MACHINE LEARNING COURSEWORK

UNIVERSITY OF LONDON

Dataset from Kaggle

Customer Propensity to purchase

LWIN MOE AUNG

Table of Contents

1. Unsupervised Learning

1.1 Introduction and Objective

The primary objective of this analysis is to perform customer segmentation through clustering methods, specifically utilizing K-Means and Gaussian Mixture Models (GMM). The dataset used for this analysis comprises 607,056 observations and includes 25 features capturing detailed customer interaction behaviors, such as clicks, basket activities, and device preferences.

1.2 Data Preparation and Feature Engineering

The dataset used for clustering analysis consisted of over **607,000 observations** and **25 behavioral features**, combining customer interactions from both training and testing samples. These features covered a wide range of digital behaviors, such as basket interactions, promotional clicks, account and delivery actions, and device usage. After merging the datasets, a quality check confirmed that there were no missing values across any columns, indicating a complete and reliable dataset for modeling purposes.

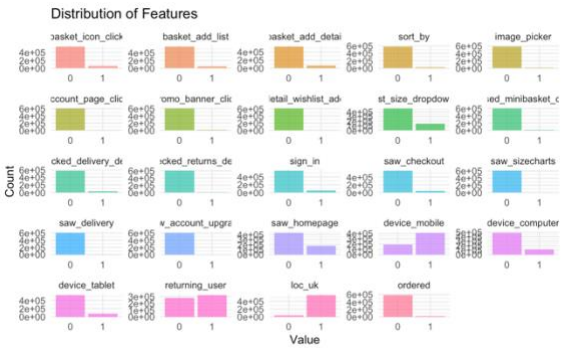

Figure 1: Distribution of Features

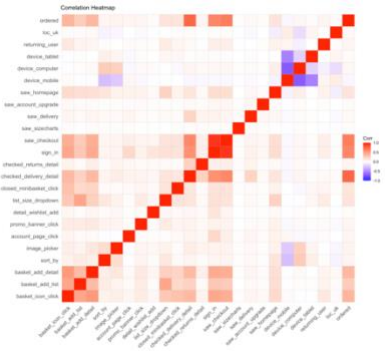

Figure 2: Correlation Heatmap

To begin exploratory analysis, numeric and binary features were identified. The binary features, consisting of actions encoded as 0s and 1s, were visualized to better understand the customer behavior distribution across key events. As shown in **Distribution of Features (Figure 1)**, most users did not perform high-intent actions such as basket_add_detail or ordered, highlighting significant behavioral imbalance and sparsity in conversion-related activities.

Next, low-variance features were removed to enhance the quality of correlation analysis. A **correlation heatmap (Figure 2)** was generated to identify multicollinearity among variables. The heatmap revealed several strong positive correlations between basket-related actions and between sign-in, checkout, and order completion events. These findings motivated the next step: feature engineering through behavioral score aggregation.

Three composite behavioral scores were created—**engagement_score**, **intent_score**, and **conversion_score**—each aggregating related features into a single dimension to summarize customer engagement, shopping intent, and conversion behavior. Redundant features used to compute these scores were dropped, along with optional device-type indicators, which were deemed less relevant for the clustering task.

1.3 Clustering Methodology

To segment customers based on behavioral scores, clustering was performed using both **K-Means** and **Gaussian Mixture Models (GMM)**. Prior to clustering, all engineered features were standardized to eliminate the influence of scale discrepancies, ensuring fair treatment of each variable during distance-based computations.

For K-Means, the optimal number of clusters was determined using two evaluation techniques: the **Elbow Method** and the **Silhouette Method**. **The Elbow Method (Figure 3)** evaluated the within-cluster sum of squares (WSS) for cluster counts ranging from 2 to 10. A distinct 'elbow' was observed at the point where adding more clusters yielded diminishing returns in reducing WSS, indicating an appropriate cluster number.

