Used Car Price Modeling for a California Dealership

Caleb Dimenstein, Huaiyuan Fan, Chen Sun, Hourui Guo

Frame the Problem and Look at the Big Picture

In the competitive landscape of California's used car market, dealerships face the dual challenge of accurately pricing vehicles and doing so efficiently. Various factors, including mileage, year of manufacture, accident history, and the condition of the vehicle, significantly influence used car prices. For a dealership entering this market, knowing the market and having effective pricing strategies is the key to its business success. For car price prediction, traditional methods tend to source data from local sales and average the prices of many similar cars. This method works well if you already have a common car with a common set of features. The condition of the car is judged very roughly, typically on a scale of one to three. In addition, the process of evaluating a car for trade-in is both financially and temporally demanding.

To address these challenges, we embarked on the development of a predictive pricing model designed to streamline the evaluation process and enhance pricing accuracy. By harnessing data mining and machine learning techniques, this model analyzes historical sales data to identify patterns and predict prices based on the specific attributes of each car. Our work utilized a comprehensive dataset from Cargurus.com, which contains information on over 3 million used cars, making it a robust foundation for training and evaluating various machine learning algorithms, including linear regression and random forests.

California is one of the largest used car markets in the US with over 9,211 dealerships and a $17.9 billion market size. The necessity of such a model stems from the dealerships' need to quickly offer competitive and fair prices for trade-ins, conserving valuable resources such as labor and time. This is particularly crucial in California's competitive market, where the ability to make swift, accurate pricing decisions can provide a significant competitive edge.

The performance of our model is measured using a range of metrics, from traditional ones like RMSE and MAE to more advanced analyses such as Q-Q plots, ensuring that our approach is both comprehensive and robust. While our predictive model offers a faster and more cost-effective solution for entering the market, it does not negate the value of human expertise. Insights from industry professionals are crucial for identifying key features that impact a car's value and for interpreting model outcomes. We will check the market reports and real market trends to have a blend of machine learning and human judgment ensuring that our model remains both accurate and relevant.

Our approach is based on several assumptions, including the stability of market dynamics over time and the sufficiency of the chosen features to capture the full spectrum of influences on a car's price. Despite the ever-changing nature of the used car market, our extensive dataset provides a solid basis for understanding the factors that significantly affect car prices. Moreover, our model is designed to be flexible, and capable of adapting to new data to reflect current market conditions accurately.

This structured approach to developing a used car price prediction model illustrates the strategic integration of machine learning with domain expertise to address the unique challenges of the California used car market. By prioritizing efficiency, accuracy, and adaptability, we aim to provide dealerships with a powerful tool that enhances their competitiveness and success.

Acquiring the Data

To acquire the used car dataset, we access a reliable data repository in Kaggle[1]. After accessing the dataset, we evaluate the required data columns and ascertain the necessary volume for storage. This dataset

[1] https://www.kaggle.com/datasets/ananaymital/us-used-cars-dataset

comprises various attributes including car specifications, pricing, and seller information. The space required depends on the size of the dataset, which consists of approximately 500,000 entries. We verified the legal obligations associated with the dataset, ensuring compliance with data usage regulations. There is no need for authorization to access the dataset. We create a workspace with sufficient storage capacity to accommodate the dataset. Subsequently, we download the data and convert it into a format suitable for manipulation, preserving the original data integrity. It's crucial to handle sensitive information with care, either by deleting or anonymizing it to protect privacy. We inspect the dataset to determine its size, type, and attributes, ensuring it aligns with our analytical objectives. Before any analysis, we randomly sample a test set and set it aside to prevent data snooping biases. This rigorous process ensures data integrity, legal compliance, and ethical handling throughout the analysis.

Methodology Behind Modeling Strategy

In this project, we stepped on creating a predictive model designed to transform the methodology of setting used car prices within California's competitive market. Utilizing a dataset comprising approximately 20,000 records, our team employed a variety of machine-learning techniques to achieve precise price predictions. This initiative directly addresses the critical challenge faced by our dealership: to price vehicles both competitively and efficiently.

At the heart of our effort was the objective to develop a model that not only enhances the efficiency of the pricing process but also supports the broader business objectives of operational efficiency and market competitiveness. Through the automation provided by our predictive model, the dealership has a chance to offer fair, data-informed prices for trade-ins, thereby bolstering customer trust and satisfaction.

