INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND MANAGEMENT (ISJEM) ISSN: 2583-6129
VOLUME: 04 ISSUE: 08 | AUG – 2025 DOI: 10.55041/ISJEM04985
AN INTERNATIONAL SCHOLARLY || MULTIDISCIPLINARY || OPEN ACCESS || INDEXING IN ALL MAJOR DATABASE & METADATA

Feature Engineering for Building Machine Learning Models in Automotive Industry

Vaibhav Tummalapalli
Atlanta, GA
vaibhav.tummalapalli21@gmail.com

***Abstract*— Feature engineering is crucial for predictive analytics in the automotive industry, where customer behavior is complex and influenced by multiple factors. This paper presents a framework for developing impactful features for vehicle repurchase and service propensity models using cohort-based analysis. By structuring data into observation and performance windows, it establishes clear cause-and-effect relationships. Techniques such as aggregating service histories, dealer interactions, and loyalty patterns extract actionable insights from sales, service, and campaign data. Practical examples, including partial dependence plots, highlight how features like service intervals, dealer proximity, and purchase histories enhance model accuracy and interpretability. The approach captures temporal patterns, optimizing targeting strategies and improving model performance, engagement, and marketing ROI. Future directions include integrating external data, automating feature updates, and real-time deployment**

Keywords—Feature engineering, Propensity modeling, Machine Learning, Feature creation, Automotive, temporal features, cohort analysis.

I. INTRODUCTION

The automotive customer journey encompasses two pivotal aspects: vehicle repurchase and aftersales services. These milestones form the core of the customer lifecycle, reflecting both their ongoing engagement with the brand and their likelihood to return for future purchases or services. Ensuring an effective communication strategy for each phase of this journey is critical, as it allows businesses to address customer needs through targeted marketing initiatives. Whether encouraging customers to repurchase a vehicle or reminding them to return for regular service, precision in targeting the right audience significantly influences the success of these campaigns.

For such marketing initiatives to succeed, it is essential to identify the best targets. This is where predictive modeling plays a vital role. By leveraging historical and behavioral data, predictive models enable businesses to uncover the key factors that drive customer decisions. However, the effectiveness of these models heavily depends on the quality and relevance of the features used as inputs. Feature engineering, the process of transforming raw data into meaningful variables, is a critical step in creating a proper representation of the data for the problem at hand. Mapping predictors effectively to the event or target behavior ensures that the model captures the nuances of customer behavior.

In the automotive domain, the events of interest—such as vehicle repurchase or service visits—depend on a multitude of factors. These include historical customer engagement, recency and frequency of interactions, monetary value of transactions, past communication or marketing efforts, and external influences like economic conditions and competition. Variables such as recency, frequency, and monetary value (RFM), well-established in customer lifetime value (CLV) modeling [1], were central to feature engineering in our framework, capturing the intensity and timing of customer engagement.

In this paper, we present a framework for creating variables that maximize the predictive power of sales and service data. By demonstrating various feature engineering techniques, we illustrate how to extract actionable insights to improve the accuracy and relevance of predictive models in the automotive industry. This approach not only enhances the understanding of customer behavior but also optimizes targeting strategies, driving higher returns on marketing investments.

II. PROBLEM STATEMENT

The automotive industry generates vast amounts of data from sales transactions, service records, and marketing campaigns. While this data holds significant potential for predicting customer behavior, its raw form is often unstructured and lacks the clarity required for effective analysis. Traditional modeling approaches can falter when the input data is poorly represented, leading to suboptimal predictions and missed opportunities for customer engagement.

A core challenge lies in transforming this raw data into features that capture the underlying patterns and relationships driving customer decisions. For instance, identifying which customers are likely to respond to a vehicle repurchase campaign requires the synthesis of variables such as purchase history, service intervals, and prior campaign interactions. Similarly, predicting service visits depends on factors like recency of service, type of previous services, and communication effectiveness.

Another layer of complexity arises from the multifaceted nature of customer behavior, which is influenced by internal factors (e.g., past engagement, monetary transactions) and external conditions (e.g., economic trends, competitive activities). Capturing these dynamics in a way that aligns with the target event is critical for building robust predictive models. Without a systematic approach to feature engineering, valuable insights can remain hidden, and the predictive models may fail to achieve their full potential.

This paper addresses the gap by presenting a structured methodology for feature engineering in the automotive domain. By leveraging historical sales, service, and campaign data, we aim to demonstrate how to create meaningful variables that enhance predictive performance and improve targeting strategies across marketing initiatives.