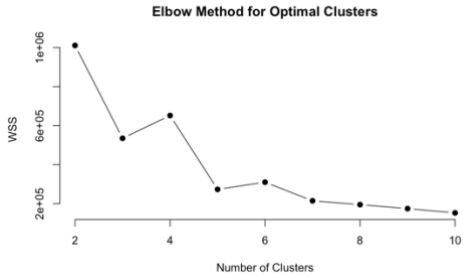

Figure 3: Elbow Method for Optimal Clusters

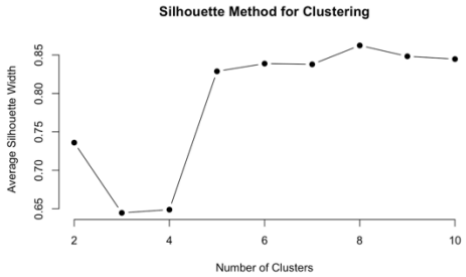

Figure 4: Silhouette Method for Clustering

The Silhouette Method (Figure 4) validated optimal cluster selection by measuring cohesion and separation. Unlike K-Means, Gaussian Mixture Models (GMM) flexibly model clusters of varying shapes, sizes, and orientations using probabilistic labels, effectively capturing overlapping or non-spherical clusters.

1.4 Results and Visualization

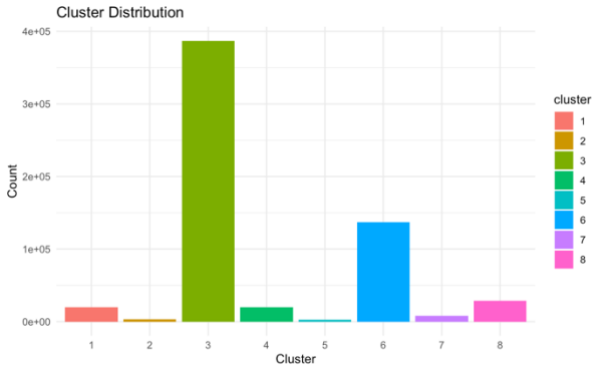

Figure 5: K Means Cluster Distribution

Following the application of K-Means clustering with the optimal number of clusters set to 8, the model segmented users into distinct behavioral groups. **K-Means Cluster Distribution (Figure 5)** revealed a heavily skewed structure, with Cluster 3 representing the majority of users—categorized as passive or silent users—while other clusters varied in size, capturing more specific behavioral profiles.

```
Average Silhouette Score: 0.8430507
```

The clustering quality was quantitatively evaluated using the average silhouette score, which was **0.843** , indicating well-separated and cohesive clusters. Further insights into the structure and spread of these segments are visible in the **3D Cluster Plot (Figure 6)** across the three behavioral scores: engagement, intent, and conversion.

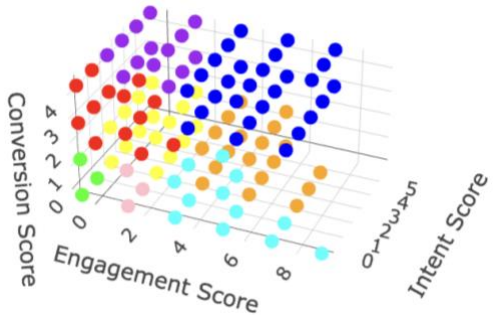

Figure 6: 3D Cluster Plot

To understand the composition of each segment, a dual-axis visualization was created to display the number of customers per cluster alongside their average feature scores. As shown in **Cluster Distribution with Average Feature Values (Figure 7)**, each cluster exhibited a unique behavioral profile, ranging from “Low-Engagement Browsers” and “Window Shoppers” to “Committed Buyers” and “Indecisive Visitors.”

