INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 7, July 2026 | ISSN: 2456-4184 | IJNRD.ORG

An Automated Framework for Unbiased Consumer Segmentation: Balancing Statistical Precision and Business Key Performance Indicators

A Distributed Enterprise Approach Using Apache Spark and Multi-Dimensional Feature Reduction

[1]Anvesh Reddy Minukuri,

[1] ***MBA Graduate, UIUC, Gies School of Business VP, Sr Lead, GenAI/ML Engineer, JPMorgan Chase***

Abstract: Consumer segmentation remains a foundational imperative for optimizing go-to-market strategies, product development, and organizational sales alignment. Nevertheless, a persistent structural challenge within enterprise analytics lies in the inherent conflict between unsupervised statistical clustering methodologies—which prioritize intra-cluster variance minimization—and strategic business objectives, which necessitate actionable, unbiased segments strictly aligned with Key Performance Indicators (KPIs). Traditional market segmentation initiatives frequently yield heavily skewed, human-biased models attributable to computational constraints, limited exploratory iterations, and the erroneous application of continuous distance metrics to ordinal survey instruments. This paper introduces a highly scalable, automated consumer segmentation architecture deployed on distributed systems via Apache Spark. The proposed framework autonomously generates, evaluates, and audits over 15,000 distinct segmentation combinations. The technical pipeline integrates a rigorous multi-dimensional data reduction sequence, executing variable clustering and Principal Component Analysis (PCA) across every major domain—including customer demographics, firmographics, census data, internal telemetry, consumption patterns, digital interactions, product usage, and service attributes—to ensure robust variance coverage while mitigating information loss in high-dimensional spaces. Furthermore, survey and customer telemetry are dynamically reweighted to correct for sample-to-population divergences. Beyond standard Silhouette validation, the architecture incorporates Calinski-Harabasz and Davies-Bouldin indices, executes temporal stability assessments via bootstrap Jaccard similarity, incorporates SHAP (SHapley Additive exPlanations) for explainable feature attribution, implements Population Stability Index (PSI) monitoring for concept drift, and provides rigorous financial ROI impact modeling and feature weight sensitivity analyses. Candidate models are filtered via automated business profiling thresholds and statistical delta metrics, with final curated solutions surfaced through an interactive Streamlit dashboard.

IndexTerms **-** Consumer Segmentation, Market Analytics, Bias-Variance Tradeoff, Apache Spark, Distributed Computing, Multi-Dimensional Feature Reduction, Principal Component Analysis (PCA), Tetrachoric Correlation, Polychoric Correlation, Sample-to-Population Reweighting, K-Means++ Clustering, Algorithmic Gating & Threshold Disqualification, Calinski-Harabasz Index, Davies-Bouldin Index, Temporal Stability Analysis, Bootstrap Jaccard Similarity, Explainable AI (XAI), SHAP (SHapley Additive exPlanations), Population Stability Index (PSI), Concept Drift Governance, Financial ROI Modeling, Key Performance Indicators (KPIs), Interactive Dashboards, Streamlit

1. Introduction: The Variance-Bias Dilemma in Market Segmentation

Understanding consumer attitudes, behavioral intents, and latent motivations is vital for cross-functional alignment across modern enterprises. As consumer mindsets dynamically evolve—particularly across digital media consumption, mobile interactions, and product expectations—organizations must segment their subscriber base to tailor strategic planning effectively. However, the execution of market segmentation initiatives frequently falters due to an inability to harmonize clustering solutions with both empirical statistical rigor and practical business imperatives.

This tension engenders a recognized dichotomy within enterprise data science:

The Statistical Perspective (Variance Optimization): Statisticians and machine learning engineers naturally gravitate toward algorithms that generate the tightest, most mathematically distinct clusters. While these models successfully optimize intra-cluster variance, they frequently yield segments devoid of practical storytelling capability or strategic interpretability for business stakeholders.

The Business Perspective (KPI Fitting): Conversely, executive and marketing leadership often advocate for segmentations that corroborate preconceived operational narratives or specific tactical mandates. This introduces severe human bias, forcing complex multidimensional data into unnatural partitions that fail rigorous empirical validation.

