**What drives the price of a car?**

The main objective of this application is to analyse and identify the factors that impact car prices. By conducting this analysis, we will provide valuable recommendations to our client, a used car dealership, regarding the features and qualities that consumers value in a used car. This will help the dealership make more informed decisions and improve their understanding of customer preferences.

The application uses Vehicle Dataset from Kaggle, which contains around 426K samples of used cars. In this application, we have explored various aspects of the dataset, such as the car's manufacturer, model, year, mileage, condition, and other relevant attributes. It is important to keep in mind the specific needs and objectives of the client to ensure that the final recommendations are relevant, practical, and aligned with their goals.

We used industry standard process, named CRISP-DM framework, to work through a data problem.

1. **Business Understanding**

In the "What Drives the Price of a Car" project, we have analyzed a subset of the Vehicle Dataset from Kaggle, which contains around 426K samples of used cars. The main objective of this analysis is to identify the factors that impact car prices. By conducting this analysis, we will provide valuable recommendations to our client, a used car dealership, regarding the features and qualities that consumers value in a used car. This will help the dealership make more informed decisions and improve their understanding of customer preferences.

Throughout the analysis, we will explore various aspects of the dataset, such as the car's manufacturer, model, year, mileage, condition, and other relevant attributes. It is important to keep in mind the specific needs and objectives of the client to ensure that the final recommendations are relevant, practical, and aligned with their goals.

Thank you Kaggle for allowing us to use your dataset <https://www.kaggle.com/datasets> .

1. **Data Understanding: Exploratory Data Analysis (EDA)**

* Load and Read DataSet
* Understand Features and Datatype
* Analyse Data and Statistics

About Dataset

The vehicles dataset contains total 18 features/variables including target variable named price. Here's brief about each feature:

* id: Unique identifier of each sample
* region: Region where vehicle belongs to
* price: Price of a Vehicle
* year: Year of a vehicle
* manufacturer: Manufactures of a vehicle i.e. Ford, Toyota etc.
* model: Model of the vehicle i.e. Elantra, Camry etc.
* condition : Condition of a Vehicle
* cylinders : No. of Cylinders in a vehicle
* fuel : Fuel type of a vehicle i.e. Gas, Electric etc.
* odometer : No. of Kms/Miles vehicle has driven
* title\_status : Title of a vehicle
* transmission : Transmission type of a vehicle i.e. manual/auto etc.
* VIN : Unique Identifier of a vehicle
* drive : Drive type of a vehicle i.e front wheel drive etc.
* size : Size of a car i.e. full-size, compact etc.
* type : Vehicle type
* paint\_color : Color of a Vehicle i.e. white/black etc.
* state : State of a vehicle

1. **Data Preparation**

Following steps are done for Data Preparation:

- Break down the data to cat\_col(categorical columns) and num\_col (numerical columns)

- Handle Missing Values and Feature Encoding:

- encode the categorical values (ignoring null values)

- imputing the missing values of the numerical columns with a BayesianRidge estimator

- Removing the records which doesn’t have the VIN

- imputing the missing values of the categorical columns with a BayesianRidge estimator

- Handle Outliers, correct data types

- Handling inconsistent data (example: price shouldn't be less than 1)

- Visualize cleaned Data

- Drop unnecessary features that don't have impact on the price (i.e. 'VIN', 'paint\_color','state','id') & duplicates

1. **Modeling**

We have analysed the data against various regression models through different hyperparameters and perform cross validation to determine the best suited model for a vehicle dataset. The models used in current applications are:

* Linear Regression Model with Polynomial Degree
  + Linear Regression (Best Polynomial Degree = 2) .

A graph with blue and white bars

Description automatically generated

A graph with numbers and lines

Description automatically generated

* Lasso Regression Model with Feature Selection =3

A graph with blue bars

Description automatically generated

A graph with numbers and lines

Description automatically generated

* Linear regression model using feature selection using GridSearchCV

A graph with blue bars

Description automatically generated

A graph with blue and orange bars

Description automatically generated

* Ridge Regression Model using best alpha: Best Alpha = 0.001

A graph with blue bars

Description automatically generated

A graph of blue and orange bars

Description automatically generated

RMSE and R2 for the above models are summarized in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Top Feature Importances | RMSE | R2 |
| Linear Regression with polynomial(degree=2) | odometer, year, drive, fuel | 3428.0065 | 0.3599 |
| Lasso with Feature Selection=3 | Year, cylinders, transmission | 3911.6323 | 0.1665 |
| Linear Regression with feature selection using GridSearchCV | odometer, year, transmission, cylinder | 3773.9272 | 0.2242 |
| Ridge regression Model using best alpha | odometer, year, transmission, cylinder | 3773.9272 | 0. 2242 |
|  |  |  |  |

1. **Evaluation**

RMSE gives you an absolute measure of prediction error, whereas R2 gives you a relative measure of how well the model explains the variability in the data.

