Elective Project:

Using Dimensionality Reduction to Analyze Vintage Car Market Opportunity

Table of Contents

Executive Summary

The automotive industry is undergoing significant transformations due to shifting market conditions, globalization, and the integration of data-driven technologies. This has prompted SecondLife, a prominent used car dealership in the US, to adapt its business strategy, particularly in the realm of vintage cars. SecondLife has collected extensive data on vintage car sales and aims to leverage this information to optimize its targeting strategies and better cater to the diverse preferences of potential customers as a means to pivot from traditional used car sales to capture a growing market.

SecondLife has recognized the potential of focusing on vintage cars in response to market dynamics and opportunities. To refine their approach to the market, they have been meticulously gathering data on vintage car sales, encompassing various attributes such as displacement, horsepower, and acceleration. This wealth of information provides an opportunity to employ data science methodologies, particularly machine learning, to identify distinct groups of vintage cars based on specific configurations of features. By doing so, SecondLife can streamline its marketing efforts and tailor its approach to different customer segments more effectively by deriving actionable insights from vintage vehicle data.

The primary objective for SecondLife is to identify meaningful combinations of vintage car features that can distinguish specific groups within their inventory. These distinctions will serve as a foundation for developing targeted marketing strategies, allowing SecondLife to appeal to a more diverse and nuanced audience. The end goal is to enhance customer engagement and increase vintage car sales by aligning marketing efforts with the unique preferences and characteristics of distinct customer segments.

Business Problem Overview

SecondLife, a prominent used car dealership with an extensive presence across the United States, is strategically shifting its focus to vintage cars. In response to this market shift, the organization has been diligently collecting data on the vintage cars they have sold over the years. This initiative aligns with the broader industry trend of

leveraging data to derive valuable insights and drive informed decision-making. These groups will be defined based on various attributes such as displacement, horsepower, and acceleration. The goal is to find combinations of these features that can effectively distinguish one group of vintage cars from another.

The identification of these distinct groups is critical for SecondLife's strategic initiatives. By understanding the unique characteristics and configurations that define each group of vintage cars, the dealership can tailor its marketing and sales strategies more effectively. This targeted approach allows SecondLife to connect with a diverse audience, optimizing its efforts to meet the preferences and needs of different customer segments interested in vintage cars.

Key Challenges:

1. **Data Complexity**: The vintage car dataset is complex, given the numerous attributes associated with each vehicle type. Managing and analyzing this data to extract meaningful insights may pose a challenge.
2. **Model Interpretability**: Implementing machine learning models to identify distinct vintage car groups requires careful consideration of model interpretability. The ability to explain and understand the basis for grouping is crucial for informed decision-making, which can be quite challenging when justifying a clustering method to a business audience.
3. **Operational Integration**: Translating data-driven insights into actionable marketing strategies necessitates seamless integration with SecondLife's operational processes. Ensuring that the identified vintage car groups align with practical marketing initiatives is essential.

To address these challenges, SecondLife will engage its data science team in a comprehensive analysis of the vintage car dataset. This involves employing machine learning algorithms for clustering vintage cars based on their attributes. The results can then be translated into actionable marketing strategies, guiding SecondLife in tailoring its approach to different customer segments.

Solution Approach

To address this challenge, this project will employ dimensionality reduction techniques, such as Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE). These techniques will help uncover patterns and relationships within the vintage car dataset, revealing combinations of features that are indicative of distinct groups. The insights derived from this analysis will serve as a foundation for informed decision-making, enabling SecondLife to enhance its vintage car sales strategy and engage with the market more efficiently.

Why PCA and t-SNE? How does it aid our approach?

- **Dimensionality Reduction**: The vintage car dataset at SecondLife likely contains a multitude of features that contribute to the characterization of each vehicle. However, the sheer number of features can make it challenging to discern meaningful patterns and relationships. PCA and t-SNE serve as powerful dimensionality reduction techniques to address this challenge.

Uncovering Latent Patterns with PCA:

- **Variance Retention**: PCA identifies the principal components that capture the maximum variance in the data. By retaining a subset of these components, we reduce the dataset's dimensionality while retaining the essential information. This aids in simplifying the interpretation of the vintage car dataset.
- **Global Structure Understanding**: PCA excels in preserving global structures within the data. It helps identify the major trends and variations across vintage cars, allowing SecondLife to grasp the overarching characteristics that define different groups.

Capturing Local Similarities with t-SNE:

- **Preserving Local Relationships**: Unlike PCA, t-SNE focuses on preserving local similarities in the data. It is particularly effective in capturing intricate, non-linear

relationships between vintage cars. This is valuable for identifying subtle distinctions that might be critical for grouping similar vehicles.