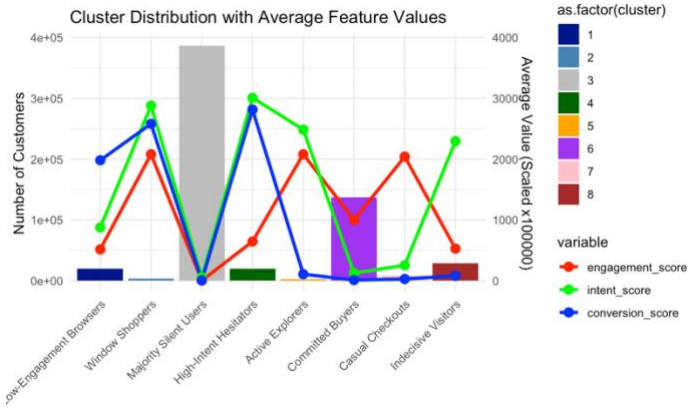

Figure 7: Cluster Distribution with Average Feature Values

The GMM model was then applied to the same scaled data. The number of optimal clusters and model shape parameters were determined using the Bayesian Information Criterion (BIC), as visualized in **BIC for GMM Model Selection (Figure 8)**. A GMM model with 9 clusters was selected based on this criterion. The **3D visualization of GMM clusters (Figure 9)** showed more fluid boundaries between groups, with each customer assigned to a cluster probabilistically.

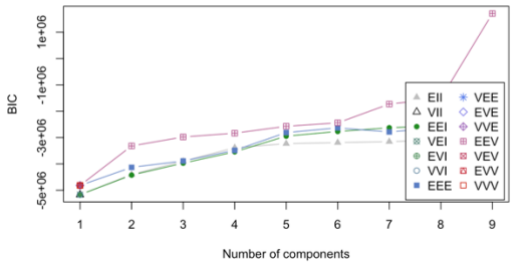

Figure 8: BIC for GMM Model Selection

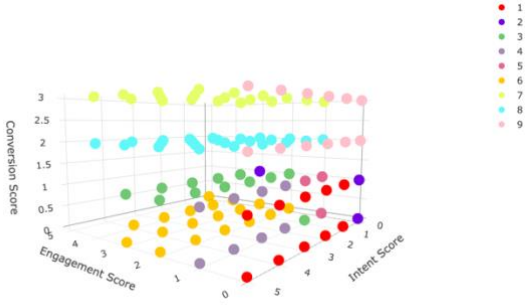

Figure 9: 3D GMM Clusters

```
Average Silhouette Score for GMM (Sampled): 0.8911595
```

The average silhouette score for the GMM approach was **0.891**, slightly outperforming K-Means in terms of cohesion and separation.

Finally, the combined visualization in **GMM Cluster Distribution with Average Feature Values (Figure10)** effectively captured the size and behavioral characteristics of each segment. Together, the results from both clustering methods offered a robust foundation for personalized marketing strategies and customer engagement planning.

