S&DS 425/625 Capstone Report

Jiayi Chen, Qianmeng Chen, Kaifeng Gao

12/15/2024

Abstract

An overview of your report, including one or so sentences on each of these:

- a non-technical description of the problem you are trying to solve or the question you are trying to answer, and why you are trying to answer that question
- a non-technical description of the data, where it came from, and what it contains, including possibly the predictors, the outcome, and the observations
- a non-technical description of what kind of analysis you did, including high-level description of what the predictors were, what the outcome was, and how to interpret the results of the model
- a brief summary of the models that are used
- a non-technical description of the results of the model and main takeaways.

An abstract is one paragraph with text only and is aimed at a technical audience. This appears at the beginning of the report.

-Our content start from here

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.

Executive Summary

An executive summary is typically longer than the abstract, up to a page, could possibly contain key visualizations, tables, or other figures that help communicate either the raw data or the results of the model, and is intended for someone outside of the data science/analytics team of an organization. It is important to be as concise as possible, and describe each of those points above without using language that is overly technical and not part of commonly used English. The executive summary is a separate document.

Note that in the abstract, executive summary, and throughout the report you should avoid using first-person singular pronouns like "I" and "me", even if you are the only author. Use "we" or use passive voice.

-Our content start from here

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.

Key Insights

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.

Implications

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.

Introduction

A few paragraphs that contain the following:

- background on the topic you are studying, including the motivation behind the project and the problem statement you mentioned in the abstract, but in more detail. This could include a brief literature review of related work
- a description of the data
- a few sentences or a paragraph describing what is contained in the rest of the paper, including make steps and main takeaways.

One possible way to do this is to include roughly one sentence per section, that describes what is in the section and the main takeaways from each section. For example,

> "Section 2 contains data exploration and visualization, which reveals that *****. In Section 3, we build several different predictive models and find that *****. We discuss the results of the model, including *****, in Section 4. Finally, we discuss conclusions, recommendations and ideas for future work in Section 5."

You will see this in many research articles. It can seem formulaic, so if you have an alternative way to summarize the remainder of the article, go for it.

There are similarities between the introduction and the abstract. The introduction is longer and more detailed, especially in terms of background, previous work, and motivation of the problem, and contains a brief outline of the contents of the rest of the paper.

You might find these resources useful as well, courtesy of the Poorvu Center for Teaching and Learning

- Research paper writing in the Natural Sciences
- Research paper writing in the Humanities

-Our content start from here

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

This section will have descriptive statistics and visualizations of the raw data. Use this section to reveal to the reader any interesting relationships in the data, and convince the reader that the predictors are related to the outcome. Visualizations are one of the most powerful ways to communicate information to the reader, so it is important to spend time producing clear, descriptive, eye-catching visualizations.

The package `pubtheme` has a `ggplot` theme called `theme_pub` that helps with making publication-quality visualizations with `ggplot`. See https://github.com/bmacGTPM/pubtheme. There are also several templates there that you can copy, paste, and modify.

If you display a data visualization or some other summary of data, discuss the significance of what you see. What does this tell you about the data? What does it tell you that will help you with modeling? Do not simply show a visualization for the sake of showing a visualization.

Since `echo=F` is the option chosen at the top, the default will be to show the output but not the code:

```
[1] 2
```

If you don't want to show the output either, you can use `include=F`:

Nothing was shown above. If you want to force it to show the code for some reason, you can override the default by putting the options `echo=T` for this chunk.

```
1+1
```

```
[1] 2
```

However, since this is a formal report, you will likely not want to show code.

Kgod

Modeling/Analysis

Describe regression or classification model(s) used, or the analysis that was performed. For each regression or classification model, discuss

- any assumptions that are made
- the observation, the predictors, and the outcome (aka the rows of X, the columns of X, and y)
- what model you are using, and write out the model
- what the coefficients mean (when applicable) and how this is related to your problem
- appropriate measures of the performance of the model, such as measures of fit and predictive ability
- whether or not you think the model is appropriate for this kind of data, and why, and

- how easy/hard it is to interpret the results and explain them to either a technical or non-technical audience.

For other kinds of analysis, what you give is highly dependent on the type of analysis. But in general, talk about assumptions, if they are appropriate, how they might not be appropriate, and why you chose this type of analysis.

Visualization and interpretation of the results

Create visualizations of the results when appropriate, focusing on visualizations that

- help describe aspects of the results that have real-world interpretation
- help the reader understand how the model addresses the problem you are studying.

Visualizations are one of the most powerful ways to communicate information to the reader, so it is important to spend time producing clear, descriptive, eye-catching visualizations.

Discuss the results of the model or models you chose, and describe how they are related to the problem statement or question that you were trying to answer in the project.

If you have built multiple models or types of analysis, compare the measures of performance and the ease of interpretability across models or types of analysis, stating which model or models performed best, and which model or models were most interpretable. Finally, decide which model or type of analysis is best for your particular problem based on some combination of performance and interpretability.

-Our content start from here

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.

Suitability and Interpretability

- Linear regression offers interpretability, making it more accessible to non-technical stakeholders but is less suitable for this complex dataset.
- BPNN provides superior predictive performance but requires visualization techniques to interpret results for non-technical audiences.

Visualizations

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.

Interpretation

- **Functional Attributes**: Consistently identified as the most important predictors of sales across both datasets and models, attributes such as fuel efficiency, engine size, and vehicle dimensions strongly influence consumer decisions.
- **Facial Expressions**: Sentiment-based probabilities derived from perceived facial expressions showed negligible influence on sales, challenging the hypothesis that visual design elements significantly impact purchasing behavior.
- **Model Comparison**: BPNN outperformed linear regression in predictive accuracy across both datasets, with the aggregated dataset yielding the best results. However, linear regression remains valuable for its interpretability.

Insights and Recommendations

- For aggregated data, BPNN provides the most accurate predictions, making it suitable for long-term strategic planning.
- Temporal factors in yearly datasets add complexity, requiring more sophisticated approaches or additional contextual features (e.g., economic indicators or seasonal trends) to improve accuracy.
- Future research should explore expanding datasets to include external factors or consumer feedback and investigating other design elements that may influence niche markets.

The comparative analysis underscores the importance of selecting models and datasets appropriate to the specific goals of the analysis, balancing predictive performance with interpretability to address both technical and non-technical audiences.

Conclusions and recommendations

One or two paragraphs stating conclusions, recommendations, and ideas for future work and improvements.

-Our content start from here

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

List any references for your data source(s), other work or results, etc.

Appendix (optional)

Any supporting information or additional information that isn't necessary to have in the main body of the paper. For example, huge tables can go here, especially if they are more than one page. Tables that are, for example, 100 pages most likely should not be included at all.