Predicting MSRP and Identifying Factors Impacting Price

By

Roma Chitale

Vamshi Gadepally

Daisuke Yagyu

Supervisor: Anil Chaturvedi, PhD

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

August 2023

Abstract

The existing method leveraged by Nissan Motor Co to determine vehicle price (MSRP) is oversimplified and inaccurate. It is not optimized to consider various macroeconomic factors, vehicle features, and target demographics, directly impacting sales, profitability, and brand perception. This work aims to solve these issues by building machine learning models that predict prices more accurately and explain what vehicle specifications, macroeconomic factors, and customer reviews impact MSRP the most. Additionally, Generalized Linear Models were developed to explain how different factors move MSRP up or down. This project developed three XGBoost models to predict car prices for three vehicle categories, as well as three Generalized Linear Models, one per vehicle category, to show the impact that different factors have on MSRP. The best-performing XGBoost models achieved an R-squared value between 0.97 and 0.99. A total of 14 models – four non-linear models, nine machine learning models, and a mixed-effect model were built and assessed before selecting the three top-performing XGBoost models.

Keywords: machine learning, Nissan, price prediction, MSRP, vehicles, random forest, boosting algorithms, generalized linear model

Executive Summary

Nissan's current method for pricing its vehicles is overly simplistic and is not comprehensive enough to consider various macroeconomic factors or changes in consumer preferences. This, coupled with a general slowdown of sales in the US automotive industry, has led Nissan's sales and profit to decline over the past several years. It hopes to leverage several machine learning techniques to implement a more robust and accurate pricing methodology. This work aims to help solve that by using various machine learning techniques to predict future MSRP, explain what features are most important in determining the price, and develop Generalized Linear Models to explain how various factors impact MSRP.

Fourteen models – four non-linear models, nine machine learning models, and a mixed effect model were trained and evaluated for their ability to predict MSRP. Eight data sources containing information related to 26 car manufacturers and 164 unique models spanning 10 model years were used to train the model. Six additional data sources relating to ten years' worth of US macroeconomic data were sourced and added to the data set.

Three models were selected at the end, each pertaining to a specific vehicle category determined by an unsupervised clustering algorithm. The three best-performing achieved an R-squared value between 0.97 and 0.99. This improved pricing model and the feature importance insights provided to Nissan can be used to determine more accurate vehicle prices and boost sales.

Table of Contents

List of Figures

List of Tables

Introduction

Problem Statement

After experiencing years of consistent growth, the automotive industry as a whole has seen a decline in sales and revenue over the past few years. In 2022, auto manufacturers reported the worst vehicle sales in over a decade. The COVID-19 pandemic and the semiconductor chip shortage are a few factors contributing to this industry-wide decline.

Nissan, in particular, experienced more challenges. Historically, it had been the third most popular Japanese auto manufacturer in the United States, only behind Toyota and Honda, respectively. Strong sales in 2016 and 2017 almost saw Nissan overtaking Honda as the second most popular Japanese auto manufacturer. However, Nissan shifted its strategy and began prioritizing sales volume over profit. Vehicles were heavily discounted, and car rental fleets were overloaded with their cars, which greatly hurt the brand image and resulted in declining sales.

The current pricing strategy takes a linear regression of historical prices to determine future prices. It is standard practice in the industry to make pricing assumptions and estimates two to three years before the vehicle hits the market. This has made it increasingly difficult for Nissan to make an accurate and effective pricing forecast with its existing methodology. With a desire to change brand perception and boost sales, Nissan wants to implement a data-driven approach to pricing its vehicles so that it can become competitive in the market again.

Analysis Goals

The primary goal of this project is to develop and implement a pricing model that will help Nissan set the correct MSRP for its fleet of vehicles, improving sales and profitability. The model will also explain which vehicle features, macroeconomic factors, and customer reviews have the most significant impact on price. The secondary goal is to develop a Generalized Linear

Model that will allow Nissan to understand what factors move MSRP up or down and by how much.

Scope

Due to differences in macroeconomic factors and car makes and models around the world, the scope of this work is solely focused on vehicles sold in the US market. As a result, the models will only be applicable to Nissan vehicles and its competitor's vehicles sold in the United States. Various macroeconomic factors are also included as features in the model. These factors are very difficult to predict with high accuracy, so there are possible deviations between what was forecasted and what economic conditions actually look like in the future.

Background

Business Partner and Industry

Nissan Motor Co. is a Japanese multinational automobile manufacturer headquartered in Yokohama, Japan. Founded in 1933, it has grown into one of the world's largest automakers, with production facilities and sales networks in more than 160 countries. Nissan's lineup includes a wide range of vehicles including SUVs, crossovers, trucks, and electric vehicles like the Nissan LEAF, which is the world's best-selling electric car. Some of Nissan's top competitors include Toyota, Honda, General Motors, Ford, and Hyundai.

After years of steady growth between 2010-2017, Nissan changed its sales strategy and prioritized sales volume over profit. To boost sales, vehicles were sold at deeply discounted prices, but this shift in strategy had unintended consequences. Rather than increase sales, heavily discounting cars created an image that Nissan's cars were second-rate products, which hurt the brand value and decreased sales, as shown in Figure 1. A medium-strong direct linear dependence exists between brand value and vehicle sales (Nadanyiova et al., 2019).

Table 1. *Nissan Vehicle Sales from 2005-2023*

Year	Sales	YOY Change	US Marketshare	Marketshare Change
2005	1,079,662	0	6.4	0
2006	1,019,249	-5.6	6.2	-3.1
2007	1,068,232	4.81	6.67	7.04
2008	949,533	-11.11	7.24	7.84
2009	769,103	-19	7.43	2.58
2010	908,570	18.13	7.89	5.84
2011	1,042,533	14.74	8.21	3.88
2012	1,249,387	19.84	8.13	-1.06
2013	1,248,420	-0.08	8.06	-0.82
2014	1,387,164	11.11	8.45	4.58
2015	1,486,091	7.13	8.54	1.08
2016	1,564,400	5.27	8.94	4.48
2017	1,593,464	1.86	9.26	3.44
2018	1,493,877	-6.25	8.62	-7.36
2019	1,345,681	-9.92	7.9	-9.09
2020	917,265	-31.84	6.24	-26.66
2021	977,645	6.58	6.54	4.54
2022	729,365	-25.4	5.34	-22.51
2023	235,813	0	6.58	0

The COVID-19 pandemic and subsequent semiconductor chip shortage proved to be additional challenges for Nissan. The pandemic negatively impacted the entire automotive industry. Automotive manufacturing facilities worldwide were temporarily shut down, and US vehicle sales declined by 15% from 2019-2020 (Coffin et al., 2022). The semiconductor chip shortage also further slowed down production. In 2021 vehicle sales in the US were 12% lower than in 2019 (Coffin et al., 2022). The industry is still recovering from these events.

Literature Review

Price Prediction Using Machine Learning

Various machine learning models can be used to predict car prices. There are simple methods, such as linear regression, which Nissan currently uses, and more complex models, such as any boosting algorithms or tree-based models. In past attempts to predict car prices, Light Gradient Boosted Machine (Light GBM) outperformed other models, having 91% accuracy on

test data (Tokakura et al., 2021). Not only could it predict price with reasonably high accuracy, but it was also effective at selecting the most impactful features that drive price.