To ensure our findings and recommendations are accessible to all stakeholders, we plan to structure our presentation into six distinct parts. This approach allows us to cater to audiences with varying levels of expertise, from those with a deep understanding of data science to those more familiar with the business aspects of dealership operations. The centerpiece of our presentation will be the models and model selection process, which constitutes the core of our project. Here, we will dissect the advantages and disadvantages of each model from both a data science perspective—focusing on aspects such as accuracy, scalability, and robustness—and a business viewpoint, considering factors like implementation cost, ease of integration into existing systems, and potential impact on the dealership's workflow and profitability.

By adopting this dual-perspective approach, we aim to provide a comprehensive overview that not only highlights the technical excellence of our predictive models but also underscores their practical applicability and potential to contribute to the dealership's strategic goals. This balanced presentation strategy ensures that all stakeholders can appreciate the value and implications of our work, fostering a collaborative environment for discussion and decision-making.

Explore the Data(Preliminary version before data cleaning)

The original dataset contains numerous data columns that have not been standardized such that we may not be able to dive deeper before data cleaning. For instance, the back_legroom has both numbers and texts like 33.2 in. Thus, before data cleaning, we created a copy of the dataset for the data exploration task, ensuring it was manageable for preliminary analysis. Subsequently, we initiated a Jupyter notebook to meticulously document our exploration journey. Each attribute was scrutinized to comprehend its nuances comprehensively. This involved delineating its name, discerning its data type (whether categorical, integer/float, bounded/unbounded, text, structured, etc.), and assessing the percentage of missing values. Additionally, we examined the attribute's noisiness and identified the type of noise it exhibited, which could range from stochastic patterns to outliers or rounding errors.

Figure 1

	Name	Type	% of missing values	Noisiness
vin	vin	object	0.0	Limited, depends on data quality and entry
back_legroom	back_legroom	object	7.442473	Limited, depends on data quality and entry
bed	bed	object	99.91691	Limited, depends on data quality and entry
bed_height	bed_height	object	91.125217	Limited, depends on data quality and entry
bed_length	bed_length	object	91.125217	Limited, depends on data quality and entry
...	...	...	...	...
wheel_system	wheel_system	object	5.838331	Limited, depends on data quality and entry
wheel_system_display	wheel_system_display	object	5.838331	Limited, depends on data quality and entry
wheelbase	wheelbase	object	7.442473	Limited, depends on data quality and entry
width	width	object	7.442473	Limited, depends on data quality and entry
year	year	int64	0.0	Outliers possible

66 rows × 4 columns

Evaluating the attribute's usefulness for our analytical task was crucial, alongside determining its distribution type. In supervised learning scenarios, we pinpointed the target attribute, price. Visualization of the data was a central component of our exploration, aiding in comprehending patterns, trends, and potential relationships within the dataset. Through systematic examination and visualization, we aimed to derive meaningful insights that informed subsequent analyses and model development effectively.

Prepare the Data

To ensure the data is usable and useful, Specifically in this case where there are numerous steps to be followed, a preparation guideline is created with the following steps:

- **Data Cleaning:**
  - Handle missing values by imputing
  - Remove Redundant columns to ensure data integrity.
  - Remove columns unrelated to the objective
- **Data Aggregation:**
  - Aggregate Data to create meaningful groups for analysis
- **Data Transformation:**
  - Create new features based on vehicle major packages
  - For columns containing data and unit measurement, extract data and create new columns for clarity
- **Feature Reduction:**
  - Drop highly correlated features to reduce multicollinearity
    - Use Correlation to drop highly correlated numerical features
    - Use Cramer's V to drop highly correlated categorical features
- **Data Transformation (continued):** *# The reason why we are separating this step is by converting categorical data after Cramer's V, we save a lot of computation time*
  - Convert categorical data into a suitable format(One-hot)
- **Feature Selection:**
  - Select the most important features using Decision Tree Regressor
- **Data Normalization:**
  - Transform numerical data to normal distribution to improve model performance

Data Cleaning:

To address missing values in existing data, a list of columns that have null values larger than 50 percent were dropped (12 columns in total: **bed, bed height, bed length, cabin, combined fuel economy, is certified, is cpo, is oemcpo, owner count, salvage, theft title, vehicle damage category**), any row of data that contained null values were dropped.