Presently, a significant deficiency exists in the availability of agnostic frameworks designed to navigate this variance-bias tradeoff in unsupervised market segmentation. Due to the labor-intensive nature of exploratory data analysis, data scientists typically examine only a handful of model iterations, selecting a solution based on a restricted subset of statistical metrics, which ultimately produces a highly subjective, human-biased outcome. Harnessing Apache Spark for distributed computing, this research delineates a scalable, fully automated analytical pipeline capable of generating, scoring, and profiling upwards of 15,000 distinct clustering solutions to definitively resolve this methodological dichotomy.

2. Survey Design, Execution, and Sample Generalization

To capture the cognitive and psychological drivers underpinning consumer actions, internal transactional telemetry must be systematically augmented with external survey data. The initial phase of the proposed architecture involves deploying a targeted survey instrument designed to capture competitor perceptions, unmet consumer needs, and granular behavioral intent.

2.1 Sample Sizing and Statistical Power

To ensure that the survey sample size generalizes robustly to the broader subscriber population, rigorous sample size estimations were conducted. Threshold parameters were benchmarked against two distinct confidence levels to guarantee representativeness: a 95% Confidence Interval (CI) with a 1.0% Margin of Error, and a 99% Confidence Interval (CI) with a 1.0% Margin of Error.
Let N represent the total subscriber population size, Z denote the critical value corresponding to the desired confidence level, p signify the estimated population proportion of a given attribute, and E indicate the specified margin of error. The required sample size n is mathematically formulated as:

$$n = [(Z^2 * p * (1 - p)) / E^2] / [1 + ((Z^2 * p * (1 - p)) / (E^2 * N))]$$

2.2 Survey Execution and Sample-to-Population Reweighting Mechanism

A fundamental challenge in empirical survey research and enterprise analytics is that collected data inherently originates from a finite sample rather than a complete census of the target population. Voluntary survey response rates and customer telemetry collection are subject to non-response and selection biases; respondents who complete surveys or active customer records rarely mirror the exact demographic, socio-economic, and behavioral distribution of the total subscriber base. To establish true statistical congruence and ensure that downstream segmentation accurately reflects the broader population, a robust reweighting mechanism was implemented.
The weighting architecture leverages comprehensive benchmark data across major strata utilized by the enterprise to manage customer portfolios. These include customer demographics, firmographics, geographic census data, internal telemetry, consumption patterns, digital activity indices, product holdings, and service interactions. By comparing the true proportion of each stratum in the broader target population against its observed proportion in our sample, dynamic sampling weights are derived.
The row weight W_i assigned to an individual respondent i belonging to a specific multi-dimensional stratum is calculated as the ratio of the target population proportion (P_i) to the observed sample proportion (S_i):

$$W\_i = P\_i / S\_i$$

Upon derivation, these dynamic sample-to-population weights are applied directly to every individual record prior to model estimation. In iterative optimization algorithms such as K-Means, these weights are passed as observation weights to ensure that centroid updates and distance calculations reflect true population distributions rather than unadjusted sample biases.

For distance metrics and partition scoring, version scores and normalized distances are evaluated. For instance, version scores are computed as $1 - c_{1,\text{Total}}$ weighted proportionally by respective segment sizes. Furthermore, internal cluster validation incorporates the Calinski-Harabasz variance ratio criterion CH and the Davies-Bouldin index DB:

$$CH = \frac{B_k/(k-1)}{W_k/(N-k)}$$

where B_k represents between-cluster dispersion, W_k represents within-cluster dispersion, k is the number of clusters, and N is the total sample size. The proposed architecture integrates K-Means++ as the foundational observational clustering algorithm, leveraging probabilistic centroid seeding to accelerate mathematical convergence and avoid suboptimal local minima [2].

3. Literature Review

The evolution of market segmentation methodologies has progressed through several distinct phases over the past two decades. Early approaches primarily relied on statistical methods such as k-means and hierarchical clustering applied to structured customer data. These foundational techniques established the importance of factors such as demographic profile, purchase frequency, and product adoption history in grouping customer segments. As computational capabilities advanced, the field witnessed a shift toward more sophisticated ensemble methods and automated exploration pipelines, which improved segmentation robustness by capturing complex non-linear relationships between variables.

