Statistical Techniques for feature selection in Machine learning Models

Vaibhav Tummalapalli[1], Kiran Konakalla[2]

[1]vaibhav.tummalapalli21@gmail.com
[2]kiran.konakalla7@gmail.com

Abstract

Feature selection is a critical step in the machine learning pipeline, particularly when working with high-dimensional datasets common in the automotive and marketing domains. Selecting the most informative predictors not only improves model accuracy and interpretability but also enhances computational efficiency and decision-making speedy factors in real-time business applications. This paper explores four foundational statistical techniques for feature selection: Information Value (IV), Chi-Square Test, Analysis of Variance (ANOVA), and Correlation Coefficients. Each method is presented with its theoretical foundation, historical significance, and mathematical formulation. Beyond academic context, we highlight practical implications of feature selection in driving operational efficiency, reducing model training costs, and improving the effectiveness of customer segmentation, campaign targeting, and vehicle sales predictions. By understanding and leveraging these techniques, practitioners can streamline model development and ensure actionable insights that translate to measurable business outcomes

Keywords: Feature Selection, Chi-Square, ANOVA, Information Value, Machine Learning, Correlation, Pearson Coefficient

I. INTRODUCTION

In an age where data-driven decision-making defines competitive advantage, the ability to extract relevant insights from large and complex datasets is paramount. Nowhere is this more apparent than in the automotive industry, where organizations routinely analyze data on customer demographics, vehicle ownership, purchase history, service patterns, lifestyle, and digital engagement to power models for marketing, sales forecasting, and campaign optimization. However, high-dimensional datasets—those with hundreds or even thousands of features—can pose serious challenges: increased training time, overfitting, model complexity, and reduced interpretability.

Feature selection addresses these challenges by identifying the most relevant and informative predictors while removing redundant or irrelevant variables. This step not only boosts model performance metrics such as precision and recall but also supports faster model deployment and easier interpretation by business users. In domains like automotive marketing, where speed-to-decision is critical—whether for launching a new campaign, forecasting demand for EVs, or identifying high-value service prospects—efficient feature selection becomes a key operational enabler.

Among the variety of techniques available, statistical methods stand out for their simplicity, transparency, and grounding in hypothesis testing and data distribution analysis. Unlike black-box methods such as LASSO or tree-based feature importance, statistical techniques offer clear interpretability, making them suitable for regulated industries and explainable AI initiatives.

This paper focuses on four widely adopted statistical techniques:

- **Information Value (IV)** – used extensively in binary classification tasks, especially in credit scoring and churn modeling.
- **Chi-Square Test** – a non-parametric method for evaluating the independence between categorical features and the outcome variable.
- **Analysis of Variance (ANOVA)** – a parametric test for comparing group means and identifying continuous features that vary significantly across target classes.
- **Correlation Coefficients** – particularly Pearson and Spearman coefficients, which assess linear and monotonic relationships among continuous features and target variables.

Each technique is introduced with a historical overview, followed by a theoretical explanation, mathematical formulation, and practical examples. We also explore how these methods contribute to **operational efficiency** by reducing feature engineering time, lowering computational costs, and simplifying deployment workflows. Moreover, we discuss their applications in automotive sales and marketing use cases—such as optimizing campaign target lists, improving service propensity models, and eliminating multicollinearity in pricing and retention analyses.

By grounding feature selection in well-established statistical principles, this paper equips practitioners with robust tools to improve model quality while aligning machine learning outputs with real-world business needs [1],[2].

II. Feature Selection Methods

A. *Information Value*

Origin & Context

Information Value (IV) is a statistical measure that originated in the domain of credit risk modeling, where it was first used to evaluate the predictive strength of variables in assessing loan default risk. The concept is closely tied to Weight of Evidence (WoE), a transformation developed in the early 20th century for logistic regression models. WoE provides a numerical expression of how a given attribute differentiates between two binary classes, typically an "event" (e.g., default, purchase, response) and a "non-event" (e.g., no default, no purchase, no response).

IV builds upon WoE by aggregating this discriminatory power across bins of a continuous or categorical feature. The resulting IV value serves as a summary metric that quantifies the usefulness of a predictor in separating the two outcome classes, making it a powerful tool for feature selection and model simplification [3]

Theoretical Basis

The central idea behind IV is that a good predictive feature will have distinct distributions for events and non-events. For instance, if a customer's income level strongly influences the likelihood of purchasing a vehicle, we would expect the distribution of income among purchasers (events) to differ significantly from that of non-purchasers (non-events). IV formalizes this intuition by measuring how much information the feature contributes to the prediction task.

