**Practical Application Assignment 11**

***What Drives the Price of a Car?***

This exercise is based on a Used Car dataset from Kaggle (seen in data folder). CRISP-DM process provides a framework for working through this problem and is employed as below.

**Business Understanding**

The objective of this task is to identify what attributes of used cars (e.g., year, condition, odometer reading, model, manufacturer, fuel type, transmission type, number of cylinders etc.) impact the price of used cars in the US. This identification would be helpful in providing recommendations to the client, i.e., a used car dealership, for what customers value most in used cars that make them more or less expensive. The used car dealership could use this information to understand customer preferences and optimize their inventory, sales, and profit.

**Data Understanding**

The dataset is used cars dataset with 426,880 rows and 18 columns. The columns are : id, region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission, VIN, drive, size, type, paint\_color, and state. Most of these columns are of type object (i.e., are categorical), and only 4 columns -id, price, year, and odometer- are of type integer or float (i.e., numerical).

It is understood the price column will be the predictand and the rest of the columns will be predictors. Thus, it is a regression problem.

The data is then inspected for number of null-values (ranging from as low as 0.28% in the column named year to 71% in the column named size), which are examined carefully in different columns and then removed or replaced.

**Data Preparation**

The data is cleaned by inspecting different columns, and removing columns that are not relevant as a predictor of price, like id and VIN. Column named region is also removed because it is superfluous with another column named state.

Numerical columns of price, year, and odometer are inspected and cleaned in terms of outliers. Price values have a huge range and a non-normal distribution, and it is decided to transform it to log\_price (for a somewhat normal histogram/distribution). Correlation matrix of these numerical variables also shows the correlations are better with log\_price than price.

Some variables/columns have very few missing values(<4%), so their corresponding rows of missing data are dropped. Size has >70% of the data missing, so the size column is dropped. Next, all columns are examined with respect to price-this helps in understanding which variables have an impact on price visually, and see if columns with 20-40% of missing values (condition,cylinders,drive,type,size) should be kept or removed or replaced. This also helps in deciding for finalizing predictors for the modelling step.

Condition, cylinders, drive, and type have >20% missing values. These missing values are imputed by filling NaNs with their respective modes.

**Modeling**

The predictors’ data primarily contains categorical variables that are believed to have major impact on price. Category encoders are considered for use, and the JamesSteins Encoder is employed. But prior to encoding, the data is split into train and test for hold-out cross validation. Since there is sufficient data, this method of cross-validation would be sufficient for model accuracy evaluation, however, at the end, additional k-fold validation is also performed.

Heatmap of encoded data showed that model, odometer, year have the highest correlations with log\_price, followed by type, transmission, and manufacturer. There is some multicollinearity as some variables are correlated with each other, so regression models that can handle multicollinearity (like ridge and lasso) are also run.

Different regression-based machine learning models were explored for modeling price of used cars as a function of different attributes. To start off, a baseline simple linear regression model was run, followed by polynomial regression of higher degrees, and then regression models that can handle multicollinearity, with their hyperparameters optimized. Here is a list of the models explored:

1. Simple Linear Regression
2. Regression with Polynomial Features

The best polynomial regression chosen was the one with degree equal to 3 (degree 4 gave an error similar to degree 3 but was way more computationally expensive)

1. SFS-selected Polynomial Regression

Was run for degree=3 polynomial features, and its optimization selected top five features.

1. Ridge Regression

Was run with degree 3 polynomial features and selected optimal hyperparameter alpha of 0.001. The top features selected are a function or some higher order or interaction term including year, odometer, model, type, drive, and condition.

1. Lasso Regression

Was run with degree 3 polynomial features and it selected an optimal hyperparameter alpha of 0.001.

1. Lasso-based SFS-selected Polynomial Regression

The Lasso-based SFS polynomial regression was run with alpha of 0.01 and it selected top five features.

**Evaluation**

Below are the Validation Mean Squared Errors (MSEs) for the 6 models run above:

1. Simple Linear Regression : 0.17

2. Polynomial Regression : 0.12

3. SFS-based Polynomial Regression : 0.18

4. Ridge Regression : 0.12

5. Lasso regression : 0.16

6. Lasso-based SFS Polynomial Regression: 0.19

Based on the hold-out validation method, validation MSEs are lowest in: 1. Polynomial regression of degree 3 model, and 2. Ridge regression with polynomial degree 3 and alpha 0.001. Of these two, Ridge regression was more computationally efficient.

Lastly, k-fold validation was performed for the Ridge model, and the k-fold cross validation MSE of 0.127 was similar to hold-out cross validation of 0.125. The statistical measure that represents the proportion of the variance for a dependent variable that is explained by independent variables in a regression model (R-squared) for this best model is 81.3%.

The permutation importance of this best model (using R-squared as a scoring metric) is:

A picture containing table

Description automatically generated

**Deployment**

From permutation importance of the best model, the following are the top variables (in order) that determine the price of a used car and can be informed to the client/used car dealer. These top features were consistently regarded as the most important features from other regression models explored as well.

- Model

- Year

- Odometer

- Fuel

- Condition

- Cylinders (number of cylinders)

- Transmission and Type

**Recommendations**

Model of the car, year that the car was from, and the odometer reading have the highest impact on the used car price. It is recommended that the client focuses on showcasing cars that are more recent, have driven lower miles, and belong to models that are considered premium like Porsche 911, Chevrolet corvette, Lamborghini huracan, etc. In terms of fuel, which is the next important feature, electric and diesel cars sell higher than other types, and in terms of condition of the car, customers pay higher for new, like new, and good condition cars. Cars with 8 and 10 cylinders are priced higher, while cars with 5 cylinders the lowest. Lastly, customers pay higher for automatic or other types of transmission than manual; and pay lowest for mini-van type of cars.

**Next Steps**

With additional data collection over the years, the models can be refined and can make better predictions over time. Other forms of imputation (for missing values) like iterative imputer could be tried in the future to see if it improves the model results. Additionally, other machine learning models like Extreme gradient boosting, decision tree regression, Random Forests etc. would be utilized in the future to evaluate their accuracy, and compare them against the regression models run in this exercise.