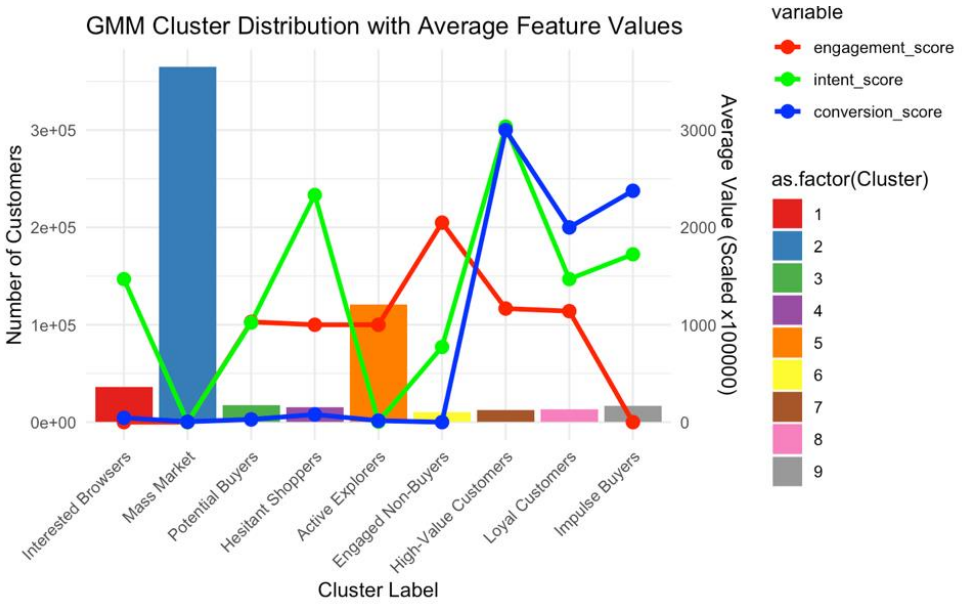

Figure 10: GMM Cluster Distribution with Average Feature Values

1.5 Insights and Business Implications

The clustering analysis yielded actionable customer segments with distinct behavioral profiles, enabling personalized marketing strategies.

For instance, **Low-Engagement Browsers** and **Window Shoppers** can be targeted with awareness campaigns or personalized offers to encourage deeper engagement. **Majority Silent Users**, being the largest group, may be better served with reactivation efforts or deprioritized in resource-intensive campaigns.

Committed Buyers and **High-Value Customers** should be prioritized for loyalty programs and upselling, while **Indecisive Visitors** may benefit from urgency tactics like limited-time discounts.

GMM clusters, including **Mass Market**, **Potential Buyers**, and **Loyal Customers**, provide additional nuance for strategic targeting. Probabilistic assignment allows marketers to tailor actions based on confidence levels in customer classification.

2. Classification

2.1 Introduction and Objective

The primary objective of the classification task was to predict the likelihood of customer purchase—formally referred to as **propensity classification**. Using customer behavioral data, this supervised learning approach aimed to assign a binary outcome (purchase or no purchase) based on observed features. This classification framework enables the business to identify high-propensity customers and strategically target them with conversion-focused campaigns, while also recognizing low-propensity users who may require nurturing or re-engagement strategies.

2.2 Data Preparation and Feature Engineering

The classification dataset included over 600,000 customer interaction records with no missing values. To enhance predictive power, several behavioral features were engineered. These include aggregated scores such as engagement_score, intent_score, conversion_score, basket_activity, and checkout_actions, alongside a derived device_usage indicator to quantify cross-device behavior.

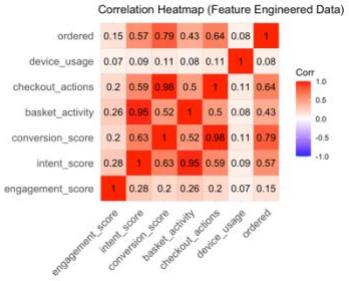

Figure 11: Correlation Heatmap

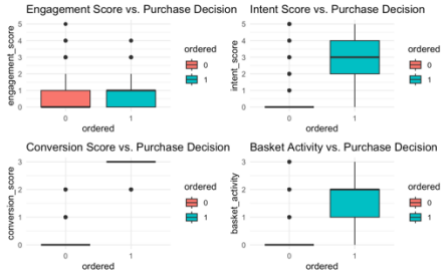

Figure 12: Relationship Between Features

A correlation heatmap (Figure 11) revealed that intent-related variables, particularly intent_score, conversion_score, and checkout_actions, had the strongest positive associations with the target variable ordered. This suggests that deeper behavioral engagement—especially around checkout stages—is a key indicator of purchase propensity.

To better understand the relationship between features and purchase decisions, Box plot (Figure 12) illustrates how the distribution of key behavioral scores varies across ordered (1) and non-ordered (0) classes. For example, users who made purchases displayed notably higher intent and conversion scores compared to non-buyers, validating their relevance in predictive modeling.

