Car Price Prediction- PAC Report

Report submitted by

Shivani Singh

Uni: ss6973

Data Exploration

The dataset initially consisted of 46 columns, including 19 numeric and 25 categorical columns. During the exploration phase, binary character variables like 'is_cpo,' 'fleet,' 'frame_damaged,' 'has_accidents,' 'isCab,' 'is_new,' 'salvage,' and 'franchise_dealer' were identified among the 46 columns.

Data Preprocessing

- Numerical Data
  Missing values in columns such as 'fuel_tank_volume_gallons,' 'highway_fuel_economy,' 'city_fuel_economy,' 'engine_displacement,' 'torque,' and 'power' were imputed by grouping the data based on model names. Remaining missing values were imputed with the median of their respective columns.

- Categorical Data
  Columns like 'interior_color' and 'exterior color' with more than 60 levels were grouped into the five most common levels. Label encoding was applied to other categorical variables. Columns with around 50% missing data were dropped, and the remaining missing values were imputed with "Not known." Binary 'True/False' data types were converted to numeric variables (1 and 0).

Dropped variables with more than 45% missing values and those less correlated with price (correlation < |0.3|). Dropped columns included 'id,' 'owner_count,' 'exterior_color,' 'interior_color,' 'power,' 'torque,' 'description,' 'make_name,' 'model_name,' and 'trim_name.'

Feature Engineering

The 'listed_date' column was processed to extract the month, and categorical columns like 'power' and 'torque' were further processed to separate values into numeric columns.

Categorical columns like 'power' and 'torque,' which had values like “170 hp @ 5,600 RPM” and “203 lb-ft @ 2,000 RPM,” were split into 'hp' and 'RPM' columns, treating them as numeric columns.

Encodings

Categorical data columns were target-encoded, replacing values with the average price.

Exploratory Data Analysis

Correlations among independent variables and their relationships with the target variable ('price') were examined. Highly correlated independent variables (above 0.7) were dropped to address multicollinearity issues.

Training and Validation Data Split

The dataset was divided into training and validation sets to evaluate the root mean squared error (RMSE) score.

Feature Selection

Lasso regression was employed for feature selection. The important variables that was used in my model were make_name', 'model_name', 'trim_name', 'body_type', 'fuel_tank_volume_gallons', 'fuel_type', 'highway_fuel_economy', 'city_fuel_economy', 'transmission', 'transmission_display', 'wheel_system', 'wheel_system_display', 'wheelbase_inches', 'back_legroom_inches', 'front_legroom_inches', 'length_inches', 'width_inches', 'height_inches', 'engine_type', 'engine_displacement', 'horsepower', 'daysonmarket', 'major_options', 'maximum_seating', 'year', 'franchise_dealer', 'franchise_make', 'is_cpo', 'is_new', 'listing_color', 'mileage', 'seller_rating', 'ext_color_category', 'int_color_category', 'month', 'has_accidents 'isCab_, 'salvage, 'fleet_ 'frame_damaged.

Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce dimensionality, but the final model excluded PCA as it did not improve the test error RMSE.

Models Used

Various models, including multivariate linear regression, random forest, regression trees, and XGBoost, were employed. The linear regression model served as a baseline, achieving an RMSE of 16000 with all variables. XGBoost gave RMSE of 2865 in my 13th submission in Kaggle.

Analysis

The baseline linear regression model yielded an RMSE of 16434 with all variables. After multiple submissions and employing models like random forest and XGBoost, the final RMSE achieved was 2865. Addressing multicollinearity and exploring ensemble methods, specifically Random Forest, further improved the RMSE to 2865, which was the best on the test data.

To enhance model accuracy, techniques such as regularization, cross-validation, hyperparameter tuning, learning rate adjustment, and ensemble methods were explored. Multicollinearity was addressed by dropping highly correlated variables, significantly improving model performance.

Areas of Improvement

Utilize the “description” column by applying NLP techniques to extract important features from text data reviews.

Implement multiple cross-validation iterations to prevent overfitting and reduce test error.

Address the positive skewness of the target variable (price).

Reconsider handling columns with more than 40% missing values, aiming for better preprocessing and retention.

In conclusion, model building approach starting from data preprocessing, feature engineering, and iterative model selection submissions, along with analysis, led to improved predictive models for car precision.