1. Executive Summary

This project presents the development of a churn prediction system for FoodCorp, based on transactional customer data and guided by the framework established in the ConsultingCorp report. Churn was defined as an inactivity period of over 30 days, balancing early detection with practical actionability. Initial data cleaning addressed anomalies such as outliers and negative values, ensuring data quality for reliable analysis. Exploratory analysis revealed distinct purchasing patterns among customer segments, with regularity and frequency serving as key behavioral indicators.

A comprehensive set of temporal and non-temporal features was engineered to reflect both short- and long-term engagement patterns. A tumbling window approach captured weekly trends in spending, frequency, and diversity, while features such as recency and average days between visits summarized broader behavioral tendencies. Feature importance analysis identified lagged spend and product diversity as the most relevant predictors, while frequency and recency contributed minimally and were excluded.

Three models were evaluated using an out-of-time validation strategy. The Random Forest Classifier demonstrated the best performance, with a tuned version achieving a ROC AUC of 0.80. When tested the model correctly classified 74% of churners, demonstrating a strong ability to identify churners.

Behavioral insights derived from the model highlighted that churners typically exhibit declining spend, low product variety, and irregular visit patterns, while non-churners show the opposite. These insights can support targeted marketing strategies by identifying at-risk customers and enabling proactive engagement before full disengagement occurs.

2. Analysis of Churn

Following the approach outlined in the ConsultingCorp report, churn is defined as a customer not making a purchase within a set number of days, denoted as β. In this project, β is set to 35 days. That is, a customer is considered to have churned if they go more than 35 days without buying.

This threshold is supported by the report's "Distribution of times between visits" section, which shows typical shopping frequencies. A 35-day gap indicates a clear deviation from normal behavior, suggesting the customer may be disengaged or shopping elsewhere.

Additionally, the report's graph "Target class (have churned) proportions vs. churn definition" shows that using a 35-day threshold classifies around 30% of customers as churners when looking back with full future data, a proportion that is neither too low nor too high, making it a practical target for intervention.

Overall, this definition offers a balanced, actionable way to detect churn early, aligning with the report's recommendation to use a global and operationally simple churn definition.

3. Technical Report

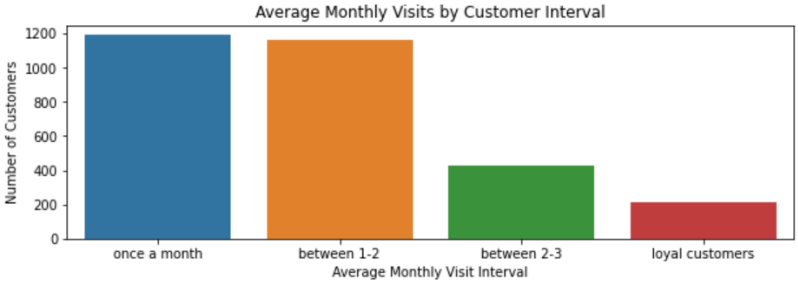
3.1. EDA & Data Cleaning

During the initial data inspection, certain anomalies were identified in the **receipt_lines** table. Notably, certain transactions contained unusually high values that were likely due to data entry errors, while others had negative amounts, which are presumed to reflect money refunds or promotional prize redemptions.

To ensure that these anomalies did not distort the representation of customer behavior, appropriate adjustments were applied. Negative transaction values were converted to their absolute values to preserve legitimate customer interactions, such as refunds or reward collections. Similarly, outlier values were replaced with the average product value (excluding the outliers themselves). These corrections were implemented to retain meaningful visit information while minimizing the impact of data irregularities. Given the low frequency of such anomalies, the adjustments had a minimum effect on the overall dataset.

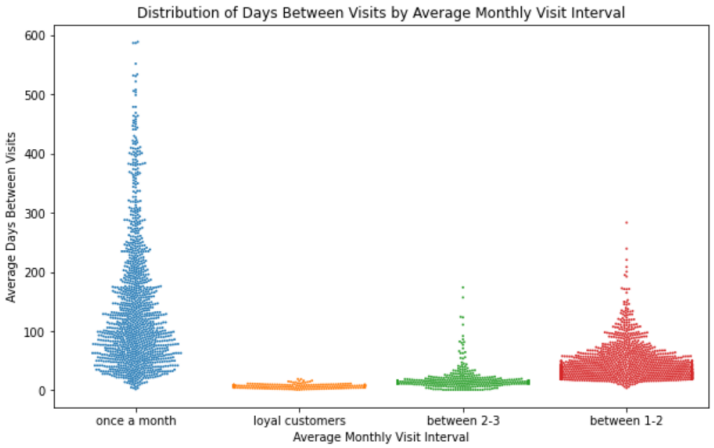
Following the preprocessing phase, an exploratory analysis was conducted on customer purchase frequency. The dataset comprises 4,938 unique customers, of whom 1,947 (approximately 39%) made only one purchase throughout the observed period. Overall, most customers purchased either once or between one and two times per month, as illustrated in Figure 1.

