TransVision

Pullabhotla Vijay[1], Koti Vennela Khushi[2], Gaddam Advitha[3], Allada Nagasai Varaprasad[4], Dr. Yamarthi Narsimha Roa

[1, 2, 3, 4]School of Computer Scixence and Engineering, VIT – AP University, Inavolu, Beside AP Secretariat, Amaravati AP, India

[5]Department of Artificial intelligence and Machine Learning, School of computer science and engineering, VIT – AP University, Inavolu, Beside AP Secretariat, Amaravati AP, India

Abstract*—In this ambitious financial landscape, customer satisfaction is the sharp-edged dagger that cuts through the intense competition that credit card companies face whose sole aim is to retain and expand their user base. This paper – "TransVision" presents an information-centric substructure that leverages data analytics[2] and machine learning techniques to optimize end to end transaction journey for credit card users of a hypothetical financial organization. TransVision provides insights into user behaviours by analyzing transaction data across key sectors, enabling targeted customer segmentation based on socio-economic attributes. Through robust fraud detection algorithms, it safeguards customer interactions by identifying anomalous patterns, thereby reducing transaction risks. Additionally, sentiment analysis[7] of customer feedback provides actionable insights into user satisfaction, enabling continuous improvement of the customer experience. Our results demonstrate that TransVision not only enhances transactional security but also improves the company's ability to proactively address customer needs, fostering both customer retention and acquisition.*

Keywords*—Transaction Journey Optimization, Customer Behavior Analysis, Customer Segmentation, Fraud Detection, Sentiment Analysis, Machine Learning, Data Analytics*

I. Introduction

Credit card companies operate in an intensely competitive financial environment where retaining and expanding a usership hinges on delivering exceptional customer experiences while setting the seal on transactional security. However, these companies face significant challenges that threaten their competency to flourish in this landscape. Understanding diverse user behaviours is a critical yet complex task, as customer preferences and spending habits vary widely across demographic, socio-economic and cultural lines. Without deep insights into these demeanours, companies risk delivering generic experiences that fail to resonate with their audience. Mitigating transaction fraud is another pressing concern. The rise of avant-garde fraud schemes, especially transaction tampering not only jeopardizes financial security but also hinders user confidence. Even a single breach can lead to lasting damage to a company's reputation and lead to significant financial losses and the company will part ways with the loyalty of the customers. Addressing customer dissatisfaction is equally challenging in an era where consumers demand fast, secure, and personalized services. Negative experiences- such as— delayed transactions, unresponsive support, or unresolved complaints— can propel users to competitors, Moreover, dissatisfaction often stems from a lack of proactive engagement, as companies struggle to anticipate customer needs and resolve issues before they escalate. Together, these challenges create a high-stakes environment where any environment where any failure to meet user conjectures or ensure undertaking security can erode trust and loyalty, ultimately impacting a company's ability to retain and grow its customer base.

This paper introduces *TransVision*, a comprehensive framework that leverages technological advancements to optimize the end-to-end transaction tourney for credit card users. This substructure employs advanced machine learning models that analyse transaction patterns to identify anomalies indicative of fraudulent behaviour. By using these algorithms, the framework not only detects potential fraud in real-time but also minimizes false positives, ensuring legitimate transactions are not disrupted. It incorporates segmentation models based on socio-economic attributes, allowing credit card companies to group users by income level, spending habits, geographic location, and other relevant factors. This targeted segmentation enables organizations to design customized marketing campaigns, recommend tailored product offerings, and engage with users in a manner that meets their specific needs. Such precision fosters deeper customer relations and enhances loyalty. *TransVision* leverages sentiment analysis techniques to process and interpret customer feedback at scale. By evaluating sentiment scores and identifying patterns of dissatisfaction, the framework provides actionable insights into areas for improvement. This allows credit card companies to prioritize high-impact changes, resolve complaints more effectively, and enhance the overall customer experience.

I. Methodology

I.A. Dataset

The dataset used in this study was synthetically generated using the **Faker** library, designed to simulate realistic transaction data while adhering to real-world financial constraints. The dataset mirrors typical transaction environments faced by credit card companies, ensuring relevance and applicability in a competitive financial landscape. The **Faker** tool was utilized to generate data fields that reflect actual attributes encountered in credit card transaction datasets. These fields include:

Transaction_ID: Randomly generated unique identifiers for each transaction.
Customer_ID: Unique customer identifiers ensuring no overlap between individual users.

Transaction_DateTime: Randomized timestamps distributed to mimic daily and seasonal transaction trends.
Transaction_Type: Categories such as purchases, refunds, or cash advances.
Category: Items or services purchased, modeled on real-world consumer behavior patterns.
Merchant_Location: Geographic data representing diverse merchant locations.
Payment_Type: Payment methods, including credit cards, debit cards, and digital wallets.
Age and **Gender**: Demographic attributes that simulate user diversity.
Customer_Loyalty_Score: Scores calculated to emulate loyalty metrics based on synthetic purchase behaviors.
Review_Text and **Review_Rating**: Simulated customer feedback, incorporating a range of sentiments and ratings to represent realistic user experiences.
Amount: Transaction values following a distribution mimicking typical spending patterns.
Old_Balance and **New_Balance**: Account balances reflecting financial activity.
IsFraud: Flags for fraudulent transactions, generated to represent anomaly patterns.