Random Forest and Support Vector Machine (SVM) have also been used. SVM can predict car prices with more precision than multivariate regression or any simple multiple regression (Listiani, 2009). It is better at handling data sets with more dimensions and is less prone to overfitting and underfitting. While Random Forest does not have as high of an accuracy as SVM, it has been shown to be effective when dealing with large data sets and can estimate missing values in the data set better than SVM (Gegic et al., 2019).

While applying a single machine learning algorithm might be sufficient for many prediction problems, given the complexity of the data sets used specifically for vehicle price prediction, it does not usually perform well. A single machine learning algorithm provided an accuracy of less than 50%. When multiple machine-learning models were combined and used, the accuracy was around 92.38% (Gegic et al., 2019). The increased model accuracy comes at the expense of computational efficiency. It will require more significant computational resources and time.

Forecasting Macroeconomic Factors

Forecasting macroeconomic factors that impact MSRP is crucial for predicting future vehicle prices. Many approaches can be taken, including various time series analytical methods. Examples include moving average, exponential smoothing, and Box-Jenkins. It is important to note that when attempting to forecast complex macroeconomic factors, a single method might not be able to give a reliable result. Using several different methodologies can improve forecasting results (Szilágyi et al., 2016).

Data

Data Sources

Nissan provided eight data sources that contained information on: 1. vehicle specifications, 2. customer evaluations from Edmunds.com, 3. customer research from the New Car Buyer Survey (NCBS), 4. residual value of used vehicles, 5. vehicle sales volume, 6. sales invoice, 7. direct competitors of Nissan vehicles, and 8. brand power. Each data source includes information on Nissan vehicles sold in the US market and vehicles sold by its top competitors. The time frame of all eight data sets is from 2012-2023.

Each data source had different levels of granularity and different reporting cadences. For example, the vehicle specifications data set included data on a car's make, model, year, and trim. In contrast, many other data sets were aggregated at the make, model, and year. Additionally, some data sets had numbers aggregated monthly while others were yearly. This provided some challenges when attempting to combine data into a single clean table that could be used for analysis and modeling. The specific details of each data set can be seen in Table 1.

Additional data was collected from multiple public online resources around various macroeconomic factors such as GDP, inflation, Consumer Confidence Index (CCI), iron ore and metal price, semiconductor price index, and various other price indexes. All indexes selected are limited to data only from the United States. The time frame of the data sets is from 2011-2023, and numbers are reported monthly.

Table 2. *Data Sources and Descriptions*

Data Source	Description	Time Granularity	# of Features
Vehicle Specifications	All internal and external specifications and features of a car.	Yearly + Ad hoc Update	2,682
Edmunds.com	Reviews made on Edmunds.com website. Includes people who have purchased the car or have done a test drive.	Yearly	1

NCBS	Customer research was conducted using the New Car Buyer Survey after someone purchased a car.	Yearly	21
Residual Value	Percent of MSRP a car can hold after 12, 24, 35, 48 and 60 months.	Yearly	5
Sales Volume	Actual monthly sales	Monthly	1
Sales Invoice	The actual price paid by the customer when purchasing a car. The price paid is usually different than the MSRP.	Monthly	1
Nissan Competitors	List of Nissan vehicles in the US market and the direct competitors. This is at a make-and-model level.	Yearly	1
Brand Power	A numeric score of brand power for Nissan and its competitors.	Quarterly	1

Descriptive Analysis

A majority of vehicle prices fell between $15K and $75K, which led the price distribution to be right skewed. This can be seen in Figure 1. To address the skewed distribution, a log of prices was taken, which led to a more normal distribution. This can be seen in Figure 2. It was essential to have the dependent variable (price) be normally distributed since most machine learning models assume a normal distribution of data.

Figure 1. *Vehicle price distribution before taking the log of price.*

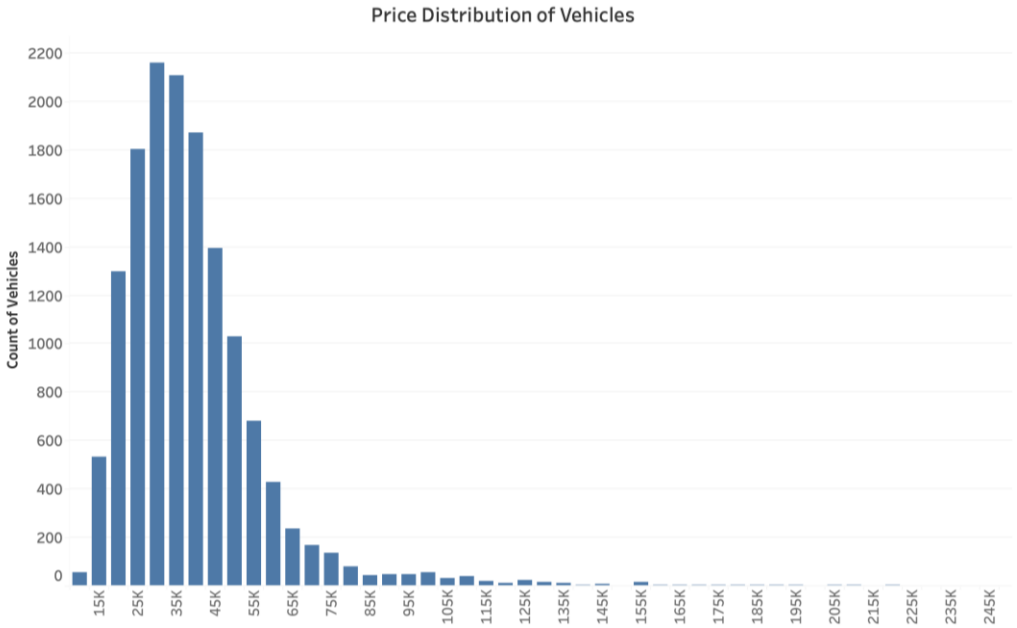
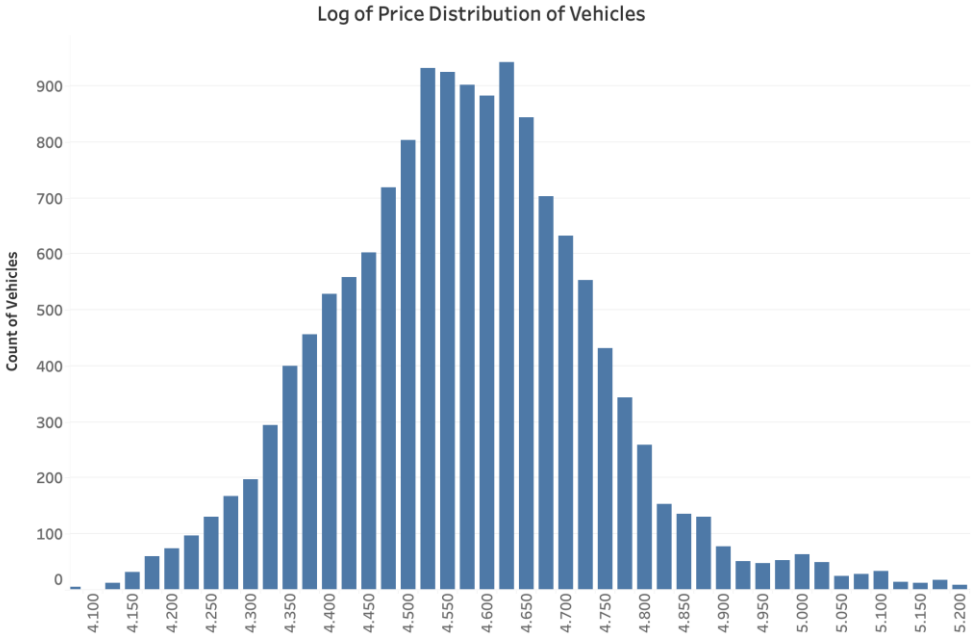


