Task Brief : This report is an iteration over previously submitted one with additional workings for linear regression model to confirm the explainability of identified variable.

Business Problem

We're an Indian automobile firm "JaiHind Auto". We're planning to enter/ launch into the US markets which is primarily[1] split across automobile manufacturer/ designers from US, Europe and Japan. From a product development aspect i.e., automobile manufacturing, we're in the design phase, and looking at design iterations based on market feedback. Context being, we're working towards a product launch in US market. While the core performance aspects in the design is our USP, we're interested in understanding price sensitivity in US markets, specifically what aspects of car impacts its price, since it is quite different from Indian customers where we have been doing fairly well, and now we aim to have 'a foot in the door' in a ~$80 billion [2] auto market.

We've tasked a data analytics firm to find answer to the problems:

- What features impact car price?
- How well do these features explain the price?

Background

Its been 5 years since our first car rolled out of the shop floors into customer garages, and we have been profitable for last 4 quarters now. While we continue to double down and cater to 4 of the 30 Indian states and preferences in a ~$32 billion market[3]. Based on investor and promoter confidence, we do have expansionary plans which primarily focuses on making design adaptations and leveraging the current supply chain issues through inhouse smart logistics solutions. Given lack of visibility into growth, and launch aspects, we focus on executing this assignment through an external automotive consultant.

We kind of have good rapport with the data analytics firm and its processes since they are a global firm and have been our partners since a decade - the day we got our first round of funding. Given a problem statement, we typically expect a market survey, subsequent analysis and recommendations as the scope of deliverables. One of the task tenet is to consider using the available independent variables for modeling the price. Management wants to control the levers with each such independent variable – which are candidates for business strategy, design manipulation, etc. Understanding the pricing dynamics in a new market is one of the key operational objectives of this analytics/ modeling project. We have received details of the market survey conducted by the firm.

[1] https://en.wikipedia.org/wiki/Automotive_industry_in_the_United_States
[2] https://www.grandviewresearch.com/industry-analysis/us-automotive-aftermarket
[3] https://www.ibef.org/industry/india-automobiles

We hear Binay has gathered the datasets from across the American market and cleansed them a bit. Binay P Jena is the lead consultant deployed on this task and subsequent sections of this project report is compiled and consolidated by Binay/ firm.

Data Description (Preparation)

Based on market survey, we had the compiled dataset[4] (**`CarPrice_Assignment.csv`**) with various fields capturing car features, dimensions, its price and other brand/ identification details. 22 unique auto brands were captured in the dataset.
In most cases the dataset rows were sourced directly from the dealership pos machines and registration records, through some instances of typos/ spelling errors, it does seem like surveyor used manual methods of data entry (unless we want to speculate that pos machines or vehicle catalogs having typos which have very remote possibility).
Data cleansing by force mapping or hardcoding to correct automaker names was followed for cleaning up. We didn't notice any duplicates or NULL values, so dataset was fairly clean.

Methods

Following techniques are used for analyzing the the dataset for its intended objectives:

- Exploratory analysis
  - visualizing the data for price distribution i.e., plot the price distribution for car in a histogram and smoothen the best fit curve
  - visualizing categorical data i.e., a histogram plot of categories like cars by brand names, by fuel type and type
  - decode/ symbolling i.e., categorizing cars based on prices into numerical decode
  - assessing price of cars against various features like car specification/ performance (like engine type), brand, type, fuel type, aspiration, number of doors, enginelocation, number of cylinders, fuel system, drive wheel
- Feature Engineering
  - We try to derive new features (using the variables) based on their correlation and relevance. This is by attempting to aggregate the variables (for example fuel consumption is a combination of `citympg` and `highwaympg` which we try to assign similar weights but based on customer usage patterns. Weights of 0.45 and 0.55 were chosen for both of these fuel usage patterns)
  - Assessing fuel economy vs price for correlation aspects.
- Based on exploratory / visual analysis, we come up with list of significant variables and find the correlation amongst them for model development
- Model Development
  - For model development first convert categorical variables to numeric ones (like `cyclindernumber`, `carsrange` i.e., *brand_category*, `enginetype`, `fueltype`, etc)

