Student Name: John Anokye

Business Intelligence and Machine Learning Project for Predicting Vehicle Selling Prices

In a rapid evolving automative industry, car manufacturing company, All American Motors Corp. (AAMC) and its partnered dealership all over the country need to find ways to maintain their competitive in the automotive market. With the exponential growth in data within the last decade, AAMC can combine AI and business intelligence on the rigorous existing sales data that has been collected by the company over years to predict vehicle sales and improve revenue.

Predicting vehicle sales encompasses analyzing historical data, market trends, economic indications, consumer behavior, consumer feedback, surveys and rating to analyze current sales trends to understand consumer behaviors and preferences when it comes to purchasing a vehicle. The data can also be used to forecast future demand as well as to improve services and manufacture products(cars) to suit consumer preferences. This will intern help to continuously grow the business.

In this project, I will be leveraging advanced analytics and artificial intelligence techniques in a probably research case use-case, to discover trends and patterns, actionable insights on sales trends of AAMC car sales over the last decade and also predict vehicle selling prices based on conditions of vehicles to help in sales forecasts into the future.

Data is chaotic and entropic! A sound and actionable decision intelligence will lie heavily on an accurate and organized data source. As a first step in this project, data will be collected and cleaned. Data will be gathered from systems that contain historical sales records, economic indicators, market trends and consumer behavior data. To ensure quality and accuracy, the data will undergo series of preprocessing steps. For example, this involves handling missing values in the data set, outlier detections to prevent skewing of the analysis and data normalization (leveling the playfield). Relevant features that will significantly impact vehicle sales predictions will be identified from the dataset. Models will then be developed to for predictions, evaluated and optimized for improved prediction accuracy. Some of the machine learning models that will be considered include Linear regressions, decision trees and neural networks.

To provide actionable insights to business partners and stakeholders, interactive dashboard will be developed using Power BI. This will display key performance indicators (KPI), key business metrics, trends, and predictions. For example, visualizations such as real-time sales forecast by state, time period, visualization of consumer sentiments can be all be developed which will allow users and business stakeholders to make informed business decisions.

In conclusion, this project will aim in revolutionize how vehicle sales are predicted by All American Motors Corporation by using advanced analytics and machine learning. This will be achieved by integrating comprehensive data, developing accurate models and creating interactive BI dashboards to provide valuable insights to assist business stakeholders to make data-driven decisions. This approach is envisioned to improve sales strategy, help in managing inventory, enhance customer satisfaction and most important help All American Motors Corporation to grow its market share in the competitive automotive market.

BI/AI/MI Story Board

	Set Up	Actions	Outcomes	Results
Current Implementation	-Relies on traditional methods of inventory management. Done manually at the dealers through physical inventory -Selling prices are predicted with limited access to diverse and comprehensive dataset	-Accountants and business leaders make vehicle prices and production prediction, manufacturing, and sales predictions directly from previous year. -Dealers call in to place new orders of vehicles when their inventory goes down	- Company consistently overestimate or underestimate sales targets for present fiscal year -Underproduction, dealerships under-stocked with inventory of best-selling cars, dealership overstocked with vehicles in less selling regions/states	- Customers wait on long lead times before vehicles arrive at the dealership lot - Company misses sales and revenue targets.
Future Implementation	-Collect comprehensive historical sales, vehicle conditions, demographics and consumer data -Supply chain team, dealership should use real-time data for inventory management	-Predict selling prices of vehicles accurately base on vehicle features/conditions, inventory, demand with comprehensive collected sales, demographics and consumer data with ML/AI models -Develop BI dashboards to for inventory management, vehicle price predictions	- Role out developed models to accountants and business leaders so to make accurate calls on revenue, production and sales targets for fiscal year. - Dealerships makes intime order for new vehicles	- Services and purchases delivered to customers at a faster rate - Sales and manufacturing grow over time

Data Source: The dataset for this project will be sourced from Kaggle data repository (Link). The "Vehicle Sales and Market Trends Dataset" provides a comprehensive collection of information pertaining to the sales transactions of various vehicles. This dataset encompasses details such as the year, make, model, trim, body type, transmission type, VIN (Vehicle Identification Number), state of registration, condition rating, odometer reading, exterior and interior colors, seller information, Manheim Market Report (MMR) values, selling prices, and sale dates. The dataset has about 600,000 records.