After removing sparse columns and records containing null values, 15,009 records and 53 columns remained. The next step was to remove columns unrelated to the project objective, and a list of 18 irrelevant columns was dropped (**vin, main picture URL, city, dealer ip, description, engine_type, latitude, listed date, listing id, longitude, model name, sp id, sp name, transmission_display, trimId, trim_name, wheel system display, franchise make**).In addition, redundant columns (**exteior_color**) were removed

The last step in data cleaning was to aggregate data, the column's interior color contains inconsistencies and redundancy, such as formatting errors, typos, and the same colors with various names. 'Interior Color' was aggregated into 13 unique values (**black, other, brown, white, gray, cloth, red, blue, metallic, beige, yellow, orange, green**)

Data Transformation:

In this step, new columns were created based on the most popular vehicle options contained within the major_options column. In addition, columns that contained data and its unit of measurement(**back legroom, front legroom, height, length, wheelbase, width**) were dropped, and the data was extracted and placed in new columns with the unit of measurement in the header for clarity.

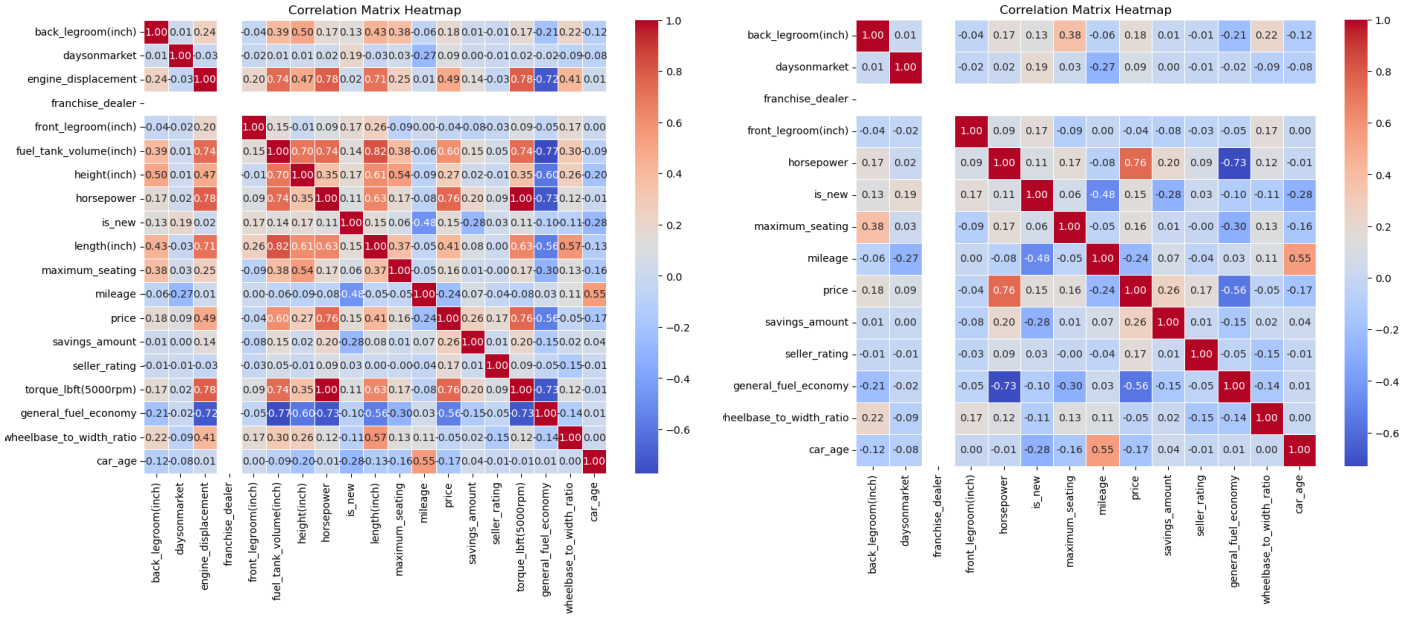
Feature Reduction:

After the removal of unnecessary columns and the addition of other useful ones, the correlation between numerical and categorical features needed to be calculated to help determine the usefulness and redundancy of each column. Because the current data has 47 columns with mixtures of numerical and categorical variables, separate methods should be used for numerical and categorical data.

Numerical data correlation was determined as high if their correlation was higher than 0.6, and a heat map was constructed for visualization for better interpretation, based on the correlation heatmap five numerical columns with high correlation were dropped (**engine displacement, torque lb-ft, fuel tank volume, length, height**).

