Executive Summary

This report explains a weekly churn prediction model for FoodCorp, designed to enable timely and interpretable churn detection across its customer base. Churn was defined as 60 days of inactivity after a reference date, validated through visit gap distributions and churn rate trade-offs. This captured gradual disengagement patterns while remaining actionable for weekly deployment.

To ensure real-world deployment conditions, features were engineered using SQL with strict temporal cutoffs. Each week, features were derived from a 60-day lookback window segmented into short-, mid-, and long-term tumbling windows, and churn labels were assigned using a 60-day lookahead period. Customers were only considered eligible if they had at least two prior purchases and were active at the time of prediction.

Two predictive models were developed using Python: a Logistic Regression pipeline with undersampling and a tree-based XGBoost pipeline with SMOTE balancing. Logistic Regression was used as a baseline classifier, offering speed and early-stage insights of predictor behavior. In contrast, XGBoost was selected as the final deployment model due to its better performance in handling nonlinear feature interactions, maintaining calibration, and achieving more stable F1 scores across volatile test periods.

Both pipelines were tuned using time-aware cross-validation, calibrated using sigmoid scaling, and evaluated under a rolling weekly sliding window strategy. Dynamic thresholds were used to maximize F1 score, preserving predictive performance across changing churn rates and seasonal behaviors. Across nine test periods (December 2021 to February 2022), XGBoost consistently outperformed Logistic Regression, achieving accuracy as high as 0.83, precision between 0.60-0.70, recall between 0.71–0.84, and F1 scores above 0.66 in all weeks, peaking at 0.75.

SHAP analysis revealed several consistent predictors of churn: long-term inactivity (*visits_31_60*), irregular visit timing (*stddev_days_between*), and disengagement signals (*recency_days, recency_ratio*), alongside monetary proxies like *CLV*. Both slow decline in engagement and abrupt behavioral change were picked up by these features.

The insight report confirms that churn at FoodCorp is driven by visit irregularity and recent engagement drop-offs than by spend or basket size alone. For instance, even high-value customers were shown to churn when their engagement pattern decreases. A typical churner profile included rising recency, increased irregularity, and a decrease in recent visit frequency, all of which were validated through both calibrated predictions and SHAP analysis.

This system provides FoodCorp proactive churn detection, weekly model deployment, and interpretable results for decision-making. Future improvements may include external signal use (e.g., holidays or promotions), automated retraining pipelines, and customer segmentation overlays to personalize interventions and enhance marketing efficiency.

Churn Definition

Churn was defined as no purchases in 60 days following a reference date (β = 60). This definition was used globally to balance analytical rigor with business practicality, and was selected based on:

- Customer Visit Gap Distribution: 60 days was more than the median inter-visit period (~75%) of customers, which showed a change from typical behavior.
- Target Class Proportion (PCCC): This definition produced an estimated base churn rate of ~10.4%, ensuring manageable targeting of churners while maintaining business impact.
- Strategic Relevance: The period is long enough to filter out sudden drop-offs, and short enough to allow timely, preemptive engagement.

This definition was applied on a rolling window of twelve weekly reference dates, ranging from November 2021 to February 2022. Churn labels were generated after each reference date with cautious to avoid data leakage. Customers were only considered eligible for churn prediction if they had at least two purchases and were active (i.e., not already churned) before the reference point. Weekly churn rates ranged from 7.4% to 11.3%, with higher churn levels in late January which is expected after the holiday season. This definition allowed proper temporal evaluation and was consistent with FoodCorp's weekly prediction interval.

Feature Engineering

A rolling SQL-based feature construction was done to simulate real-time deployment environment for the model used. For each reference date, features were constructed using strictly historical data to ensure a strict temporal cutoff between inputs and labels to avoid leakage and enable real time-aware evaluation **(Figure 1).**

Temporal Windowing Design

A fixed 60-day tumbling window was used, split into three lookback windows:

- 0–14 days: Short-term customer engagement
- 15–30 days: Medium-term behavioral trends
- 31–60 days: Longer-term purchase patterns

These segments helped in capturing temporal changes in behavior. Additionally, for each reference date, churn was labeled based on inactivity over the next 60 days, while features were derived from the past 60 days to maintain clear separation.

