**Project Car Prices**

**Rowland Savage June 25th 2024**

**Business Understanding**

The goal of this analysis is to understand what factors make a car more or less expensive. A number of clear recommendations are made in the “Results” section below for the client, a used car dealership, as to the key factors consumers value in a used car.

In addition, we have developed a model to predict the price of a any given car given the other factors available. Several techniques were used, and the various outcomes discussed below (Random Forrest produced the most accurate prediction)

**Data Understanding**

The datafile provided (Vehicles.csv) had 426,880 rows of data on used car sales, and contained the following information;

* id int64
* region object
* price int64
* year float64
* manufacturer object
* model object
* condition object
* cylinders object
* fuel object
* odometer float64
* title\_status object
* transmission object
* VIN object
* drive object
* size object
* type object
* paint\_color object
* state object
* dtype: object

**Data Preparation**

In order to clean the provided data in order to make the most effective use of it during the modelling stage the following actions were performed.

-

- Rows removed by removing top 1% price outliers: 4265

- Rows removed where price is zero: 32895

- Rows removed where odometer is missing: 2244

- Rows removed where year is missing: 994

- Rows removed where transmission is unknown 1496

- Rows removed where manufacturer is unknown: 14546

- Rows removed where fuel is unknown: 1888

- Rows removed where title status is missing: 6530

- Rows removed where model is missing: 4396

- Set missing type values to 'unknown'

- Removing rows where 'cylinders' equals 'other'

- 'Cylinders' values, removed text and converted to numeric

- Filled missing paint color values with 'unknown'

The initial data file was had 426,880 rows, and after the cleaning we were left with 211,345 to complete the analysis.

**Results**

After analyzing the attached data the factor most correlated with the vehicles price is transmission type “other” and the car’s year. Vehicles with more cylinders are more valuable as are pickups more than trucks and trucks, in turn are more expensive than coupes.

Dealers should pay particular attention to cars that are in “Good” condition as this impacts price even more than vehicles that are “like new” or “new”. White is the only color that stands out as having a positive impact on the price.

As for manufacturers, dealers should favor these in the following order

1. Ram
2. Ford
3. Chevrolet
4. Audi
5. Jeep
6. Infiniti
7. Porsche
8. Lexus
9. Mercedes-Benz

Conversely, dealers should avoid hatchback vehicles, and Honda as a manufacturer. Higher odometers are correlated with lower prices as are gas vehicles, sedans and vehicles with front wheel drive.

The below table shows all the major factors;

|  |  |
| --- | --- |
| Factors | Correlation |
| transmission\_other | 0.39529311 |
| Year | 0.381936 |
| cylinders | 0.3323686 |
| type\_pickup | 0.25855806 |
| type\_truck | 0.179219 |
| manufacturer\_ram | 0.164772 |
| type\_other | 0.16218604 |
| condition\_good | 0.13177163 |
| paint\_color\_white | 0.11871704 |
| fuel\_other | 0.11417593 |
| manufacturer\_gmc | 0.1006774 |
| drive\_rwd | 0.07409509 |
| size\_full-size | 0.07175712 |
| type\_coupe | 0.06179007 |
| manufacturer\_ford | 0.05763246 |
| manufacturer\_chevrolet | 0.05428235 |
| manufacturer\_audi | 0.04820479 |
| manufacturer\_jeep | 0.04692925 |
| manufacturer\_infiniti | 0.04132107 |
| manufacturer\_porsche | 0.03926437 |
| manufacturer\_lexus | 0.03513549 |
| manufacturer\_mercedes-benz | 0.03305714 |
| type\_hatchback | -0.1070372 |
| manufacturer\_honda | -0.1283289 |
| condition\_fair | -0.1312539 |
| Odometer | -0.176007 |
| fuel\_gas | -0.1968016 |
| type\_sedan | -0.2330994 |
| drive\_fwd | -0.3087462 |

**Modelling Step 1 – initial analysis**

In order to evaluate the data, I began with an initial analysis, using some of the simpler techniques, the outcomes of which is shown below;

**Simple Linear Regression**

A graph with a red line and blue dots

Description automatically generated

Simple Linear Regression RMSE: 12877.72025454713

**Multiple Linear Regression**

A blue dotted line graph

Description automatically generated

Multiple Linear Regression RMSE: 10764.930126003925

**Singular Value Decomposition**

A graph of blue dots

Description automatically generated

**KMeans**

A chart with different colored dots

Description automatically generated

**Principle Component Analysis**

A graph of blue dots

Description automatically generated

**Modelling Step 2 – More complex analysis**

However, I wasn’t satisfied with the results of the initial analysis and employed the following techniques to build a more accurate model to predict a vehicle’s price.

**Ridge Regression**

A graph of blue dots and a black line

Description automatically generated

* Ridge Regression - Mean Absolute Error (MAE): 7798.717586680765
* Ridge Regression - Mean Squared Error (MSE): 108620008.3525037
* Ridge Regression - Root Mean Squared Error (RMSE): 10422.092321242588
* Ridge Regression - R-squared (R2): 0.3498023322790099

**Lasso Regression**

A graph showing the difference between a line and a line

Description automatically generated with medium confidence

* Lasso Regression - Mean Absolute Error (MAE): 7805.578444765261
* Lasso Regression - Mean Squared Error (MSE): 108142903.88667749
* Lasso Regression - Root Mean Squared Error (RMSE): 10399.178038993154
* Lasso Regression - R-squared (R2): 0.3526582721343332

**Random Forrest**

A graph showing a line of blue dots

Description automatically generated with medium confidence

* Random Forest - Mean Absolute Error (MAE): 3574.082382874411
* Random Forest - Mean Squared Error (MSE): 35694741.27733562
* Random Forest - Root Mean Squared Error (RMSE): 5974.507617982893
* Random Forest - R-squared (R2): 0.7863318381167035

**Support Vector Regression**

A graph with a line and a dotted line

Description automatically generated

* Support Vector Regression - Mean Absolute Error (MAE): 9616.171497544146
* Support Vector Regression - Mean Squared Error (MSE): 158330499.45352983
* Support Vector Regression - Root Mean Squared Error (RMSE): 12582.944784649173
* Support Vector Regression - R-squared (R2): 0.05223611160391073

**Gradient Boosting Regression**

A graph showing a blue line

Description automatically generated with medium confidence

* Gradient Boosting Regressor - Mean Absolute Error (MAE): 4086.0741694742915
* Gradient Boosting Regressor - Mean Squared Error (MSE): 41838296.0718468
* Gradient Boosting Regressor - Root Mean Squared Error (RMSE): 6468.252938146961
* Gradient Boosting Regressor - R-squared (R2): 0.7495566153976638

**Conclusion**

The most effective technique for predicting a cars price from the information provided was using a Random Forest with a Mean Absolute Error (MAE) of 3574.082382874411.