The dataset was carefully crafted to replicate:

1. **Real-Time Constraints**:
   - Temporal distributions of transactions were modeled to reflect busy periods, such as weekends or holidays.
   - Spending categories were aligned with sector-specific trends, ensuring relevance.
1. **Demographic Variability**:
   - Age, gender, and loyalty scores were distributed to match expected customer demographics in credit card user bases.
1. **Fraudulent Patterns**:
   - Fraudulent transactions were seeded using behavioral irregularities, such as unusually high amounts, improbable locations, or inconsistent transaction frequencies.
1. **Transaction Trends**:
   Time-series patterns were embedded, ensuring the dataset could support analyses like seasonal trends, customer lifetime value prediction, and churn modeling.

I.

I.A. Behavioral Analysis

Credit card companies grapple with the intricacies of decoding customer behaviors to deliver services that resonate deeply with their user base. Understanding transaction patterns and identifying spending trends is a daunting yet necessary task. Without robust data analytics, deriving meaningful insights from the vast volumes of transaction data becomes nearly impossible. Data analytics plays a pivotal role in unraveling these complexities by processing and interpreting user activities, highlighting key trends, and identifying actionable opportunities. This enables organizations to predict customer needs and preferences accurately, enhancing engagement and fostering long-term loyalty.

1. CLTV Prediction Formula:

$$CLTV = (\text{Average Transaction Value} \times \text{Purchase Frequency}) \times \text{Customer Lifespan}$$

where:

- Average Transaction Value = $\frac{\text{Total Revenue}}{\text{Total Number of Transactions}}$
- Purchase Frequency = $\frac{\text{Total Number of Transactions}}{\text{Unique Customers}}$

Equations :

Outliers were flagged using Z-scores:

$$Z = \frac{(X - \mu)}{\sigma}$$

Transaction Pattern Analysis: Patterns were identified by analyzing variables such as Transaction_Type, Category, and Merchant_Location, providing insights into user spending behavior.

Cohort Analysis: Users were grouped by their Signup_Date to examine retention trends over time.

Customer Lifetime Value (CLTV) Prediction: CLTV was estimated using user-specific metrics, such as Transaction_ID, Amount, and Old_Balance.

Time-Series Analysis of Transactions: Temporal trends in Transaction_DateTime were analyzed to identify seasonal variations and predict future transaction volumes.

Distributions and Skewness of Numerical Variables: Variables such as Amount, Old_Balance, and New_Balance were checked for normality and skewness to guide preprocessing steps.

Outlier Detection: Statistical methods and visualization tools were used to identify and handle outliers in the dataset.

Bivariate Analysis: Relationships between pairs of variables, such as Age vs. Amount, were explored to uncover significant correlations and trends.

I.

I.A. Customer Segmentation

The heterogeneous nature of credit card users necessitates a precise approach to segmentation. Grouping customers based on spending patterns, socio-economic attributes, and geographic data is essential for targeted marketing and personalized engagement. However, achieving this granularity is an arduous task. Clustering methods alleviate this challenge by categorizing users into distinct, data-driven segments. This empowers companies to tailor their offerings and campaigns effectively, ensuring each customer feels understood and valued, thereby bolstering retention and driving growth.

To enable precise segmentation of customers based on their transaction behaviors, socio-economic attributes, and spending patterns, the following clustering techniques were employed. Each method provides unique advantages in understanding data distributions and uncovering latent group structures, facilitating targeted strategies for customer engagement:

1. **Gaussian Mixture Model (GMM) Clustering**

GMM[5] clustering was used to model customer data as a mixture of multiple Gaussian distributions. This probabilistic approach helps in assigning customers to clusters based on the likelihood of their data points belonging to specific distributions, offering a nuanced understanding of overlapping customer groups.

$$P(x_i) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k)$$

Where:

- π_k: Mixing coefficient (prior probability of cluster k), $\sum_{k=1}^{K} \pi_k = 1$
- $\mathcal{N}(x_i|\mu_k, \Sigma_k)$: Gaussian distribution with mean μ_k and covariance matrix Σ_k:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

1. Hierarchical Clustering

Hierarchical clustering was utilized to create a tree-like structure (dendrogram) representing nested groupings of customers. This method is particularly effective in identifying hierarchical relationships among customer segments and determining optimal cluster numbers based on domain-specific thresholds.

- Single linkage:

$$d(A, B) = \min_{a \in A, b \in B} ||a - b||$$

- Complete linkage:

$$d(A, B) = \max_{a \in A, b \in B} ||a - b||$$

- Average linkage:

$$d(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} ||a - b||$$

1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN[3] was applied to identify clusters of customers with dense transaction patterns while labeling sparsely populated areas as noise. This method is robust in handling outliers and non-linear cluster shapes, making it suitable for detecting atypical customer behavior.

For a feature x, normalized value x_{norm} is given by:

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where:

- $\min(x)$ = minimum value of feature x
- $\max(x)$ = maximum value of feature x

$$N_\epsilon(p) = \{q \in D : ||p - q|| \leq \epsilon\}, \quad |N_\epsilon(p)| \geq MinPts$$

Where:

- $||p - q||$: Distance between points p and q.
- $N_\epsilon(p)$: Neighborhood of point p.

1. K-Means Clustering

K-Means clustering, a widely adopted partition-based technique, grouped customers into distinct clusters based on their proximity in the feature space. It provided a straightforward yet powerful method for segmenting users into well-defined categories for marketing and personalization efforts.

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2$$

Where:

- C_k: Set of points assigned to cluster k.
- $\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$: Centroid of cluster k.