Categorical data correlation was determined using Cramer's V correlation, no heat map was constructed due to the high volume of categorical variables. Instead, functions were created returning pairs of columns that had high correlation, and based on that four more columns were dropped(**fuel type, fleet, wheel system, Third Row Seating**).[2]

Figure 2



Data Transformation (continued):

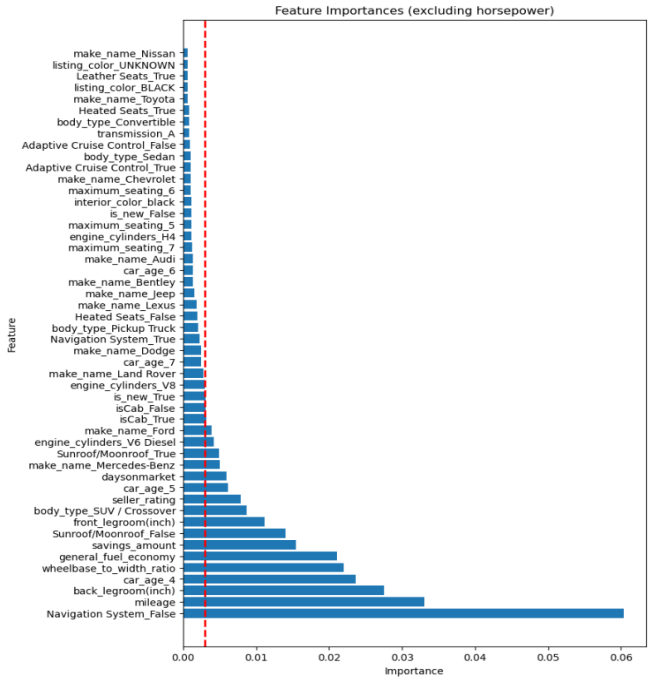
In this step, one-hot encoding was used to give a dummy variable to each variable for categorical features. After one hot encoding of 28 categorical features, our dataset contained 177 columns with a mixture of numerical and categorical variables (encoded).

Feature Selection:

Because our current dataset contains 177 columns, feature selection was essential to reduce computational complexity and remove non-important features, to identify the most important features a Decision Tree Regressor was used, the threshold was drawn at 0.004 and any feature with an importance score higher than 0.004 is considered important. Therefore,17 Features were selected for our prediction model(encoded columns were put back together and seen as one feature from the original dataset, Categorical columns: **is cab, is new, sunroof/moonroof, navigation system, body type, engine cylinders, make name, car age.** Numerical columns: **horsepower, back legroom, days on market, front legroom, mileage, savings amount, seller rating, general fuel economy, wheelbase to width ratio**).

A feature important graph was constructed and placed below for visualization and better interpretation (notice, the graph does not contain horsepower's importance score. Horsepower had a very high important score on our objective, if we constructed the graph containing horsepower, it would be hard to draw a threshold).[3]

Figure 3

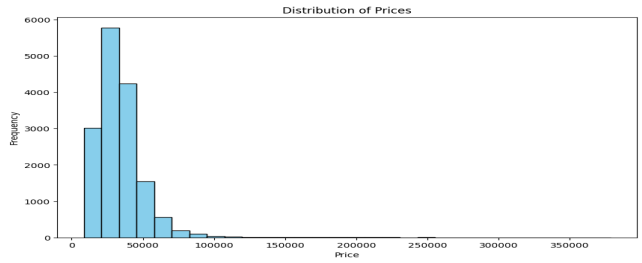


Visualization before modeling:

The distribution of car prices is left skewed[4], indicating the presence of a small number of extreme values that will impact the final modeling outcomes. The EDA section will examine the reason for this fact and provide some key insights for modeling by exploring the relationship between variables.

[3] Figure 3
[4] Figure 4

Figure 4



To begin with, the make of cars significantly influences the price distribution, especially with luxury brands such as Rolls-Royce, Bentley, and Aston Martin, which boast average car prices exceeding $150,000. Also, the most expensive cars, with a top 0.1% price, are mostly Rolls-Royces. Therefore, dropping some extremely large outliers in the model part will help increase precision.[5]

