S&DS 425/625 Capstone Report

Jiayi Chen, Qianmeng Chen, Kaifeng Gao

12/15/2024

Abstract

Understanding the factors that influence car sales is essential for the automotive industry. In this project, we analyze various variables, including the reputation of the car brand (rate) and the "facial expressions" of cars, to evaluate their impact on sales performance. Leveraging data from Deep Visual Marketing, we applied random forest, linear regression, backpropagation neural network, and recurrent neural network to identify the key determinants of car sales. Contrary to our hypothesis, we found that while variables like fuel consumption, car model, and size significantly influence sales, the perceived facial expressions of cars do not. These findings provide valuable insights for manufacturers aiming to optimize sales strategies.

Introduction

Understanding the determinants of car sales is a critical area of research within the automotive industry. As consumer preferences evolve, both functional attributes such as fuel efficiency and aesthetic elements such as design play essential roles in purchasing decisions. Recent advancements in machine learning have enabled researchers to analyze large datasets, uncovering nuanced insights into what drives sales performance. Motivated by these developments, this study aims to investigate the impact of various factors, including car brand reputation, physical dimensions, and perceived "facial expressions," on car sales. While prior research highlights the importance of functional characteristics like fuel consumption and engine performance, the role of visual design elements remains underexplored. This project seeks to address this gap by combining traditional statistical methods with state-of-the-art machine learning models.

The dataset used in this study was obtained from the Deep Visual Marketing platform and consists of 5,269 car records, each containing detailed attributes. These include physical specifications (e.g., engine size, dimensions), performance metrics (e.g., fuel consumption, average sales), and sentiment-based probabilities derived from visual features resembling emotions (e.g., happiness, sadness). Preprocessing steps such as imputation for missing values, feature scaling, and one-hot encoding were applied to prepare the data for modeling.

The remainder of this paper is organized as follows. Section 2 describes the methodologies employed, including random forest, linear regression, backpropagation neural networks, and recurrent neural networks, to analyze the relationships between the variables and car sales. In Section 3, we present the results of the models, highlighting the dominance of functional attributes and the lack of influence from perceived facial expressions on sales performance. Section 4 discusses the implications of these findings, emphasizing recommendations for manufacturers and marketers to optimize their strategies. Finally, Section 5 concludes with a summary of the main insights and suggestions for future research directions. **I am not sure if we need have section XXX, we can revise this later**

By examining both practical and aesthetic factors, this study provides a comprehensive analysis of car sales determinants, offering valuable insights for the automotive industry to refine its design, production, and marketing strategies.

Data exploration and visualization

Modeling/Analysis

This project employed two datasets to investigate the factors influencing car sales and evaluate the predictive performance of different modeling approaches. The first dataset contained aggregated data, where sales for each car model were averaged across all years. The second dataset included yearly sales records for each car model, providing a more granular view of temporal sales patterns. Both datasets were analyzed using linear regression and backpropagation neural networks (BPNN), enabling a thorough comparison of model performance and insights across different data structures.

Assumptions

1. **Linear Regression**:
   - Assumes a linear relationship between predictors and the outcome (sales).
   - Assumes predictors are independent and residuals are normally distributed with constant variance.
   - May struggle to capture non-linear relationships and interactions in the data.
2. **Backpropagation Neural Network (BPNN)**:
   - Assumes that relationships and patterns in the data can be captured through non-linear transformations using hidden layers and activation functions.
   - Assumes features are properly scaled for efficient gradient descent optimization.

Observations, Predictors, and Outcome

- **Aggregated Dataset**:
  - Observations: Rows represent unique car models.
  - Predictors: Physical attributes (e.g., engine size, fuel efficiency, dimensions), categorical variables (e.g., fuel type, body type), and sentiment-based probabilities (e.g., happiness, sadness).
  - Outcome (y): Average sales of each car model across all years.
- **Yearly Dataset**:
  - Observations: Rows represent yearly sales data for each car model.
  - Predictors: Same as the aggregated dataset, with the addition of the year variable to capture temporal trends.
  - Outcome (y): Annual sales for each car model.