1. Mean Shift Clustering

Mean Shift[9] clustering was employed to identify clusters by iteratively shifting data points toward the densest regions in the feature space. Its ability to adapt to the number of clusters dynamically makes it valuable in uncovering natural customer groupings without predefined assumptions.

$$m(x) = \frac{\sum_{x_j \in N_\epsilon(x)} K(||x_j - x||^2) x_j}{\sum_{x_j \in N_\epsilon(x)} K(||x_j - x||^2)} - x$$

Where:

- $K(||x_j - x||^2)$: Kernel function, typically a Gaussian kernel.
- $N_\epsilon(x)$: Neighborhood of x within a radius ϵ.

I.

I.A. Fraud Detection

Fraudulent activities pose a persistent threat to the integrity of financial transactions, eroding customer trust and exposing organizations to substantial financial risks. Identifying such activities amidst high transaction volumes is akin to finding a needle in a haystack. Fraud detection methods streamline this process by analyzing transaction data to detect suspicious patterns, enabling organizations to respond in real time. These techniques enhance security measures, reduce the likelihood of breaches, and restore user confidence in the safety of their transactions.

1. Data Preprocessing

The dataset contains a diverse range of features, including transaction details, customer profiles, merchant information, and review scores. To streamline the analysis, irrelevant columns such as Transaction_ID, Customer_ID, and Merchant_ID were excluded, focusing only on features that directly influence fraud prediction.

2. Feature Scaling (Normalization)

Normalization ensures that numeric features like Amount, Old_Balance, and New_Balance are scaled to a range [0, 1], preventing features with larger magnitudes from dominating the model. Continuous features such as Amount, Old_Balance, and New_Balance were normalized using MinMaxScaler. This scaling ensures that all numeric features have a uniform range, preventing dominance by features with larger magnitudes during the learning process.

3. Handling Class Imbalance with SMOTE (Synthetic Minority Oversampling Technique)

Synthetic Minority Oversampling Technique (SMOTE) generates synthetic samples for the minority class (fraudulent transactions) by interpolating between existing minority class samples. Fraudulent transactions constituted only a small proportion of the dataset (0.12%), creating a significant class imbalance. Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for the minority class, effectively balancing the dataset and reducing the model's bias towards non-fraudulent transactions.

Given a minority class sample x and one of its k-nearest neighbors x_{nn}, a synthetic sample x_{new} is generated as:

$$x_{\text{new}} = x + \delta \cdot (x_{\text{nn}} - x)$$

Where:

- δ = random value in the range [0, 1]

1. One-Hot Encoding

One-hot encoding converts categorical variables (e.g., Transaction_Type, Category) into binary vectors, making them suitable for numerical computation.

Mathematical Representation:

For a categorical variable X with n unique categories, one-hot encoding transforms X into n binary variables $\{x_1, x_2, \ldots, x_n\}$:

$$x_i = \begin{cases} 1 & \text{if category } i \text{ is present} \\ 0 & \text{otherwise} \end{cases}$$

1. Model Selection and Training

1.A.

1.A.1)

1.A.1.a) Logistic Regression

Logistic regression models the probability of a transaction being fraudulent by applying a logistic (sigmoid) function to a weighted sum of the input features. It is widely used as a baseline model due to its simplicity and interpretability, particularly when features have a linear relationship with the target variable. However, it may not perform well on complex, non-linear datasets.

Mathematical Formula:

$$P(y = 1 \mid x) = \sigma(w^T x + b)$$

Where:

- $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function
- w = weight vector
- x = feature vector
- b = bias term

I.

I.A.

I.A.1)

I.A.1.a) Random Forest Classifier

Random Forest[6] is an ensemble learning method that builds multiple decision trees during training and combines their predictions to improve accuracy and generalizability. It excels in handling high-dimensional data and capturing non-linear patterns, making it highly effective for fraud detection. The randomness in feature selection reduces overfitting compared to single decision trees.

$$Prediction = \frac{1}{n}\sum_{i=1}^{n} Tree_i(x)$$

Where $Tree_i(x)$ represents the prediction of the i-th tree, and n is the total number of trees.

I.

I.A.

I.A.1)

I.A.1.a) Decision Tree Classifier

Decision Trees partition the dataset into subsets based on feature values, creating a tree structure that predicts outcomes by following the path from root to leaf. They are intuitive and easy to visualize but prone to overfitting on smaller datasets, which can be mitigated through pruning techniques.

$$Gini\_Index = 1 - \sum_{i=1}^{C}(p_i)^2$$

Where p_i is the proportion of samples belonging to class i, and C is the total number of classes.

I.

I.A.

I.A.1)

I.A.1.a) XGBoost

Extreme Gradient Boosting (XGBoost)[6] is an advanced ensemble method that iteratively builds decision trees by optimizing a differentiable loss function. It incorporates techniques like regularization and shrinkage to enhance performance and prevent overfitting, making it ideal for large and complex datasets in fraud detection.

$$L = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

I.

I.A.

I.A.1)

I.A.1.a) K-Nearest Neighbors (KNN)

KNN classifies a transaction by comparing it to its kkk-nearest neighbors in the feature space and assigning the majority class label. While effective for smaller datasets, it can be computationally expensive for large datasets, making it less feasible for real-time fraud detection.

$$d(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x_i')^2}$$

I.

I.A.

I.A.1)

I.A.1.a) Gaussian Naive Bayes

Gaussian Naive Bayes[8] assumes that features follow a Gaussian distribution and calculates the probability of a transaction being fraudulent using Bayes' theorem. Despite its simplicity, it may underperform in datasets with dependent or non-Gaussian features.