- Reference dates: 12 weekly windows from 2021-11-10 to 2022-02-01
- Feature window: 60 days before each reference date
- Label window: 60 days after each reference date

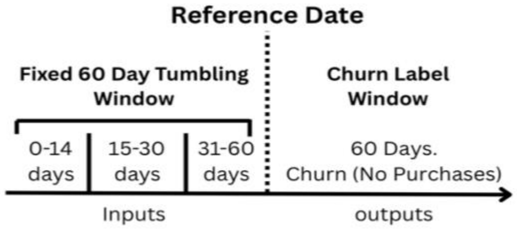

Figure 1 Temporal Feature Engineering & Label Design

SQL Features

All features were generated in SQL based on lagged behavioral summaries, using strictly historical data relative to each prediction point. Features focused on five behavioral aspects:

- **Recency & Timing**
  - *recency_days:* Days since last purchase
  - *avg_days_between, min_days_between, max_days_between, stddev_days_between:* consistency of visit timing
  - *recency_ratio:* Recency normalized by average gap between visits

- **Frequency & Engagement**
  - *frequency_0_14, frequency_15_30, frequency_31_60:* Count of transactions per window

- *recency_weighted_frequency:* Weighted activity metric (3:2:1 scheme) emphasizing recency over raw counts

- **Monetary Value**
  - *spend_0_14, spend_15_30, spend_31_60:* Spend per time window
  - *clv (Customer Lifetime Value):* Approximate lifetime value (total spend / distinct receipts)

- **Basket Data**
  - *basket_size_0_14, basket_size_15_30, basket_size_31_60:* Average basket size over time
  - *visits_0_14, visits_15_30, visits_31_60:* Visit frequency

Feature Selection

The final feature set for the chosen model (XGBoost) was selected based on:

- **Quantitative Importance:** SHAP values were used to identify strong predictors.
- **Stability Over Time:** Features that consistently ranked highly across SHAP plots were prioratized (e.g., *visits_31_60, stddev_days_between, and frequency_31_60*).
- **Business relevance:** Variables like *clv, recency_ratio, and spend_15_30* provided interpretable insight into customer value, timing, and decay.

Inital explorations with ratio/delta features (e.g., spend trend slopes) revealed instability across windows due to high correlation with the churn definition. These were removed to reduce label leakage risk and improve generalizability. Instead, the final feature set focused on interpretable lagged metrics, binned into the lookback windows, capturing recency and gradual drop-off patterns. Some volatile predictors, such as clv, were subject to ablation testing, but removing it caused a drop in recall and F1-score. Similarly, features like *frequency_0_14 and recency_weighted_frequency* were not among the top-ranked features, but they demonstrated strong localized (situational) importance for short-term churners. The Final Set of Feature included in Figure 2.

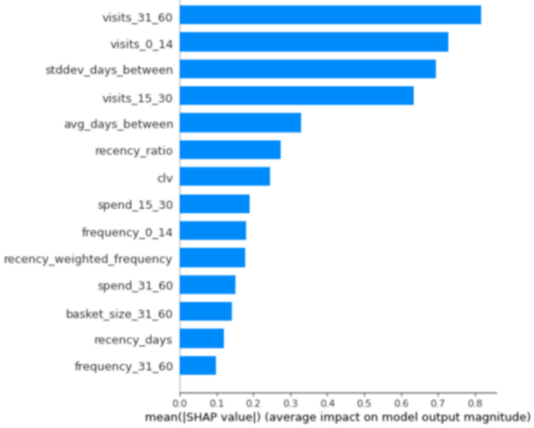

Figure 2 Final Feature Set + Importance

Prediction Approach

XGBoost as well as Logistic Regression was trained using SQL-based features and tested using a rolling time-aware strategy. These models were selected because Logistic Regression is mainly used for its interpretability, speed, and robustness in high-dimensional structured data (Hosmer et al., 2013). Which provided a strong and explainable baseline which was particularly useful for early-stage deployment. On the other hand, XGBoost is a gradient boosting algorithm known for its ability to capture non-linear relationships and handle feature interactions effectively which makes it a strong option for churn prediction (Chen & Guestrin, 2016).