4V Model Analysis

Dimension	Score
Volume	Significant amount - 4
Variety	Desired Variety – 4 Representative – 3
Velocity	Medium Velocity – 3
Veracity	High data quality – 5 Labels correctly labeled – 5 Good fit for use - 5

PHASE 2 – DATA AND MODEL PREPARATION

The goal of this this study is to predict the selling price prices of vehicles sold by AAMC depending on the features/condition of the vehicle so as to help the company forecast sales and revenue properly. An AI model will need to be trained on historical dataset collected on sold vehicles. The model will then be used to predict the selling prices of vehicles in a new dataset which primarily consists of the new inventory of cars to be sold by the company.

1. Historical Data Collection

Historical dataset was collected so as to learn the patterns, trends and what essentially drive sales for AAMC. These trends will help in predicting future sales when it is trained in a ML algorithm to help make future predictions. The dataset consists of 16 variables and about 100,000 records, with each record representing an observation of vehicle sold. The dataset has to be in a clean and tidy format to train an AI model. Few preprocessing steps were performed to improve the dataset quality. Missing values in the dataset constituted about 1% of the dataset and hence all rows with missing values were dropped. The 'salesdate' column was a string and was converted to the proper datatype – datetime. The 'condition' attribute consisted of a wide range of values between 1 to 50. This attribute was binned and grouped into groups of 1 to 5. All numeric categorical features were also converted to strings.

2. Features and Lables Identification

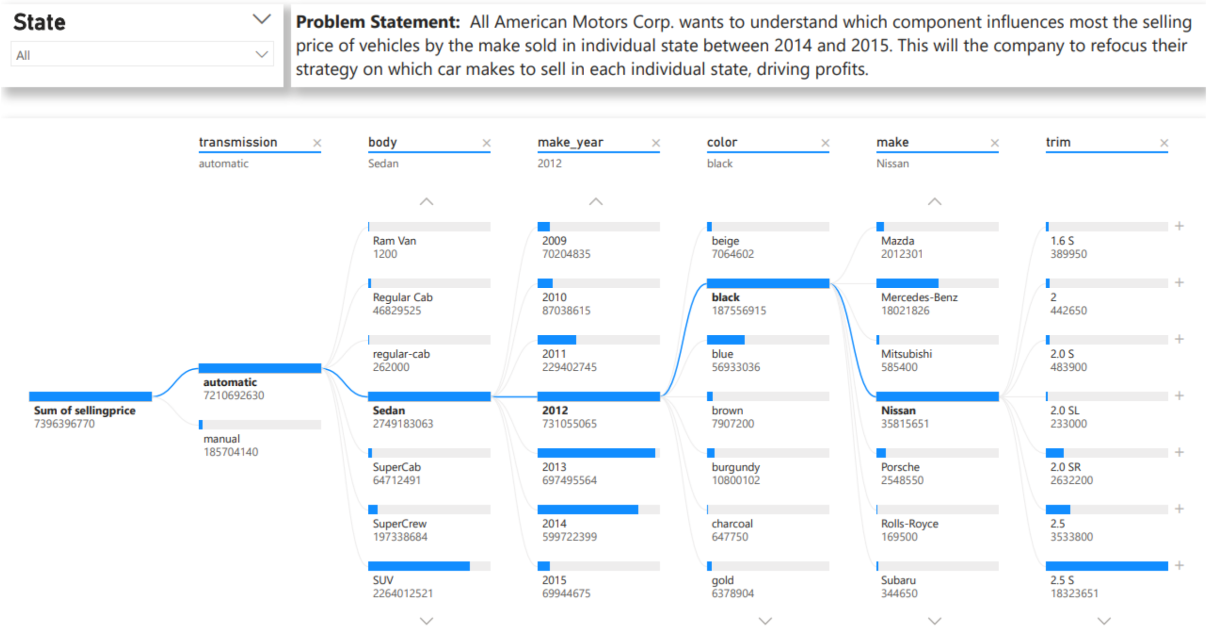
Features and labels in dataset were identified to help train the AI model. The label to be predicted and used in the AI model training is the 'sellingprice' attribute. This is a continuous/numerical data values which will also determine the model selection. There are 15 other attributes in the dataset. Not all 15 attributes in the dataset will have direct influence in predicting the target variable. Techniques such as correlation analysis, principal component analysis (PCA) and lasso regression were performed to narrow down the important predicting features. The features that were selected after this analysis model were 'year', 'make', 'model', 'transmission' and 'condition'.