Deep learning and automated machine learning applications in consumer behavior modeling have emerged as a promising frontier for enterprises seeking to understand increasingly complex customer journeys. Research has established that interpretable machine learning approaches can effectively balance the seemingly competing objectives of model performance and explainability in segmentation contexts. By employing techniques such as SHAP (SHapley Additive exPlanations) values, attention mechanisms, and feature attribution models, researchers have demonstrated that complex models can provide comprehensible explanations for their classifications.

Gaps in existing approaches remain substantial despite these advancements, particularly regarding the reconciliation of statistical variance optimization with business KPI alignment. Most current implementations still operate on manual, time-consuming model iterations with significant human bias, limiting the exploration of optimal feature combinations. Additionally, while individual studies have demonstrated the value of specific data types, few have successfully integrated heterogeneous enterprise streams—spanning demographics, firmographics, census data, telemetry, and digital interactions—into unified automated frameworks.

4. Feature Engineering and Data Reduction Pipeline

Following the collection, cleaning, and reweighting of data, internal telemetry and survey repositories were merged to construct approximately 1,500 engineered features capturing the multi-dimensional spectrum of consumer profiles. Managing this high-dimensional feature space without incurring severe information loss or the curse of dimensionality necessitates a rigorous, multi-stage data reduction pipeline combining variable clustering and Principal Component Analysis (PCA).

4.1 Multi-Dimensional Variable Clustering and PCA across Enterprise Domains

High-dimensional enterprise datasets inherently contain multi-collinearity and redundancy across disparate data streams. To prevent information loss while distilling the feature space, a structured sequence of variable clustering and Principal Component Analysis (PCA) was executed independently and conjointly across each major analytical dimension:

- **Customer Demographics:** Encompassing age, gender, ethnicity, marital status, and household size to capture foundational consumer attributes. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Firmographics:** Incorporating business-to-business organizational scales, industry classifications, revenue bands, and enterprise hierarchies where applicable. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Census Data:** Integrating localized geographic, median income, housing density, and regional socioeconomic indicators from official census repositories. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Internal Data:** Aggregating historical customer master data, account tenure, billing status, and profile updates. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Consumption Patterns:** Capturing transactional volumes, average purchase magnitude, spending velocity, and category preferences. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Digital Data:** Quantifying web session frequency, mobile app engagement metrics, clickstream behaviors, and digital channel preferences. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Product Usage:** Analyzing active subscriptions, feature adoption rates, hardware/device types, and utilization intensity. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.
- **Services Info:** Evaluating customer service inquiry frequency, support ticket resolution times, escalation history, and service tier adherence. Variable clustering grouped highly correlated items within each dimension, while PCA extracted orthogonal latent components, ensuring robust variance coverage without redundancy.

4.2 Tetrachoric and Polychoric Correlations

Subsequent to feature reduction across dimensions, remaining variables must be mathematically normalized. A critical flaw in conventional segmentation is the uncritical application of continuous distance metrics (e.g., Euclidean distance) to ordinal survey data and Likert scales with varying directional orientations. Treating ordinal integers as continuous linear variables introduces substantial measurement error.

To rigorously preprocess this data, Tetrachoric correlations (for binary variables) and Polychoric correlations (for multi-category ordinal variables) were deployed. These psychometric methods assume that observed discrete item responses manifest from underlying, unobserved continuous bivariate normal distributions, mapping raw ordinal ratings into robust latent correlation matrices [1].

```
import numpy as np
import statsmodels.api as sm

def tetrachoric(x, y):
    # Calculate the tetrachoric correlation between two binary survey variables
    # to handle ordinal data scales effectively and prevent distance distortion.
    crosstab = np.c_[x, y]
    table = np.zeros((2, 2))
    for i in range(len(crosstab)):
        table[crosstab[i][0], crosstab[i][1]] += 1
```

```
    r = sm.stats.tetrachoric(table)
    return r
```

4.3 Monotonic Feature Engineering via Isotonic Regression and IV/WOE Calculation

To further optimize high-dimensional continuous and categorical attributes prior to variable clustering and downstream model ingestion, a robust and scalable framework integrating Isotonic Regression with concurrent Information Value (IV) and Weight of Evidence (WOE) calculation was implemented [6]. This methodology ensures monotonic relationships and optimal binning, enhancing numerical stability and linear compliance across predictive features while avoiding information loss [6].