Both models were optimized for recall and precision via class balancing, calibration, and dynamic thresholding, XGBoost was picked as the final deployment model due to its higher F1 score, strong calibration, and more stable precision across test weeks, especially in January and February. XGBoost consistently outperformed the Logistic regression model.

Logistic Regression Pipeline

- **Feature Standardization**: Features were scaled using StandardScaler() to ensure comparability which is essential for linear models.
- **Class Balancing**: Majority class was undersampled to balance the training set.
- **Hyperparameter Tuning**: Hyperparameter tuning was performed using RandomizedSearchCV with TimeSeriesSplit cross-validation to ensure temporal integrity during model selection.
- **Probability Calibration**: Applied CalibratedClassifierCV (sigmoid method, prefit) to improve probability alignment. Although originally introduced for SVMs (Platt, 1999), this approach is also used across classification models, including Logistic Regression to correct probability distortions caused by class imbalance or regularization (Niculescu-Mizil & Caruana, 2005).
- **Threshold Optimization**: The F1-maximizing threshold was selected dynamically for each test week from the range 0.1–0.9.

XGBoost Pipeline

- **Class Balancing**: SMOTE (Synthetic Minority Over-sampling Technique (k=3)) was used during the training process within an ImbPipeline to To address the natural class imbalance (~10% churners on average), without data loss.
- **Hyperparameter Tuning**:
  - RandomizedSearchCV with TimeSeriesSplit cross-validation ehich explored combinations of: n_estimators, max_depth, learning_rate, subsample, reg_alpha, reg_lambda
  - Tuned using 3-fold StratifiedKFold, evaluated with ROC AUC
- **Probability Calibration**: The best pipeline was calibrated using CalibratedClassifierCV with sigmoid scaling (`cv='prefit'`).
- **Threshold Optimization**: Weekly F1-maximizing thresholds were computed from the 0.1–0.9 range.
- **Model Interpretation**: Weekly SHAP analysis was conducted to ensure explainability.

Evaluation Strategy

To adapt to evolving customer behavior and preserve temporal causality, a sliding window evaluation strategy was used and predictions are made weekly using only data available **before the prediction date**, with no lookahead leakage.

- **Train**: 3 consecutive reference weeks (e.g., 2021-11-10 to 2021-11-24)
- **Test**: The next unseen reference week (e.g., 2021-12-01)
- **Repeat**: Across 9 rolling test weeks, from 2021-12-01 to 2022-02-01

This approach guaranteed that training data always preceded the test week, features were constructed using only the 60-day lookback window prior to each reference date, and labels were generated based on churn defined as 60 days of inactivity following the reference date. This setup avoided data leakage and adhered to best practices for time series model validation, as emphasized in Bergmeir & Benítez (2012).

Evaluation Metrics & Threshold Optimization

Each weekly test set was evaluated using a comprehensive set of classification metrics: **AUC (Area Under the ROC Curve)**, **accuracy**, **precision**, **recall**, and **F1 score**. The F1 score was prioritized as the primary optimization metric, as it balances precision (minimizing wasted interventions) and recall (capturing true churners).

Additionally, given behavioral drift and fluctuating churn rates across weeks, a fixed probability threshold was not suitable. Instead, churn probabilities were evaluated from 0.1 to 0.9 and the F1-maximizing threshold was selected for each test week. These thresholds were logged to assess stability and support calibration insights. This dynamic thresholding preserved performance in volatile periods such as the post-holiday phase in January.

Model Performance Summary

Logistic Regression

The Logistic Regression model emphasized simplicity and recall, making it well-suited for early churn detection. Across the nine test weeks:

- **Recall** ranged from **0.72 to 0.89**, with higher sensitivity to churners
- **Precision** remained modest (**0.51–0.58**)
- **F1 scores** peaked at **0.66**
- **AUC** consistently exceeded **0.82**

This model performed best in identifying customers with sudden behavioral changes, such as sharp increases in *recency_days* or erratic patterns (*stddev_days_between*). However, it produced a higher false positive rate, limiting its practical precision for weekly campaign targeting.

XGBoost

The XGBoost model delivered balanced and consistent performance:

- **AUC** ranged from **0.86 to 0.91**
- **Precision** reached **0.70**
- **Recall** ranged from **0.71 to 0.84**
- **F1 scores** consistently exceeded **0.66**, peaking at **0.75**.

