**Practical Application Assignment 11.1**

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

The practical application assignment targets identifying key drivers for used car prices.

1. Business Understanding

The Business Understanding focuses on finding out the objectives and requirements of the project. The objective of this capstone project is to identify the key factors that drives used car prices, which helps to develop the criteria of business success by providing recommendations to used car dealership.

Achieving this objective requires applying the CRISP-DM reference model to the available resources, in this case the used car dataset. From a data mining perspective, the data mining project definition means to build a model that can be adopted by used car dealership to predict used car prices based on these key factors identified. This fosters the establishment of strong business understanding that is essential for a winning business.

1. Data Understanding

The data understanding is closely linked with Business Understanding. This phase starts with initial data proceeds to get familiar with the data, to discover first insights into the data, and to detect interesting subsets to form hypotheses.

After I explored the dataset, I found that that many data columns have missing items or unrealistic values. Some of the rows can be dropped if the missing items are relatively smaller, but if a feature contains too many missing items, dropping every rows may significantly reduce the size of the dataset which makes the analysis not as accurate as it should be.

Based on these findings, I decide to apply a multi-step strategy. The first step is to handle the missing items in several of the columns. The second step is to use *LabelEncoder()* to encode features from non-numerical values to numerical ones. The last step is to remove some unrealistic data as well as features that are not closely related to the prediction of the used car prices.

One final thought is that this dataset is not time series data, because it does not carry a sequence of data points indexed in time order. Because these data points do not consist of successive measurements made from the same source over a fixed time interval, so they can't be used to track change over time.

Before starting the modeling, I also created the visualization of the heatmap of the used car’s feature correlation as shown below. The heatmap shows that the top 6 features that are highly correlated with ‘price’ are ‘odometer’, ‘transmission’, ‘year’, ‘size’, ‘condition’ and ‘model’.

A chart with numbers and a number of numbers

Description automatically generated with medium confidence

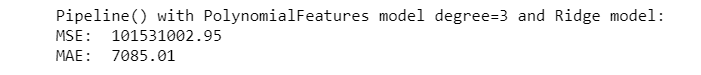
1. Modeling

For modelling, I first performed the linear regression using the six features of 'year', 'manufacturer', 'condition', 'odometer', 'model', 'size' to predict 'price'. The MSE and MAE produced from the linear regression model is shown below.

A close-up of a text

Description automatically generated

Next, I built *Pipeline()* with PolynomialFeatures model degree=3 and Ridge model, still using the six features of 'year', 'model', 'condition', 'odometer', 'transmission', 'size' to predict 'price'. The MSE and MAE produced from the pipeline with ridge model is shown below. The results show that the Ridge regression produces slightly lower MSE and MAE than the linear regression model does.



Finally, I splitted data into training and testing in order to perform cross-validation. For modeling, I iterated the ‘k’ value from 1 to 8 in order to identify the best value for degree complexity. The plot shows the best complexity that minimized Test Error is degree=4.

A graph of a line

Description automatically generated with medium confidence

1. Evaluation

Applying the degree=4 identified using Ridge model, I built the best model and fitted the training data. The best model scores 0.48 using test data for cross validation.

After that, I conducted permutation importance and found out the top 5 features that drives the used car price are 'year', 'odometer', 'size', 'transmission' and 'condition' as shown below.

A number with numbers and symbols

Description automatically generated with medium confidence

Lastly, I examined the relationships of the 'price' with ''odometer' and 'model' respectively which display the plots below.

A blue and green lines

Description automatically generated with medium confidence

1. Deployment

The price of used cars can be affected by many different factors. In this practical application assignment, I identified that the top 5 features that drives the valuation of used vehicles are 'year', 'odometer', 'size', 'transmission' and 'condition'.

Based on these findings, I would advise the used car dealership to treat the age of the vehicle and mileage on the vehicle as the most important factors, followed by size of the vehicle, the transmission type of the vehicle as well as the vehicle condition.

In addition, other factors that are not included in this study may affect the used car price as well, with supply and demand being one critical factor.