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

I.

I.A. Sentiment Analysis

Customer feedback, while abundant, often remains underutilized due to the challenge of interpreting sentiments at scale. Ignoring this wealth of information results in missed opportunities to address dissatisfaction and refine services. Sentiment analysis bridges this gap by evaluating customer sentiments expressed in reviews, complaints, and surveys. This enables organizations to uncover areas of concern, improve service quality proactively, and create an experience that resonates positively with users, fostering satisfaction and loyalty.

1. **Data Collection and Preprocessing**

Input Data: The dataset consists of customer reviews and feedback data, which includes Review_Text and Review_Rating columns. The Review_Text contains textual feedback from users, while Review_Rating indicates the numerical rating given by users.

Text Cleaning: The text data was preprocessed by removing unnecessary characters such as punctuation marks, numbers, and special symbols. Tokenization was performed to break the review text into individual words (tokens), and stopwords (commonly used words such as "the", "is", etc.) were eliminated to focus on meaningful words.

Text Normalization: Lowercasing all text and stemming or lemmatization were applied to reduce words to their root forms (e.g., "running" becomes "run"), making it easier for VADER to analyse sentiment effectively.

1. **Applying VADER for Sentiment Analysis**

VADER is designed to compute sentiment scores based on the predefined lexicon of words and their associated valence (sentiment strength). The sentiment scores for each review text are calculated based on positive, negative, neutral, and compound scores.

Positive Score: Represents the percentage of positive sentiment words in the text.

Negative Score: Represents the percentage of negative sentiment words.

Neutral Score: Measures the portion of text that is neutral in sentiment.

Compound Score: A normalized score that combines the positive, neutral, and negative scores, resulting in an overall sentiment score that ranges from -1 (extremely negative) to +1 (extremely positive). A compound score closes to 0 indicates neutral sentiment.

$$\text{Compound Score} = \frac{\sum(\text{valence of each word})}{\sqrt{(\text{sum of squared valence of each word})}}$$

Where:

- The valence of each word is derived from a predefined lexicon of words and their respective sentiment values.
- The compound score ranges from -1 (extremely negative) to +1 (extremely positive), with 0 indicating a neutral sentiment.

1. **Sentiment Classification**

Based on the compound score:

Positive: If the compound score is greater than 0.05 (indicating a predominantly positive sentiment).

Negative: If the compound score is less than -0.05 (indicating a predominantly negative sentiment).

Neutral: If the compound score is between -0.05 and 0.05 (indicating a neutral sentiment).

1. **Sentiment-Based Insights**

After computing the sentiment scores, the results were visualized using various charts and plots to better understand the sentiment distribution across reviews.

I. RESULTS

I.A. Behaviourial Analysis

1. Churn Prediction Analysis: The Random Forest Classifier was employed to predict customer churn based on the features of Recency, Frequency, Monetary contributions, and Customer Loyalty Score. Churn was defined as customers whose recency exceeded a specified threshold. The model demonstrated exceptional performance with the following results:

Accuracy: 1.0

Precision, Recall, F1-Score: 1.0 for both churned and active customers

These results provide a robust predictive model for churn, allowing the organization to take proactive measures to retain at-risk customers, ultimately reducing churn rates and improving customer lifetime value.

1. Behavioural Analysis Based on Spending

Customers were further segmented based on their spending levels, using quantiles to categorize them into Low, Medium, High, and Very High spending brackets. Key findings include:

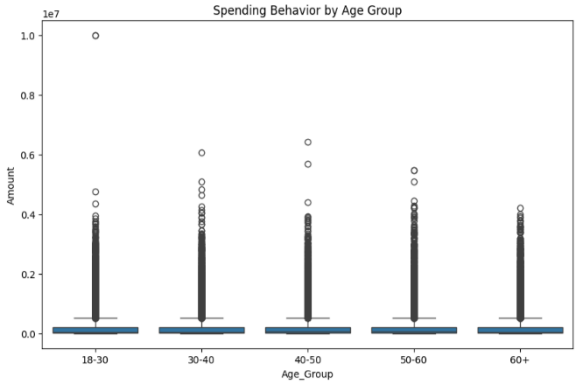
Low Spend: Customers with infrequent transactions and low household income.

Medium Spend: Average spenders with moderate transaction frequency.

High Spend: Consistent spenders with higher monetary contributions and income.

Very High Spend: Premium customers who make the highest contributions and have substantial income levels.

This segmentation aids in resource allocation, enabling tailored loyalty programs and VIP management strategies to engage high-value customers.



Peak Transaction Hours: Transactions show a uniform pattern with slight peaks around midday and evening. This information is useful for optimizing staffing and inventory management during peak hours.

Category Preferences: "Travel" is the most popular category, followed by "Food" and "Shopping." These insights can help tailor marketing efforts and prioritize product offerings.

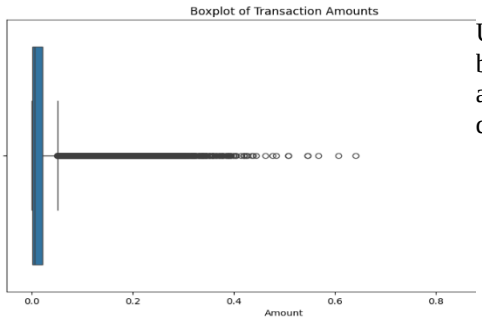
Average Income by Loyalty Level: A slight decrease in income is observed as loyalty levels increase, suggesting that loyalty programs attract a broader range of customers, not just high-income ones.