III. PROPOSED SOLUTION

To address the challenges outlined, we propose a cohort-based framework for feature engineering in the automotive industry. This framework focuses on structuring data into two distinct windows—the Observation Window and the Performance

Window—to derive predictive features tailored to specific events of interest, such as vehicle repurchase or service visits .

A. Cohort Definition

- **Observation and Performance Windows**: Each cohort is defined by a specific time point (cohort date) that divides the data into an Observation Window and a Performance Window.

o The **Observation Window** includes all historical data available up to the cohort date, such as customer demographics, transaction history, and behavioral metrics (e.g., recency, frequency, monetary value).

o The **Performance Window** captures whether the event of interest (e.g., service visit or vehicle purchase) occurred after the cohort date within a specified time frame.

- **Seasonality and Trend Capture**: Using multiple cohort dates allows for a deeper understanding of seasonal patterns and long-term trends in customer behavior, ensuring the features account for temporal dynamics in the automotive market.

- This division ensures a clear cause-and-effect relationship, with predictors derived from past behavior and outcomes observed in the future.

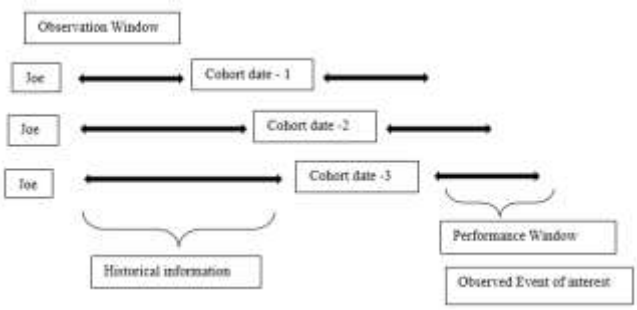

Fig 1. Cohort Set up

- The above image provides a visual representation of how cohort-based [5] analysis evaluates customer behavior at various time points. Here's a breakdown of the concept:

- Here the customer "Joe," is observed at several specific time points or cohorts (e.g., Time Point-1, Time Point-2, Time Point-3) and the event of interest is assessed in the performance window. This is done for every customer in the population.

- This approach captures temporal changes in customer behavior, providing insights into patterns and trends over time [3].

B. Feature Derivation

- **Aggregate Metrics**: Compute cohort-level aggregates such as average spend, frequency of service visits, types of repairs performed and purchase history.

By employing this structured framework, businesses can transform raw data into a rich set of features that accurately reflect customer behaviors and external influences. The combination of Observation and Performance Windows, along with dynamic tracking across multiple time points, ensures robust predictive and Machine learning models tailored to the automotive industry. This approach not only enhances modeling capabilities but also provides actionable insights for marketing strategies, ultimately improving customer engagement and retention

IV. SERVICE FEATURE DERIVATION TECHNIQUES

Feature engineering is pivotal in transforming raw sales, and service data into actionable variables for predictive modeling. Below, we outline several key techniques used to extract meaningful features tailored to specific events.

A. Service Type Aggregation:

Service type aggregation captures the diversity and frequency of services a customer has utilized leading up to a cohort date, offering valuable predictors for customer behavior and engagement.

- **Service Type Counts**: Metrics such as the number of brake repairs, engine services, or tire replacements reflect specific service needs and vehicle health. These can indicate potential future service visits or vehicle replacement considerations.
- **Diverse Service Categories**: Aggregating the number of distinct service types highlights overall customer engagement with the service department, providing insights into loyalty and usage patterns.
- **Temporal Trends**: Tracking the evolution of service types over time helps predict recurring needs and declining engagement, supporting proactive retention strategies.

These features enable models to forecast service visit likelihood, segment customers based on service behaviors, and design targeted campaigns for maintenance plans or upselling opportunities. By integrating service type data, predictive models gain a richer context for understanding customer preferences and lifecycle stages.

B. Dealer Distance and Interaction Features:

Dealer distance and interaction metrics provide critical insights into customer behavior and preferences, making them valuable for predictive modeling.

- **Maximum and Minimum Dealer Distance**: Reflect a customer's travel willingness and proximity to service centers, influencing retention and service visit frequency.

- **Total Unique Dealers Visited**: Indicates loyalty or shopping behavior, with higher counts suggesting less loyalty. This is useful for retention models and cross-dealer marketing strategies.
- **Average Distance Between Dealers**: Captures geographic spread in dealer interactions, highlighting convenience and potential disengagement risks.