3. Training Data Split

In machine learning, data is split into training and testing data. The model will be trained with the training dataset and then evaluated with the testing dataset which is unseen by the model to determine and evaluate the performance of the model. AutoML in Azure was used in this study. However, a train-test split option was selected with a 90% training dataset and 10% testing dataset.

4. Algorithm Selection

Before building the model a diagnostic technique, key influencer visualization, was conducted to help determine which variable/feature influences most the selling price of vehicles in individual states with the range of date of the dataset. Refer to the figure 1 below for the diagnostic analysis. The feature that will be predicted in this training will be the 'sellingprice' attribute. This is a numerical field and hence a Regression Model was selected in train the model.

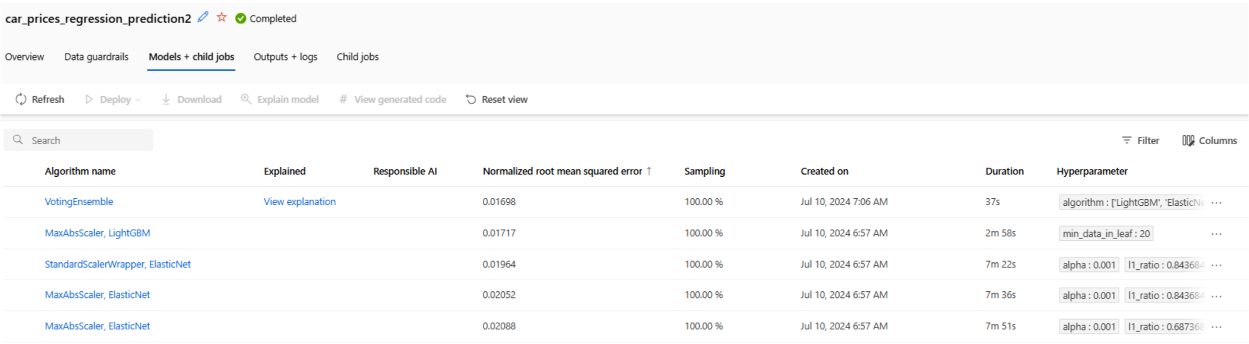


Problem Statement: All American Motors Corp. wants to understand which component influences most the selling price of vehicles by the make sold in individual state between 2014 and 2015. This will the company to refocus their strategy on which car makes to sell in each individual state, driving profits.

Figure 1: Diagnostic analysis

5. Model Evaluation

Regression model training task was selected to train the dataset and some of the algorithm that were considered in the model training were linear regression, voting ensemble, LightGBM, ElasticNet and decision tree. The training performance is show in the figure 2 below. The best model selected was VotingEnsemble which had the lowest normalized root mean squared error of 0.01698.



car_prices_regression_prediction2 ✓ Completed

Algorithm name	Explained	Responsible AI	Normalized root mean squared error ↑	Sampling	Created on	Duration	Hyperparameter
VotingEnsemble	View explanation		0.01698	100.00 %	Jul 10, 2024 7:06 AM	37s	algorithm : ['LightGBM', 'ElasticNe
MaxAbsScaler, LightGBM			0.01717	100.00 %	Jul 10, 2024 6:57 AM	2m 58s	min_data_in_leaf : 20
StandardScalerWrapper, ElasticNet			0.01964	100.00 %	Jul 10, 2024 6:57 AM	7m 22s	alpha : 0.001 l1_ratio : 0.843684
MaxAbsScaler, ElasticNet			0.02052	100.00 %	Jul 10, 2024 6:57 AM	7m 36s	alpha : 0.001 l1_ratio : 0.843684
MaxAbsScaler, ElasticNet			0.02088	100.00 %	Jul 10, 2024 6:57 AM	7m 51s	alpha : 0.001 l1_ratio : 0.687368

