What Drives the Price of a Car?
Project Summary
Shan Gao

CRISP-DM

Problem Statement	**What drives the price of a car?** Deciding car prices is **a regression problem**, aimed at **predicting a continuous value** based on provided car features.
Business Understanding Summary	**Why?** Conceptually, car prices are determined by crucial **car specifications** such as cylinder count, fuel type, drive type, model, and **manufacturer brands**, complemented by aesthetic factors like paint colors. It also encompasses indicators of **car condition and transferability** such as the general condition, year, and odometer reading, and title status. **What**: Available key features such as the manufacturer brand, year, odometer, model, and car condition are provided by **the dataset** to be significant predictors of a car's market price. These features **largely capture the mechanical state and desirability of a car.** **How**: Insights from predictive models enable car dealerships to **tailor inventory** and marketing strategies to buyer preferences, enhancing business outcomes by **aligning offerings, optimizing sales, and increasing profitability**.
Data Understanding summary	EDA steps (Basic Statistics Check, Report Profiling, and Data Visualization Plotting) helped me identify the following data issues: 1. **Missing VINs:** About 30% of the records lack Vehicle Identification Numbers (VINs), which points to potential issues in data capture, data quality, or the application of privacy filters. This significant gap necessitates extensive data cleanup or imputation, which could potentially skew the insights derived from predictive models. 2. **Prevalence of Missing Values:** A significant number of records have missing entries for attributes such as size, type, paint color, etc. These missing values are often correlated, suggesting systematic gaps in data collection.

3. **High Cardinality:** Attributes like model and manufacturer exhibit high cardinality, presenting challenges for modeling due to the vast number of unique entries.
4. **Imbalance in Categorical Data:** Attributes like car title, etc. might cause model bias, inadequate learning for minority classes, or overfitting to the majority class.
5. **Incorrect feature type:** The number of cylinders
6. **Zero Pricing Issues:** A substantial portion of the dataset includes car records listed with prices of zero (33,000 records), which further indicates the messiness or inaccuracy of the dataset.
7. **Numerical Outliers:** There are extreme outliers present within the numerical variables (price, odometer, etc.), which could potentially skew the analysis and model predictions.
8. **Distribution Skewness:** The distributions of year, odometer, and price are not perfectly normal, indicating potential biases or a need for data transformation in predictive modeling.

Key Data Preparation and Method Used Before Modelling

Consequently, 9 data preparations

1. **Initial Feature Removal:**
   a. Eliminate the VIN, state, region, and ID as they do not impact car prices or demand and display high cardinality.
   b. Remove the 'Size' field due to its high rate of missing values (over 70%).
2. **Handling Missing Values:**
   a. After unsuccessful attempts to improve model performance through recode and regrouping to create better models, I decided to drop all records with missing values, except for cylinder and car condition, which can be imputed using their median. (after splitting the data set)
3. **Addressing High Cardinality:**
   a. Discard the 'Model' field with over 20,000 distinct values. Reorganize manufacturers into a new 'Manufacturer Group' based on value distributions. My rationale is streamlining brand representation without excessive model detail.
4. **Reduce Features with Categorical Imbalance:**
   a. Drop columns with unbalanced categories like transmission and car title.
5. **Adjust Numerical Features:**
   a. Numerically code the 'Cylinder' feature and ordinally encode 'Car Condition' to better fit regression analysis.
6. **Remove Numerical Outliers:**

   a. Exclude records where the price is zero.
   b. Use quantile-based methods to eliminate extreme outliers in prices, odometers, and years.

7. **Feature Transformation:**
   a. Create a 'Log Odometer' feature to correct skewness, handle outliers, and reinforce linear relationships and model performance.
8. **Data Encoding and Standardization:**
   a. Apply one-hot encoding to categorical variables and standardize numerical variables for modeling.
9. **Feature Engineering:**
   a. Employ polynomial transformations to create new features aimed at enhancing model performance.

Section Notes:

I understand now why over 80% of a data scientist's time is dedicated to data cleaning and transformation. The process is an truly challenging and often overwhelming process that nearly had me losing my mind to build a better model.

Modeling Techniques

I applied simple Linear Regression, Ridge with or without Polynomial features, optimizing parameters through grid search to find the best prediction model.

By comparing three different models (price vs. logged price target variables)

1. Simple Linear Regression with Polynomial (5-fold validation)
2. Ridge (simple validation)
3. Ridge with polynomials. (simple validation)

Best Model

My best model is Ridge with Polynomials with logged Target Variables

Model Evaluations

1. **Performance Metrics and Visualization**: Apply residuals and QQ plot viz to verify model reliability and accuracy.
2. **Error Analysis**: Utilize key metrics such as Mean Squared Error and Mean Absolute Error to evaluate model accuracy.
3. **Validation Techniques**: Both Cross-Validation and Holdout Methods

Conclusions/Business Insights

1. **Model Performance**: Polynomial features significantly enhance model accuracy. Log transformations improve data distribution for more reliable predictions.
2. **Key Influencers**: Manufacturer, vehicle type, year, odometer, and drive are pivotal in pricing.
3. **Inventory Trends**: High demand for Lexus, Toyota, trucks, and offroad vehicles. Trucks and pickups maintain premium market positioning.
4. **Consumer Preferences**: High-end vehicles are favored over popular lower-end models. Yellow, often linked to sports cars, commands higher prices.
5. **Risk Management**: Newer American vehicles with four-wheel drive are less risky. Japanese and Korean brands are preferred for their value retention in the secondary market.
6. **Surprisingly**, car conditions and the number of cylinders have shown to be insignificant factors in the dataset. Is it largely because the condition of the car positively correlates with its year and negatively with the odometer readings? Additionally, the imputation of a significant portion of missing data (over 34%) for cylinders and car condition may explain their limited significance to some extent.

Next Steps

1. **Data Collection and Cleaning:** Reduced the dataset from 400,000 to less than 200,000 records by eliminating extensive missing values to enhance model performance.
2. **Feature Removal:** This might cause the model overfitting issue. More business research could help.
3. **Market Demand and Economic Factors:** Plans to integrate market demand and economic data aim to refine valuations by including demand-supply dynamics.
4. **Computational Limitations:** Limited computational resources curtailed extensive model testing; utilized Lasso and polynomial features.
5. **Feature Selection and Model Expansion:** The initial categorization of cars into five price groups underperformed; future models will focus on granular feature selection and robust data sources to boost accuracy.

Final Thoughts/ Project Takeaways

Machine Learning and AI meld the science of data with its artistry. Much time is spent on data gathering, cleaning, and transformation, which require both creativity and technical skill. Though modeling appears straightforward, data preparation is an art, involving randomness and nuanced decision-making.