Figure 13: Class Distribution

Given the initial class imbalance—where non-purchasers far outnumber purchasers—a downsampling strategy was employed to create a more balanced dataset with a 10:1 ratio (majority to minority class). The final distribution is visualized as in **Class Distribution (Figure 13)** , ensuring that models trained on this data would not be biased toward the majority class. This step was critical for fair performance evaluation in classification tasks.

2.3 Methodology

To model the binary purchase outcome, two classification algorithms were applied: **Logistic Regression** and **Random Forest**. Logistic Regression provided a baseline interpretable model, while Random Forest offered an advanced ensemble method capable of capturing complex, nonlinear relationships.

Both models were trained and evaluated on the balanced dataset using 10-fold cross-validation to ensure robustness and avoid overfitting. Hyperparameter tuning for the Random Forest model was conducted via grid search, optimizing parameters such as the number of trees and maximum depth for improved predictive performance and generalizability.

2.4 Model Evaluation and Results

The performance of the Logistic Regression (**glmnet**) and **Random Forest** models was evaluated based on accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC).

Logistic Regression (glmnet): **Feature importance (Figure 14)** analysis highlighted checked_delivery_detail, saw_delivery, basket_add_detail, and checked_returns_detail as the most significant predictors.

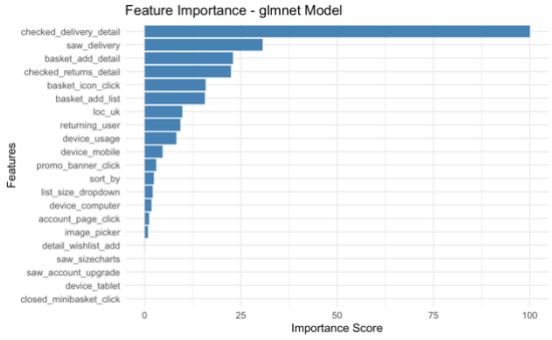

Figure 14: Feature Importance glmnet Model

The ROC curve (Figure 15) indicated strong predictive capability, with an AUC near 1, demonstrating excellent discrimination between classes.

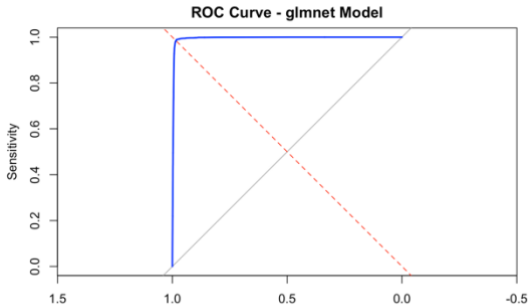

Figure 15: ROC Curve glmnet Model

The confusion matrix (Figure 16) showed good performance in correctly identifying non-purchasers (high specificity) but revealed challenges in accurately capturing purchasers (moderate sensitivity).

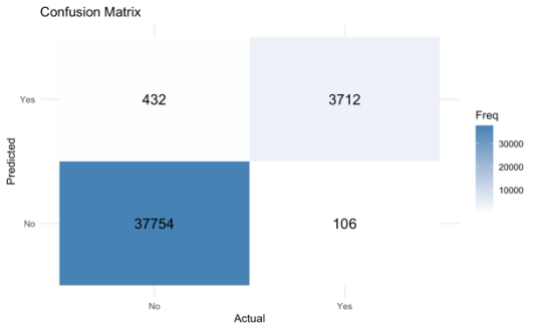

Figure 16: Confusion Matrix

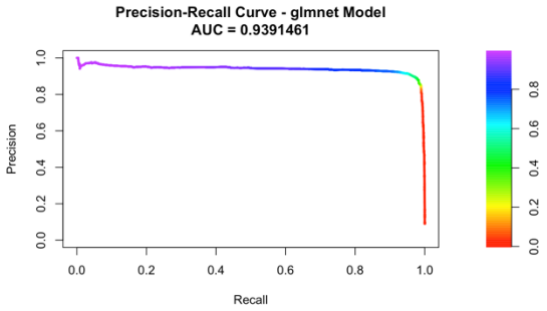

Figure 17: Precision Recall Curve- glmnet Model

The Precision-Recall curve (Figure 17) further validated the model's strong capability to balance precision and recall effectively, with an AUC of approximately 0.94.