**Summary of the Best Model**

* **Model**: Linear Regression with Polynomial (degree=2)
* **RMSE**: 3428.0065
* **R2**: 0.3599
* **Top Features**: odometer, year, drive, fuel

This model outperforms the others based on the evaluation metrics provided.

Once we have determined the best suited model (Linear Regression with Polynomial (degree=2)), we are going to perform following:

* Predict the Price against Test Dataset (Actual vs Predicted)
* Model Indicators i.e., MSE, RSME, R2 Score
* Coefficients and Intercept
* Visualize Linear Regression on Actual Vs Predicted
* Visualize Price against each Independent Variables

A red and blue line graph

Description automatically generated

**Interpret the Coefficients and Y-Intercept**

* Coefficient Value: The value of each coefficient represents the expected change in the target variable for a one-unit change in the corresponding feature, assuming all other features remain constant.
* Intercept: The intercept represents the expected value of the target variable when all features are zero.

Output: Coefficients: [ 1.66567916e+02 -1.96145542e+04 2.16293311e+03 -3.82171282e+00

-4.87551262e+04 -1.67054386e+04 3.66772993e+04 2.00596658e-01

3.05811019e+03 8.12985632e+03 2.11986307e+04 3.15431355e+04

-7.98459748e+03 9.92538529e-03 -8.52970063e-02 -1.91223616e-02

-2.69490799e-06 8.30536699e-01 -3.30579079e-01 9.75831916e-01

1.62969633e-06 2.74532539e-01 -2.12260468e-02 -9.89477677e-02

-2.02462616e-01 -2.89033056e-02 4.93961513e+00 -1.08420185e+00

1.82306181e-03 2.42699769e+01 8.28137532e+00 -1.96774510e+01

-1.00135609e-04 -1.45330525e+00 -4.06628480e+00 -1.20287379e+01

-1.54674878e+01 4.04554005e+00 3.66692083e-01 -9.09812545e-04

1.49716563e+00 5.87097990e+00 1.47101686e+00 2.00555187e-05

1.14364702e+00 -1.02949177e+01 1.27430102e+01 -4.76131290e+00

-1.00701647e+00 3.06986476e-06 -2.37636114e-03 8.59221182e-03

1.08771947e-02 -2.41379894e-09 -1.91024115e-03 -3.02892475e-03

-7.84090498e-04 4.26379818e-03 3.19121334e-03 2.72567504e+01

3.73413257e+01 -1.64374071e+02 -2.95855349e-04 -5.61747521e+01

5.56022484e+02 -5.13206625e+01 3.48083375e+01 -8.06387776e-01

8.38087040e+01 1.65511840e+01 -1.49148371e-03 2.64781940e+01

6.11455597e+01 -1.55374462e+02 -6.52617801e+01 -9.24328372e+00

3.31782397e+02 -2.70074323e-03 1.66831823e+01 1.42600147e+02

1.10500478e+02 6.30541079e+02 -3.91168859e+00 1.04773790e-09

-5.28714554e-06 -1.81008711e-03 -3.29684528e-04 -2.35452659e-03

4.62219770e-04 -6.77152252e+01 -1.10875100e+02 -1.01317996e+01

-1.14820272e+01 -1.26617022e+01 -4.13415592e+02 9.95646547e+00

5.60287701e+02 -2.27320740e+01 1.93395885e+03 -8.25278948e+01

-5.11560207e+01 -4.18140470e+02 -3.53657694e+00 -1.34711465e+01]

Intercept: 19477294.935316425

1. **Deployment**

* Save the trained model to a file for future use.
* Load the saved model and use it to make predictions on new data.

**Top Features Influencing Car Price**:

* **Year**: The model suggests that the year of the car is the most significant factor influencing its price. This is expected as newer cars tend to be more expensive.
* **Drive**: This feature also plays a significant role in determining the price. Different drive types (e.g., 2WD, 4WD) can affect the price.
* **Odometer**: The mileage of the car is another crucial factor. Cars with lower mileage typically have higher prices.
* **Fuel**: The type of fuel the car uses is an important factor. Cars that use cheaper or more efficient fuel types may have different pricing.
* **Cylinders and Transmission**: These features also contribute to the price. The number of cylinders and the type of transmission (manual or automatic) can affect the vehicle's price.

**Next steps and Recommendations**

* Year and Odometer plays significant role in Vehicle price. Keep vehicles of recent year and less Odometer value vehicles
* Continue to gather new samples for recent years data, evaluate & tune model to improve price prediction and determines the factors that drives Vehicle Price.
* Analyze the residuals (differences between actual and predicted prices) to identify any patterns that might indicate areas where the model is underperforming.