Category Preference by Loyalty Level: Customer preferences remain consistent across loyalty levels, indicating that loyalty programs may not significantly impact category choices.

Spending Behavior per Customer: Significant variation exists in spending behavior, with high-value customers showing much higher spending. Identifying these customers helps with targeted marketing and retention efforts.

Cohort Analysis: Fluctuations in spending over time point to seasonal trends or marketing impacts, which can guide future marketing strategies and product launches.

Customer Lifetime Value (CLTV) Prediction: CLTV insights help prioritize high-value customers for retention efforts, ensuring more efficient use of resources for personalized marketing.



I.

I.A. Customer Segmentation

Using KMeans[3] Clustering, customers were segmented based on Recency, Frequency, Monetary contributions, Age, and Household Income. The segmentation revealed four distinct customer groups:

Segment 0: Younger customers with moderate spending and average transaction frequency.

Segment 1: Older customers with high transaction frequency and significant monetary contributions.

Segment 2: Mid-age customers with consistent spending and high transaction amounts.

Segment 3: High-income younger customers with substantial spending but low transaction frequency.

These segments offer opportunities for targeted marketing and personalized strategies to enhance customer engagement and loyalty.

The Silhouette Score, which quantifies the quality of clustering by measuring the separation and cohesion of clusters, was computed for each clustering algorithm. The results are as follows:

KMeans Clustering: Achieved a silhouette score of 0.6710, indicating that the clusters are well-separated and compact. This makes KMeans suitable for customer segmentation, where clear and distinct groupings are beneficial.

Hierarchial Clustering: Scored 0.6710, showing similar performance to KMeans, indicating that the clusters are well-separated and compact, which is suitable for effective customer segmentation.

Gaussian Mixture Model (GMM): Received a lower silhouette score of 0.2575. This score suggests that the algorithm struggles with overlapping clusters, leading to confusion and poor separation. Hence, GMM was not effective for this dataset.

DBSCAN (Density-Based Spatial Clustering): Achieved a perfect silhouette score of 1.0000, excluding noise points. This indicates that DBSCAN is particularly effective in identifying dense clusters and accurately separating outliers. This makes DBSCAN a strong candidate for clustering datasets with noise or irregular patterns.

MeanShift Clustering: Produced a maximum valid silhouette score of 0.6816, demonstrating its effectiveness in identifying density-based clusters without requiring a predefined number of clusters. This makes it particularly useful for capturing subtle patterns in the data.

These results validate the performance of the clustering algorithms for this dataset. MeanShift stands out as the most appropriate algorithm, effectively capturing customer behavior patterns without the need for predefined cluster numbers. KMeans and Agglomerative Clustering also performed well, providing interpretable and robust results, making them reliable options for customer segmentation. DBSCAN is well-suited for outlier detection, while GMM performed poorly due to the overlapping nature of the customer groups, making it less effective for this particular dataset.

I.

I.A. Fraud Detection

a) Logistic Regression

Accuracy: Logistic Regression performed well with an accuracy of 98.10% on the imbalanced dataset. However, after applying SMOTE to balance the dataset, the accuracy improved slightly to 98.93%.

True Positives (TP): In the imbalanced case, the model correctly identified 54,360 fraudulent transactions, which is a good number. After balancing, the TP dropped to 13,471, but this was expected since SMOTE adds synthetic instances, which may impact the accuracy of fraud detection.

True Negatives (TN): The model also correctly identified 64,647 non-fraudulent transactions in the imbalanced case, which decreased to 16,154 after SMOTE.

False Positives (FP) & False Negatives (FN): Othe imbalanced dataset, the model made 15,227 false positive predictions and 25,446 false negative predictions. These numbers decreased significantly after applying SMOTE, with FPs reducing to 3,772 and FNs to 6,523, showing an improvement in fraud detection accuracy with the balanced dataset.

Logistic Regression performed well both on the imbalanced and SMOTE-balanced datasets. The SMOTE balancing increased the accuracy and decreased both false positives and false negatives, highlighting the importance of balancing the dataset for improving model performance. However, the drop in TP after SMOTE might suggest a trade-off between precision and recall in this scenario.

a) Random Forest Classifier

Accuracy: Random Forest performed exceptionally well on both the training and testing datasets. The model achieved perfect accuracy (100%) on the training dataset and 99.41% on the testing dataset, suggesting it generalizes well without overfitting.

True Positives (TP): The model detected 79,806 fraud cases in the training dataset, and 19,958 in the testing dataset. This shows that the Random Forest model is very effective at identifying fraudulent transactions.

True Negatives (TN): Similarly, 79,874 non-fraudulent transactions were correctly identified in training, and 19,726 were identified in testing, reflecting a strong ability to detect non-fraudulent cases.

False Positives (FP) & False Negatives (FN): The false positive rate was very low with only 200 FPs in the testing dataset, indicating a low rate of incorrectly classifying non-fraudulent transactions as fraud. Furthermore, the model only had 36 false negatives, meaning it rarely missed fraudulent transactions.

The Random Forest model performed very well, showing high accuracy and low error rates, both for fraudulent and non-fraudulent transactions. It demonstrated a robust ability to handle both the training and testing datasets without overfitting. The model's high performance can be attributed to its ensemble nature, which aggregates predictions from multiple decision trees.

a) Decision Tree Classifier

- **Accuracy**: Decision Tree performed slightly worse than Logistic Regression and Random Forest on the imbalanced dataset, with an accuracy of 97.52%. After applying SMOTE, the accuracy improved to 98.75%, showing that balancing the dataset had a positive effect on performance.