Random Forest: Similar to the **glmnet model**, Random Forest identified checked_delivery_detail as the most influential feature, followed by basket_add_detail and basket_icon_click, emphasizing their relevance to predicting purchases as shown in **(Figure 18)**.

Figure 18: Feature Importance (Random Forest)

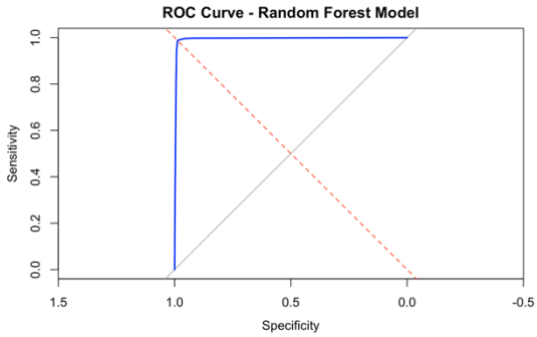

Figure 19: ROC Curve-Random Forest Model

The **ROC curve (Figure 19)** for Random Forest was equally impressive, indicating a highly reliable model for class separation with a high AUC value.

The **Confusion matrix (Figure 20)** demonstrated slightly improved performance in identifying true positives compared to Logistic Regression, suggesting a better capacity to detect potential buyers.

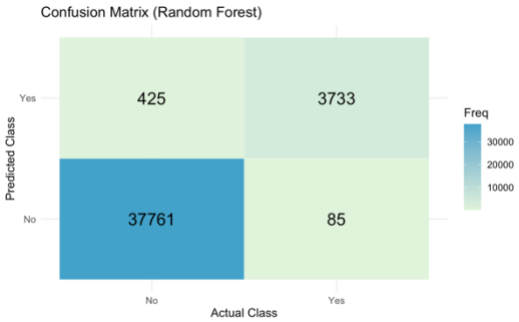

Figure 20: Confusion Matrix (Random Forest)

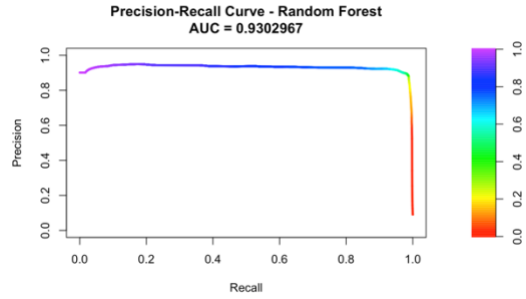

Figure 21: Precision Recall Curve (Random Forest)

The Precision-Recall curve (Figure 21) for Random Forest indicated strong overall performance (AUC approximately 0.93), reinforcing its robustness in classification tasks, particularly useful in targeting strategies.

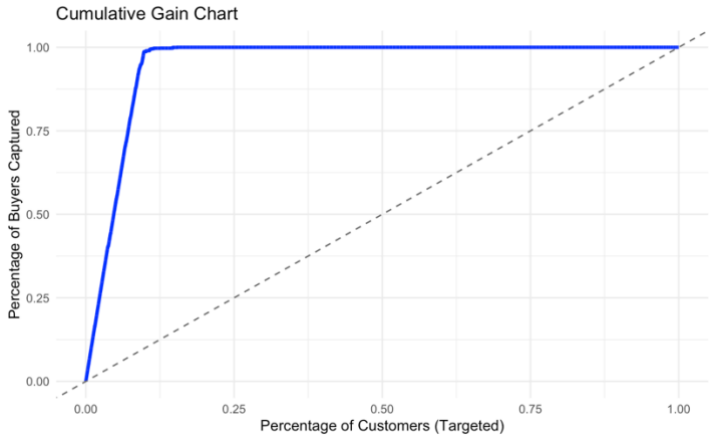

Figure 22: Cumulative Gain Chart

Additionally, the **Cumulative Gain Chart (Figure 22)** demonstrated the models' substantial effectiveness in identifying high-propensity customers. It showed that targeting a small fraction of the highest-scored customers could capture most of the potential buyers, significantly optimizing marketing efforts and resource allocation.