Figure 2: Trained model performance

6. Deployment

Model endpoint from the trained model will be deployed to an end point and employed in a power bi dashboard that can be used to predict the selling price of new vehicles in the inventory before they are sold. This will help in accurate estimation of revenue for forecasting purposes for AAMC. The regression model that was deployed in can be seen below in figure ???.

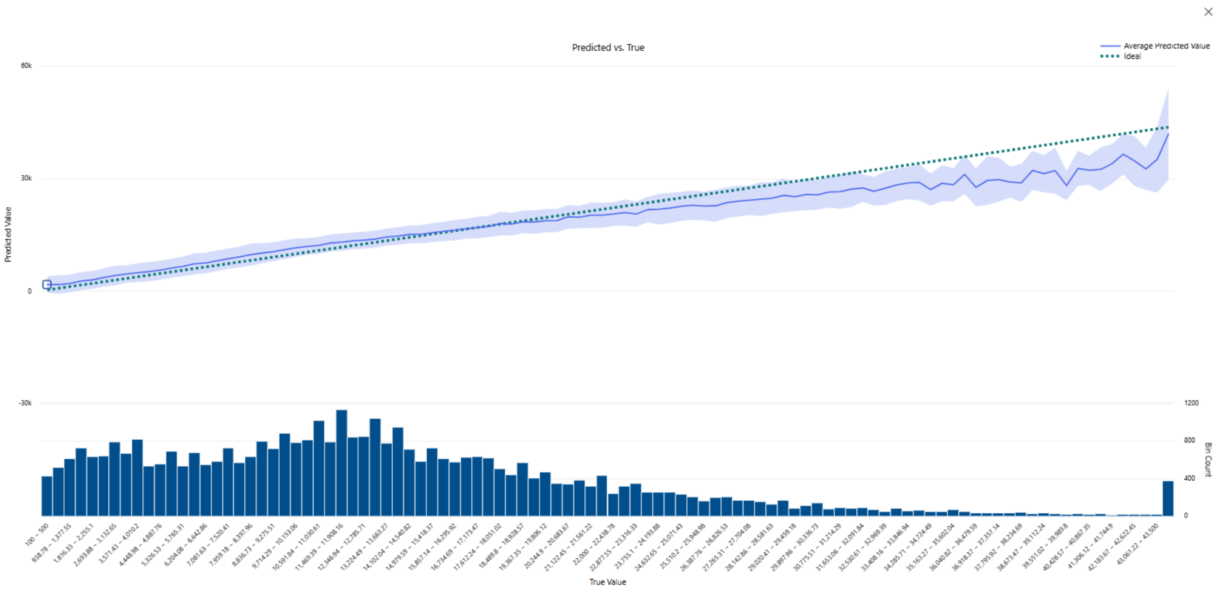


Figure 3: Regression model obtained from the trained dataset in Azure Auto ML

4V Framework Assessment of Current and Future Data Needs

Volume:

Currently, the dataset comprises about 100,000 records, providing a significant volume of data that supports robust model training and ensures comprehensive coverage of vehicle sales transactions. For future needs, the volume is expected to grow as more transactions are recorded, which will further enhance the model's accuracy and reliability by incorporating a wider range of data points.

Variety:

The dataset includes a diverse range of 16 variables such as year, make, model, transmission type, condition rating, and selling prices. This variety captures multiple aspects influencing vehicle sales. Future data collection should aim to maintain or increase this variety by possibly including new variables such as customer demographics or market trends, enriching the dataset and potentially improving model performance.