Figure 5

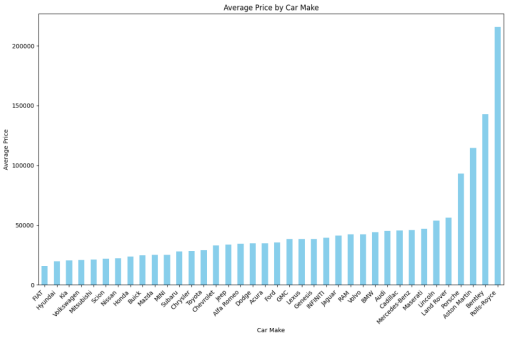
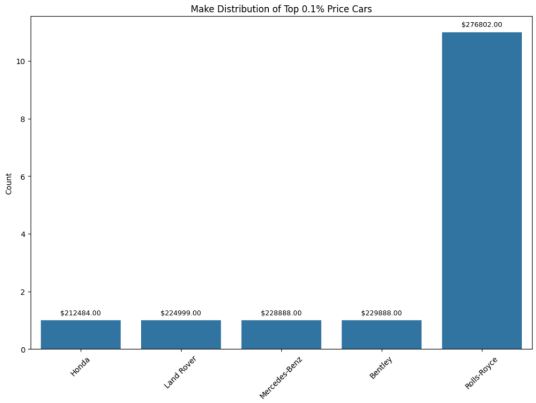
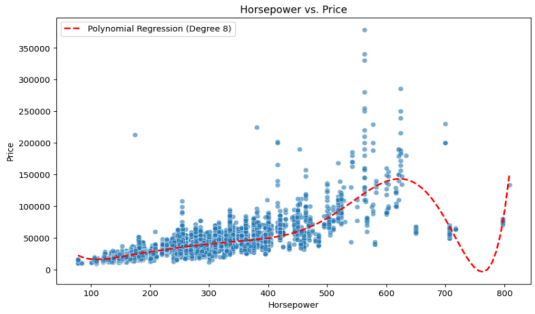


Figure 6



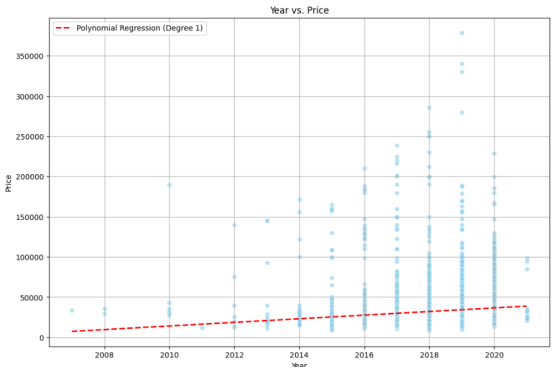
In addition to that, based on the horsepower vs price chart[6], it indicates that transformation and standardization need to be done to find a linear relationship.

Figure 7



It is hard to find a relationship between year and price, but in general, the price of a used car is lower if it is older. The outliers are concentrated between the years 2018 to 2019.[7]

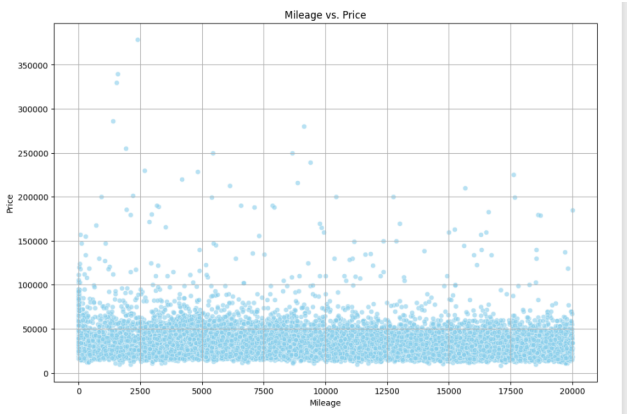
Figure 8



[5] Fi
[6] Fi
[7] Fi

We only have the mileage of used cars between 0 - 20,000 miles, which is almost as new as a new car, we cannot explore a clear pattern between mileage and price now.[8]

Figure 9



Data Normalization[9]:

To run a prediction model using selected features we looked at the distribution of numerical features and applied the proper transformation method to each. A combination of log scaling and exponential scaling was applied to various features regarding how to normalize them properly

- Log Scaling: Price, wheel base-to-width ratio, horsepower, front legroom, general fuel economy, and savings amount
- Exponential scaling: Seller rating

The remaining features were all categorical so no normalization was required.

The Goodness of Fit of Linear Regression Model

After running linear regression, it is important to check the model's goodness of fit. In this case, a scatter plot[10] was constructed to test linearity and identify potential outliers, and a QQ plot was constructed to check the distribution of residuals.[11]

If data points in the scatter plot form a diagonal line, it indicates a good fit. As seen in Figure 10, the actual price and predicted price form a diagonal line that indicates a linear relation and a good fit.