To apply IV, the feature is first binned—either into equal-sized quantiles, business-driven intervals, or using supervised binning methods—and then the proportion of events and non-events is calculated within each bin. The larger the divergence between these distributions, the higher the IV, indicating greater predictive power

Mathematical Details

1. Divide the feature X into k bins or intervals.
2. Compute the proportions of events ($\%E_i$) and non-events ($\%NE_i$) in each bin i:

$$\%E_i = \frac{E_i}{E}, \quad \%NE_i = \frac{NE_i}{NE}$$

where E_i and NE_i are the counts of events and non-events in bin i, and E and NE are the total counts.

3. Calculate the Weight of Evidence (WoE):

$$\text{WoE}_i = \ln\left(\frac{\%E_i}{\%NE_i}\right)$$

4. Computing the IV:

$$\text{IV} = \sum_{i=1}^{k}(\%E_i - \%NE_i) \cdot \text{WoE}_i$$

Interpretation Guideline

Table 1 - IV Guideline

IV Value	Predictive Strength
< 0.02	Not useful
0.02 – 0.1	Weak predictive power
0.1 – 0.3	Medium predictive power
0.3 – 0.5	Strong predictive power
> 0.5	Suspiciously strong (may indicate overfitting or data leakage)

Applications

While IV was originally developed for credit scoring, its applicability extends well beyond finance into domains like automotive marketing and sales, where binary outcomes (e.g., purchase vs. no purchase, campaign response vs. non-response) are common.

- **Marketing Campaign Optimization:** IV can help identify customer attributes—such as lifestyle indicators, digital engagement metrics, or historical service interactions—that best predict response to marketing campaigns. This enables more precise targeting, reducing outreach costs and improving campaign ROI.
- **Sales Funnel Efficiency:** For predicting conversion at different sales stages (e.g., test drive scheduled vs. no-show), IV can rank features such as lead source, time of inquiry, or vehicle model preference, enabling sales teams to prioritize high-quality leads.
- **Operational Streamlining:** In aftersales service models, IV can identify high-utility features like mileage bands, warranty status, or prior service timing, helping businesses prioritize proactive outreach to customers with high service propensity. This leads to better technician scheduling and inventory planning.
- **Feature Reduction for Faster Model Deployment:** By eliminating low-IV variables early, data scientists reduce training complexity and streamline deployment pipelines. This is especially important in real-time systems or constrained environments (e.g., mobile apps, dealership tablets).

B. Chi-Square Test

Origin & Context

The Chi-Square Test, introduced by Karl Pearson in 1900, is one of the most widely used non-parametric statistical methods. Originally designed to evaluate the goodness-of-fit between observed and expected distributions, it has evolved into a fundamental tool in hypothesis testing, particularly for evaluating associations between categorical variables. In the context of machine learning and predictive modeling, the Chi-Square Test plays a vital role in feature selection, especially when dealing with categorical predictors or binned numerical features [2].

Its simplicity and versatility make it a preferred method for determining whether a feature has a statistically significant relationship with the target variable. This is especially valuable in early-stage exploratory data analysis (EDA) or automated feature screening pipelines.

Theoretical Basis

The Chi-Square Test of Independence examines whether two categorical variables are statistically independent—i.e., whether the distribution of one variable is unrelated to the distribution of the other. In machine learning, this is typically used to evaluate whether a categorical feature has a meaningful association with a binary or multi-class target variable.

Under the null hypothesis (H_0), the test assumes that there is no association between the feature and the target (they are independent). The alternative hypothesis (H_1) posits that a relationship does exist. The key idea is to compare the observed frequencies of data points falling into each category combination to the expected frequencies assuming independence. A large deviation between the observed and expected frequencies suggests dependence, and the null hypothesis may be rejected

Mathematical Details

1. Construct a contingency table with observed frequencies O_{ij} for each category i of the feature and each class i of the target.
2. Calculate expected frequencies E_{ij} :

$$E_{ij} = \frac{\text{Row Total}_i \times \text{Column Total}_j}{\text{Grand Total}}$$

3. Compute the Chi-Square Statistic:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

4. Compare $\chi 2$ to the critical value from the Chi-Square distribution table at a chosen significance level (α).

Interpretation

- A high Chi-Square value (or low p-value) suggests that the feature and target variable are not independent, i.e. the feature is statistically associated with the target and may be valuable for predictive modeling.
- A low Chi-Square value implies that the observed distribution is close to the expected distribution under independence, indicating the feature has limited predictive relevance.

