Price Forecasting Model for Nissan Vehicles and Identification of Factors Impacting Price

Roma Chitale
Vamshi Gadepally
Daisuke Yagyu

Advisor: Anil Chaturvedi, PhD

Table of Contents

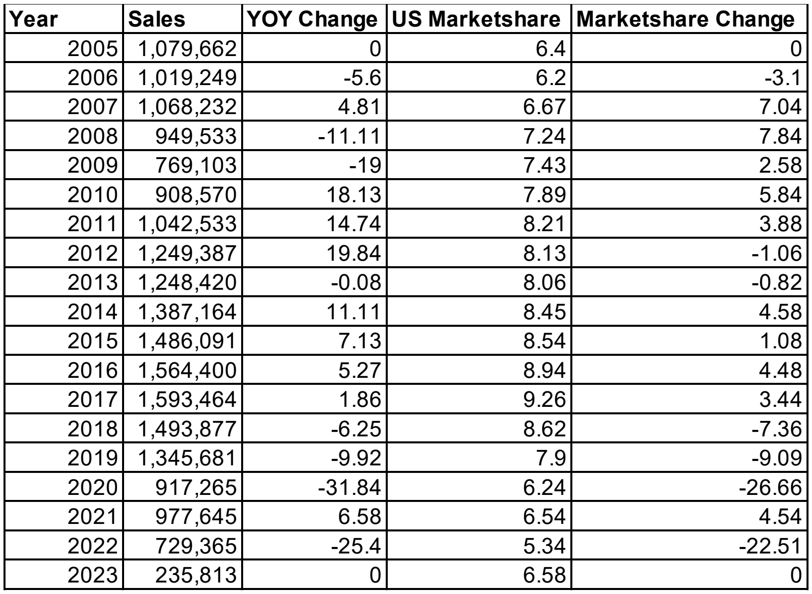
Business Partner

Nissan Motor Group is a Japanese multinational automobile manufacturer headquartered in Yokohama, Japan. Founded in 1933, it has since 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 electrics 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. We are working with Nissan's Market Intelligence department, which is tasked with transforming the product planning process using data management and data science.

Background

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 numbers in over a decade. There are a multitude of factors that have contributed to this, including the COVID-19 pandemic, which popularized work-from-home culture, and the semiconductor chip shortage.

Nissan in particular has been experiencing more challenges as of late. In 2021 global revenue was ~$71B, while revenue for the US market was ~$19.5B. Historically, Nissan has been the third most popular Japanese car manufacturer in the United States, only behind #1 Toyota and #2 Honda. It was close to overtaking Honda as the #2 Japanese auto manufacturer in 2016 and 2017. However, Nissan began prioritizing volume over profit by discounting their cars and overloading car rental fleets with their vehicles. This greatly hurt Nissan’s brand image and resulted in declining sales.



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

Figure 1: Nisan's vehicle sales in the US market from 2005 - 2023

Business Problem

Nissan has been struggling to set the right MSRP (Manufacturers Suggested Retail Price) for their vehicles, which directly affects sales and profitability. The current pricing strategy uses a simple linear regression model that is based on historical pricing trends. It is not optimized to consider various macroeconomic factors, vehicle features, and the target audience, which is impacting revenue and brand perception. As evident by the decline in sales after discounting its vehicles, setting the right price is critically important. Nissan needs to establish a more effective pricing strategy to address these challenges and improve their overall business performance.

Opportunity Statement

We have an opportunity to leverage various data science and analytics techniques to develop a better pricing model that Nissan can use to set the right MSRP for their vehicles. By understanding what vehicle features and external factors affect price, we can build a more accurate and effective pricing model. This gives Nissan the opportunity to increase sales and revenue and enhance brand perception. An optimized pricing strategy will help Nissan position their vehicles competitively, target the right audience, and reduce inventory costs. With an effective pricing strategy in place, Nissan can improve their market share, increase profitability, and strengthen their position in the automotive industry.