5. Observational Clustering Automation on Apache Spark

Upon completing multi-dimensional data reduction and normalization, the core of the agnostic segmentation architecture was executed. Rather than manually tuning a single model, distributed automation scripts were deployed on Apache Spark [3] to stochastically sample feature combinations across dimensions and autonomously construct over 15,000 distinct clustering solutions. Feature combinations incorporated dynamic partitioning splits, allocating 60% of feature weight to survey data and 40% to internal telemetry. Within the survey data partition, feature subsets were further randomized across demographics, behaviors, attitudes, and perceptions.

5.1 Comprehensive Algorithm Testing and Computational Benchmarks

The distributed processing pipeline systematically evaluated diverse observational clustering methodologies. Empirical evidence compiled across the 15,000 iterations revealed pronounced performance divergences across algorithmic classes, as summarized in Table 1.

From an infrastructure and execution standpoint, processing over 15,000 iterations across ~1,500 engineered features and millions of customer records was provisioned on an enterprise Apache Spark cluster comprising 64 worker nodes (each configured with 16 vCPUs and 64 GB RAM). Total end-to-end execution time for feature reduction, model generation, scoring, and profiling averaged 4 hours and 18 minutes, demonstrating exceptional distributed scalability.

Algorithm Type	Avg. Silhouette Score	Convergence Rate	Business KPI Pass Rate	Methodological Conclusion
Hierarchical (Agglomerative)	0.62	Very Slow	4.8%	Computationally heavy; failed size distribution constraints.
DBSCAN (Density-Based)	0.41	Slow	1.2%	Inadequate handling of varying cluster densities; extreme outlier generation.
K-Means (Standard)	0.68	Fast	8.4%	Frequent entrapment in suboptimal local minima; moderate business fit.
K-Means++	0.74	Fast	14.6%	Optimal centroid initialization [2]; superior empirical alignment with KPIs.

As demonstrated in Table 1, K-Means++ consistently outperformed alternative approaches by accelerating mathematical convergence and avoiding poor local minima through optimal probabilistic centroid seeding [2].

```
from sklearn.cluster import KMeans
import random

# Automate iterations for specific feature combination strategies
for i in range(0, 15000):
    clus_vars = random.sample(atti_vars, random.randint(25, 35)) + \
                random.sample(behave_vars, random.randint(10, 20)) + \
                random.sample(subscriberdata_vars, random.randint(15, 20)) + \
                random.sample(demographics_vars, random.randint(5, 7))
    X = data_scaled[clus_vars]
    for kclus in [5, 6, 7]:
        model = KMeans(n_clusters=kclus, init='k-means++', n_init=15, max_iter=300,
```

```
    random_state=seed)
            model.fit(X, sample_weight=weights)
            yhat = model.predict(X)
```

6. The Variance-Bias Tradeoff: Delta Metrics, Internal Validation, and Profiling

While generating 15,000 solutions is computationally tractable within a distributed framework, the primary methodological innovation resides in the automated selection and curation mechanism. A dedicated profiling module was engineered to compute distinct statistical delta metrics and evaluate inter-segment separation across every attribute.

6.1 Custom Delta Metrics and Internal Validation Indices

All generated model iterations were scored to quantify partition distinctiveness across functional categories:

- **Version Score:** A distance-based metric defined as 1 - c_1(Total), weighted proportionally by segment size.
- **Weighted Z-Score:** Standardizes operational performance across enterprise KPIs, scaled by relative segment weight.
- **Normalized Score:** Raw cluster distance weighted by respective segment proportions.
- **Range Score:** Evaluated as the maximum-minimum spread across all clusters, weighted by segment size.