Figure 2. *Vehicle price distribution after taking the log of price*



As seen in Figure 3, the price of Nissan's vehicles and its top Japanese competitors exhibited a slow and steady increase over the past ten years. On the other hand, their German counterparts had more volatility. Even when comparing price changes during major economic events like the COVID-19 pandemic and the semiconductor chip shortage that started in 2019, the price of Japanese vehicles remained reasonably stable. German brands like Audi, BMW, and Mercedes had more significant price increases. These manufacturers passed the higher material costs over to the customer, while Japanese manufacturers absorbed the additional costs.

Figure 3. *Vehicle price changes for Nissan and some competitors from 2012-2023*

There was little to no correlation between price and the variables from any of the eight data sets provided by Nissan or the macroeconomic factors. Correlation heatmaps were created to display the findings visually. The variable with the strongest correlation was the Power and Pickup rating from the NCBS research (0.41). Most of the macroeconomic factors had a weak positive correlation with price. All the correlation heatmaps can be found in Appendix A.

Methodology

Feature Engineering

Handling Categorical Variables

The vehicle specifications data contained 19 categorical variables. These variables are related to various aspects of the internal and external features of Nissan and competitors' vehicles. One hot encoding was utilized to convert the categorical variables into dummy

variables with a numerical representation to address this issue. However, this came with a downside. The vehicle specifications data already contained over 2000 variables. One hot encoding increased the dimensionality, making the data set even larger.

Using NLP on Customer Reviews

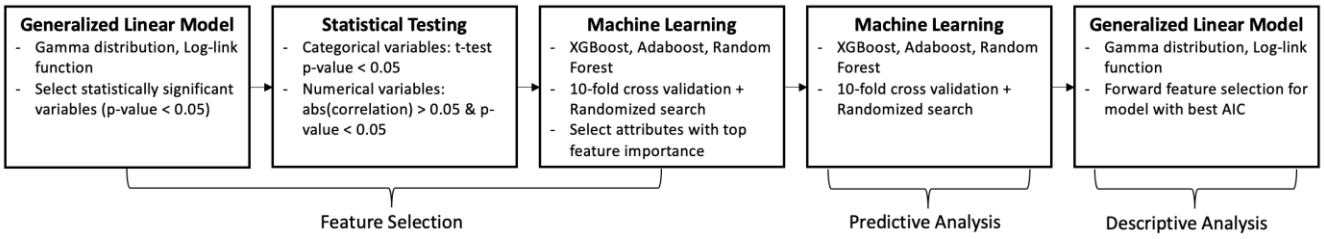
Customer evaluation data from Edmunds.com came in long chunks of text, making the data unusable for modeling in its current state. Each was broken into individual sentences to make the reviews consumable, resulting in roughly 500,000 sentences. Based on a manual glance at the reviews, ten pre-defined categories addressing different aspects of the car that were commonly reviewed were created. The ten categories were: advanced safety features, comfort and quietness, dynamic performance, ride comfort, safety, utility, interior design, exterior design, fuel efficiency, and infotainment.

A training data set was created with 2,000 to 4,000 sentences per 10 categories. Using an NLP transformer on the training data set, each sentence was categorized and given a probability score of how likely the sentence addressed one of 10 pre-defined categories related to the car. Additionally, a sentiment score ranging from -1 and 1 was assigned to each sentence. Each category was manually reviewed to determine which probability threshold had the highest accuracy without compromising on the volume of data.

Each sentence in the review was assigned to one of the ten categories using these thresholds. Once each sentence was categorized, a weighted average of the sentiment score per review was taken. Another average of those sentiment scores was taken to get the sentiment score at the model level. The final output of the NLP model resulted in a sentiment score for each of the ten categories at the model level.

Modeling Frameworks

Figure 4. *Overall Modeling Framework*

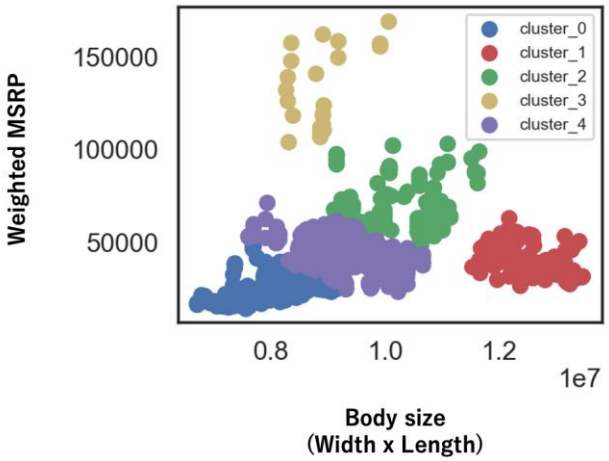


Feature Selection

The first modeling framework developed and tested was used for feature selection. The data set contained 1,105 variables and 12,847 rows of data. Careful feature selection was necessary to prevent the model from overfitting. Based on EDA work conducted and prior industry knowledge, important features would likely differ based on vehicle segments since there is a difference in the customer profile. Due to small data volumes, electric vehicles, and sports/performance vehicles were excluded. All remaining vehicle segments were kept and included in the feature selection process.

Segments were combined into clusters by using unsupervised clustering algorithms. Various distance-based, density based, and hierarchical clustering methods were built and considered. K-means clustering was finally selected due to the clean clustering partitions. Vehicle body size (Body Width x Body Length- and MSRP weighted by each vehicle grade's volume were used as variables. Both variables needed to be standardized as distance-based algorithms are sensitive to scale. The k-means cluster created five clusters, as seen in Figure 5.

Figure 5. *K-Means cluster by MSRP and Body Size*



Each segment was categorized into either one of the five clusters. Given the number of data points, three clusters were selected. Clusters 2 and 3 were dropped, and the remaining three were used for modeling. The clusters were named as follows: Cluster 0 was Low-price Small size, Cluster 1 was Mid-price Large size, and Cluster 4 was Mid-price Mid size. The category each vehicle segment was added to can be seen in Table 3. There were 3,492 data points in the Low-price Small size cluster, 3,841 in Mid-price Large size, and 4,556 in Mid-price Mid size.