- **True Positives (TP)**: The model correctly identified 48,965 fraudulent transactions on the imbalanced dataset. After SMOTE, this dropped to 15,924, which again reflects the impact of synthetic data generation in balancing the dataset.

- **True Negatives (TN)**: The Decision Tree identified 70,123 non-fraudulent transactions on the imbalanced dataset, and 15,833 after SMOTE. The

decrease in TN is consistent with the trend of SMOTE impacting the classification of both classes.

 False Positives (FP) & False Negatives (FN): The false positive rate on the imbalanced dataset was 13,751, and after SMOTE, it reduced to 4,789. Similarly, false negatives reduced from 33,174 to 7,378, reflecting the impact of SMOTE in reducing errors.

The Decision Tree model's performance improved when the dataset was balanced using SMOTE, with the accuracy increasing and the number of false positives and false negatives decreasing. However, it still lags behind Random Forest in terms of overall accuracy and error reduction, possibly due to overfitting or a lack of ensemble techniques.

a) K Nearest Neighbours (KNN)

 Accuracy: KNN[8] achieved an accuracy of 98.14% on the imbalanced dataset, and improved to 98.88% after applying SMOTE, demonstrating a similar improvement as Logistic Regression and Decision Trees when balancing the dataset.

 True Positives (TP): The model identified 55,152 fraudulent transactions correctly on the imbalanced dataset. With SMOTE, this dropped to 14,822, reflecting the addition of synthetic data.

 True Negatives (TN): The number of correctly identified non-fraudulent transactions was 64,744 for the imbalanced dataset, and 16,025 after balancing.

 False Positives (FP) & False Negatives (FN): KNN had 15,039 false positives on the imbalanced dataset, which decreased to 4,114 after SMOTE. False negatives decreased from 25,609 to 7,496 after balancing, indicating better fraud detection after applying SMOTE.

KNN showed a good improvement in performance after balancing the dataset, similar to other models like Logistic Regression and Decision Tree. The model showed strong accuracy and a significant reduction in errors after SMOTE balancing, although it still has relatively high false positives.

a) **Gaussian Naïve Bayes Classifier**

Accuracy: Gaussian Naive Bayes had an accuracy of 97.06% on the imbalanced dataset, and improved to 98.25% after applying SMOTE, showing that balancing the dataset positively impacted the model's performance.

True Positives (TP): The model correctly identified 51,744 fraudulent transactions in the imbalanced case, dropping to 13,589 after SMOTE. The decrease in TP after balancing is expected due to the synthetic data introduced.

True Negatives (TN): Naive Bayes correctly identified 68,214 non-fraudulent transactions in the imbalanced case, and 16,000 after SMOTE.

False Positives (FP) & False Negatives (FN): The false positives decreased from 16,308 to 4,927 after SMOTE, while false negatives dropped from 34,030 to 7,845, indicating that balancing the dataset improved both the precision and recall of the model.

a) XG Boost Classifier

 Accuracy: XGBoost performed very well with an accuracy of 98.91% on the imbalanced dataset, and achieved 99.27% accuracy after applying SMOTE, the highest among all models after balancing.

 True Positives (TP): XGBoost identified 57,463 fraudulent transactions correctly on the imbalanced dataset. After SMOTE, this dropped to 15,798, but this is still a strong performance.

 True Negatives (TN): The model correctly identified 64,299 non-fraudulent transactions on the imbalanced dataset, and 16,161 after balancing.

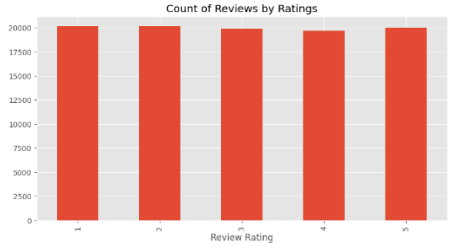
 False Positives (FP) & False Negatives (FN): The false positive rate was 12,351 on the imbalanced dataset, which decreased to 3,607 after SMOTE. False negatives decreased from 26,102 to 5,421, demonstrating a strong improvement after balancing.

XGBoost demonstrated the best overall performance in terms of both accuracy and error reduction, especially after balancing the dataset with SMOTE. Its higher accuracy compared to other models suggests that XGBoost is highly effective for this fraud detection problem, even when dealing with class imbalance.

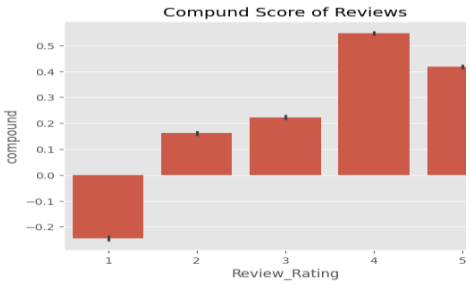
In conclusion, the AI-based fraud detection system demonstrated that **XGBoost** outperformed other models with the highest accuracy, achieving 99.27% after applying SMOTE for balancing the dataset. While **Logistic Regression**, **Random Forest**, and **KNN** also showed significant improvements after balancing, **XGBoost** consistently provided the best results in terms of both overall accuracy and error reduction, making it the ideal choice for fraud detection tasks. The application of **SMOTE** across all models notably reduced false positives and false negatives, emphasizing the importance of addressing class imbalance in the dataset. Ultimately, XGBoost's superior performance establishes it as the most effective model, although other models like **Random Forest** and **Decision Trees** can be considered for specific cases depending on computational constraints or interpretability requirements.

