Business Intelligence and Machine Learning Project for Predicting Vehicle Selling Prices

Student Name: John Anokye

PHASE 1

Introduction

In the rapidly evolving automotive industry, All American Motors Corp. (AAMC) and its partnered dealerships across the country need to maintain their competitive edge in the market. With the exponential growth in data over the last decade, AAMC can leverage AI and business intelligence to analyze its extensive sales data to predict vehicle sales and improve revenue. This project aims to combine advanced analytics and artificial intelligence techniques to discover trends, provide actionable insights, and predict vehicle selling prices based on vehicle conditions, ultimately aiding in sales forecasts and revenue enhancement.

Project Overview

Predicting vehicle sales involves analyzing historical data, market trends, economic indicators, consumer behavior, feedback, surveys, and ratings to understand current sales trends and consumer preferences. This data can also be used to forecast future demand and improve services and product offerings to align with consumer preferences, thereby supporting business growth. In this project, advanced analytics and AI techniques will be used in a practical research case to uncover trends, actionable insights, and predict vehicle selling prices to aid in future sales forecasts.

Data Collection and Preparation

Data is chaotic and entropic; therefore, accurate and organized data sources are crucial for sound decision-making. The first step in this project involves collecting and cleaning data from systems containing historical sales records, economic indicators, market trends, and consumer behavior data. The data will undergo preprocessing steps to handle missing values, detect outliers, and normalize the data to ensure quality and accuracy. Relevant features that significantly impact vehicle sales predictions will be identified for model development, evaluation, and optimization.

Machine Learning Models

Various machine learning models, including linear regression, and decision trees, will be considered for this project. The models will be trained and optimized for improved prediction accuracy. The chosen models will be evaluated based on their performance metrics, and the best-performing model will be selected for deployment.

Data Visualization and Insights

To provide actionable insights to business partners and stakeholders, an interactive dashboard will be developed using Power BI. This dashboard will display key performance indicators (KPIs), business metrics, trends, and predictions. Visualizations such as real-time sales forecasts by state, time period, and consumer sentiments will enable users and stakeholders to make informed business decisions.

Impact on Business

This project aims to revolutionize how vehicle sales are predicted by AAMC by integrating comprehensive data, developing accurate models, and creating interactive BI dashboards. This approach will improve sales strategies, manage inventory more effectively, enhance customer satisfaction, and help AAMC grow its market share in the competitive automotive market.

Phase 2 – Data and Model Preparation

Historical Data Collection:

The historical dataset consists of 16 variables and about 100,000 records, representing vehicle sales observations. The data was cleaned and preprocessed to ensure it is in a usable format for AI model training.

Features and Labels Identification:

The label to be predicted is the 'sellingprice' attribute, a continuous numerical value. Important features were identified through correlation analysis, PCA, and lasso regression, narrowing down to 'year,' 'make,' 'model,' 'transmission,' and 'condition.'

Training Data Split:

The data was split into training (90%) and testing (10%) datasets. AutoML in Azure was used for this process.

Algorithm Selection:

A diagnostic technique, key influencer visualization, was conducted to determine the most influential variables. A regression model was selected for training, considering algorithms like linear regression, voting ensemble, LightGBM, ElasticNet, and decision tree.

Model Evaluation:

The voting ensemble model was identified as the best performer with the lowest normalized root mean squared error of 0.01698.

Deployment:

The trained model's endpoint will be deployed and integrated into a Power BI dashboard for predicting selling prices of new vehicles in inventory, aiding in accurate revenue estimation for AAMC.

4V Framework Assessment of Current and Future Data Needs

Volume:

The current dataset's volume about 100,000 records, supports robust model training. Future data growth will enhance model accuracy.

Variety:

Current data variety captures multiple aspects of vehicle sales. Future data collection may include additional variables for improved model performance.

Velocity:

Regular updates provide timely insights. Increasing update frequency will allow quicker adaptation to market changes.

Veracity:

Maintaining high data quality is crucial. Implementing rigorous data validation and automated cleaning tools will preserve data integrity as the dataset expands.