[4] https://www.kaggle.com/code/shaistashaikh/carprice-assignmentexample/data

  - Split data into 70-30 as train and test datasets, and the dataset is scaled
  - For linear regression, we use the OLS ordinary least square method (just for its simplicity) to identify and eliminate variables. Each iteration we look for the correlation coefficient and p-value to come up with variable to eliminate, and we continue to iterate cycles until normalization of variables

Analysis

Data Exploration/ Visualizations

- The price distribution plot is skewed to right, so most prices are in the lower ranges (< $15,000)
- Mean and median of the prices seems considerably different, which hints at the spread of price points being uneven or likely higher variance
- Price points are spread out farther from the mean since 80% of cars have prices <$20000 and the rest 20% are between $20000 and $45400
- `gas` fueled cars are more in use than `diesel` ones
- `Sedan` is the most preferred car type
- From the car symboling exercise based on prices, we convert each row of the dataset to a numerical decode
  - Cars marked as -1 symbols seems price high (with insurance rating -1 good, it does seem fine). It looks like symbolling with 3 has price similar to -2 value.
  - Price dip is noticed at symbolling 1
- If we're to observe Engine type `ohc` is the most preferred engine type
- `ohcv` is typically the priciest, while `ohc` and `ohcf` have a low price range (there is just one row for `dohcv`)
- `Toyota` seems to be the most purchased car brand
- `Jaguar` and `Buick` seem to have the highest average price
- `diesel` fueled cars are higher priced than `gas`
- There is a higher average price for `hardtop` and `convertible`
- Number of doors doesn't seem to impact price, the categories don't exhibit as much differences
- Aspiration aspect of cars suggests `turbo` has higher price range than `std` (through there seem to be some outliers here)
- Its inconclusive to derive anything based on `enginelocation`, due to insufficient data
- While `eight` cylinders are the most priced, most common types are `four`, `five`, `six` cylinders
- `drivewheel` category seems to have a significant difference on price. Most high ranged cars have `rwd` drivewheel
- the variables `carwidth`, `carlength` and `curbweight` seem to have positive correlation (0.84), while `carheight` shows no correlation trend with price
- `citympg` and `highwaympg` seem to be negatively correlated with price, and `enginesize`, `boreratio`, `horsepower` and `wheelbase` have a high positive correlation with price

Feature Engineering

- fuel usage i.e., `fueleconomy` has significant negative correlation with price
- fuelsystem, drivewheel and car range i.e., market categorization of the car as medium, budget or premium/highend cars vs prices leads us to understanding that premium priced cars prefer `rwd` drivewheel with `idi` and `mpfi` fuel system
- from the correlation matrix – heat plot and observing the values we notice a highly correlated prices with the `curbweight`, `enginesize`, `horsepower` and `highend` variables

Model Development

- we try to playaround with the variables through multiple iterations by dropping and updating the linear regression runs, until we identify variable through which we can predict the price of a car most accurately. Since it's a linear regression, we are looking to get to an equation of the form y = mx+ nz….+ c form.
    - By chopping changing `carwidth`, `horsepower`, `carsrange` and `enginetype` variables are the residual variables and the equation for line of best fit is
    - *price = 0.3957 carwidth + 0.4402 horsepower + 0.2794 carsrange -0.0414 enginetype -0.0824*

Conclusion

The scope of this phase of the assignment is limited to exploratory analysis and feature engineering alone. Since its necessary to get a raincheck/ a gut feel with the variables in scope, and have it reviewed with stakeholders, I have intentionally de-scoped model development from this phase of analysis.

In the next iteration, I will be randomly splitting the dataset for train/test and subsequent model development using linear regression.

From the analyses conducted visually based on exploratory analysis and correlation plots , its observed that the features that impact price the most are `curbweight`, `enginesize`, `highend` , and `horsepower` aspects of the car.