If the sample vs. theoretical quantiles form a relatively straight line then we can say the data is normally distributed. As seen in Figure 11 the smaller observations, cars that are much less expensive than the average, are predicted to be less expensive than their actual values meaning that the lower tail's distribution has been extended, relative to the normal distribution. The larger observations however are predicted to be more expensive than they are, as seen from the residuals on the right-hand side of the graph falling above the perfect normally distributed line. In summary, cars that are much less expensive than the average car price are slightly undervalued, and cars that are much more expensive than the average are slightly overvalued.

---

[8] Figure 9
[9] Appendix Table 2
[10] Figure 10
[11] Figure 11

Figure 10

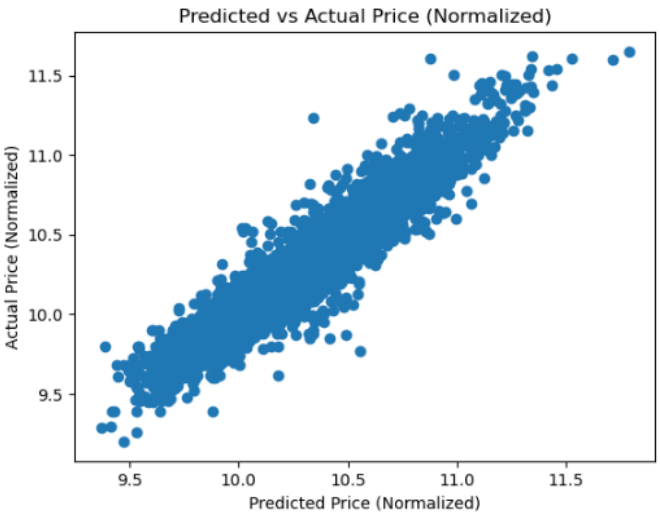
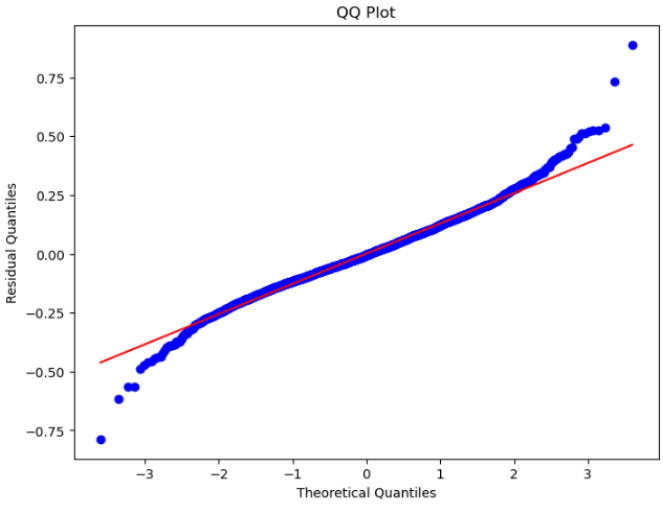


Figure 11



Model Selection

To begin the model selection process a linear regression model was run on the entire dataset once data cleaning was performed. Given the number of features involved before encoding (17), which was brought down from 121 once data cleaning was performed, an adjusted R2 score of 0.8945 was captured. While the features in this linear regression were able to account for 89.45%[12] of the variation in the target variable, price, the model's numerical linearity is concerning. The linearity scatter plot showed that linear regression was a relatively poor fit compared to other potential models. After demonstrating how a simple model may not be the best fit, several machine learning models were selected including **Random Forest**, **Gradient Boosting**, and a **Voting Regressor** taking the average of all model results.

A Random Forest Regressor was chosen to deploy. There are quite a few benefits of utilizing this form of modeling. Since Random forest Regressors are made up of many decision trees, 200 in our analysis, provides a higher level of accuracy than individual models by aggregating the predictions of multiple base learners (decision trees). Since our data has many features, overfitting must be taken into account. Random Forest models are less prone to overfitting than other complex models like deep neural networks. The implantation of this model helped us remove features as well.

Next, Gradient Boosting was chosen to deploy. Gradient Boosting Regressor is known for high accuracy because it makes predictions sequentially, which means the next prediction is based on correcting the errors of the previous one. In addition, the Gradient Boosting Regressor could handle non-linear relationships and various types of data including numerical and categorical data, which is a perfect fit for our data set, and it eliminates the concern of non-linearity of numerical variables vs target variables.