Project Outcome

The project goal is to develop and implement a pricing model that will help Nissan set the right MSRP for their fleet of vehicles and in turn increase sales and profitability. This is of the utmost importance because setting the MSRP too high can deter customers from purchasing a vehicle, which impacts sales and revenue. If set too low, it diminishes brand value and profit margins.

Root Cause/Symptoms

A major reason why predicting MSRP is a difficult task is because car manufacturers set the MSRP based on many internal and external factors. Some factors that are taken into consideration are manufacturing, material and processing costs, vehicle specifications, competitor pricing, and various macroeconomic factors like inflation, GDP, unemployment, and vehicle subsidies. As a result, car price prediction can be a challenging task due to the high number of attributes that should be considered.

Another limitation on any predictive model's performance is the dataset. If the dataset does not include features that are strongly correlated to the price, the algorithm might not have access to enough information to accurately infer price. On the flipside, having irrelevant features also affects the model's accuracy and can give inaccurate predictions. This makes feature selection a very important pre-processing step, that needs to be performed. This is a step that Nissan isn't currently performing with their current methodology.

Nissan's existing method takes a simple historical trends approach and considers only a few factors, making it less comprehensive than other methods. It doesn't consider customer evaluations of older models, brand power or macroeconomic factors that impact the industry. An excessively simplistic price forecasting method like the existing one, will unsurprisingly be inaccurate. For these reasons, having robust datasets, doing all of the relevant pre-processing work, and evaluating multiple machine learning algorithms will allow us to develop better and more accurate models.

Goals of Analysis

The overall goal of the project is to help Nissan understand the factors affecting the price of cars in the American market and build supervised model(s) to predict the best MSRP using the most important internal and external features.

This will be accomplished by first sourcing and preparing datasets on which all the algorithms can be trained and evaluated on. As mentioned earlier, most datasets rarely include only relevant features. Including irrelevant and redundant features impacts performance by reducing the model’s ability to generalize on data it hasn’t seen before, and it can impact accuracy. For these reasons, it’s highly important to select and include only relevant in the model.

The next step is to build and provide Nissan with supervised learning model(s) that can be used to set MSRP. There are several supervised machine learning models that can be used for price prediction. Some of the models up for consideration are Ridge and Lasso regression, boosting algorithms like Adaboost and XGBoost, as well as Random Forest, Support Vector machine, and K-Nearest Neighbors. We will use evaluation metrics like mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) to determine the best performing model.

The objective of supervised models is to use historical data to train the model and use that to make predictions on new and unseen data. In the context of our project, we will select the model that is best at predicting MSRP for new and upcoming vehicles in future years that are not included in the training dataset. In order to do this, we also need to be able to forecast macroeconomic factors.

Lastly, the model we provide to Nissan will also deepen their understanding of the most important vehicle features, and how various macroeconomic factors affect price. This will help Nissan make more informed decisions when launching new vehicles.

Data and Sampling

Since the main objective of the project is to build a better pricing model for Nissan, the response variable is MSRP for new cars. The frequency of data update depends on the vehicle model. In the US market, it is typical that a vehicle model updates its price yearly, unless there are some specification changes that facilitate the need to update the price before then. The unit of analysis is car version, which includes model name, model year and grade name.

Nissan has provided us with many datasets that can be used to build our supervised ML model. The following is a brief description of each dataset and how we can use it in out model to forecast MSRP:

- Features application data: Application of various technologies and functionalities that a vehicle is equipped with. Not all, but some key features may increase the product's MSRP. For example, if a vehicle has an LED or halogen headlamp, those differences can be shown in this data.
- Web evaluation data: Customer’s word-of-mouth information on a car review website called Edmunds.com, including customer review stars, are collected using web-scraping. When customers are satisfied with the product, the vehicle can be sold more, even with a high price.
- Customer research data: Market research to understand how customers are satisfied with the product can be collected for Nissan and its competing vehicles. When customers are satisfied with the product, the vehicle can be sold more, even with a high price.
- Used car values: Resale prices in the used car market can also be used as used car values. When a used car's price is higher, the new car's MSRP would also be higher.
- Sales volume: The number of vehicles sold in a year. The prior year's sales volume can impact the pricing of the next model year.
- Invoice (Transaction price at dealers): Car prices customers pay at dealers also may influence MSRP. The price gap between the Invoice and MSRP would be influenced by brand power and the attractiveness of the models.
- Segment: MSRP and features relationship differ depending on the customer type. A categorical variable that shows the vehicle is compared within what vehicles are another aspect to care for.