Table 3. *Categorization of Segments into K-Mean Clusters*

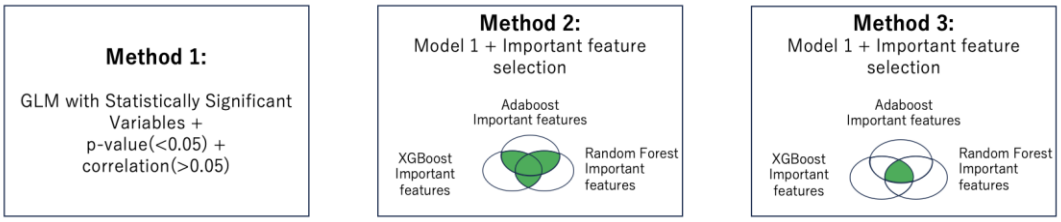
Segment Used by Nissan	Category determined by K-Mean Cluster
Compact	LowPrice_Smallsize
Compact LCV Van	LowPrice_Smallsize
Compact SUV	LowPrice_Smallsize
Entry	LowPrice_Smallsize
Entry SUV	LowPrice_Smallsize
Lower Mid SUV	LowPrice_Smallsize
Lower Midsize	LowPrice_Smallsize
D Coupe	MidPrice_Midsize
Large SUV	MidPrice_Midsize
Mid Luxury	MidPrice_Midsize
Midsize Pickup	MidPrice_Midsize
Midsize SUV	MidPrice_Midsize
Near Luxury	MidPrice_Midsize
Upper Midsize	MidPrice_Midsize
Full-size Pickup	MidPrice_Largesize
Full-size LCV Van	MidPrice_Largesize

After creating the three clusters, a Generalized Linear Model (GLM) was created, and statistically significant variables with a p-value < 0.05 were selected. Considering that the target variable (MSRP) had a positive distribution and a long tail, a Generalized Linear Model using gamma distribution and a log link function was used to determine the best features for the predictive modeling process. This resulted in 247 features being selected for Low-price Small size, 282 for Mid-price Large size, and 267 for Mid-price Mid size.

Further statistical testing was performed on the selected features. For categorical attributes, a t-test was conducted. If the mean values of the two groups had a statistically significant difference (p-value < 0.05), the features were selected. A correlation analysis was conducted for numerical features. The features were selected if the correlation was greater than 0.05 and the correlation was significant (p-value < 0.05). The final features that could be selected for modeling were 216 for Low-price Small size, 249 for Mid-price Large size, and 232 for Mid-price Mid size.

XGBoost, Adaboost, and Random Forest were also used to determine important features since the methodology above still provided a large number of attributes. Two additional feature selection methods were implemented, using a combination of important features from the GLM with statistical testing, and the tree-based algorithms. In the first method, statistically significant variables and variables that were the top 200 features in at least two of the three tree-based models were selected. The second method selected statistically significant variables and the top 200 features from all three tree-based models. In the end, there were three different feature selection methods. The three methodologies can be seen in Figure 6.

Figure 6. *Feature Selection Methodologies*



Predictive Modeling

The second modeling framework developed and tested various machine-learning models and a mixed-effect model that could be used for vehicle price prediction. Since the objective of the project was to provide Nissan with insights into important factors that impact MSRP, four tree-based algorithms (Decision Tree, XGBoost, Random Forest, and Adaboost) and a mixed-effect model were under consideration. However, upon determining the complexity of the features in the data set, three ensemble models (XGBoost, Random Forest, and Adaboost) and a mixed effect model were finally selected for predictive analysis.

For the tree-based models, the data set was split so that 70% of the data was used for training and 30% for testing. Hyperparameter tuning was performed on each model to get optimized hyperparameter values and maximize the model's predictive accuracy. Each model's performance was validated using 10-fold cross-validation, and the best-performing model was selected. The test data set was used to evaluate model accuracy. R-squared and Root Mean Square Error (RMSE) were used as model performance metrics. A balance between train and test accuracy was also considered so that the model was not over or underfitted. Three feature selection techniques were used for each of the three clusters, resulting in nine machine-learning models being developed and assessed.

XGBoost is an ensemble boosting algorithm that is highly efficient and scalable, has regularization techniques such as L1 and L2 regularization, which help prevent overfitting, and

has parameters such as max_depth that can be fine-tuned to make the model good at generalizing on unseen data. Additionally, it is good at handling missing values in the data set and can effectively capture non-linear relationships between features and the target variable. Based on hyperparameter tuning, the best-performing model for Low-Price Small size included 605 estimators, a learning rate of 0.1, a max depth of 7, and 0.0 gamma. The best-performing model for Mid-price Large size included 750 estimators, a learning rate of 0.1, max depth of 5, and 0.0 gamma. The best-performing model for Mid-price Mid size included 850 estimators, a learning rate of 0.1, max depth of 5, and 0.0 gamma.

Random Forest is an ensemble learning method that combines multiple decision trees to create a predictive model. Random Forest was a strong candidate because of its robustness to overfitting and outliers, its ability to handle numerical and categorical features well, and its generally fast compute times. Based on hyperparameter tuning, the best-performing model for Low-Price Small size included 300 estimators, 103 max features, 7 min samples leaf, and a max depth of 8. The best-performing model for Mid-price Large size included 500 estimators, 140 max features, 2 min samples leaf, and a max depth of 9. The best-performing model for Mid-price Mid size included 100 estimators, 165 max features, 2 min samples leaf, and a max depth of 10.

AdaBoost is an ensemble learning technique that combines the prediction of weak learners to create a strong learner. It starts by assigning equal weight to each data point, then iteratively gives higher weights to misclassified samples in each iteration, making the weak learners focus on the most difficult classification problems in the training data set. The final prediction is a weighted sum of the weak learner's predictions. Based on hyperparameter tuning, the best-performing model for Low-price Small size included 500 estimators and a learning rate

of 1. The best-performing model for Mid-price Large size included 500 estimators and a learning rate of 1. The best-performing model for Mid-price Mid size included 1000 estimators and a learning rate of 1.

A mixed-effect model was also developed and tested to predict "Target Weighted RetailPrice" in the test data set, which is the weighted logarithm of MSRP. A mixed-effect model is a statistical modeling technique that is used to analyze data with a nested or hierarchical structure. It is useful when working with clustered data or data with varying sample sizes. A mixed-effect model is similar to a traditional linear regression, except it considers both fixed and random effects. Fixed effects signify variables that are hypothesized to exert a direct influence on the dependent variable. They are treated as constant across all observations in the dataset. Random effects signify the influences of variables that exhibit a fluctuating impact on the dependent variable across distinct groups. A total of 112 features were selected for the model. These features were the fixed effects. The "PROFILE_MAKE" (car manufacturer) and "CATEGORY" (price and body size cluster) variables that came from the Vehicle Specification data set were the random effects.

For the mixed-effect model, the data set was split so that 80% of the data was used for training and 20% for testing. This resulted in 8,500 data points in the test data set and 2,125 data points in training. The test data set was used to evaluate model accuracy. Model evaluation metrics used were Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared for the train and test data sets.