In addition to Silhouette scores, internal cluster validation was fortified by computing the Calinski-Harabasz index (variance ratio criterion) and the Davies-Bouldin index (ratio of within-cluster scatter to between-cluster separation). Models exhibiting subpar density ratios were automatically filtered prior to business profiling.

6.2 Automatic Model Disqualification (Business Thresholds)

An automated gating mechanism was implemented to rigorously filter out models failing to balance statistical variance optimization with operational constraints. Candidate solutions were automatically disqualified unless they satisfied strict governance criteria:

1. **Cluster Size Distribution:** Solutions were rejected if segment sizes fell outside pre-defined thresholds. A minimum segment proportion of >= 0.05 (5%) and a maximum proportion of <= 0.30 (30%) ensured that no single segment dominated the population, nor was any segment overly granular.
2. **Sensitive Feature Bias Audit:** Models were systematically audited for demographic parity on sensitive attributes (e.g., maintaining balanced male-to-female ratios within segments) to guarantee algorithmic fairness.
3. Models satisfying both statistical separation benchmarks and business KPI thresholds were designated as 'Qualified' and ingested into the visualization layer. Table 2 outlines the statistical delta metrics for the final qualified 5-segment architecture.

Segment ID	Size (%)	Weighted Z-Score	Norm. Distance	Bias Audit
Segment 1	22.4%	+1.42	0.81	Passed
Segment 2	18.1%	-0.65	0.44	Passed
Segment 3	28.5%	+0.94	0.67	Passed
Segment 4	14.2%	-1.20	0.31	Passed
Segment 5	16.8%	+0.15	0.52	Passed

7. Supervised Validation, Explainable AI (SHAP), and Temporal Stability

To empirically validate the predictive robustness of the unsupervised clusters, all qualified segmentation solutions were subjected to supervised Machine Learning Algorithm (MLA) classifications utilizing Random Forest and XGBoost classifiers.
By treating cluster assignments as target class labels, this supervised validation step measured the generalization accuracy of the feature space. High classification accuracy confirmed that the generated segments exhibited strong structural separability and generalizability for scoring incoming subscriber cohorts.

7.1 Explainable AI (XAI) via SHAP Integration

To bridge the transparency gap inherent in complex classification models predicting cluster memberships, SHAP (SHapley Additive exPlanations) was integrated into the validation pipeline [4]. By leveraging cooperative game theory principles, SHAP assigns exact feature attribution values to individual customer assignments. This enables enterprise stakeholders to visualize precisely which behavioral or demographic features drive a customer's classification into a specific target segment, ensuring total algorithmic transparency [4].

7.2 Temporal Stability Analysis via Bootstrapping and Jaccard Similarity

To ensure operational robustness and prevent overfitting to static historical snapshots, temporal stability was evaluated using bootstrap resampling and Jaccard similarity coefficients across rolling monthly cohorts. Subsamples of the customer database were iteratively clustered, and cluster membership stability was measured against the baseline taxonomy. Segments demonstrating a

mean Jaccard similarity score >= 0.85 across bootstrap iterations were classified as structurally stable over time, confirming that the derived personas reflect enduring consumer behavior rather than transient noise.

8. Production Governance: Concept Drift Monitoring (PSI) & Dashboard Deployment

Maintaining the long-term integrity of a consumer segmentation model requires continuous governance against macroeconomic and behavioral shifts.

8.1 Population Stability Index (PSI) Governance

To detect concept drift and telemetry degradation post-deployment, the Population Stability Index (PSI) was instituted across all analytical dimensions [5]. By continuously comparing baseline segment population distributions against incoming scoring batches, enterprise data engineering pipelines automatically trigger alerts when PSI exceeds established governance thresholds (PSI > 0.25), signaling that consumer behavior has evolved sufficiently to warrant re-running the 15,000-iteration Spark pipeline [5].

8.2 Deployment via Interactive Streamlit Dashboard

To eliminate the bottlenecks associated with manual data science reporting and streamline ongoing metric governance, an interactive analytics application was engineered utilizing a Flask API backend paired with a Streamlit frontend.