2.5 Business Insights and Practical Applications

The results from the classification models have direct practical implications for targeted marketing and CRM strategies:

- **High-Propensity Customers**: Customers predicted with high purchase likelihood should be targeted immediately with personalized incentives or time-sensitive promotions to capitalize on their readiness to convert.
- **Moderate-Propensity Customers**: Deploy nurturing campaigns with tailored messaging and moderate incentives to increase purchase motivation over time.
- **Low-Propensity Customers**: Initiate engagement or reactivation campaigns, emphasizing content-driven communications or highlighting brand value to stimulate renewed interest.

By leveraging classification insights, the business can optimize marketing efficiency, reduce acquisition costs, and enhance overall customer engagement and retention.

3. Regression

3.1 Introduction and Objective

The primary objective of the regression task is to predict continuous customer propensity scores, enabling precise customer ranking by their likelihood to purchase. Unlike binary classification, regression provides nuanced insights by assigning a numerical probability (ranging from 0 to 1) to each customer. This continuous score facilitates sophisticated targeting, allowing businesses to prioritize customers based on predicted purchasing potential and to tailor marketing interventions accordingly.

3.2 Data Preparation and Feature Engineering

To predict continuous propensity scores for customer ranking, key interaction terms capturing nuanced behaviors (e.g., basket interactions, delivery returns, device-specific actions) were created. Numerical features were scaled with Min-Max normalization to maintain balanced model influence. Continuous propensity scores from a Random Forest model were selected as the target variable, enabling more precise customer segmentation and targeted marketing compared to binary outcomes.

3.3 Methodology

Two regression models were used to predict continuous propensity scores:

- **Linear Regression** provided interpretability and confirmed linear relationships through residual analysis.
- **Gradient Boosting Regressor** effectively modeled complex, non-linear interactions, optimized via cross-validation.

3.4 Results and Model Evaluation

```
Linear Regression - RMSE: 0.04198229 MAE: 0.01954676
Gradient Boosting - RMSE: 0.02084499 MAE: 0.007517819
```

The models were evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The Linear Regression model achieved an RMSE of 0.04198 and MAE of

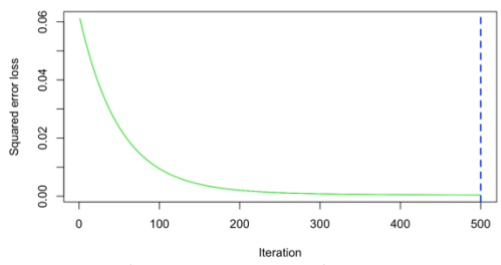

Figure 23: The Learning Curve

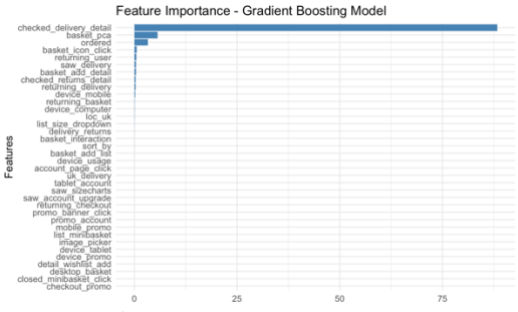

Figure 24: Feature Importance

0.01955, whereas the Gradient Boosting model outperformed significantly with an RMSE of 0.02084 and MAE of 0.00752, indicating greater predictive accuracy.

The Learning Curve (Figure 23) demonstrated how error rates decreased and stabilized through iterative model training, indicating effective learning and convergence.