Descriptive Analysis

The third modeling framework developed and tested a Generalized Linear Model (GLM) to accomplish the secondary objective of this project. The second objective was to develop a

model that would allow Nissan to understand what factors moved MSRP up or down and by how much. Generalized Linear Models show the relationship between the dependent variable and the independent variable. MSRP was the dependent variable and features like vehicle specifications, customer reviews, etc. were the independent variables. The model would allow Nissan to see how much MSRP would increase or decrease based on the application of the features that were selected through the feature selection process highlighted above. Three Generalized Linear Models were built—one for each of the three categories.

The Generalized Linear Models used Gamma distribution with a log link function because MSRP was right skewed and need to undergo a log transformation to normalize the distribution. Forward stepwise selection was used as the variable selection technique. It helped identify the most relevant independent variables (predictors) to include in the GLM by iteratively adding predictors to the model one at a time until AIC stopped decreasing.

The Akaike Information Criterion (AIC) was used to evaluate the model. AIC considers the model's goodness of fit and its complexity. It looks at two values – the number of independent variables used to construct the model (k), and the maximum likelihood estimate of the model. The best-performing model that was chosen had the lowest AIC value and explained the greatest amount of variation while using the fewest possible independent variables.

Findings

Predictive Modeling

XGBoost was the best-performing model across all three categories. It had the highest R-squared and the lowest RMSE of all the models built. The XGBoost models gave different feature importance for each category, and the feature selection methodology that provided the best results also varied by category. Random Forest was also a strong contender since it also

performed well, with R-squared values between 0.91 and 0.96 and RMSE between 1,611 and 2,624.

The best-performing XGBoost model for Low-price Small size had an R-square of 0.97, and RMSE of 868. Feature selection Method 3 was used to determine which features should be included in the model. This included overlapping features that were deemed to be important in the XGBoost, Adaboost, and Random Forest machine learning models, along with statistically significant features (p-value < 0.05 & correlation > 0.10) from the Generalized Linear Model. This resulted in 198 features being used for modeling. The XGBoost model also provided the most important features in determining MSRP. The top 10 important features can be seen in Table 4.

Table 4. *Top 10 Important Features for Low-price Small size*

Feature Name
1. Manual Air Conditioning
2. Secondary Ventilation Control
3. LED High Beam
4. Pedestrian/Cyclist Avoidance System
5. Heated Front Passenger Seat
6. Start Stop System
7. Plastic Steering
8. Overtaking Sensor
9. Subwoofer
10. Navigational System

The best-performing XGBoost model for Mid-price Large size had an R-square of 0.99, and RMSE of 805. The feature selection process that led to the highest model performance was Method 3, the same as the one used for the Low-price Small size XGBoost model. A total of 204

features were selected and used for this model. The top 10 important features can be seen in Table 5.

Table 5. *Top 10 Important Features for Mid-price Large size*

Feature Name
1. Overtaking Sensor
2. Subwoofer
3. LED Low Beam
4. Compressor Turbo
5. Crew Cab
6. Rear Drive Wheels
7. Privacy Glass
8. Regular Cab
9. Plastic Steering
10. Glass Roof

For the Mid-price Mid size category, the best-performing XGBoost model had an R-square of 0.98 and RMSE of 1,814. The feature selection process that led to the highest performance was Method 2. Instead of choosing features deemed important by all four feature selection models, features that had the highest importance in at least two of the models were selected. This resulted in 205 features that would be used in the model. The top 10 important features can be seen in Table 6.

Table 6. *Top 10 Important Features for Mid-price Mid size*

Feature Name
1. Passenger Side Door Mirror Tilt for Reverse
2. Auto Dimming Door Mirror
3. Front and Passenger Seat Electric Lumbar Controls
4. Maximum Torque
5. Rain Sensor

6. Secondary Ventilation Control
7. Previous Year's MSRP Weighted by Volume
8. Screen Size Inches
9. Plastic Steering
10. Wheel Diameter Inches

The mixed-effect model performed well but was not able to outperform the XGBoost models. All but 36 features were statistically significant and had an effect on TARGET_Weighted_RetailPrice. The non-significant features were still kept in the model since they did not diminish the model's predictive power. The model results did come with a convergence warning, which suggests that the model might not have converged to the best solution and that results should be interpreted with caution. The Mean Squared Error was 4420.86, which means that on average, the model's predictions are approximately $4420.87 off from the actual values. The test RMSE was 6349.52, which means on average, the model's predictions are around $6349.52 off from the actual values when considering both underpredictions and overpredictions. The R-squared on the test data set was 0.9016, which suggests that the model explains approximately 90.16% of the variance in the test set target variable.

Descriptive Analytics

Generalized Linear Models (GLM) were built and tested for each category. The best-performing GLM for Low-price Small size included 121 features and had an AIC of 56,183.9. The model for Mid-price Large size included 133 features and had an AIC of 65,382.8, and the model for Mid-price Mid size included 125 features and had an AIC of 81,287.2. The three GLMs performed well and were able to help quantify the impact that the application of these features would have on MSRP.

Discussion

The primary goal of this project was to develop and implement a pricing model that would help Nissan set the right MSRP for its fleet of vehicles, improving sales and profitability. The model would also explain which vehicle features, macroeconomic factors and customer reviews have the most significant impact on price. The secondary goal was to develop a Generalized Linear Model that would allow Nissan to understand what factors moved MSRP up or down and by how much. Both objectives were achieved using the XGBoost models and Generalized Linear Models developed for each vehicle category.

The XGBoost models for each of the categories outperformed Nissan's existing forecasting methodology. When predicting MSRP for vehicles with model years between 2012-2023, Nissan's existing approach had an RMSE of 1,207 for Low-price Small size, 1,569 for Mid-price Large size, and 4,883 for Mid-price Mid size. The XGBoost models developed for this project had an RMSE of 825 for Low-price Small size, 929 for Mid-price Large size, and 1,316 for Mid-price Mid size. This marked a 32% improvement for Low-price Small size, and 41% and 73% for Mid-price Large size and Mid-price Mid Size respectively.

Although the mixed-effect model was a strong candidate, some shortfalls hindered it from outperforming any of the tree-based models. The MAE, RMSE, and R-squared values all demonstrate an acceptable level of performance, but it still could not beat any of the tree-based models, especially XGBoost. A mixed-effect model requires a sufficient number of groups or clusters to effectively estimate the random effects. The model we developed might not have had a sufficient number of groups, which could have contributed to suboptimal performance. In fact, the model had a convergence warning that indicated that the model might not have converged to the best solution.

There were also challenges encountered along the way when building the mixed-effect model. It was initially difficult to add features to the model and a 'LinAlgError: Singular matrix' error would appear each time. This was due to many features in the vehicle specification data set being sparsely populated across most car models. Alternate approaches were used to circumvent the issue. Any feature that introduced the error was systematically excluded from consideration, which meant that 112 features could still be included in the model.

XGBoost proved to be the best-performing model. Although the Random Forest models performed well, XGBoost was better due to some inherent properties. By virtue of being a boosting algorithm, the sequential manner in which the model is built, each new model corrects mistakes made by previous models. This optimization reduces bias and variance, thus helping improve model performance. XGBoost also provides several regularization techniques that help prevent the model from overfitting and improves the model's ability to generalize.

