Predictive Modeling of Used Car Prices: An Application of the SEMMA Methodology

Shawn Chumbar (generated via ChatGPT

September 26, 2023

Abstract

In this study, we employ the SEMMA (Sample, Explore, Modify, Model, Assess) methodology to build a predictive regression model for used car prices. Utilizing a dataset of 4,009 used car listings, we navigate through the stages of data sampling, exploration, preprocessing, modeling, and assessment. A baseline linear regression model serves as our primary predictive tool. The research highlights the importance of each SEMMA phase and offers insights into the intricacies of used car price determinants.

1 Introduction

Used car pricing remains a complex domain influenced by various factors, from brand prestige and vehicle age to mileage and fuel type. Predicting the price of used cars can benefit both buyers and sellers in the market, ensuring fair pricing and informed purchasing decisions. This research delves into the dataset's characteristics, applies data preprocessing techniques, and constructs a predictive model to estimate used car prices.

2 SEMMA Explained

SEMMA stands for Sample, Explore, Modify, Model, and Assess. It is a systematic methodology for data mining and predictive modeling. The process begins with obtaining a representative sample of data, followed by exploring and visualizing the data to understand its patterns and relationships. Next, the data undergoes modifications such as cleaning, transformation, and encoding to prepare it for modeling. The modeling phase involves applying algorithms to the preprocessed data to generate predictive models. Finally, the models are assessed using appropriate metrics and visualizations to evaluate their accuracy and usefulness.

3 Methods (SEMMA Methodology)

3.1 Sample

The dataset was sourced and loaded for analysis. The dataset encompasses 4,009 entries and 12 features. Here's a snapshot of the data:

brand	model	model_year	milage	fuel_type
Ford	Utility Police Interceptor Base	2013	51,000 mi.	E85 Flex Fuel
Hyundai	Palisade SEL	2021	34,742 mi.	Gasoline
Lexus	RX 350 RX 350	2022	22,372 mi.	Gasoline
INFINITI	Q50 Hybrid Sport	2015	88,900 mi.	Hybrid
Audi	Q3 45 S line Premium Plus	2021	9,835 mi.	Gasoline

Table 1: Snapshot of the dataset's top 5 rows (Part 1)

engine	transmission	ext_col	int_col	accident	price
300.0HP 3.7L V6	6-Speed A/T	Black	Black	1 accident	$10,300
3.8L V6 24V GDI DOHC	8-Speed Automatic	Moonlight Cloud	Gray	1 accident	$38,005
3.5 Liter DOHC	Automatic	Blue	Black	None	$54,598
354.0HP 3.5L V6	7-Speed A/T	Black	Black	None	$15,500
2.0L I4 16V GDI DOHC Turbo	8-Speed Automatic	Glacier White	Black	None	$34,999

Table 2: Snapshot of the dataset's top 5 rows (Part 2)

3.2 Explore

Data exploration is a fundamental step in the SEMMA methodology, aimed at understanding the underlying structure, characteristics, and patterns in the dataset. For our used car dataset, we conducted a series of exploratory analyses:

- **Summary Statistics**: We began by computing basic summary statistics to get a sense of the central tendencies and spread of our key numeric variables like `price`, `milage`, and `model_year`. This gave insights into the average values, spread, and potential outliers in the data.
- **Distribution Visualizations**: Histograms and KDE plots were generated for features like `price`, `milage`, and `model_year`. These visualizations helped in identifying the distributions, spotting any skewness, and understanding the range of values these features take.

- **Categorical Data Analysis**: We examined categorical features such as `brand`, `fuel_type`, and `transmission`. This helped us understand the variety of brands available in the dataset, the popularity of various fuel types, and the distribution of transmission types among the cars.
- **Missing Values and Anomalies**: During exploration, we identified missing values in some columns and anomalies in the `fuel_type` column. This initial identification was crucial for the subsequent data preprocessing phase.
- **Relationships and Correlations**: We also looked at how various features correlate with the target variable `price`. Scatter plots and correlation matrices were used to identify potential predictors and understand multicollinearity in the dataset.

Through these exploration steps, we gained a comprehensive understanding of our dataset, which guided our decisions in the subsequent phases of the SEMMA process.

3.3 Modify

The modify phase in the SEMMA methodology pertains to data preprocessing, which involves cleaning, transforming, and making necessary adjustments to the dataset to prepare it for modeling. For our used car dataset, we took the following steps:

- **Handling Missing Values**: We identified and addressed missing values in the dataset. Some of the columns with missing values were `clean_title`. We took a strategy of imputing these missing values based on the mode of the respective columns.
- **Addressing Anomalies**: During our exploration, we discovered anomalies in the `fuel_type` column, where some values were labeled ambiguously. We standardized these labels to ensure consistency across the dataset.
- **Encoding Categorical Variables**: The dataset contained several categorical variables like `brand`, `model`, `fuel_type`, and `transmission`. To make these features usable for our linear regression model, we applied one-hot encoding, which converted these categorical variables into a format that can be provided to machine learning algorithms to better understand and generate meaningful predictions.
- **Feature Scaling**: While we didn't apply feature scaling in this analysis, it's worth noting that it's a common step, especially for algorithms sensitive to feature scales. Techniques like Min-Max scaling or Standardization (Z-score normalization) can be employed to ensure all features have a similar scale.
- **Feature Engineering and Selection**: Based on the exploration phase, new features can be derived, or some of the existing features can be transformed to better represent the underlying patterns in the data. Additionally, irrelevant or redundant features can be dropped to simplify the model and potentially improve its performance.