- Brand Power: Nissan conducts a tracking survey to know how customers evaluate each brand. High-brand power products can be sold more, even with a higher price.

Although we have many datasets that contain a lot of useful information, there are still some data gaps. The first thing we are missing is data on macroeconomic factors. As mentioned in earlier sections, things like inflation, subsidies, and material cost have an impact on MSRP. This means that including these factors is key to being able to build the best performing model. Additionally, car age is a critical factor in determining the car price. It is essential to know the number of years since the last vehicle model refresh because it is too old, the technology and design becomes outdated, and the value depreciates.

Analysis Plan

This project will consist of two analyses:

1. A pricing model to help set the right MSRP
2. Time-series analysis of macroeconomic factors

Pricing Model for MSRP

1. Data Collection: The first step is to collect the right data. This should include a wide range of features that impact MSRP. Examples include—historical MSRP, transaction/invoice data, sales volume, and customer feedback.

2. Data Pre-processing and Transformation:

   Important steps include dealing with:

   - Missing values and outliers
   - Splitting the data into train/test/validation sets
   - Feature scaling using methods like normalization, standardization, or scaling to a specific range
   - Converting categorical variables using encoding methods like one-hot encoding

A critical step here is feature selection. We intend to use relevant statistical tests and feature ranking techniques. Methods under consideration are:

- Using a Correlation matrix, which involves calculating the correlation coefficient. between the target variable (MSRP) and each of the predictor variables.
- ANOVA F-test to determine if there is a significant difference in the means of the predictor variables for different levels of the of the target variable (MSRP).
- Recursive feature elimination technique that involves recursively removing the least important predictor variables until the desired number of variables is obtained.

Another crucial step is to incorporate the results of the impact analysis of the macroeconomic factor time series model. This involves estimating the impact of each macroeconomic factor on the MSRP and feeding that information into the pricing model. We will discuss this in greater detail in a later section.

The last data pre-processing step is using Natural Language Processing (NLP) to transform customer reviews into quantitative values. The Edmunds dataset includes customer evaluations on vehicle specifications and other features for both Nissan and competitor vehicles. We will use NLP techniques extract useful information from customer reviews and to get an insight into customer sentiment regarding vehicles.

3. Model Training: We intend to experiment with multiple algorithms to find the one that works best for our problem. Each method has its own advantages and disadvantages. Most can capture complex interactions between features and can handle high-dimensional data.

   Up for consideration are:

   - Random Forest – resistant to overfitting and deals well with missing data.
   - K-Nearest Neighbour – can capture nonlinear relationships and no assumptions about the underlying data distribution are required.
   - Gradient boosting algorithms like LightGBM and XGBoost – can handle missing data well and are generally efficient in terms of memory usage and computation time.

- Ridge and Lasso Regression – can effectively deal with multicollinearity and Regularization using the penalty term can prevent overfitting. They also can handle large datasets well.
- Support Vector Machines (SVMs) – can handle nonlinear relationships and can prevent overfitting using the C parameter.
- Artificial Neural Networks (ANNs) – can handle missing data well and can learn from unstructured data too.