The EDA process showed non-linear relationships between the features in the data set and the target variable (MSRP). Additionally, even though features used in the models were significantly reduced from what was available in the data set, each model still had at least a minimum of 100 features. XGBoost is good at capturing non-linear relationships between features and the target variable and can effectively handle high-dimensional feature spaces. These qualities further make XGBoost's performance stronger than Random Forest and Adaboost.

Despite the XGBoost model's strong performance, there are some limitations. Hyperparameters need to be carefully tuned, and finding the best combination of these hyperparameters can be time-consuming and challenging. If hyperparameters are not tuned correctly, it can lead to overfitting or underfitting, which can affect the model's performance.

XGBoost can also be sensitive to outliers and other noise in the data. If there are outliers or noise in the training data, it can significantly impact the model.

Despite strong modeling performance, there is one overall limitation to all the work done in this project. The issue Nissan is facing concerning a drop in sales and profit is not only caused by inaccurate MSRP predictions. It is just as much a result of marketing and brand image problems. Although the XGBoost and Generalized Linear Models provided to Nissan can address and tackle one side of the problem, it is up to Nissan to fix its brand image and convey the quality of the vehicles.

Conclusion

This project explains and showcases three XGBoost models that can predict the MSRP of Nissan vehicles as well as that of its competitors and also explain which vehicle features and macroeconomic factors have the most significant impact on price. Additionally, the project provides Nissan with three Generalized Linear Models that allow decision-makers at Nissan to change strategy to obtain the most optimal MSRP, by assessing the unit impact that various features have on price. Each model addresses a specific vehicle category (Low-price Small size, Mid-price Large size, and Mid-price Mid size).

Nissan can use the models when determining the price of future vehicles two to three years before they hit the market. It will improve prediction accuracy over the methodology currently in use. It is important to note that both the data and the model should be updated regularly since there are rapid developments in vehicle technology and specifications. The data should be reflective of that if the model is to have the best predictive accuracy. Since price has an impact on manufacturer brand value and sales, a more precise pricing model will boost the company’s sales and profitability. It is estimated that had Nissan been using the XGBoost

models to determine MSRP, it could have earned an additional $168 Million USD from 2019 to 2023. The methodology used to calculate the financial impact was verified and approved by Nissan.

The next steps of this project would be to expand upon this work and build an optimization model to find a balance between sales volume and MSRP. Additional next steps include deploying a pipeline to automate all the necessary tasks. This includes clustering, feature selection, and running the machine learning models for each cluster.

References

Coffin, David & Downing, Dixie & Horowitz, Jeff & LaRocca, Greg (2022). The Roadblocks of the COVID-19 Pandemic in the U.S. Automotive Industry. United States International Trade Commission.

Listiani, Mariana (2009). *Support Vector Regression Analysis for Price Prediction in a Car Leasing Application* (Masters Thesis). Hamburg University of Technology.

Nadanyiova, Margareta & Gajanova, Lubica & Moravcikova, Dominika & Oláh, Judit. (2019). The Brand Value and its Impact on Sales in Automotive Industry. LOGI. 10. 41-49. 10.2478/logi-2019-0005.

Szilágyi, Roland & Varga, Beatrix & Geczi-Papp, Renata. (2016). Possible Methods for Price Forecasting. 10.26649/musci.2016.135.

Totakura, Sri Sai Ganesh Satyadeva Naidu & Kosuru, Harika (2021). *Comparison of Supervised Learning Models for Predicting Prices of Used Cars* (Bachelor Thesis). Blekinge Institute of Technology.

Appendix A: Correlation Between MSRP and Other Variables

Correlation heatmaps to determine the correlation between the target variable (MSRP weighted by each vehicle grade's volume) and various features from six of the data sets provided by Nissan, as well as macroeconomic factors in the US economy. This was used as a first attempt to understand which features should be selected for the models that were going to be developed.