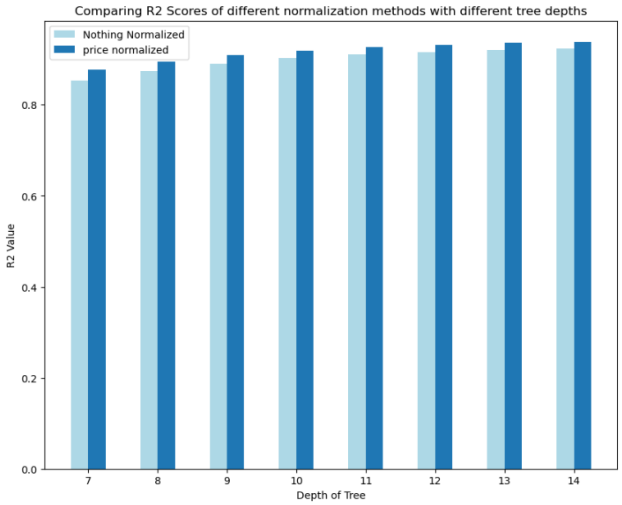
In the end, a Voting Regressor was deployed to combine previous models and take an average of the prediction result to create a more stable and robust model that improves performance. A weight was assigned to each of the models based on their performance and adjusted R^2 score (Linear Regression: 0.2, Random Forest: 0.3, Gradient Boosting: 0.5)

[12] Table 1

Model Tuning

If no hyperparameters were provided, the random forest regressor would run until each leaf node only contained a few samples. This can lead to various issues such as overfitting and creating a model too complex to interpret. Max depth was the primary hyperparameter tuned for our model. Utilizing the test R2 score to measure the results, max depths of the trees ranging from 7 to 15 were sampled.[13] To further show the impact of normalization two different models were run. One model with no normalization and one model where each numeric variable was normalized. As shown below there is a slight difference in R2 performance between no normalization and normalization.

Figure 12



Random Search Analysis

To further tune our model, a random search was performed to Random Forest and Gradient Boosting to find the optimal hyperparameters that maximize performance.

Random Forest Hyperparameters

- **Max Depth**: range from 5 to 15
- **Min Samples Split:** 20, 30, 40, 50, 60
- **N Estimators:** 100, 200, 300, 400, 500
- **Max Left Nodes:** range from 2 to 100

Gradient Boosting Hyperparameters

- **N Estimators**: 50, 100, 200
- **Learning Rate**: 0.01, 0.1, 0.2
- **Max Depth**: range from 1 to 100
- **Min Samples Split**: 20, 30, 40, 50, 60
- **Min Samples Leaf**: range from 1 to 100
- **Max Features**: None, sqrt, log2

The optimal hyperparameters for Random Forest are at estimators: 400, min samples split: 50, max_leaf_nodes: 71, max_depth: 10. For the final model the hyperparameters optimized in the grid search analysis were deployed.

[13] Figure 10

The optimal hyperparameters for Gradient Boosting are at n estimators: 200, min samples split: 30, min samples leaf: 33, max features: sqrt, max depth: 27, and learning rate: 0.7.

Final Model Performance

After training a Linear Regression model, an adjusted R^2 score of 0.8994 was obtained. To improve performance, a Random Forest model was used with a random search for hyperparameter optimization, resulting in an adjusted R^2 score of 0.9396. Subsequently, a Gradient Boosting Regressor model was implemented, achieving an adjusted R^2 score of 0.9478. Finally, the use of a Voting Regressor method further improved the model, yielding an adjusted R^2 score of 0.9508.

Future Implementations

As seen in figure 11 our model overprices more expensive cars, such as Rolls Royces and Mercedes Benz. The model created in this research can be specialized further to solely predict the value of high end cars. This specialized model would allow us to restrict our initial model to cars under a certain price threshold, eliminating the instances of outliers, and in the end, make our original model more accurate. Another way our model can be expanded is to use the same data cleaning, feature engineering, and feature selection techniques but apply it to motorcycles and other types of road approved vehicles. The concept would remain almost identical, collect the price of motorcycles, obtain all the features of the vehicles, and train models to determine which features are most important in predicting price. By doing so the dealership would be able to expand their product line not just to cars but to other road approved vehicles.