I. *I.A. Sentiment Analysis*

Count of Reviews by Ratings: The bar chart displays the distribution of review counts across ratings from 1 to 5. The distribution is nearly uniform, indicating that customers have given reviews across all rating levels. This provides a balanced view of customer sentiments associated with different ratings and supports comprehensive analysis.

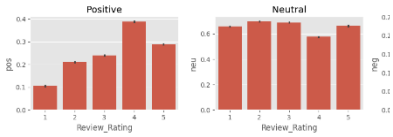


Compound Score of Reviews: This bar chart illustrates the average compound sentiment score for each review rating. A clear positive correlation is observed, where higher ratings (4-5) tend to have stronger positive compound sentiment scores. In contrast, lower ratings (1-2) are associated with more negative compound sentiment scores, which align with customer dissatisfaction.



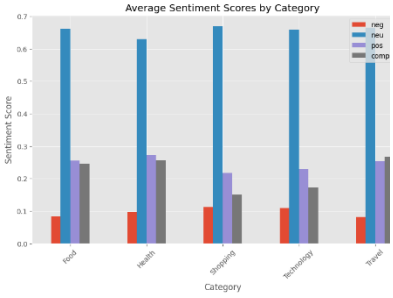
Positive, Neutral, and Negative Sentiment Scores by Review Ratings: Bar charts depicting sentiment breakdown for each review rating show the following trends:

a. Positive Sentiment: Increases as ratings rise, indicating that higher ratings correlate with more positive feedback.
b. Neutral Sentiment: Remains fairly constant across all ratings, showing that many reviews are neutral, regardless of rating.
c. Negative Sentiment: Higher for lower ratings, particularly ratings of 1 and 2, highlighting dissatisfaction in those reviews.

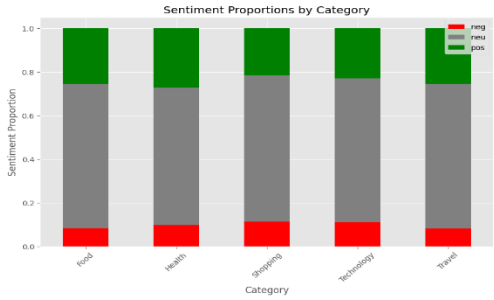


Average Sentiment Scores by Category: This grouped bar chart represents average sentiment scores (negative, neutral, positive, and compound) for different transaction categories:

a. Travel and Health categories show relatively higher positive and compound scores, reflecting more favorable sentiment.
b. Shopping has the lowest sentiment scores across all categories, with more negative feedback compared to others.

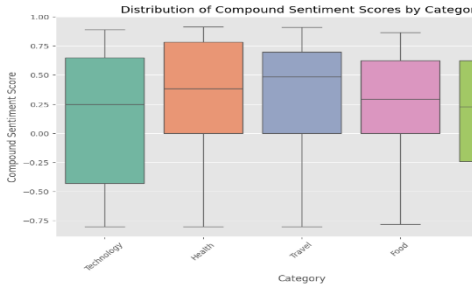


Stacked Bar Chart for Sentiment Proportions by Category: The stacked bar chart visualizes the relative proportions of negative, neutral, and positive sentiment scores across different categories. Neutral sentiment dominates in all categories, but Shopping and Technology exhibit higher negative sentiments, while Travel and Health show relatively more positive sentiments.



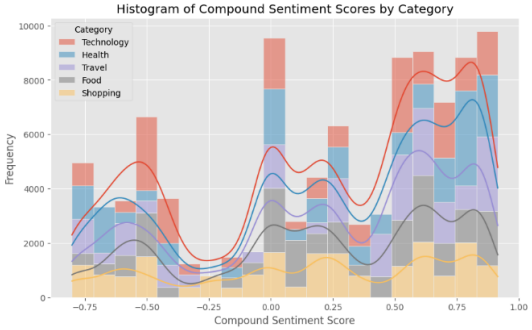
Boxplot of Compound Sentiment Scores by Category: The boxplot reveals the distribution of compound sentiment scores for each category:

a. Health and Travel exhibit higher median scores, indicating stronger positive sentiment.
b. Technology shows a broader range of scores, suggesting more varied feedback from customers.

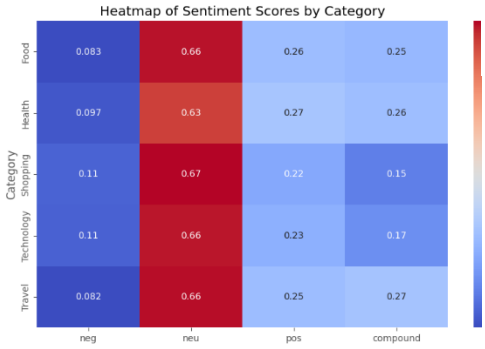


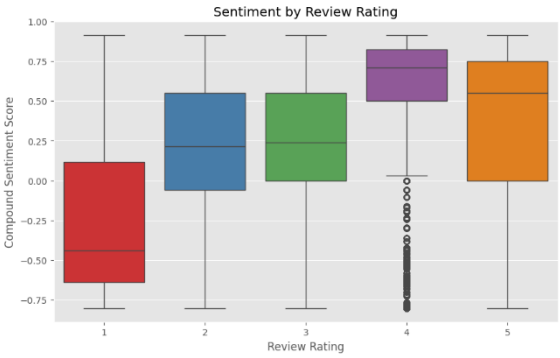
Histogram of Compound Sentiment Scores by Category: The histogram represents the distribution of compound sentiment scores across categories:

a. Health and Travel have stronger peaks at higher compound scores, indicating a greater proportion of positive reviews in these categories.
b. Shopping and Technology show wider distributions, with more variation in sentiment.