Applications

The Chi-Square Test has a broad range of practical applications in business analytics, especially for feature selection in classification tasks involving categorical variables. In the automotive industry, it is particularly useful for identifying key demographic, behavioral, or categorical predictors that influence purchasing or engagement decisions.

- **Marketing Campaign Targeting:** Categorical attributes such as region, household type, income bracket, or customer segment can be evaluated for their association with response to marketing campaigns. A significant Chi-Square score indicates that certain categories (e.g., urban dwellers or high-income groups) respond differently to specific offers, informing segmentation and message customization.

- **Sales Funnel Optimization:** Attributes like lead source (website, referral, dealership), campaign channel (email, direct mail, social) or vehicle interest category (SUV, sedan, electric) can be tested for association with progression through the sales funnel (e.g., inquiry → test drive → purchase). Features with significant association are prioritized for lead scoring models and sales strategy refinement.

- **Feature Screening in Rule-Based Systems:** In operational systems where explainability is crucial—such as CRM rules or eligibility criteria for finance offers, the Chi-Square test can help select categorical inputs that have meaningful, statistically grounded relationships with key decisions.

- **Inventory and Service Planning:** For aftersales operations, features like service type, warranty status, dealership zone, and vehicle brand can be tested for their impact on service engagement, helping optimize promotions or staffing at specific service locations.

Advantages For Operational Efficiency

- **Low computational cost:** Easy to implement even on large datasets.
- **Non-parametric:** No assumptions about the distribution of data, making it robust in noisy real-world environments.
- **Model-agnostic:** Applicable before selecting a modeling algorithm, improving upstream data preparation.
- **Supports automation:** Can be embedded into AutoML workflows for categorical feature screening.

C. ANOVA (Analysis of Variance)

Analysis of Variance (ANOVA) was developed by Sir Ronald Fisher in the early 20th century as a method to statistically compare the means of three or more groups and assess whether any of those group means differ significantly. Originally introduced in agricultural experiments, ANOVA has become a cornerstone of statistical analysis across disciplines, including psychology, economics, and increasingly, machine learning.

In predictive modeling, ANOVA is used for feature selection when the independent variable is continuous, and the target variable is categorical (typically binary or multi-class). It helps determine

whether the distribution of a numerical predictor differs across categories of the target variable, making it especially valuable for classification models.

Theoretical Basis

ANOVA evaluates the extent to which variation in a continuous feature X can be explained by differences between the groups defined by a categorical target variable Y. It does this by partitioning the total variance of X into two components:

- **Between-Group Variance (SSB):** Measures how much group means differ from the overall mean.
- **Within-Group Variance (SSW):** Measures the variance within each group (noise or unexplained variation).

The central hypothesis tested by ANOVA is:

- **Null Hypothesis (H_0):** The means of the groups are equal; the feature X does not vary significantly across target classes.
- **Alternative Hypothesis (H_1):** At least one group mean differs significantly; the feature X may help distinguish between target categories.

A **high F-statistic** suggests that the between-group variance dominates within-group variance, implying the feature may be predictive.

Mathematical Details

1. Compute the Total Sum of Squares (SST):

$$SST = \sum_{i=1}^{N}(x_i - \bar{x})^2$$

2. Compute the Between-Group Sum of Squares (SSB):

$$SSB = \sum_{j=1}^{k} n_j \left(\bar{x_j} - \bar{x}\right)^2$$

where n_j is the group size and $\bar{x}_j$ is the group mean.

3. Compute the Within-Group Sum of Squares (SSW):

$$SSW = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x_j}\right)^2$$

4. Calculate the F-statistic:

$$F = \frac{SSB/(k-1)}{SSW/(N-k)}$$

The calculated F-Statistic is then compared to a critical value from the F-distribution, or its p-value is tested against a significance level α. If $p < \alpha$, the feature is deemed statistically significant.

Interpretation

- A high F-value and low p-value indicate that the continuous variable shows significantly different values across target groups, suggesting predictive utility.
- A low F-value means the feature does not explain much variance between groups, and might be excluded from modeling

Applications [2]

ANOVA is especially useful in identifying numeric predictors that have meaningful differences across response categories, helping business teams prioritize high-impact variables in their models.

- **Customer Purchase Propensity Models**: Features like monthly payment, household income, or average service cost can be evaluated for their variance across vehicle buyer types (e.g., new vs. used, SUV vs. sedan). For example, if monthly payment amounts differ significantly between EV purchasers and non-EV purchasers, ANOVA would flag this feature as valuable.