Summary

Once we obtained our final model the results were interpreted such that a business that is not familiar with machine learning can easily understand our results. Once we obtain our predicted results from the Random Forest model the prices have to be exponentiated since we took the log price during our feature engineering portion. In our findings, the Random Forest model accurately predicts car prices up to $75,000 however, past this threshold the model tends to overprice the cars.[14]

To highlight the model, Table 1 depicts the frequency at which we over-price cars at certain price thresholds. Overall 15% of cars are either over or underpriced (with a threshold of $1,000). This gives us a pricing accuracy of 85%. Over time, however, as more cars are brought into the dealership, we expect this accuracy to increase as the model is trained further and an ever-growing dataset. For cars that are overpriced by $15,000 or more, half of them are sedans. Our model tends to overprice more high-end engine cylinders as well. Half of the cars greatly overpriced have V8 cylinders which are typically deployed in high-performance cars. Cars overpriced by this much, $15,000, would have been greater if the data set was not limited at the beginning of this data exploration process. Since we limited the price of a car to $150,000 we limited the amount of cars to be greatly overpriced as our model tends to over-priced more expensive cars. Despite this limitation, the implementation of our predictive model can still help with the pricing strategy, giving a reference to the dealership about fair prices for used cars.

To implement this model we would sell it to the car dealership and through them inputting all the features of the car in question our model would produce a predicted price for the car. To ensure that our findings and methodologies are accessible and beneficial to all stakeholders, we have structured our presentation into six comprehensive segments. This approach is designed for both individuals with a background in data science and those with expertise in the business domain. By doing so, we aim to bridge the gap between technical model development and practical business applications, ensuring a holistic understanding of our project's impact.

[14] Figure 11

The focal point of our presentation will be the exploration of the models and the critical process of model selection, which constitutes the core of our project. We intend to analyze the advantages and disadvantages of each model from both the perspectives of data science and business implications. This dual-side analysis will provide a balanced view, enabling stakeholders to understand the technical robustness of the models and their strategic relevance to our dealership's business objectives.

Appendix

Table 1

Amount Overpriced/ Underpriced	Number of prediction (cumulative)	Percentage of prediction (cumulative)
$15,000	30	0.69%
$12,500	58	1.34%
$10,000	110	2.54%
$7,500	233	5.38%
$5,000	548	12.66%
$2,500	1,509	34.86%
$2,000	1,879	43.40%
$1,500	2,345	54.17%
$1,000	2,937	67.84%
$500	3,625	83.74%
$10	4,313	99.63%

Table 2

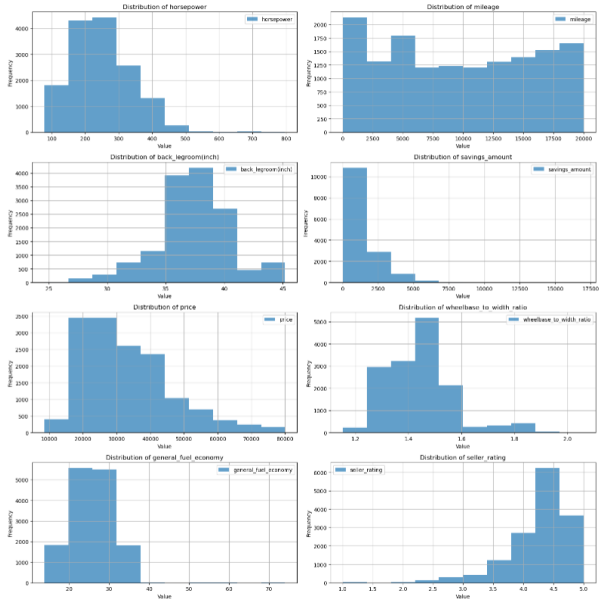
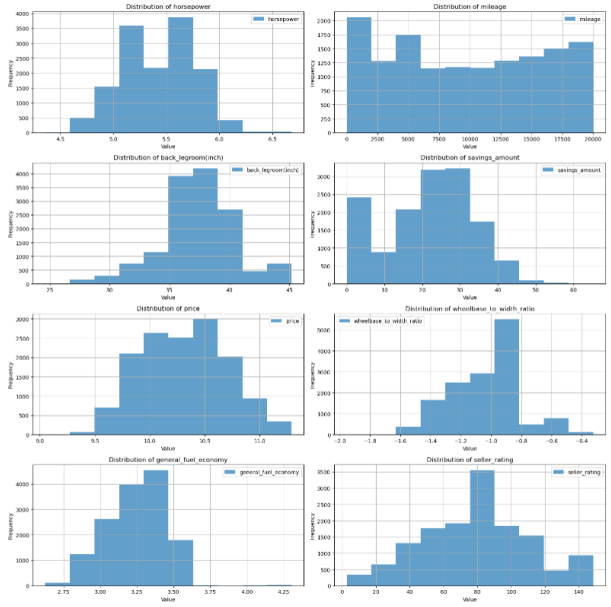
Data Pre Normalization	Data Post Normalization
	

Figure 13: Model actual price compared to predicted price