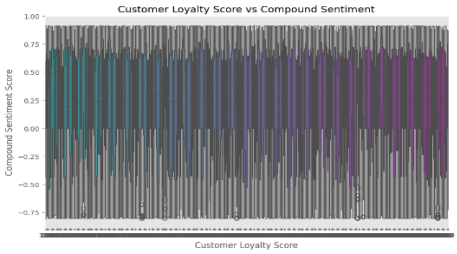


Heatmap of Sentiment Scores by Category: This heatmap compares average sentiment scores across categories for negative, neutral, positive, and compound sentiment. Categories such as Travel and Health are associated with higher positive and compound sentiment scores, highlighted in warm colors, while Shopping and Technology show lower sentiment scores.





Customer Loyalty Score vs. Compound Sentiment: The boxplot explores the relationship between customer loyalty scores and compound sentiment scores: Higher loyalty scores tend to correlate with stronger positive sentiment. Customers who are more loyal exhibit more positive sentiment in their reviews, suggesting that positive experiences lead to greater customer retention.



Sentiment by Review Rating: This boxplot shows the relationship between review ratings and compound sentiment scores:

a. Ratings 4-5 have more consistent distributions of stronger positive sentiments.
b. Ratings 1-2 show a wider distribution of compound sentiment scores, indicating a mix of feedback, with some reviews being positive despite low ratings.

Sentiment by Customer Gender: The bar chart compares compound sentiment scores between customer genders: Female customers have slightly more positive sentiments than male customers, suggesting that gender might play a role in sentiment tendencies.

Sentiment by Customer Age: The boxplot visualizes the variation in sentiment scores by age group:

a. Older customers (55+) show a trend toward higher positive sentiments.
b. Younger customers (18-25) exhibit a broader range of sentiments, indicating more mixed feedback among this age group.

I. CONCLUSION

To further enhance the research, incorporating real-world data and advanced techniques like association rule mining, sequential pattern mining, and social network analysis can provide deeper insights[10]. Time series analysis can identify trends and anomalies, while ensemble learning can boost model performance. Explainable AI can increase transparency and trust in model decisions. Dynamic segmentation and CLTV modeling can improve customer engagement and retention. Real-time fraud detection, adversarial machine learning, and network analysis can enhance security[11]. Aspect-based sentiment analysis and topic modeling can provide granular insights from customer feedback. Blockchain technology can revolutionize the industry by improving security, transparency, and trust. Ethical considerations, such as data privacy and fairness, must be prioritized. Future research directions include behavioral economics, personalised financial advice, and leveraging blockchain technology.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Prof Yamarthi Narasimha Rao for their invaluable guidance, support, and encouragement throughout this research.We acknowledge the support of VITAP for providing the necessary resources and facilities. We are also grateful to the anonymous reviewers for their constructive feedback, which significantly improved the quality of this paper.

REFERENCES

1. T. R. George, "Unveiling Customer Segmentation Patterns in Credit Card Data using K-Means Clustering: A Machine Learning Approach," in IEEE Xplore, 2024.
2. Jain, S., & Sharma, M., "Data Analytics for Credit Card Fraud Detection using Machine Learning Algorithms," International Journal of Machine Learning and Computing, vol. 10, no. 2, pp. 129-136, 2024.

3. Mishra, M., & Singh, A., "Application of K-Means and DBSCAN Clustering for Customer Segmentation in Credit Card Transactions," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 538-549, 2023.
4. Patel, A., & Patel, D., "Deep Learning for Customer Behavior Analysis in Credit Card Transactions," Journal of Computational Intelligence and Neuroscience, vol. 2023, Article ID 7493921, 2023.
5. Singh, R., & Jain, A., "Techniques for Clustering and Classification of Credit Card Data: A Comparative Study of K-Means, DBSCAN, and Gaussian Mixture Models," IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 4, pp. 1015-1027, 2023.
6. Rath, S., & Yadav, A., "Random Forest and XGBoost for Credit Card Fraud Detection: A Study of Feature Importance," IEEE Access, vol. 11, pp. 7804-7815, 2023.
7. Bhatia, S., & Singh, P., "Sentiment Analysis in Credit Card Transactions: Using NLP for Customer Feedback," International Journal of Data Science and Analytics, vol. 8, no. 1, pp. 56-70, 2023.
8. Ghosh, S., & Sharma, K., "Gaussian Naive Bayes and K-Nearest Neighbors for Credit Card Fraud Detection," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 6, pp. 1121-1134, 2024.
9. Agrawal, R., & Garg, R., "Hierarchical Clustering and Mean Shift Clustering Algorithms for Segmentation of Credit Card Users," Journal of Financial Data Science, vol. 9, no. 2, pp. 45-59, 2023.
10. Chakraborty, S., & Banerjee, D., "Future Trends in Credit Card Analysis: Leveraging AI and Deep Learning for Enhanced Fraud Detection," Journal of Financial Technology, vol. 2, no. 3, pp. 210-227, 2024.
11. Sharma, A., & Singh, P., "Impact of Socio-Economic Factors on Customer Behavior and Credit Card Usage," IEEE Access, vol. 11, pp. 2314-2326, 2023.