These features are essential for predicting service visit frequency, identifying churn risk, targeting regional promotions, and optimizing service network operations. By incorporating these variables, models gain a nuanced understanding of customer engagement and dealer-level dynamics.

C. Repair Order and Mileage Features

Repair order (RO) and mileage features provide essential insights into customer service patterns and vehicle usage, making them highly valuable for predictive modeling.

- **Repair Order Metrics**:
  - **Total Repair Orders**: Reflects the frequency of service interactions, serving as a key indicator of customer engagement with the service department.

  - **Months Since Last and First Repair Order**: Capture temporal gaps in service history, which can indicate customer loyalty or potential disengagement.

  - **Average Time Between Repair Orders**: Highlights service regularity, offering insights into recurring maintenance patterns or gaps that may signal customer churn.

- **Service Due Factor**:
  - A normalized measure of how overdue a customer is for their next expected service based on historical service intervals. This helps predict the likelihood of upcoming service visits and supports proactive outreach efforts.

- **Mileage Metrics**:
  - **Average Mileage Between Service Visits**: Indicates vehicle usage intensity, which correlates with service needs and potential wear and tear.

  - **Estimated Current Mileage**: A predictive estimate of the vehicle's mileage at the time of the cohort, useful for targeting services tied to mileage thresholds (e.g., oil changes, tire rotations).

D. Loyalty Index Features

- **Service Loyalty Index**:
  - **Definition**: Tracks customer service engagement over specific intervals (e.g., 12, 12-24, 24-36, 36-48, 48-60 months) by flagging their activity during these periods.

  - **Predictive Utility**: Helps identify recurring service patterns, indicating loyalty and consistent engagement with the service department.

- **Dynamic Repair Order Counts**:
  - **Definition**: Calculates the total number of repair orders within defined timeframes (e.g., last 12 months, 12-24, 24-36, 36-48, 48-60 months).

  - **Predictive Utility**: Highlights short-term engagement trends, identifying spikes or drops in service interactions.

E. Payment Type Features

- **Revenue Metrics by Payment Type**:
  - **Definition**: Aggregates customer payments into categories like customer pay, warranty pay, and internal pay, capturing sums, averages, and extremes over different time periods like we did for loyalty indices and dynamic repair order counts.

  - **Predictive Utility**: Provides insights into the cost structure of customer interactions, helping model customer profitability. Helps identify patterns in payment types, such as frequent warranty claims, which might indicate dissatisfaction or vehicle reliability concerns.

- **Labor and Parts Metrics**:
  - **Definition**: Separately aggregates labor and parts revenue metrics, such as average labor cost or total parts quantity.

  - **Predictive Utility**: Provides granular insights into service cost drivers. Enables predictive models to segment customers based on the complexity or type of service needs, such as high-labor repair jobs vs. routine part replacements.

- **Distinct Dealer and Parts Counts**:
  - **Definition**: Tracks the number of unique dealers visited or parts used within a specific timeframe.

  - **Predictive Utility**: Distinct dealer counts reflect customer loyalty or shopping behavior. High diversity in parts indicates significant vehicle repair history, potentially signaling aging vehicles or reliability issues.
- **Percentage Metrics**:
  - **Definition**: Calculates the proportion of visits or revenue attributed to specific service types or payment categories.

  - **Predictive Utility**: Useful for normalizing customer data across different engagement levels. Identifies dominant patterns, such as a high percentage of warranty repairs, which may inform retention or satisfaction strategies

V. SALES FEATURES DERIVATION & UTILITY

A. Dealer Distance Metrics

- **Variables**: Maximum and minimum dealer distances, total unique dealers, and average distance between dealers.

- **Predictive Utility**:
  - Helps predict **customer retention**; customers traveling greater distances may churn due to convenience issues.
  - Useful in models estimating **purchase propensity** at nearby dealers.
  - Identifies optimal dealership locations for regional marketing campaigns or network expansion.

B. New vs. Used Vehicle Metrics

- **Variables**: Historical counts of new and used vehicle purchases.
- **Predictive Utility**:
  - Distinguishes customers in different lifecycle stages, aiding in **repurchase likelihood models**.
  - Helps in segmenting customers for **targeted campaigns**, such as promotions for certified pre-owned vehicles.

C. Luxury vs. Mass-Market Purchases

- **Variables**: Historical counts of luxury vs. mass-market vehicle purchases.
- **Predictive Utility**:
  - Highlights customers likely to purchase premium models, supporting **upselling strategies**.
  - Aids in **market segmentation**, predicting preferences for future purchases.

