**Understanding Factors Affecting Used Car Prices**

**Executive Summary**

This report outlines findings from an analysis of a dataset including 426K used cars. Factors that significantly influence the prices of used cars are identified and can be used to fine-tune prices of dealers’ inventories.

**Data Analysis**

The dataset includes some numerical features and categorical features. Numerical features include ID, year, odometer, and price. Their distributions are visualized below.

A graph of a diagram

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

**Data Preparation**

Outlier prices and odometer readings are filtered out.

A graph with blue and black text

Description automatically generated

A diagram of a diagram

Description automatically generated

Pearson correlation for numerical features are calculated. Ranked Numerical Features by Correlation Strength indicating a degree of association between the variable and price:

* odometer 0.414961
* year 0.240999

ANOVA for each categorical feature is calculated. Ranked Categorical Features by ANOVA p-value, indicating whether observed differences among group means could be due to specific variable being studied or if they are just due to chance:

* region: F-Statistic = 17.11, p-value = 0.00
* model: F-Statistic = 8.37, p-value = 0.00
* condition: F-Statistic = 1305.45, p-value = 0.00
* cylinders: F-Statistic = 1198.19, p-value = 0.00
* fuel: F-Statistic = 2747.30, p-value = 0.00
* drive: F-Statistic = 5651.68, p-value = 0.00
* size: F-Statistic = 1999.38, p-value = 0.00
* type: F-Statistic = 1719.88, p-value = 0.00
* paint\_color: F-Statistic = 304.11, p-value = 0.00
* state: F-Statistic = 56.00, p-value = 0.00

**Modeling**

Three models are built using different training data or algorithms.

1. Model 1 – linear regression model trained using only two numerical features, odometer and year.

Model Evaluation Metrics:

Mean Squared Error: 98024874.99423555

R^2 Score: 0.21155398282392646

A graph showing the difference between price and price

Description automatically generated

A graph showing a plot of residuals

Description automatically generated with medium confidence

1. Model 2 – linear regression model trained using the two numerical features and the above listed categorical features.

Model Evaluation Metrics:

Mean Squared Error (MSE): 17041510.93282232

R-Squared (R2): 0.8629295735145145

A graph showing a line of blue dots

Description automatically generated with medium confidence

A graph showing a plot of residuals

Description automatically generated with medium confidence

1. Model 3 – random forest regressor model trained using the two numerical features and the above listed categorical features.

Model Evaluation Metrics:

Mean Squared Error (MSE): 17041510.93282232

R-Squared (R2): 0.8629295735145145

A graph showing the difference between price and price

Description automatically generated

A graph showing a plot of residuals

Description automatically generated with medium confidence

1. Model 4 – grid search for the random forest regressor model.

Model Evaluation Metrics:

Mean Squared Error: 20940511.22238466

R^2 Score: 0.8315686434500327

A graph showing the difference between the price and the price

Description automatically generated

A diagram of a plot

Description automatically generated

**Evaluation**

R-squared ranges from 0 to 1, where 0 means that the model does not explain any of the variability of the response data around its mean, and 1 means it explains all the variability. As shown above, model 3 is much better than model 2, which is much better than model 1. Interestingly, model 4’s R-Squared is decent, but it does not match up to model 3.