From the analyses conducted through statistical modeling of the variables, `carwidth`, `horsepower`, `luxury` and `hatchback` are the most significant variables.

While these are the result of an academic study based on market survey, the implications of this assignment cascade to a business financing decision and potential launch of auto in a newer geographic. These very real business scenarios results in monetary risk, which each of the decision-making stakeholder 'signs up for with own blood'.

Assumptions

This dataset is retrieved from third party sources, authenticity of the data points is based on the credibility of the consulting/ analytics firm conducting the exercise. A first-hand right source

would be dealership sales data over couple of years across the country. We assume no seasonality in car purchase decisions, and the underlying tenet we have assumed is it is features captured as data columns that impact price. Non explicit attributes like brand affinity, brand satisfaction/ repulse etc aspects are ignored from scope (due to data collection limitations).
Price in the dataset is assumed to be the customer monetary cost. It is relevant assumption since in automotive sales there is MSRP (i.e., manufacturer suggested retail price) and dealer out-of-door-showroom price (which is what the end customer pays for the auto). While the former is an outcome of manufacturing costs, processes and overall business model, the latter is impacted with supply chain dynamics. (As an example, today, hypothetically or to some extent realistically too, one may very well find a brand new 2019 model which is exorbitantly priced at dealership, for the same MSRP that it has been stagnant at for over 2+ years)

Limitations

Scope of this assignment is for academic purpose only. In a real world there would be iterations with approach, code, and one consolidated presentation once all the workings are reviewed.

The dataset seems to be as of a given point in time i.e., it doesn't have the time trend aspect. It would have been ideal if we could also know the exact purchase date of any car along with its price point i.e., it was pos data than an aggregated market survey. We have ignored the seasonality aspect with car purchases for this task.

Challenges

Missing data, incorrect records, null values are some of the typical challenges we run into. However fortunately we didn't have to deal with these. Data quality related aspects were not a concern with the dataset. However, the challenge is deploying resources/ investments/ personnel based on external consultant survey wherein ownership of the decision is a business stakeholder call.
From an assignment aspect, the analytic firm focus is of academic nature, while from an operational standpoint it translates to business strategy, leadership stake/ business call, actual financial decisions and its associated risk.

Future Uses/Additional Applications

In the scope of this task, the goal is to study pricing dynamics with the intention of launching a product from a cross border player. However, this study can also be used for product development i.e., pivoting/re-design of auto features based on customer preferences and market dynamics.

Recommendations

The core of the task was to identify the variables /dimesnions that impact price and to identify the explainability of these variables. The key variables of impact are identified as the features `curbweight`, `enginesize`, `highend` , and `horsepower`. These variables seem to explain the price to 0.7 correlation score.

From an academic aspect, I believe a staggered rollout of business with enough feedback from the market would be a good way to pilot these feature set for an automotive. In a personal capacity, I would recommend staggered, controlled rollout of variants/ models with direct-dealership based partnerships until a full blown marketing campaign (which should be probably be no sooner than 2-3 years from the first car on road criteria).

Implementation Plan

Implementation plan at this stage is incomplete, since it needs a modeling exercise to confirm if the identified feature set explain the price differentials completely. The modeling exercise is typically done by splitting the dataset into test-train and assessing the dependent variable (i.e., price point) efficacy of the model development/ sampling technique.

Ethical Assessment

Dataset received, though not public. doesn't have any PII details. It would have been ideal if the dataset were directly sourced from auto dealership point-of-sale records. However, these are sourced from a market survey, so repurposing of this data for product development/ reverse engineering of design would raise questions with ethical contexts.
Dataset ignores timing/ seasonality aspect of the auto purchase decisions. This is especially relevant since today's dealership inventory/ pricing is entirely demand driven and global production/ supply chain constrained which has never happened in last few decades of manufacturing history, so this market force of supply chain constraints is unaccounted, dominant , impactful and very much real, in spite of knowing this there isn't much that can be done given lack of data collection/ capture methods at the grain of analysis.