D. Disposed Vehicle Information

- **Variables**: Days since the last vehicle disposal.
- **Predictive Utility**:
  - Indicates readiness for replacement, informing **repurchase prediction models**.
  - Useful for targeting campaigns promoting trade-in offers or lease renewal programs.

E. Purchase Metrics

- **Variables**: Days since first/last purchase, total purchases, and average time between purchases.
- **Predictive Utility**:
  - Provides insights for **repurchase likelihood models**, especially for high-frequency buyers.
  - Identifies **loyalty trends** and churn risk among customers with long purchase gaps.

F. Cumulative Purchase Trends

- **Variables**: Total purchases within specified timeframes (e.g., last 6 or 12 months).
- **Predictive Utility**:
  - Captures recency of engagement, improving predictions for **short-term purchase intent**.
  - Enables dynamic targeting of customers with recent activity for **promotional campaigns**.

G. Sales Loyalty Index

- **Variables**: Tracks purchases over specified intervals (e.g., last 12 months, 12-24, 24-26, 36-48, 48-60 months); combined loyalty index.
- As noted in the J.D. Power Automotive Brand Loyalty Study [4], customers tend to stay within brand families, making service and loyalty variables crucial in predicting repurchase
- **Predictive Utility**:
  - Indicates long-term engagement, aiding in **customer retention models**.
  - Useful for segmenting loyal customers likely to respond to **exclusive offers** or loyalty rewards.

H. Purchase Type Metrics

- **Variables**: Historical counts of cash, loan, and lease purchases.
- **Predictive Utility**:
  - Reflects financing preferences, which can improve **customized financing offers**.
  - Supports churn prediction by analyzing **customer financing trends**.

I. Vehicle Type Metrics

- **Variables**: Historical counts of different vehicle types owned by the customer (e.g., SUV, sedan, truck).
- **Predictive Utility**:
  - Identifies preferences for specific vehicle categories, enabling **personalized vehicle recommendations**.
  - Helps forecast future purchases by **tracking shifts in preferences** over time.

J. Price and Financial Metrics

- **Variables**: Maximum, minimum, and average purchase prices; down payments; and interest rates (APR).
- **Predictive Utility**:
  - Helps model **customer lifetime value** by identifying high-value customers.
  - Predicts **price sensitivity**, aiding in price-based segmentation for targeted offers

VI. EVALUATION

Feature engineering plays a critical role in enhancing the accuracy and interpretability of predictive models. By systematically creating meaningful features from raw data, the overall model performance improves significantly. Structured and interpretable features, like counts of distinct services or dealer distances, make it easier to communicate model predictions to stakeholders and refine strategies accordingly. In line with earlier findings on the effectiveness of RFM-based modeling in classification problems [2], our models consistently ranked RFM features among the most important predictors of service behavior. Below are the examples of high-impact variables from service and purchase propensity models, accompanied by their interpretation using partial dependence (PD) plots.

A. Service Propensity Models:

Partial dependence plots provide insights into how specific features influence the likelihood of customers engaging with a service campaign. The following examples illustrate the impact of critical variables:

Months Since Last Repair Order (RO):

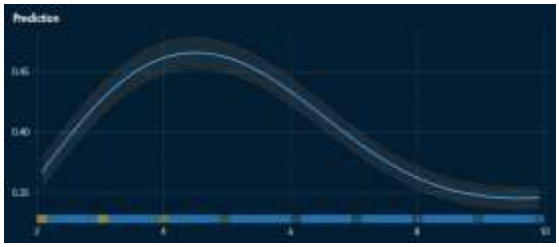

Fig 2. Partial Dependence Plot (Months since last repair order)

- **Behavior**: Customers who recently got a service (low months since last RO) are less likely to respond to a service campaign. However, the propensity peaks between 3–5 months, as this aligns with routine maintenance intervals. Beyond six months, the likelihood drops, indicating disengagement and a potential risk of defection.

- **Interpretation**: This variable helps identify the optimal timing for outreach, ensuring campaigns target engaged customers likely to respond positively.

Total Spend in the Last 12 Months:

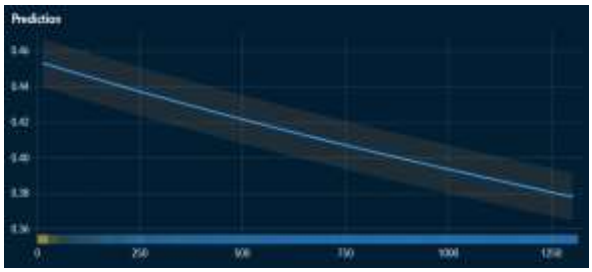

Fig 3. Partial Dependence Plot (Total Spend in Last 12 Months)

- **Behavior**: Higher spending in the last year correlates with lower service propensity in the next 60 days. This pattern suggests that customers who recently invested in substantial services are less likely to return immediately.