- **Service Campaign Targeting**: ANOVA can determine whether metrics such as mileage, time since last service, or average distance driven vary across responders and non-responders to specific service offers (e.g., tire rotation or brake service). This helps with tailoring offers to segments where the service need is more likely.

- **Lead Quality and Sales Funnel Optimization**: Continuous engagement features—like response time to lead capture, number of page views, or test drive duration—can be tested across lead outcomes (converted vs. not converted). Features with significant between-group variation can be used to rank or score lead quality.

- **Operational Efficiency and Forecasting**: Operational KPIs such as dealer visit duration, repair costs, or labor hours per vehicle type can be analyzed with ANOVA to determine whether these vary significantly by vehicle class, campaign type, or customer tier. Insights from this analysis can improve scheduling, parts stocking, and labor allocation

.

D. *Correlation Coefficients*

Correlation analysis is one of the oldest and most fundamental techniques in statistics. The concept of correlation was first explored by Francis Galton in the late 19th century and mathematically formalized by Karl Pearson in 1896. Correlation coefficients quantify the strength and direction of a relationship between two variables, making them invaluable tools in exploratory data analysis (EDA), hypothesis testing, and feature selection.

In the context of machine learning and predictive modeling, correlation analysis serves two main purposes:

- **Assessing feature relevance**: Determine how strongly a numeric feature is associated with the target variable.
- **Detecting multicollinearity**: Identify highly interrelated features that may introduce redundancy, instability, or bias into the model.

Correlation coefficients are particularly useful during preprocessing and feature engineering stages, enabling the practitioner to select variables that are both predictive and non-redundant, enhancing model performance and interpretability

Theoretical Basis

Correlation coefficients measure the degree to which two variables move together. They can indicate:

- **Positive correlation**: When one variable increases, the other tends to increase.
- **Negative correlation**: When one variable increases, the other tends to decrease.
- **No correlation**: No consistent pattern in the movement between variables.

There are two commonly used types of correlation in feature selection:

- **Pearson's Correlation Coefficient**: Measures the linear relationship between two continuous variables. Assumes normally distributed variables and a linear relationship.
- **Spearman's Rank Correlation Coefficient**: A non-parametric metric that measures the monotonic relationship between two variables using their rank orders, making it robust to outliers and non-linear relationships.

Mathematical Details

1. Pearson Correlation: Defined for two continuous variables X and Y, the Pearson correlation is calculated as:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

$r \in [-1,1]$
$r = 1$: perfect positive linear relationship
$r = -1$: perfect negative linear relationship
$r = 0$: no linear relationship

2. Spearman Rank Correlation (for monotonic relationships):

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where d_i is the rank difference for pairs i.

Applications

Feature Relevance to Target Outcomes: In predictive modeling for customer behavior, correlation analysis helps identify which numeric variables are most aligned with the business outcome. For instance:

- Vehicle price, monthly payment, or loan term might be positively correlated with EV adoption.
- Average service spend might be correlated with service campaign responsiveness.

Selecting features with high correlation to the target ensures that the model focuses on impactful predictors.

Reducing Multicollinearity in Regression Models: Highly correlated predictors can inflate standard errors, destabilize coefficients, and reduce model generalizability. For example:

- Mileage and age of vehicles might be strongly correlated. Including both could be redundant.
- Annual income and home market value may overlap in explaining financial behavior.

Using correlation matrices or heatmaps, one of each highly correlated pair can be dropped, or they can be combined into a derived feature (e.g., PCA or ratio).

Customer Segmentation and Personalization: Correlation coefficients assist in building scoring systems by identifying key relationships between behavioral signals and purchasing patterns. For example, correlation between frequency of service visits and likelihood to repurchase can inform customer value tiers or segmentation.

Campaign Efficiency and Resource Allocation: By removing redundant features and focusing on strongly correlated drivers, marketing and sales teams can simplify decision rules and deploy lighter,

faster models—ideal for real-time systems or edge applications (e.g., dealer tablets, web personalization engines).

III. CONCLUSION

Statistical techniques such as Information Value, Chi-Square Test, ANOVA, and Correlation Coefficients provide essential tools for feature selection in machine learning. By leveraging these methods, practitioners can build more efficient and interpretable models while maintaining high predictive performance. Future work could explore hybrid techniques that combine statistical methods with advanced machine learning algorithms.

REFERENCES

[1] Guyon and A. Elisseeff, *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York: Springer, 2009.

[3] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Springer, 1998.