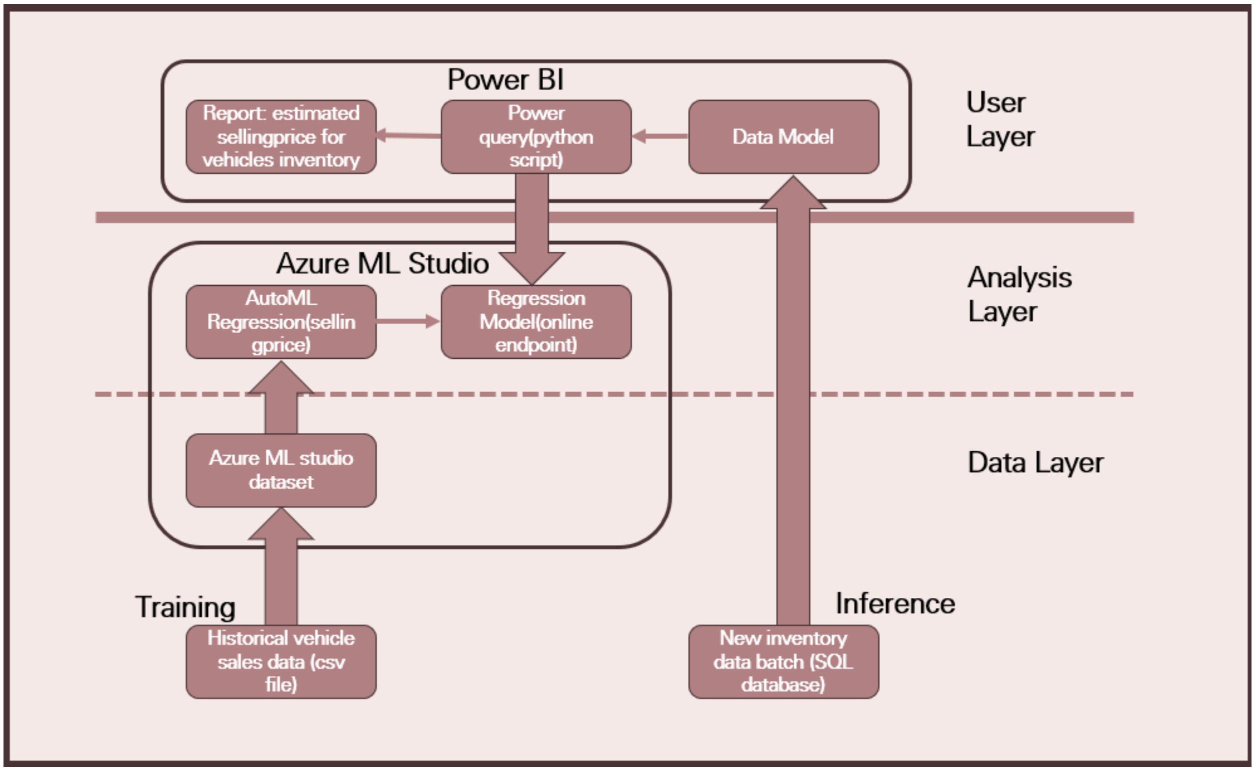
Velocity:

The dataset is updated at a medium velocity, with sales transactions being recorded regularly. This velocity is sufficient for the current model's needs. However, as the business scales, increasing the frequency of data updates could provide more timely insights and allow the model to adapt more quickly to market changes.

Veracity:

The dataset boasts high data quality, with minimal missing values and correctly labeled data, ensuring the reliability of the model's predictions. Maintaining this high level of data integrity is crucial for future needs. Implementing more rigorous data validation processes and leveraging automated data cleaning tools will help preserve and enhance veracity as the dataset expands.

Final 3 Layered Architectural Diagram


Power BI
Report: estimated sellingprice for vehicles inventory
Power query(python script)
Data Model
User Layer
Azure ML Studio
AutoML Regression(sellingprice)
Regression Model(online endpoint)
Analysis Layer
Azure ML studio dataset
Data Layer
Training
Historical vehicle sales data (csv file)
Inference
New inventory data batch (SQL database)