- **Cluster Visualization**: t-SNE produces visualizations that emphasize clusters of similar vintage cars. These visual representations aid in intuitively understanding the relationships between different groups, enabling SecondLife to make informed decisions about targeted marketing and sales strategies.

Approach to Key Objectives:

- **Group Identification**: The primary goal is to identify distinct groups of vintage cars based on attributes such as displacement, horsepower, and acceleration. PCA and t-SNE help in revealing the combinations of features that contribute most significantly to these groupings.
- **Tailored Marketing Strategies**: Understanding the unique characteristics of each group allows SecondLife to tailor its marketing strategies more effectively. By recognizing the preferences and needs of different customer segments interested in vintage cars, the dealership can refine its approach to attract and engage a diverse audience.
- **Efficient Targeting**: The insights gained from PCA and t-SNE facilitate the creation of targeted marketing campaigns. SecondLife can efficiently target specific vintage car groups to match the preferences of distinct customer demographics, thereby increasing the effectiveness of its sales efforts.

Incorporating PCA and t-SNE into the solution approach not only aids in data exploration and feature reduction but also empowers SecondLife with actionable insights. By unraveling the inherent structure within the vintage car dataset, the dealership can make informed decisions that enhance its market presence and streamline its vintage car sales strategy.

Data Overview

Here, we review the relevant data points and how they can help us to meet our objective. There are 8 variables in the dataset underlying the analysis.

1. **mpg (miles per gallon)**: This variable provides information about the fuel efficiency of vintage cars. It can help identify groups based on eco-friendliness or fuel economy preferences among customers.
2. **cyl (number of cylinders)**: The number of cylinders in the engine is indicative of a car's power and performance. This variable can assist in grouping vintage cars based on their engine specifications, catering to customers with preferences for different levels of power.
3. **disp (engine displacement)**: Engine displacement is a measure of the total volume of all cylinders in the engine. It provides insights into the vintage car's performance and can aid in categorizing cars based on their power capabilities.
4. **hp (horsepower)**: Horsepower is a crucial factor in determining a car's speed and overall performance. This variable can help in identifying vintage cars with different power levels, contributing to the grouping based on performance.
5. **wt (vehicle weight)**: Vehicle weight impacts fuel efficiency, handling, and acceleration. Grouping vintage cars based on weight can assist in targeting customers with preferences for lighter, more agile cars or those who prefer heavier, more stable ones.
6. **acc (time to accelerate)**: Acceleration time from 0 to 60 mph is a measure of a car's quickness and agility. This variable can help identify vintage cars with different acceleration characteristics, aiding in the grouping based on performance and speed.
7. **yr (model year)**: The model year provides insights into the vintage of the car. It can help identify trends and preferences for cars from specific eras, contributing to the creation of groups based on historical significance or design evolution.
8. **car name (model name)**: The model name provides information about the brand and specific make of the vintage car. This variable can assist in creating groups based on brand preferences or specific model characteristics that appeal to different customer segments.

While we can glean some great information from each of these respective points, the overall objective demands that we delve further into the minutia of the data through EDA and preprocessing to understand trends, overlapping points, correlation, redundancies, and other factors that will enable us to eliminate or combine features for the most ideal and efficient analysis outcomes.

EDA and Data Preprocessing

Initial Data Cleaning

In the preliminary analysis of the vintage car dataset, it was observed that there are 6 instances where the 'horsepower' variable is recorded as '?'. Considering these instances as missing values, a decision has been made to impute these missing values. This involves replacing the '?' entries with 'np.nan' and subsequently changing the data type of the 'horsepower' column to facilitate a more robust and accurate Exploratory Data Analysis (EDA).

Once the data type has been changed, we can impute float values in place of ‘np.nan’ by replacing each of them with the median value of the ‘horsepower’ column. This enables us to use those features while also mitigating the addition of outlier values that may impact the way cars are grouped.

Univariate Analysis

Our univariate analysis took the form of boxplots and histograms. Each visualization method can be greatly beneficial in helping us understand more about the statistical relevance of each feature in our dataset.