4. Hyperparameter Tuning: Hyperparameters are settings that control the behaviour of each model.

   There are several ways to do hyperparameter tuning including:

   - Cross-validation— it is a good technique for selecting the best hyperparameters by training and evaluating the model on different subsets of the training data. K-Fold Cross-Validation and Leave-One-Out Cross-Validation are up for consideration. By tuning the hyperparameters of the model and evaluating its performance using cross-validation, we can choose the set of hyperparameters that result in the best overall performance on the data.
   - Grid search involves specifying a range of hyperparameters for each model and then training and evaluating the model for all possible combinations of these hyperparameters.
   - Random Search: Random search involves specifying a range of hyperparameters for each model and then randomly sampling a set of hyperparameters to train and evaluate the model.

5. Model Evaluation: We will evaluate the performance of selected models based on certain evaluation metrics like mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). This will be done on the test dataset.

   - Cross-Validation using bootstrapping can be used to get the average performance metrics across all bootstrap samples.

- Plotting the residuals (difference between the predicted values and actual values) against the predicted values to identify patterns in the errors. The model with the smallest and random residuals is preferred.
- Learning Curves: Plot the training and validation error against the size of the training data to check if the model is overfitting or underfitting. The model with the lowest validation error and a small gap between the training and validation error is preferred.

6. Feature Importance Analysis: The choice of the feature importance technique will depend on the best model that we select in the previous step. Below are analyses that we can use depending on the model that is selected:
   - Random Forest – measure the mean decrease in Gini impurity of a node when a variable is selected as a split in a tree. The importance of a feature is then the average of the total reduction in impurity over all trees.
   - K-Nearest Neighbor – Recursive feature elimination where the importance of a feature is then based on the drop in model performance when the feature is removed
   - LightGBM – we can use the SHAP values to give us importance scores for each input feature.
   - XGBoost – importance of each feature based on the number of times it is used to split the data and the average gain of the splits. The higher the gain, the more important the feature.
   - Ridge and Lasso Regression, SVM, ANN – we use their model Coefficients / Weight values or Recursive feature elimination to indicate the strength of the relationship between each feature and the target MSRP.

7. Model Comparison with Existing Forecasting Method and Deployment: We will compare our best performing model to the existing forecasting method used by Nissan, and then use it to make predictions on new data.

   Since our end goal is to develop an improved model that can replace the existing methodology, we are leaving the comparison step for the end of the modelling process, after the new model has been developed, evaluated, and deemed ready for deployment. We will use appropriate performance metrics and statistical tests to help determine if the new model actually has better performance than the existing model and whether the improvement is statistically significant.

Analysis Plan: Macro-Factor Time Series Analysis

1. Data Collection: First, we will define the macroeconomic factors that may impact MSRP. Examples include, inflation, GDP, interest rates, and consumer confidence.

2. Data Pre-processing and Transformation: Since this is a time series analysis, stationarity is a key assumption. For this reason, we want to create a stationary and well-behaved time series dataset that is suitable for analysis. Some steps we will take are:
   - Check for stationarity using the Augmented Dickey-Fuller (ADF) or the KPSS tests.
   - Check for any trends and seasonality in the data that could be contributing to non-stationarity. We can detrend and de-seasonalize the data by using seasonal differencing or seasonal decomposition.
   - Box-Cox transformation to stabilize the variance.
   - Standardize the data to ensure that each variable has equal weight in the analysis.

3. Model Selection and Estimation: Need to identify and select the best model that can be used to identify and analyze the relationship between MSRP and various macroeconomic factors.
   - Autocorrelation and partial autocorrelation analysis to identify the order of the autoregressive (AR) and moving average (MA) components in the model.

   Once the order has been identified, various models can be fitted to the data. A few models up for consideration are ARIMA, SARIMA, VAR and ARIMAX:
   - ARIMAX is a good candidate because it extends the basic ARIMA model by incorporating the effects of exogenous variables (in this case the macroeconomic factors) that may have a causal relationship with the dependent variable (MSRP).
   - Vector Autoregression (VAR) is a multivariate time series model that can be used to analyze the relationship between multiple time series variables. It is particularly useful when analyzing the relationship between the price and multiple macroeconomic factors.

