Intro

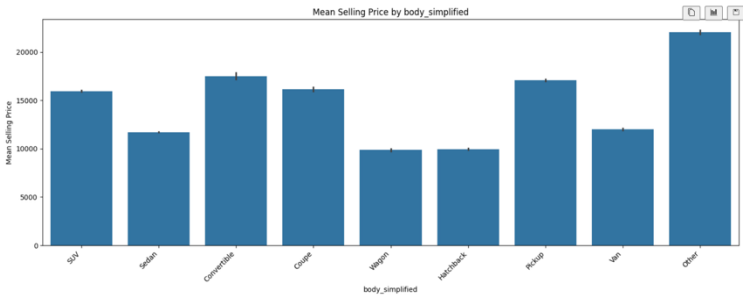
The goal of this analysis was to predict car selling prices using statistical modeling and to identify which features most influence a car's value. Being able to predict the price of a car is essential to the car marketplace, especially for online marketplaces to provide accurate prices on their products. It is also imperative to understand what makes a car expensive or cheap, so being able to discern which car features have the greatest impact on the price is very relevant. This analysis of car sales was guided by three essential questions: can we accurately predict car prices, which features are most informative of the costs, and which model provides the most effective predictions and complex feedback.

The Dataset

The dataset originates from the publicly available "Vehicle Sales Data" on Kaggle, which aggregates a wide variety of vehicle-sale records from across the United States over the past couple years. Each row corresponds to a sold vehicle and provides information through several numeric and categorical variables. Among the numeric variables are year (the year the car was manufactured), odometer (total mileage driven), and the target variable selling price. Categorical variables include features such as make (manufacturer), model, body type (e.g., sedan, SUV), transmission, state (location of sale), color, and interior.

Before modeling, I preprocessed the dataset by encoding categorical variables and split the data into training and testing sets in accordance with typical modeling standards to test the validity of the models. I also removed extraneous or potentially biasing columns (e.g., estimated market value) to ensure that predictions rely only on observable vehicle characteristics and provide a wealth of informative analysis.

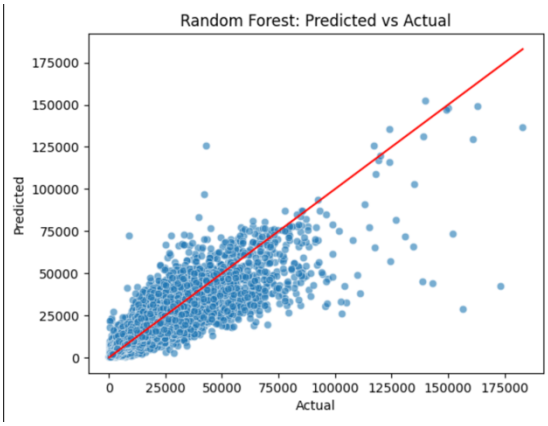
The dataset was relatively well maintained, containing over 440000 usable entries out of 560000 total datapoints. Any rows with missing data were removed, along with any extreme outliers. Additionally, for convenience, the body category was simplified into ten distinct styles: SUV, Sedan, Convertible, Coupe, Wagon, Hatchback, Pickup, Van, and Other. These ten groups effectively encompassed every body type without overbearing any single category.



Potential biases within the dataset could be the data may not fully represent all car markets in the U.S., let alone the entire world, potentially introducing sampling bias. Additionally, the missing or inconsistent entries, which were handled during preprocessing, could affect generalizability.

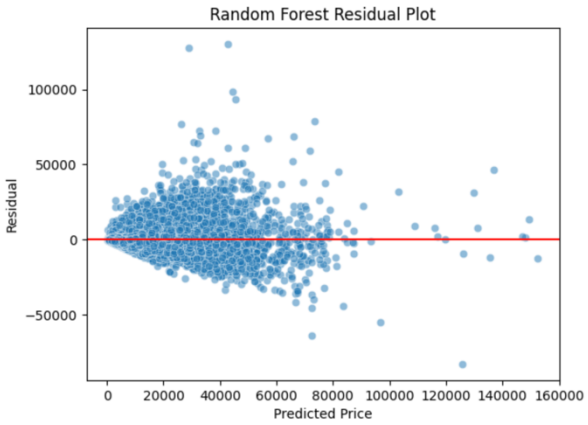
Results

I trained two models: Linear Regression and Random Forest. I decided on Linear Regression as a baseline analysis of the data and used Random Forest to account for non-linear relationships between the car features and selling price. Random Forest performed much better than Linear Regression, getting a much lower RMSE and higher R^2. However, both models did predict general trends well despite the low amount of provided variables, although extreme high- or low-priced cars were more difficult to estimate accurately in both models. Because the Random Forest Model was the better choice overall, all future analysis was done with this model in mind.



Feature analysis revealed that odometer, make, and body type were the most informative variables in predicting price, while transmission, interior material, and color had minimal impact. This aligns with intuition: mileage and brand significantly influence value, whereas cosmetic attributes are less important. However, it was interesting to see that material was so low given the high variability in interior material cost (i.e. foam vs leather seats)

Looking at the residuals for the Random Forest model, the mean residual was close to zero, meaning that there was little bias in the predictions. However, the high standard deviation also showed that there must be large errors present. Thus, by looking at the MAE, which was lower than the standard deviation, I can conclude that the model performs poorly at extreme costs, accounting for the high standard deviation, while performing very well in the middle range, indicating that average predictions were quite accurate.



Discussion

In conclusion, car prices can be predicted with reasonable accuracy, even with simpler models like Log Regression, demonstrating that there is a significant amount of data present within just a few variables. Second, essential mechanical factors (odometer, make, and body type) seem to influence a majority of the price, while cosmetic (color and interior) details have much less effect. Finally, Random Forest proved to be the most effective model, outperforming Linear Regression. This makes sense conceptually, as the simplicity of the log regression model cannot deal with non-linear relationships between variables as random forest can.

Overall, this study illustrates that modeling car features can provide a fairly accurate assessment of the potential price of a car. Future work could include exploring additional features, using different types of models, or trying to determine the most profitable future car.