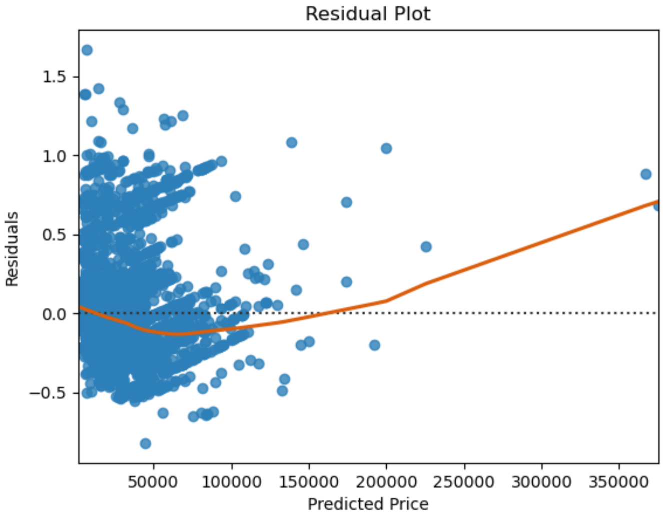

Figure 4. Residual Plot against Predicted Prices