Let's assess the summary statistics for each variable through a univariate lens:

1. **mpg (miles per gallon)**:
    a. Mean: The average fuel efficiency is approximately 23.51 miles per gallon.
    b. Standard Deviation: The variability in fuel efficiency is moderate, with a standard deviation of 7.82.
    c. Min/Max: The range is from 9 to 46.6 miles per gallon.
    d. Percentiles: The majority of cars have a fuel efficiency between 17.5 and 29 miles per gallon.
2. **cylinders**:
    a. Mean: The average number of cylinders is approximately 5.45.

   b. Standard Deviation: The variability in the number of cylinders is moderate, with a standard deviation of 1.70.
   c. Min/Max: The range is from 3 to 8 cylinders.
   d. Percentiles: The majority of cars have 4 or 8 cylinders.
3. **displacement**:
   a. Mean: The average engine displacement is approximately 193.43 cubic inches.
   b. Standard Deviation: The variability in engine displacement is substantial, with a standard deviation of 104.27.
   c. Min/Max: The range is from 68 to 455 cubic inches.
   d. Percentiles: The majority of cars have an engine displacement between 104.25 and 262 cubic inches.
4. **weight**:
   a. Mean: The average vehicle weight is approximately 2970.42 lbs.
   b. Standard Deviation: The variability in vehicle weight is substantial, with a standard deviation of 846.84.
   c. Min/Max: The range is from 1613 to 5140 lbs.
   d. Percentiles: The majority of cars weigh between 2223.75 and 3608 lbs.
5. **acceleration**:
   a. Mean: The average acceleration time is approximately 15.57 seconds from 0 to 60 mph.
   b. Standard Deviation: The variability in acceleration time is moderate, with a standard deviation of 2.76.
   c. Min/Max: The range is from 8 to 24.8 seconds.
   d. Percentiles: The majority of cars have an acceleration time between 13.83 and 17.18 seconds.
6. **model year**:
   a. Mean: The average model year is approximately 76.01.
   b. Standard Deviation: The variability in model years is low, with a standard deviation of 3.70.
   c. Min/Max: The range is from 70 to 82.
   d. Percentiles: The majority of cars in the dataset span model years 73 to 79.

What did we learn?

- The dataset contains a diverse range of vintage cars in terms of fuel efficiency, engine specifications, weight, acceleration, and model years.
- Variability in engine displacement and vehicle weight is relatively high, which makes sense given the diversity in makes, models, and engine sizes.
- The majority of cars have 4 or 8 cylinders.
- Acceleration times vary moderately, with most cars falling within the 13.83 to 17.18 seconds range.
- The dataset spans model years 70 to 82, with the majority of cars concentrated between 73 and 79. This tells us that the majority of vehicles are around the same age, which could mean a couple of things:
    1. The 70s was the era for the most popular vintage car models.
    2. People collectively dislike most cars made in the 70s and want to get rid of them, thus the wealth of models on the market.

Multivaraite Analysis

While valuable, a univariate analysis only provides a surface-level outlook on our data. Getting to a more effective, practical answer will require a comprehensive understanding of the relationship between each feature. We will explore these relationships below:

mpg (miles per gallon)

- Strong Negative Correlation with:
    - Cylinders: This indicates that cars with more cylinders tend to have lower fuel efficiency.
    - Displacement: Larger engine displacements are associated with lower fuel efficiency.
    - Horsepower: Higher horsepower is linked to lower miles per gallon.
    - Weight: Heavier cars are negatively correlated with fuel efficiency.
- Positive Correlation with:
    - Model Year: Suggests that more recent models tend to have higher fuel efficiency.

cylinders

- Strong Positive Correlation with:
    - Weight: Heavier cars tend to have more cylinders.
- Negative Correlation with:
    - Miles Per Gallon (mpg): More cylinders are associated with lower fuel efficiency.

displacement:

- Positive Correlation with:
    - Weight: Heavier cars tend to have larger engine displacements.
- Negative Correlation with:
    - Miles Per Gallon (mpg): Larger engine displacements are associated with lower fuel efficiency.

horsepower:

- Positive Correlation with:
    - Weight: Heavier cars tend to have higher horsepower.
- Negative Correlation with:
    - Miles Per Gallon (mpg): Higher horsepower is associated with lower fuel efficiency.
    - Acceleration: Higher horsepower is linked to slower acceleration.

weight:

- Strong Positive Correlation with:
    - Cylinders: Heavier cars tend to have more cylinders.
    - Displacement: Larger engine displacements are associated with heavier cars.
    - Horsepower: Heavier cars tend to have higher horsepower.
- Negative Correlation with:
    - Miles Per Gallon (mpg): Heavier cars are associated with lower fuel efficiency.

acceleration:

- Negative Correlation with:
    - Horsepower: Faster acceleration is associated with lower horsepower.

model year:

- Positive Correlation with:
    - Miles Per Gallon (mpg): Suggests that more recent model years are associated with higher fuel efficiency.

What did we learn?