These modification steps were crucial in ensuring our dataset was in an optimal format for the modeling phase. Proper preprocessing not only makes the data compatible with modeling algorithms but can also enhance the predictive performance of the models.

3.4 Model

In the model phase of the SEMMA methodology, we aimed to build a predictive regression model to estimate used car prices. Based on the nature of our target variable (continuous `price` values), we selected a linear regression approach.

- **Data Partitioning**: Before training the model, we partitioned the dataset into training and test sets. This allows us to train the model on a subset of the data and validate its performance on unseen data. Typically, around 70% to 80% of the data is used for training, with the remainder reserved for testing.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
```

- **Model Training**: We employed the linear regression algorithm from the scikit-learn library. After defining the model, we fit it using our training data.

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()
model.fit(X_train, y_train)
```

- **Model Prediction**: With the trained model, we made predictions on our test data to assess the model's performance in subsequent steps.

```
y_pred = model.predict(X_test)
```

The chosen linear regression model provided a foundation for our initial predictions. In real-world scenarios, it's often beneficial to explore multiple models and algorithms, tuning hyperparameters and employing techniques like cross-validation to optimize performance.

3.5 Assess

The assessment phase is pivotal in the SEMMA methodology as it quantifies the performance and accuracy of the developed model. For our linear regression model predicting used car prices, we employed the following evaluation techniques:

- **Quantitative Metrics**: To assess the model's performance quantitatively, we computed the Root Mean Squared Error (RMSE) and the R-squared (coefficient of determination) values. The RMSE provides an understanding of the model's error in terms of the target variable's unit (price in this case), while the R-squared value measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

```
from sklearn.metrics import mean_squared_error, r2_score

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r2 = r2_score(y_test, y_pred)
```

- **Visual Assessments**: Visualization is a powerful tool for model assessment. We plotted the actual prices against the predicted prices. A perfectly accurate model would result in all points lying on a 45-degree line (a line of perfect prediction). Deviations from this line indicate prediction errors.

```
import matplotlib.pyplot as plt

plt.scatter(y_test, y_pred)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='r
# 45-degree line
plt.xlabel('Actual Prices')
plt.ylabel('Predicted Prices')
plt.title('Actual vs. Predicted Prices')
plt.show()
```

- **Residual Analysis**: Analyzing residuals (the differences between observed and predicted values) can provide insights into model performance and potential biases. We plotted a histogram of the residuals to see if they followed a normal distribution, which is an assumption of linear regression.

```
residuals = y_test - y_pred
plt.hist(residuals, bins=30, edgecolor='k')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.title('Histogram of Residuals')
plt.show()
```

Through these assessment techniques, we gained insights into the strengths and limitations of our linear regression model. The findings can guide potential model refinements and alternative modeling strategies for future research.

3.6 Results

Our comprehensive application of the SEMMA methodology on the used car dataset led us to several key findings and insights:

- **Descriptive Insights**: The dataset provided a diverse array of used cars spanning multiple brands, models, and years. Brands such as Ford, Hyundai, and Lexus were prominently featured, with vehicles primarily using gasoline. The majority of cars had automatic transmissions, with a smaller percentage employing manual or other transmission types.

- **Data Quality**: The initial exploration and assessment highlighted some quality issues in the data. Missing values in features such as `clean_title` and ambiguous labels in the `fuel_type` column were identified and rectified during the modify phase.

- **Modeling Outcomes**: Our linear regression model served as a foundational step in predicting used car prices. While the model captured general trends, there were discrepancies between predicted and actual prices, suggesting potential areas for model refinement.

- **Performance Metrics**: The RMSE value provided an indication of the model's prediction error in terms of the price unit. The R-squared value suggested that a significant portion of the variance in car prices was explained by the features in our model, but there's still room for improvement.

- **Residual Analysis**: The distribution of residuals from our model was examined. While there was a general trend of normal distribution, some deviations suggested potential non-linear relationships or influential outliers in the data that might benefit from more advanced modeling techniques.

- **Feature Importance**: While we didn't delve into more advanced models that provide feature importance metrics, the coefficients from our linear regression model can offer insights into how different features influence the predicted car price. For instance, brand prestige, car age, mileage, and fuel type are likely significant determinants in the pricing of used cars.

In conclusion, the application of the SEMMA methodology on the used car dataset illuminated both the complexities and potential avenues for predictive modeling in the used car market. The insights obtained offer a foundation for further research, model refinement, and more advanced analytical techniques.

4 Conclusion

The SEMMA methodology offers a structured approach to data mining and predictive modeling. Further research could delve into advanced algorithms, feature engineering, and hyperparameter tuning.

References

1. OpenAI. ChatGPT. Available at: `https://www.openai.com/`.

2. Taeef Najib. Used Car Price Prediction Dataset. Kaggle, 2022. Available at: `https://www.kaggle.com/datasets/taeefnajib/used-car-price-prediction-dataset`.

3. Academic English Now. 10 Ways To Use ChatGPT To Write Research Papers (ETHICALLY) In 2023. YouTube, May 24, 2023. Available at: `https://www.youtube.com/watch?v=IqfYYxmbTuM`.

4. Dr. Asma Jabeen. ChatGPT for Scientific Research Paper Writing. Available at: `https://drasmajabeen.com/chatgpt-for-scientific-research-paper-writing/`.