- **Interpretation**: This feature assists in excluding over-serviced customers from short-term campaigns, focusing instead on those more likely to visit.

Average Days Between Repair Orders:

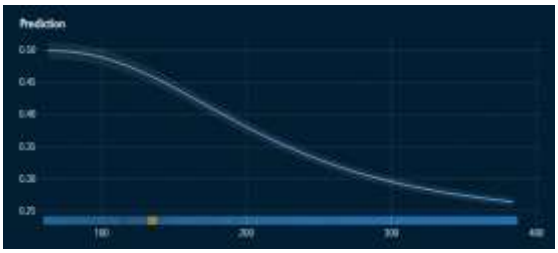

Fig 4. Partial Dependence Plot (Average Days Between ROs)

- **Behavior**: Customers with shorter intervals between visits are more engaged, and their propensity to respond is highest at around 62 days. As the average interval increases, the likelihood of a response declines, reflecting lower engagement.

- **Interpretation**: This feature aids in identifying highly engaged customers and tailoring campaigns accordingly.

B. Purchase Propensity Models

For vehicle purchase models, PD plots reveal how behavioral patterns influence the likelihood of purchasing a new vehicle:

Time Since Last Purchase:

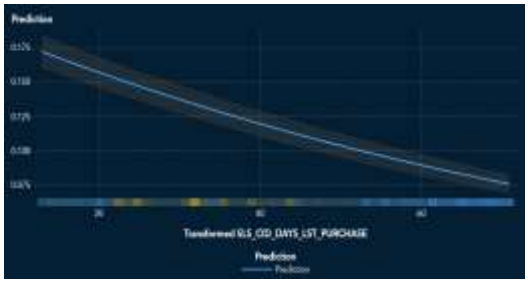

Fig 5. Partial Dependence Plot (Time Since Last Purchase)

- **Behavior**: Recent purchasers exhibit higher propensities to buy another vehicle soon, reflecting the potential for additional purchases within households or an affinity for upgrading. Conversely, individuals with extended gaps since their last purchase may either have switched brands or be satisfied with their current vehicle.

- **Interpretation**: This feature helps differentiate between active shoppers and disengaged customers, enabling targeted marketing.

VII. CONCLUSION

This paper demonstrates the transformative role of feature engineering in predictive analytics within the automotive domain. By aggregating and enriching raw data from sales and service, we created features that provide better representation of the data for the problem at hand. The structured approach of using observation and performance windows, combined with temporal tracking, enabled robust modeling of customer behavior over time. Features capturing recency, frequency, and diversity of interactions consistently provided actionable insights for business strategies.

Future Directions:

- **Dynamic Feature Updates**: Automating the creation and refresh of time-sensitive features, such as campaign engagement metrics, will allow real-time adaptation to changing customer behaviors.

- **Integration of External Data**: Incorporating macroeconomic indicators, competitive intelligence, and third-party data (e.g., social media or telematics) can enrich feature sets further, improving predictive accuracy and expanding use cases.

This paper underscores the pivotal role of feature engineering in unlocking the predictive potential of automotive data, paving the way for future advancements in customer engagement and business performance

REFERENCES

[1] P. S. Fader, B. G. Hardie, and K. L. Lee, “RFM and CLV: Using Iso-Value Curves for Customer Base Analysis,” *Journal of Marketing Research*, vol. 42, no. 4, pp. 415–430, 2005.

[2] I.-C. Yeh, K.-J. Yang, and T.-M. Ting, “Knowledge discovery on RFM model using Bernoulli sequence,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 5866–5871, 2009.

[3] Kabasakal, İnanç. (2020). Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing. 13. 47-56. 10.17671/gazibtd.570866.

[4] **J. D. Power and Associates,** "2019 U.S. Automotive Brand Loyalty Study," *J.D. Power*, 2019. [Online]. Available: https://www.jdpower.com/business/press-releases/2019-us-automotive-brand-loyalty-study.

[5] **K.-F. Wu and P. P. Jovanis**, "Cohort-Based Analysis Structure for Modeling Driver Behavior with an In-Vehicle Data Recorder," *Transportation Research Record*, vol. 2601, no. 1, pp. 24–32, 2016