Its ability to model nonlinear interactions and capture nuanced behavioral patterns made it highly effective at identifying disengagement. Key features contributing to predictions included:

- ***visits_31_60 & frequency_31_60*** (long-term engagement)
- ***stddev_days_between & clv*** (variability and customer value)

- ***recency_ratio & recency_days** (*disengagement indicators)

XGBoost also showed greater stability in performance across time, particularly during behaviorally noisy periods like early January, making it the preferred deployment candidate. Interpretability insights from SHAP analysis are detailed in Section 4.

Insight Report

Marketing Insights

1. Disengagement Happens Gradually

Churn does not typically occur as a sudden behavioral break, but rather as a gradual decline. Customers at risk of churn often show a sharp reduction in visit frequency within the last 14 days leading up to the prediction date, even ceasing visits altogether over this short-term period. These patterns usually follow a noticeable decline when compared to their behavior in the 31–60 day window, showing progressive disengagement. SHAP analysis and PDP confirms that drops in *visits_0_14 and visits_31_60* consistently appeared as leading indicators across multiple test windows (Figure 3 & 4), This indicates that churn is not abrupt, but identifiable at early stages, meaning that intervention is possible before full disengagement occurs.

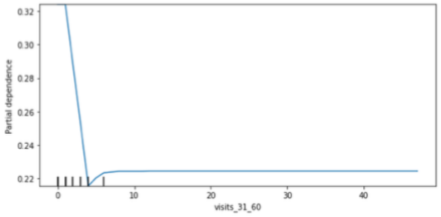

Figure 3 PDP visits_31_60

2. Irregularity Or Change in Routine

Another strong predictor of churn is the customer's habits in shopping. Customers who previously visited regularly begin to show erratic timing between visits, which is captured through high values in the stddev_days_between feature. This irregularity in visit timing serves as a crucial early signal of churn risk. Variability was the top-ranking predictor in terms of mean SHAP value (average impact of 0.89)(Figure 4), maintaining this position across all evaluated timeframes. reflecting behavioral instability and possibly points to customers exploring alternative retailers or deprioritizing FoodCorp in their routines.

3. Long-Term Inactivity Outweighs Basket Size

Although common assumption show that low spend or small basket sizes are primary churn indicators, our analysis shows that long-term inactivity is a stronger signal. Customers who had zero or very few visits in the 31–60 day window were significantly more likely to churn, even when their basket sizes were at the same range of non-churners. SHAP plots consistently ranked visit-based features above basket size, reinforcing that frequency of engagement is a stronger indicator than transaction size in churn prediction. This means that maintaining visit cadence, is usually more important than just targeting high-spend customers (figure 4).

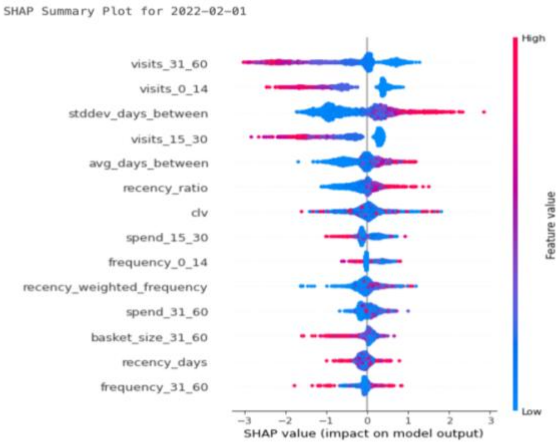

Figure 4 SHAP Analysis for the final test date

4. Recent Behavior Outweighs Historical Loyalty

While lower customer lifetime value (clv) is generally associated with higher churn risk, some high-CLV customers also disengaged when their recent engagement patterns decreased. Sudden drops in activity or a rise in recency_ratio — a measure of the recency of the customer's previous visit compared with their normal frequency — were stronger indicators. (Figure 5) reveal that when the recency ratio exceeds 2.5, even previously loyal customers become likely churners. This means that recent behavior is a more reliable churn signal than historical value, showing the need to monitor changes in individual engagement patterns over time.