- Fuel efficiency (mpg) is influenced by multiple factors, including the number of cylinders, engine displacement, horsepower, weight, and model year.
- Heavier cars with more cylinders, larger engine displacements, and higher horsepower tend to have lower fuel efficiency.
- Acceleration is negatively correlated with horsepower, indicating that cars with higher horsepower may have slower acceleration.
- More recent model years are associated with higher fuel efficiency, suggesting advancements in technology and design.

Based on this analysis, our hypothesis is to group vintage cars based on the following criteria:

1. **Fuel Efficiency Groups**:
    a. Group 1: High Fuel Efficiency (mpg)
    b. Group 2: Moderate Fuel Efficiency
    c. Group 3: Low Fuel Efficiency
2. **Performance Groups**:
    a. Group A: High Performance (Low Cylinders, Low Displacement, Low Weight, High Acceleration)
    b. Group B: Moderate Performance
    c. Group C: Low Performance (High Cylinders, High Displacement, High Weight, Low Acceleration)
3. **Vintage Appeal Groups**:
    a. Group X: Early Vintage (Model Year 70-75)
    b. Group Y: Mid-Vintage (Model Year 76-79)
    c. Group Z: Late Vintage (Model Year 80-82)
4. **Cylinder Configuration Groups**:
    a. Group I: 4 Cylinders
    b. Group II: 6 Cylinders
    c. Group III: 8 Cylinders
5. **Power-to-Weight Ratio Groups**:

a. Group P: High Power-to-Weight Ratio
b. Group Q: Moderate Power-to-Weight Ratio
c. Group R: Low Power-to-Weight Ratio

These groups can be defined based on thresholds or ranges derived from the analysis.

For example:

- High fuel efficiency might correspond to mpg values above a certain threshold.
- High performance could be characterized by low values of cylinders, displacement, weight, and high acceleration.
- Vintage appeal groups can be determined by model years.
- Cylinder configuration groups can be based on the number of cylinders.
- Power-to-weight ratio groups can be established by considering the relationship between horsepower and weight.

Dimensionality Reduction

Now we will walk through the ways that dimensionality reduction impacted our analysis. First, we explore principal component analysis. This required that we used a standard scaler to ensure that all numeric data could be cross-referenced within an equal range of values.

PCA

We applied the PCA algorithm with the number of components equal to the total number of columns in the data, then we found the least number of components required to provide the most information.

In applying this algorithm, we discovered that three components explain 90% of the variance in the data. Each component was then broken down into groups PC1, PC2, and PC3. Here's what we interpreted from these groupings:

1. **PC1**:
    a. Interpretation:
        i. Positive Loadings: cylinders, displacement, horsepower, weight.

   ii. Negative Loadings: mpg, acceleration, model year.
b. Implications:
   i. High PC1 scores correspond to cars with lower fuel efficiency (lower mpg), more cylinders, larger engine displacement, higher horsepower, and heavier weight.
   ii. This component captures the trade-off between fuel efficiency and engine specifications.

2. **PC2**:
   a. Interpretation:
      i. Positive Loadings: acceleration.
      ii. Negative Loadings: cylinders, displacement, horsepower, weight, model year.
   b. Implications:
      i. High PC2 scores correspond to cars with lower acceleration and older model years.
      ii. This component captures the trade-off between acceleration and engine specifications, with a negative influence from the model year.
3. **PC3**:
   a. Interpretation:
      i. Positive Loadings: acceleration.
      ii. Negative Loadings: mpg, displacement, weight.
   b. Implications:
      i. High PC3 scores correspond to cars with higher acceleration, lower fuel efficiency, smaller engine displacement, and lighter weight.
      ii. This component captures the trade-off between acceleration and features related to weight and engine size.

t-SNE

We applied the t-SNE algorithm to assess against PCA. In plotting the data using t-SNE and 'cylinders' as the hue, we found that we could combine several features into 3 groups separated by the count size of the engine (count of cylinders). This approach

was validated by applying multivariate analysis to the t-SNE dataset through boxplots for each feature.

The t-SNE components provided represent the coordinates of individual data points in a two-dimensional space. Each row corresponds to a data point, and the two columns represent the t-SNE components (Component 1 and Component 2) assigned to that point. The t-SNE algorithm maps high-dimensional data to a lower-dimensional space, emphasizing the preservation of local similarities between data points.