Figure A1. *Correlation Heatmap of Target Variable and Specs Data*

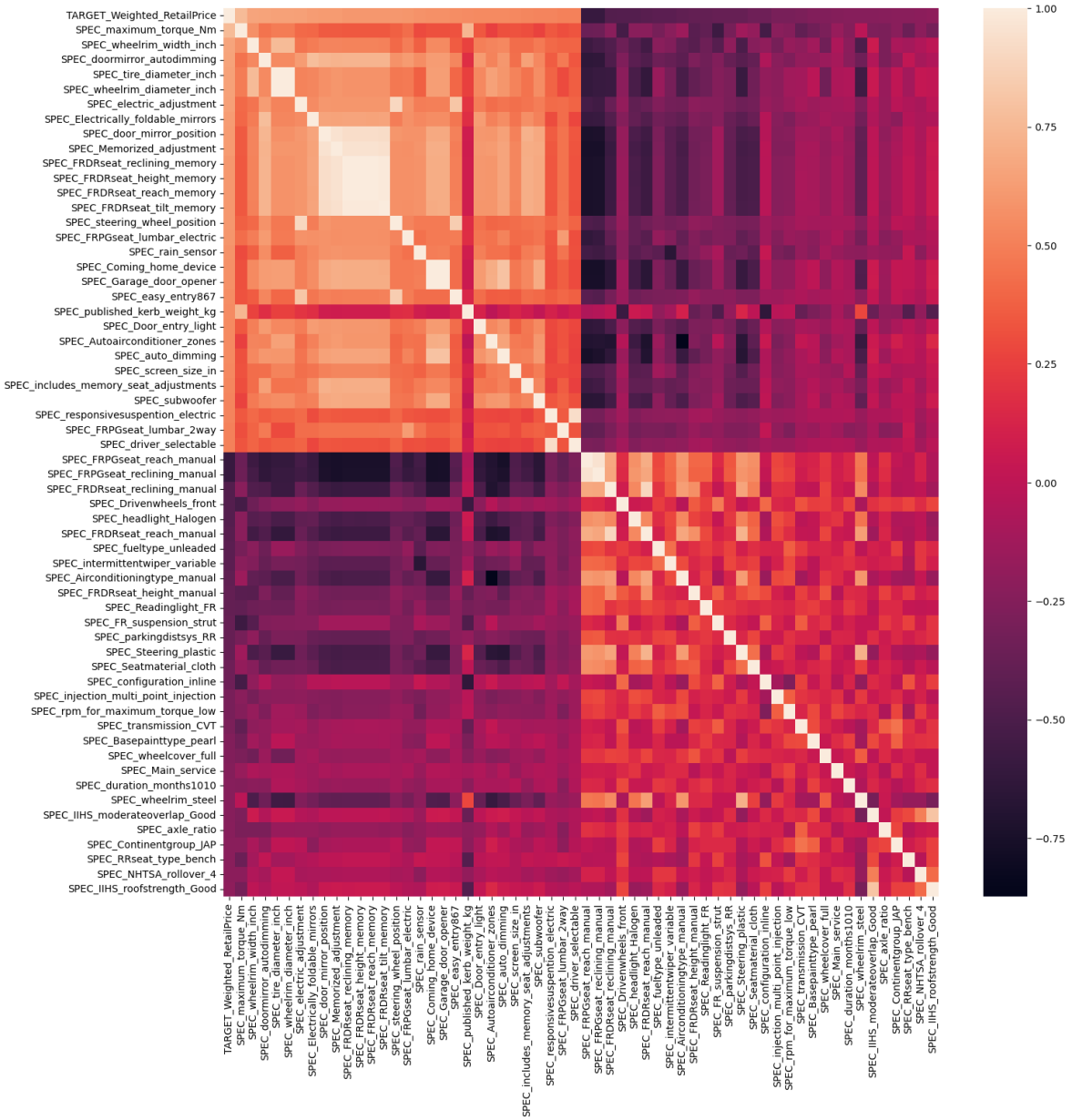


Figure A2. *Correlation Heatmap of Target Variable and Residual Value*

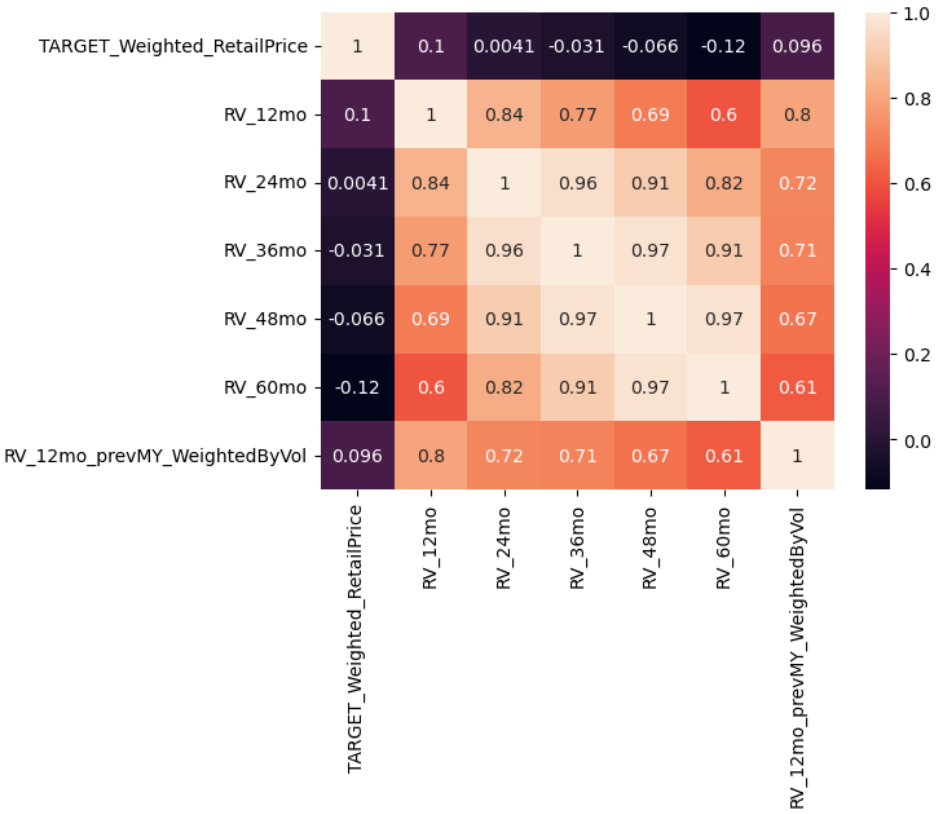


Figure A3. *Correlation Heatmap of Target Variable and Brand Value*

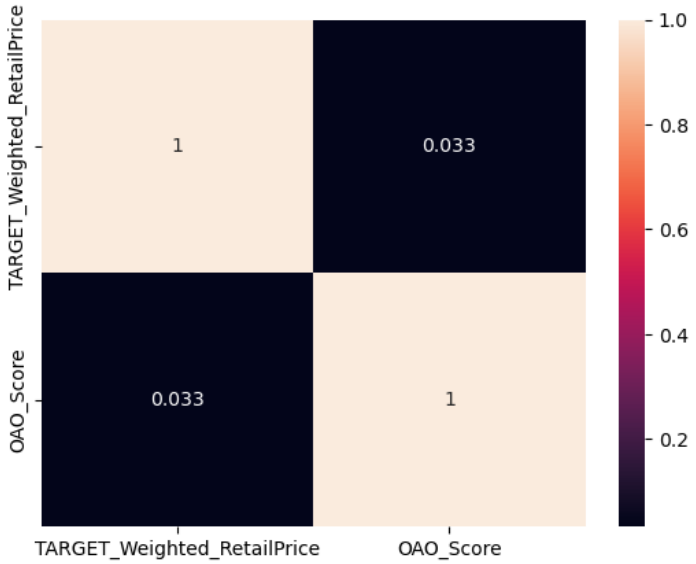


Figure A4. *Correlation Heatmap of Target Variable and Edmunds.com Reviews*