The techniques that can be used to estimate the parameters of the selected model are:

- Maximum Likelihood Estimation estimates the parameters that maximize the likelihood function of the model given the data.
- Method of Moments involves equating the theoretical moments of the model to the sample moments of the data. The parameters are estimated by solving the resulting equations.

4. Model Validation: Compare the different models based on their ability to fit the observed data and make accurate forecasts using AIC, BIC, and Mean Absolute Error (MAE).The best model can be selected based on which one has the lowest criterion values. The selection criteria consider both the goodness of fit of the model to the data and the complexity of the model, so we're seeking a good balance.

   It's also important to validate the model by checking for patterns in the residuals to ensure they are random and do not show any trend or seasonality.

   - Plot the residuals – The residuals should be randomly distributed around zero, indicating that the model has captured all the information in the data.
   - Check for autocorrelation and normality – We will use the Ljung-Box test to check for autocorrelation in the residuals and the Shapiro-Wilk or the Kolmogorov-Smirnov test to check for normality.

   If the residuals are not random, then it may indicate that the model is mis-specified or that there are other factors that are not captured by the model.

5. Refine the model: If the best model as per the selection criterion fails any of the validation tests (residuals show patterns or trends, or if there is autocorrelation or heteroscedasticity), we will re-estimate the model parameters and refit the model. This entails repeating steps 2-4 until the model passes all validation tests.

6. Impact Analysis: This is done to quantify the effect of changes in the independent variables on the dependent variable, in other words assess the effect of changes in macro variables on the

MSRP of new vehicles.

Some analysis we want to look into are:

- Granger causality test is used to identify whether the past values of the macroeconomic factors have a statistically significant effect on the future values of the dependent variable (MSRP). Those that aren't can be excluded from the model.
- Impulse response function (IRF) is used to quantify the impact of each macroeconomic factor on the MSRP over time by estimating the response of the MSRP to a unit shock in the independent variable.
- Variance decomposition determines the proportion of the variation in the MSRP that can be explained by each of the macro variables. It involves decomposing the variance of the MSRP into the contributions of each of the macroeconomic variables.

7. Interpretation and Incorporate into MSRP Model: We will interpret the results of the time series analysis to understand the impact of macroeconomic factors on the MSRP for new vehicles. We can select relevant factors and incorporate that into the pricing model. This can be done by incorporating the estimated coefficients for each macroeconomic factor in the selected pricing model.

   For example, if the time series model shows that an increase in GDP and a decrease in inflation are associated with an increase in the MSRP, we can incorporate the estimated coefficients for these two factors into the pricing model and use it to predict the MSRP of future vehicles.

   In practice, this can be done by specifying a range of values for each macroeconomic factor and using the estimated coefficients to forecast the MSRP for each scenario. The scenario that yields the highest forecasted MSRP can be selected as the best MSRP for new vehicles.

Methodology for Model Evaluation and Validation

To evaluate our model's performance, we will compare it to the MSRP predictions made by Nissan's existing methodology, using evaluation metrics like mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).

For model validation, we will split the data into a train, test, and validation datasets. Second, we will decide the best model hyperparameters based on k-fold cross-validation with randomized search. Finally, we will validate the selected model using the test dataset.

Expected Outcomes

There are two main outcomes we expect out of this project.

1. Build Forecasting Models:
    a. Pricing Model: We want to build a model(s) that can more accurately forecast MSRP for future vehicles. The scope of the model will be determined by upcoming EDA work. The goal is to have one catch-all model that can be used for all Nissan vehicles. However, if the EDA determines that macroeconomic factors have a different effect on various vehicle segments, or that the importance of vehicle features is different for different vehicle segments, we will have to develop multiple models.
    b. Time series model to forecast macroeconomic factors: We want to be able to forecast macroeconomic factors that affect MSRP and incorporate these features into the final pricing model.

2. Feature Importance: A secondary goal of this project is to be able to provide Nissan with a deeper understanding of what vehicle features have the biggest impact on MSRP and explain how various macroeconomic factors interact with and influence MSRP.