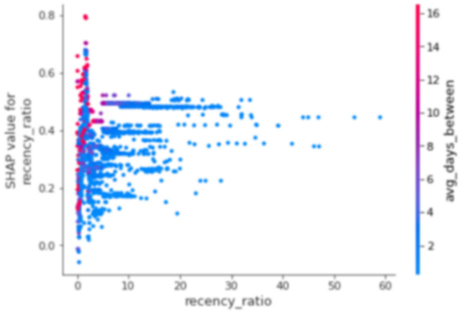

Figure 5 SHAP dependence plot showing the impact of recency_ratio on churn predictions, againsts avg_days_between.

Pen Portrait: "A Typical Churner"

Customer X had regular visit pattern, returning every 2 to 3 weeks. However, in the past 6 weeks, their behavior changed. Visit frequency dropped to just one transaction in the last 14 days, with no activity recorded before it (31–60 day window). Their recency_days is higher than usual at 13, and their stddev_days_between has increased sharply to 37.6, indicating irregularity in timing. Despite having a moderate clv of £753.93, their recency_ratio of 1.67 shows they are visiting far less frequently than usual. These changes signal disengagement, and this customer is a strong churn candidate.

Strategic Takeaways

Customers who have recently skipped visits, specially those who were previously active in the last 14 days represent an urgent case. A sudden drop-off, especially when accompanied by increased recency_ratio, is a reliable indicator of rising churn risk. Monitoring such patterns can enable timely recovery actions.

Inconsistent visit timing is a strong churn indicator. When standard deviation between visits (stddev_days_between) rises significantly, it often reflects a breakdown in routine and reduced customer loyalty. Identifying these shifts early will allow FoodCorp to initiate re-engagement in a timely manner.

Finally, cadence of engagement needs to be prioritized over transactional metrics like spend. Even customers with high lifetime value may disengage if their rhythm and visit frequency decrease. Retention strategies should focus on sustaining shopping habits, not just targeting based on purchase size.

Technical Justification of Insights

SHAP Summary Insights & Temporal drift in Feature Importance

SHAP analysis consistently highlighted a group of predictive features. Most impactful was *visits_31_60*, making long-term inactivity the highest churn signal. Next came stddev_days_between, which caught rising irregularity in visit frequency. *CLV* was moderately predictive, and sudden changes in behavior led to flagging some high-CLV churners. Recency_days and recency_ratio captured changes in visit frequency, declines in *frequency_0_14* and *frequency_31_60* reflected erosion in short- and long-term engagement. Together, these features reflect typical churn behavior: gradual disengagement, irregularity, and declining contact frequency. Moreover, it also revealed that features like *visits_31_60, stddev_days_between, and CLV* were consistently top-ranked, reflecting their strength. In contrast, features like *recency_weighted_frequency* and *frequency_0_14* varied in importance, particularly in January, due to seasonal changes (post-holiday period). This shows the need for weekly retraining and adaptive threshold tuning.

Calibration and Probability Distributions

Both models were calibrated using sigmoid. XGBoost showed a stable and well-calibrated probability estimates suitable for threshold-based targeting. On the other hand, Logistic Regression produced higher churn scores, resulting in stronger recall but more false positives due to the focus on recall. Weekly thresholds (0.1–0.9) were optimized for F1. A brief probability compression around 2022-01-01 was observed, but model calibration remained robust.

All together (SHAP scores and calibrated probabilities) enabled classification into the following segments to support different retention strategies based on behavioral context.

- **True Positives**: Inactive for a long time, irregular visits, low–mid CLV
- **False Positives**: Recent but inconsistent activity
- **False Negatives**: High-CLV customers with active short-term behavior not yet reflected in recency features

Pen Portrait Methodology
The churn persona (Section 4.1) relied on high-probability predictions with strong SHAP support. It combined disengagement signals *(e.g., recency_days, visits_0_14, visits_31_60*) with monetary value *(CLV)*, cross-validated against SHAP distributions to ensure alignment with model logic.

Limitations and Considerations

- **Model Trade-off**: The model focuses on recall with slightly lower percision.
- **SMOTE effects**: Synthetic sampling slightly reduced SHAP variation
- **External signals**: Promotions, holidays, etc., were not modeled
- **SHAP drift**: Some volatility in attribution shows the need for regular interpretability checks