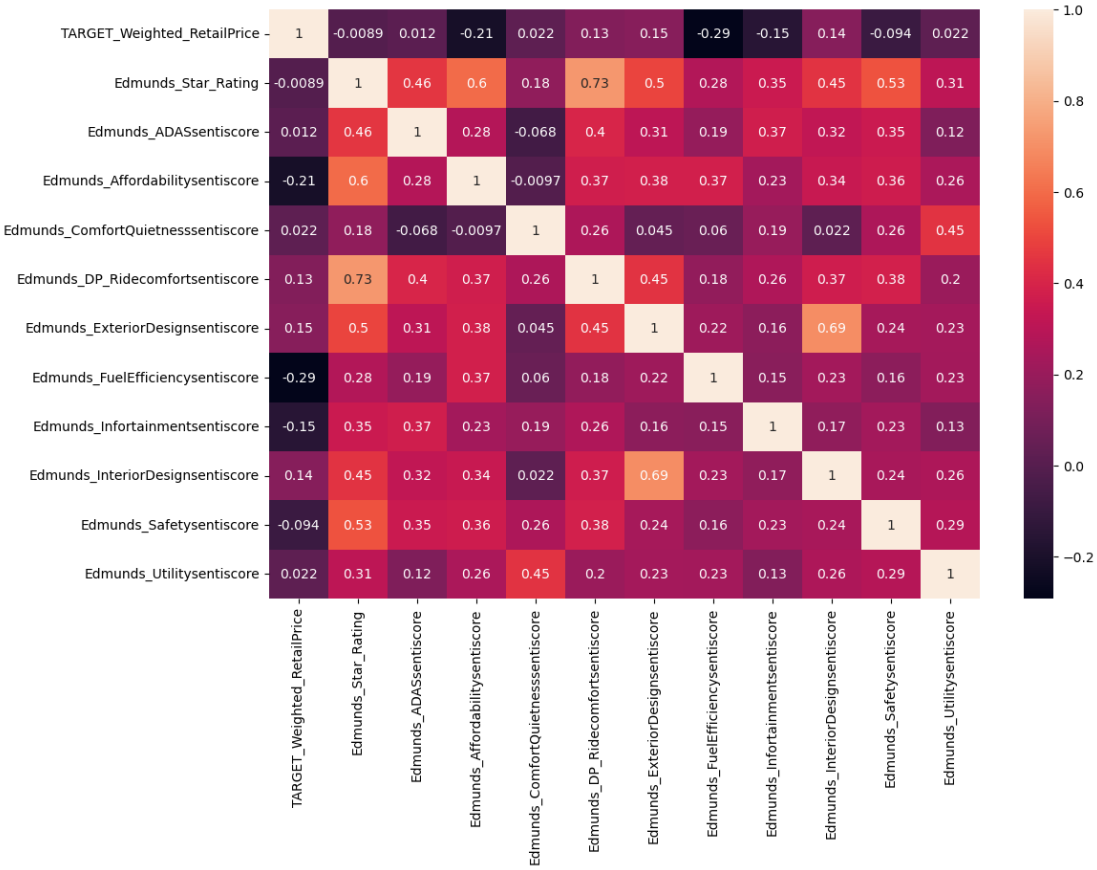


Figure A5. *Correlation Heatmap of Target Variable and Sales Volume*

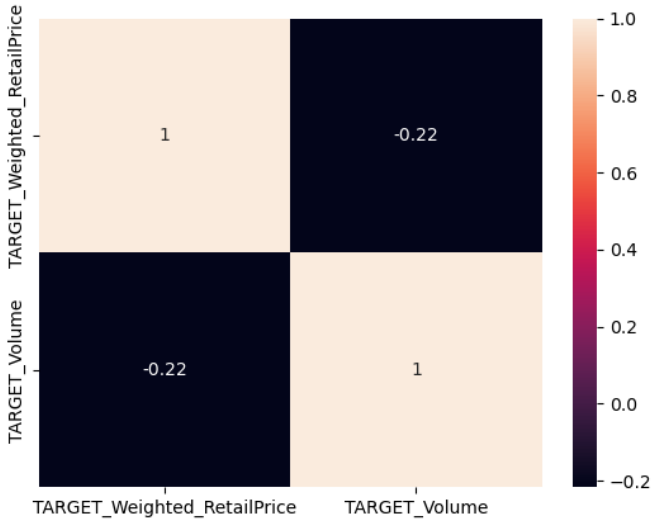


Figure A6. *Correlation Heatmap of Target Variable and NCBS Research*

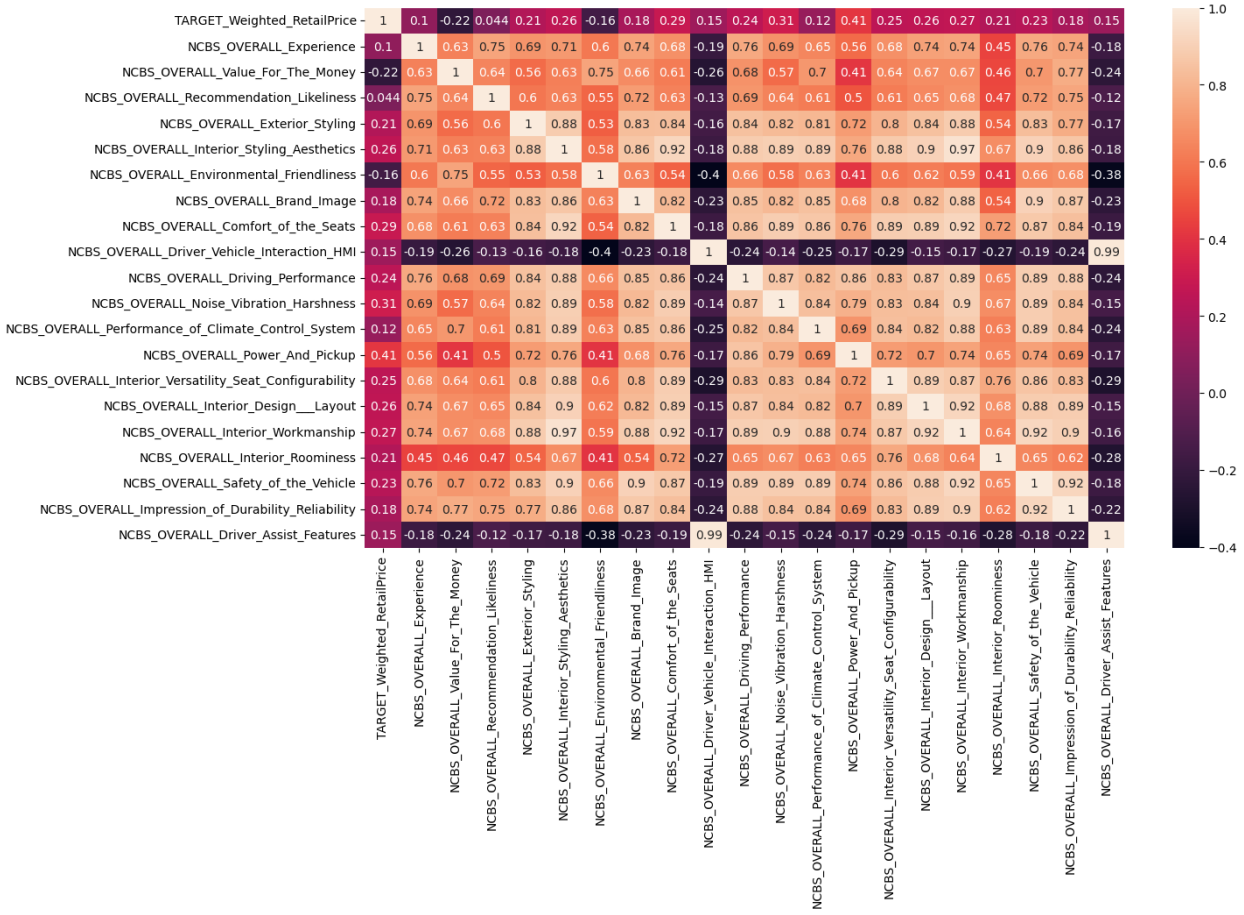


Figure A7. *Correlation Heatmap of Target Variable and Inventory*

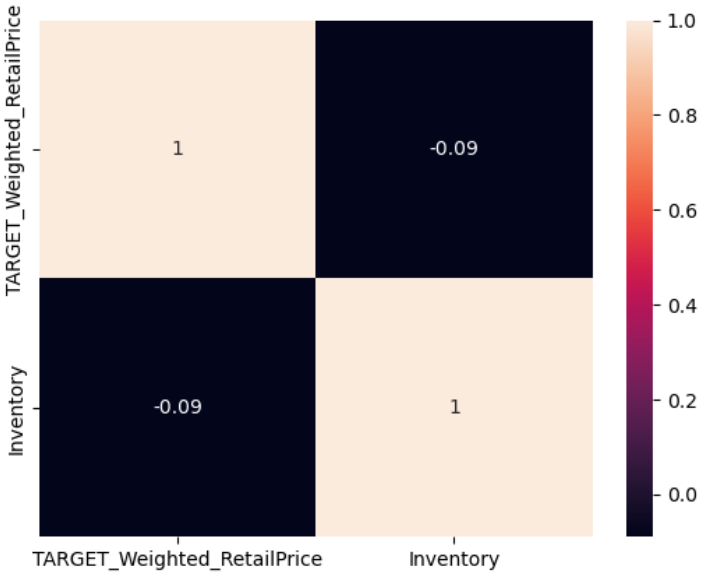


Figure A8. *Correlation Heatmap of Target Variable and Macroeconomic Factors*