Models and Their Representations

1. **Linear Regression**:
   - Aggregated Dataset: $y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$
   - Yearly Dataset: Includes an additional predictor for the year to account for temporal trends.
   - Coefficients represent the expected change in sales for a one-unit increase in the corresponding predictor, holding all other variables constant.
2. **Backpropagation Neural Network (BPNN)**:
   - A multi-layer perceptron with ReLU activation functions and dropout layers was implemented.
   - Model architecture and hyperparameters (e.g., learning rate, number of layers, and hidden units) were optimized separately for each dataset.
   - The non-linear nature of BPNN allows for capturing complex relationships and interactions among predictors.

Performance Metrics

- **Aggregated Dataset**:
  - **Linear Regression**: R-squared = 0.247, MAE approximately 3,300. Limited explanatory power due to linearity assumptions.
  - **BPNN**: MAE approximately 2,561, outperforming linear regression by capturing non-linear patterns in the data.

- **Yearly Dataset**:
  - **Linear Regression**: Lower R-squared and higher MAE compared to the aggregated dataset, reflecting challenges in modeling yearly sales variability.
  - **BPNN**: MAE approximately 3,312, with performance declining slightly due to the increased complexity and noise in yearly sales data.

Comparison of Models and Datasets

- **Linear Regression**:
  - Aggregated Dataset: Easier to interpret coefficients but struggles to model non-linear relationships effectively.
  - Yearly Dataset: Inclusion of the year variable provides insights into temporal trends but offers limited improvement in accuracy.
- **BPNN**:
  - Aggregated Dataset: Outperformed all other models in predictive accuracy, indicating suitability for summarizing long-term trends.
  - Yearly Dataset: While capable of handling complex interactions, the model's performance was hindered by the variability and noise in annual sales data.

Visualization and interpretation of the results

1. **Feature Importance (Linear Regression)**:
   - Bar plots of regression coefficients highlight the relative impact of predictors such as fuel efficiency, car size, and engine power. Differences in coefficient significance between the aggregated and yearly datasets are clearly displayed.
2. **Predicted vs. Actual Sales**:
   - Scatter plots compare predicted and actual sales for both datasets. For the aggregated dataset, BPNN predictions closely align with actual values, indicating high accuracy. In contrast, yearly dataset predictions show more dispersion, reflecting the added complexity of annual trends.
3. **Residual Analysis**:
   - Residual plots for linear regression reveal limitations in modeling non-linear patterns. For the yearly dataset, residuals exhibit heteroscedasticity, indicating that the model struggles to account for temporal variability.
4. **Learning Curves (BPNN)**:
   - Visualizations of training and validation loss across epochs for each dataset illustrate the convergence behavior of the neural network. The aggregated dataset shows smooth convergence, while the yearly dataset exhibits more fluctuations, indicative of higher complexity.

Conclusions and recommendations

This study provides a comprehensive analysis of the factors influencing car sales by leveraging machine learning and statistical methods. The results indicate that practical attributes such as fuel efficiency, car size, and engine specifications significantly impact sales performance, while perceived facial expressions in car designs have no measurable influence. These findings challenge the assumption that aesthetic features play a pivotal role in consumer purchasing decisions, emphasizing the dominance of functional attributes.

Based on these insights, manufacturers are encouraged to focus on improving fuel efficiency, optimizing vehicle dimensions, and enhancing engine performance to align with consumer priorities. Marketing strategies should highlight these practical benefits to resonate with the target audience. Although facial expressions did not emerge as a critical factor, future research could explore other design elements that may influence niche markets or specific consumer segments.

Future work could involve expanding the dataset to include additional regions or markets, allowing for a more comprehensive analysis of global trends. Incorporating temporal data, such as seasonal sales patterns or economic conditions, could also provide deeper insights into sales dynamics. Furthermore, integrating social media sentiment analysis might uncover hidden factors that contribute to consumer behavior. By addressing these areas, future studies can build on this research to provide even more actionable recommendations for the automotive industry.

References