The Streamlit dashboard streams the final qualified clustering solutions, automatically rendering multi-dimensional customer assessments. It visually contrasts customer attitudes and operational behaviors across usage intensity, service engagement, brand perception, touchpoint interactions, economic value, profitability, and churn propensity. This interface empowers executive leadership to review top-performing model iterations and seamlessly select an optimal 5-segment architecture that reconciles statistical variance minimization with strategic business execution.

9. Financial ROI Impact Modeling and Business Case Evaluation

To translate statistical separation and Customer Lifetime Value (CLV) tiers into an actionable executive business case, a financial attribution model was constructed. By mapping marketing intervention costs against expected conversion lifts and churn reduction across the finalized 5-segment architecture, enterprise return on investment (ROI) was quantified.

Specifically, targeted retention campaigns deployed on Segment 2 (Price-Sensitive Detractors, characterized by high churn risk) yielded a 28% reduction in voluntary churn, preserving approximately $14.2M in annual recurring revenue. Simultaneously, upsell campaigns directed toward Segment 1 (Premium Brand Advocates) drove a 34% increase in high-margin product adoption, validating that statistical precision directly translates into tangible enterprise profitability.

10. Feature Weight Sensitivity Analysis

To verify that the chosen segmentation model is not an artifact of arbitrary feature weighting, a rigorous sensitivity analysis was conducted around the baseline feature split (60% survey data / 40% internal telemetry). Weight allocations were systematically varied across intervals ranging from 20% to 80% survey weighting.

The evaluation demonstrated that overall cluster stability and business KPI pass rates remained highly resilient within a +/- 15% band around the baseline split. Beyond this threshold, excessive reliance on internal telemetry degraded attitudinal differentiation, whereas over-weighting survey data reduced transactional predictive accuracy. This empirical sensitivity profiling confirms the structural robustness of the selected feature weighting strategy.

11. Results and Conclusion

From the 15,000 automated distributed iterations, the selected production solution was a highly optimized 5-segment taxonomy that successfully satisfied all empirical and strategic criteria. As delineated in Table 3, the finalized segmentation achieved robust operational separation across critical business parameters while avoiding extreme, unusable partitions. Furthermore, it maintained strict demographic parity on sensitive features and exhibited exceptional classification accuracy for prospective population scoring.

Target Segment	Customer Lifetime Value	Consumption Index	Churn Risk	Core Business Persona
Segment 1	Very High	High	Low	Premium Brand Advocates
Segment 2	Low	Low	High	Price-Sensitive Detractors
Segment 3	Medium-High	High	Medium	Digital-First Pragmatists
Segment 4	Low	Medium	Very High	Disengaged Legacy Users
Segment 5	Medium	Low	Low	Stable Traditionalists

Unlike traditional market segmentation efforts—which frequently culminate in human-biased, time-constrained models driven by subjective statistical tuning—this automated, distributed methodology establishes a genuinely agnostic framework. By orchestrating a disciplined pipeline comprising multi-dimensional variable clustering, PCA across enterprise telemetry, Tetrachoric/Polychoric correlations, distributed Spark processing, internal validation indices, SHAP explainable attribution, PSI drift governance, financial

ROI modeling, and bootstrap temporal stability checks coupled with sample reweighting and automated threshold gating, modern enterprises can definitively bridge the chasm between statistical variance optimization and strategic business KPI alignment.

REFERENCES

[1] SAS Institute Inc., "Calculating Tetrachoric and Polychoric Correlations in Enterprise Analytical Environments," SAS Support Knowledge Base, KB25010, 2025. [Online]. Available: https://support.sas.com/kb/25/010.html
[2] D. Arthur and S. Vassilvitskii, "k-means++: The Advantages of Careful Seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07), pp. 1027–1035, 2007.
[3] M. Zaharia et al., "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," in Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI '12), pp. 15–28, 2012.
[4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS 2017), vol. 30, pp. 4765–4774, 2017.
[5] N. Siddiqi, Credit Risk Scorecards: Implementing and Integrating Intelligent Credit Scoring, Hoboken, NJ: John Wiley & Sons, 2006.
[6] A. R. Minukuri, "A Robust and Scalable Framework for Monotonic Feature Engineering: Integrating