Takeaways from this Method

- The t-SNE components allow visualization of the dataset in a 2D space, where each point's position is determined by its local relationships with other points in the original high-dimensional space. In this scenario, the local relationships appeared to have been dictated most by 'cylinder' variables.
- The proximity of points in the t-SNE space suggests similarity in the original high-dimensional feature space, which is consistent with the takeaways on how features correlate in our multivariate analysis.
- Clusters or patterns in the t-SNE space may indicate groups of data points with similar characteristics or relationships, which was validated by the cylinder-based grouping, as the count of cylinders is often correlated to a specific car make, which is consistent with other features like mpg, weight, and acceleration (since acceleration can also be determined by a power-to-weight relationship).

Practical Insights

We got great all-around information from this analysis using both forms of dimensionality reduction. Knowing how this data is clustered is very different from understanding how to act based on these insights. We will review a few ways that incorporating PCA and t-SNE can help be turned into practical decision-making and vintage car market-entry insights below.

1. **Targeted Marketing and Audience Segmentation**:

a. **PCA**: The principal components derived from PCA can be used to create targeted marketing strategies. For example, if PC1 highlights a trade-off between fuel efficiency and engine specifications, SecondLife can create marketing campaigns tailored to customers interested in either high-performance or fuel-efficient vintage cars.
b. **t-SNE**: The clusters identified through t-SNE can be used for audience segmentation. By understanding which vintage cars share similar characteristics, SecondLife can tailor marketing messages to specific audience preferences, increasing the effectiveness of advertising and customer engagement.

2. **Product Positioning and Market Entry Strategies**:
   a. **PCA**: PCA insights into feature importance can guide product positioning. If certain features strongly influence vintage car preferences, SecondLife can strategically position its products in the market based on these characteristics.
   b. **t-SNE**: Clusters identified through t-SNE can inform market entry strategies. If there are distinct groups of vintage cars with unique characteristics, SecondLife can choose to specialize in a particular segment or diversify its offerings to appeal to a broader audience.
3. **Optimized Inventory Management**:
   a. **PCA**: Understanding feature importance through PCA can help optimize inventory management. For instance, if certain features significantly contribute to vintage car sales, SecondLife can prioritize the acquisition and sales of cars with those features.
   b. **t-SNE**: Clustering from t-SNE can assist in grouping similar vintage cars together in the inventory. This can streamline inventory management processes and improve the overall customer experience.

Conclusion

The analysis of the vintage car market through PCA and t-SNE has provided valuable insights into the key features influencing the market landscape. These insights can serve as a foundation for making informed business decisions at SecondLife, the leading used car dealership specializing in vintage cars.

Key Insights

1. **Feature Importance**: PCA highlighted the significance of certain features, such as engine specifications, in influencing vintage car characteristics. This understanding can guide marketing strategies and product positioning.
2. **Segmentation Opportunities**: t-SNE revealed natural clusters within the vintage car dataset, suggesting distinct segments based on similarities in features. This segmentation provides an opportunity for targeted marketing and audience-specific strategies.
3. **Trade-offs and Preferences**: Both PCA and t-SNE emphasized trade-offs in vintage car characteristics, such as the balance between fuel efficiency and engine specifications. Recognizing these trade-offs helps tailor product offerings to meet diverse customer preferences.

Recommendation

Based on these insights, it is recommended that SecondLife adopts a segmented marketing approach, tailoring its strategies to different vintage car clusters identified through t-SNE. Additionally, leveraging the understanding of feature importance from PCA can inform inventory management, allowing SecondLife to prioritize cars with characteristics that strongly resonate with customer preferences.

Next Steps:

1. **Marketing Strategy Development**:
    a. Develop targeted marketing campaigns for each t-SNE cluster, emphasizing the unique characteristics and preferences within each segment.
    b. Incorporate feature-specific messaging in marketing materials based on PCA insights, aligning with customer preferences and trade-offs.

2. **Inventory Optimization**:
    a. Prioritize acquiring vintage cars that align with the most influential features identified by PCA to optimize inventory based on customer preferences.

   b. Align inventory management with t-SNE clusters to enhance customer experience and streamline operations.
3. **Customer Engagement**:
   a. Implement customer engagement initiatives, such as events or promotions, tailored to the preferences of each t-SNE cluster.
   b. Gather feedback from customers in each segment to continually refine offerings and improve customer satisfaction.
4. **Market Expansion Strategies**:
   a. Explore opportunities for market expansion by assessing the unique characteristics and preferences of different t-SNE clusters in potential new markets.
   b. Consider diversifying offerings or introducing specialized services based on identified trade-offs and preferences.

By integrating these insights into practical business decisions and strategic initiatives, SecondLife can enhance its competitive edge, deepen customer engagement, and solidify its position as a leading player in the vintage car market. Continuous monitoring and adaptation based on customer feedback and market dynamics will further contribute to sustained success in this dynamic and diverse market landscape.