Feature Importance (Figure 24) highlighted key variables influencing customer purchase likelihood, notably the features related to delivery details and basket interactions.

The Actual vs. Predicted Propensity Scores plot (Figure 25) illustrated a strong linear relationship between predictions and actual values, validating the model's predictive reliability.

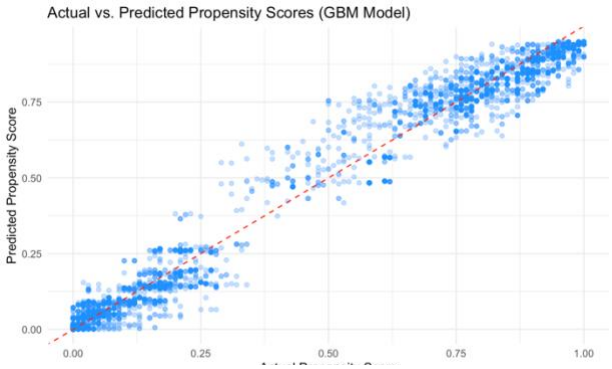

Figure 25: Actual vs Predicted Scores Plots

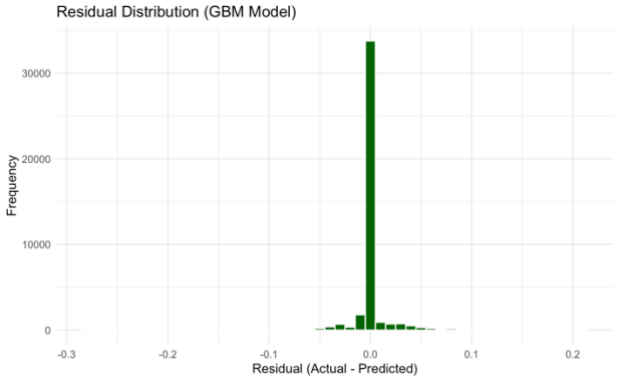

Figure 26: Residual Distribution Plot

The Residual Distribution plot (Figure 26) confirmed the model met key assumptions, displaying residuals that were largely centered around zero, indicating minimal systematic bias.

The Distribution of Predicted Scores (Figure 27) revealed clear distinctions in customer propensity, beneficial for targeted marketing strategies aimed at high-potential segments.

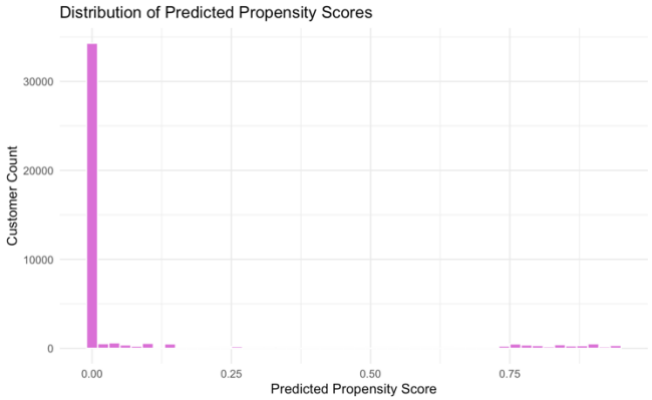

Figure 27: Distribution of Predicted Scores

3.5 Practical Implications and Deployment

The Gradient Boosting regression model enhances CRM systems like Salesforce by integrating propensity scores into customer profiles. This allows sales and marketing teams to quickly prioritize high-value prospects and automate targeted campaigns, optimizing resources and conversions.

Propensity rankings also support retention strategies, enabling personalized loyalty programs and incentives. Deploying this model within CRM platforms transforms insights into effective, revenue-driving actions.

References

Dataset – Customer Propensity to Purchase

P, B. (2018). *Customer propensity to purchase dataset*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/benpowis/customer-propensity-to-purchase-data/data [Accessed 26 Mar. 2025].