Literature Review

1. “Comparison of Supervised Learning Models for predicting prices of Used Cars” (2021, Tokakura, Kosuru)

Focus (knowledge gap to be filled):

- To find the approaches that can be utilized to forecast the car pricing.
- To find out the critical essential features from many explanatory variables and if any crucial elements would impact the pricing of the motor industry.

Hypothesis:

- Machine learning techniques are effective in forecasting consumer durable goods pricing.
- Feature importance with the tree-based algorithm would allow us to get essential features.
- Industry across important elements that can be referred to for our study.

Points to be utilized in the project:

- Aiming to detect attributes that impact the price of used cars, they tried to forecast prices using the algorithms of Linear Regression, Light Gradient Boosted Machine(LGBM), Random Forest, and Decision Tree. LGBM performed the best for their modeling with 91% on test data.
- They confirmed the LGBM models’ feature importance to determine the critical factors that would influence the price of cars. Also, heat map visualization was also effective in selecting impactful features usable for analysis.
- It turned out that Region, odometer, manufacturer, year, paint color, type, cylinders, drive, and transmission are the top 10 essential features.

2. “Car Price Prediction using Machine Learning Techniques” (2019, Gegic, Isakovic, Keco, Masetic, Kevric)

In this study published in the TEM journal researchers proposed to predict used car prices in Bosnia using typical regression models and different machine learning techniques too like Artificial Neural Network, Support Vector Machine and Random Forest.

The dataset collected in this research was historical data scrapped from a local automobile review website and several variables were considered. Respective performances of different algorithms were then compared to find one that best suits the available data set.

One of the key takeaways from this study was the researchers identifying the advantages of using certain models over others. It was concluded that the regression model that was built using SVM can predict the price of a car with better precision than multivariate regression or some simple multiple regression. This was on the grounds that SVM is better in dealing with datasets with more dimensions and it is less prone to overfitting and underfitting.

Another was the ANN model which dealt with nonlinear relations in data well which was not the case with other models that were utilizing the simple linear regression techniques. The non-linear model can predict prices of cars with better precision than other linear models.

Random forests method was shown to handle large datasets efficiently, normalization and scaling weren't required, and it estimated missing values with better accuracy which was common in this study since the data was scraped from web portals. As a result of these findings from this study, we wish to consider the mentioned ML methods for our MSRP model.

The other important takeaway was that applying single machine algorithm on the data set had an accuracy less than 50%. However, the ensemble of multiple machine learning algorithms was proposed, and this combination of ML methods resulted in a higher accuracy of 92.38%.
This is a significant improvement compared to single machine learning method approach. Therefore, we will investigate this method and its suitability to our project but one drawback of this proposed technique is that it consumes much more computational resources than single machine learning algorithm.

3. “Possible Methods for Price Forecasting” (2016, Szilágyi, Varga, Geczi-Papp)

Focus:

The paper gives an overview about the different time series methods that can be used for price forecasting. The study compares decomposition and stochastic methods and introduces a hybrid method based on creeping trend and harmonic weights.

Hypotheses:

- Determining price is important because it is a major factor of a company’s competitive advantage and has a direct impact on the future of a company.
- Decomposition and Stochastic methods have their advantages and disadvantages, but a hybrid approach will have the best results.

Review of the research:

- Goal of the study was the forecast the price of methanol,
- Initially 6 independent variables were selected but since methanol price is non-stationary and has an autocorrelation, a 7th variable had to be introduced.
- Selected method was forecasting based on creeping trend with harmonic weights.
- Determined that one method usually cannot give reliable results to it’s important to try several methods.

References

https://www.temjournal.com/content/81/TEMJournalFebruary2019_113_118.pdf

https://www.diva-portal.org/smash/get/diva2:1609361/FULLTEXT02

https://www.researchgate.net/publication/313679898_Possible_Methods_for_Price_Forecasting