S&DS 425/625 Capstone Executive Summary: Predicting Car Sales with Machine Learning Models

Jiayi Chen, Qianmeng Chen, Kaifeng Gao

12/15/2024

Summary

The automotive industry continuously seeks to understand the driving factors behind car sales to inform design and marketing strategies. This project investigates how various attributes, including a car's physical characteristics, fuel efficiency, and brand reputation, influence consumer purchasing behavior. Additionally, we explore whether perceived "facial expressions" in car designs affect sales performance, hypothesizing that visual appeal may subconsciously influence buyers.

Using data sourced from the Deep Visual Marketing platform, advanced machine learning models were developed to identify key predictors of sales. Our analysis included random forest, linear regression, backpropagation neural networks, and recurrent neural networks, each chosen for its strengths in uncovering specific patterns within the dataset.

Data Overview

The dataset contained detailed records on 5,269 vehicles, including their dimensions (length, width, height), engine characteristics, fuel consumption, brand ratings, and sentiment probabilities derived from design features resembling emotions like happiness, sadness, and surprise. Missing data was addressed through imputation, and categorical variables were one-hot encoded to ensure compatibility with machine learning models.

The dataset underwent extensive preprocessing to ensure data quality and suitability for analysis. This section details the steps taken to process the various data sources, including sales data, vehicle characteristics, advertisements, and image-based expression analysis. These steps were crucial to address missing values, handle outliers, and aggregate the data into a unified format.

1. Sales Data Preprocessing

The sales data was processed to derive the adjusted average sales for each vehicle generation model (avg_sales) using the following steps:

- Exclusion of 2020 Sales Data: Sales data for 2020 was excluded to account for market anomalies caused by the COVID-19 pandemic.
- Filtering Non-Zero Sales Values: Years with zero sales were removed to focus on meaningful sales activity.
- Exclusion of the Last Non-Zero Year: For generation models with multiple non-zero sales years, the most recent year was excluded to mitigate the influence of short-term trends.
- Outlier Removal: Unusually low sales values were identified and excluded using the interquartile range (IQR) method.
- Calculation of Adjusted Sales: The adjusted average sales were calculated as the mean of the remaining sales values. If no valid values remained after filtering, the result was set to NA.

2. Vehicle Trim Data Aggregation

The vehicle trim data was aggregated by generation model ID to compute average values and majority categories for key attributes:

- Numerical Averages:
  - Average price (avg_price)
  - Average gas emissions (avg_gas_emission)
  - Average engine size (avg_engine_size)
- Categorical Modes:
  - The most common fuel type (majority_fuel_type) was identified for each generation model.

3. Advertisement Data Preprocessing

The advertisement data was processed to standardize and aggregate key attributes:

- Unit Conversion:
  - Columns such as Average_mpg and Top_speed were cleaned by removing units (e.g., "mpg," "mph") and converting values to numeric format.
- Attribute Aggregation by Generation Model:
  - Numerical attributes:
    - Average engine power (avg_engine_power)
    - Average dimensions, including wheelbase, height, width, and length (avg_wheelbase, avg_height, avg_width, avg_length)
    - Average fuel efficiency (avg_mpg)
    - Average top speed (avg_top_speed)
- Categorical attributes:
  - Majority values for the number of seats (majority_seat_num) and doors (majority_door_num).
  - Most common engine size (majority_engine_size), body type (majority_bodytype), and gearbox type (majority_gearbox).

4. Image-Based Expression Data Preprocessing

The image data and expression analysis results were combined and aggregated to derive probabilistic and categorical features for each generation model:

- Average Emotion Probabilities:
  - The probabilities of different emotional expressions (e.g., angry, happy, neutral) were averaged across all images for each generation model: avg_prob_angry, avg_prob_disgust, avg_prob_fear, avg_prob_happy, avg_prob_sad, avg_prob_surprise, and avg_prob_neutral.
- Majority Label:
  - The most common emotional expression label (majority_label) was determined based on frequency.

5. Final Data Merging

The processed data from sales, trims, advertisements, and image-based expression analysis were merged using the generation model ID as the common key. This unified dataset served as the foundation for downstream analysis and modeling.

Methodology

1. **Random Forest**: Used to evaluate feature importance and gain insights into the most impactful variables influencing sales.
2. **Linear Regression**: Provided a foundational statistical model to quantify relationships between predictors and sales.
3. **Neural Network Models**:
   - **Backpropagation Neural Network (BPNN)**: Modeled complex interactions and non-linear dependencies between variables.
   - **Recurrent Neural Network (RNN)**: Investigated sequential patterns in the dataset, although temporal dependencies were limited.

4. **Hyperparameter Tuning**: Applied to optimize model performance, improving accuracy and minimizing errors.

Results

- **Significant Predictors**: The analysis revealed that sales are strongly influenced by practical factors such as fuel efficiency, car size, and engine specifications. These features consistently ranked as the most impactful across all models.
- **Facial Expressions**: Contrary to the hypothesis, sentiment-derived variables related to perceived facial expressions showed no measurable effect on car sales, indicating limited consumer sensitivity to these visual elements.
- **Model Performance**: Among the models tested, the backpropagation neural network exhibited the best predictive performance, achieving the lowest Mean Absolute Error (MAE). The random forest model also provided valuable interpretability, particularly in feature importance ranking.

Conclusions

The results emphasize the dominance of functional attributes like fuel economy and car dimensions in influencing purchasing decisions, while aesthetic features, such as perceived facial expressions, appear to have minimal impact. These findings challenge assumptions about the role of design in consumer choices and underscore the need for manufacturers to prioritize practical benefits in their marketing and product development efforts.

Manufacturers and marketers can draw the following actionable recommendations from this study:

- **Product Focus**: Invest in improving fuel efficiency and enhancing vehicle functionality to align with consumer priorities.
- **Marketing Strategies**: Highlight practical advantages in campaigns to resonate with buyer preferences.
- **Design Exploration**: While facial expressions lack a direct impact on sales, subtle design adjustments may still appeal to specific market segments, warranting further investigation.

This project underscores the power of combining diverse modeling techniques to analyze complex relationships in sales data, offering valuable insights for strategic decision-making in the automotive industry.