Figure 1



Building on these insights, the regularity of customer visits was further examined by calculating the average number of days between purchases for each defined customer segment. Results showed that customers in the "once-a-month" group exhibited high variability in their visit intervals, suggesting less consistent shopping habits. In contrast, more frequent buyers, such as those classified as “loyal customers,” demonstrated more stable purchasing behavior over time, this is shown in Figure 2.

Figure 2



These patterns highlight the importance of both frequency and regularity in understanding customer engagement. As a result, several features were derived to support churn prediction, including:

- The mean and standard deviation of days between visits
- The number of days since the most recent visit
- A trend metric capturing whether a customer's visit frequency is increasing or decreasing over time

Finally, the analysis of total monthly spend revealed a gradual downward trend, particularly from mid-2021 onwards. No clear seasonal patterns were identified, indicating that the decline in spending is structural rather than cyclical.

3.2. Feature Engineering

Based on the exploratory analysis of customer behavior, a combination of temporal and non-temporal features was selected to capture different aspects of customer engagement relevant to churn prediction.

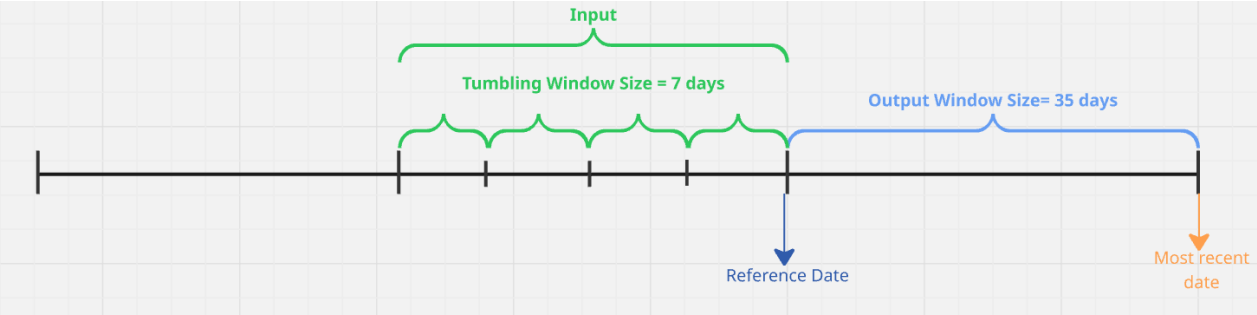
Temporal features were constructed to reflect short-term patterns in customer activity. Specifically, **spend**, **frequency**, and **diversity** were included as lagged variables measured weekly. These features allow the model to detect variations in customer behavior over time, which is particularly useful for identifying when typically, loyal customers begin to disengage. The use of multiple lags enables the model to learn dynamic trends and anticipate churn even when recent activity appears stable.

In contrast, features such as average days between purchases and recency were treated as non-temporal. These features summarize long-term patterns and are particularly informative for customers who exhibit irregular purchasing habits. For such cases, a high average interval between visits or a long time since the last visit may signal a greater likelihood of churn. Including these non-temporal features complements the lagged inputs by helping the model distinguish between temporarily inactive customers and those with historically low engagement, thus improving predictive precision.

To create the temporal features, a tumbling window approach was used with a size of 7 days per window. Each customer's behavior, such as spend, frequency, and diversity, was aggregated weekly over four consecutive weeks, forming the input features. The reference date marks the point at which predictions are made.

The output window spans 35 days after the reference date and is used to determine whether a customer has churned based on the selected definition. This structure ensures that all input features are based strictly on past behavior, preventing data leakage. This is illustrated in Figure 3:

Figure 3



The initial model selected for baseline evaluation was the Random Forest Classifier (RFC), chosen for its robustness, ability to handle nonlinear relationships, and its common use in early-stage classification tasks. RFC was also employed to assess the relative importance of features using two complementary approaches: Permutation Importance and SHAP values. Permutation importance provides a model-agnostic estimate of the impact each feature has on predictive performance, while SHAP values offer more detailed insight into how each feature contributes to individual predictions. Both methods are well-established and widely used in practice for model interpretation. In this analysis, all lagged variables were grouped and assessed as a single block to reflect their collective temporal behavior.

Results from both methods were consistent and indicated that frequency and recency contributed minimally to model performance. Based on this, these features were candidates for exclusion in subsequent model iterations.

3.3. Model Evaluation and Feature Selection

Following the feature importance analysis, a comparison of three models was conducted: Random Forest Classifier, Logistic Regression, and XGBoost. Random Forest Classifier was selected as a baseline model due to its robustness to noise and its ability to capture non-linear relationships, while Logistic Regression was included as a lightweight and interpretable alternative. Finally, XGBoost was selected for its strong predictive power and effectiveness in handling imbalanced datasets.

To ensure consistency in model evaluation, the ROC AUC score was chosen as the main metric. This metric is particularly suitable for imbalanced datasets, as it reflects the trade-off between recall and the false positive rate. This characteristic makes it especially appropriate when the priority is to identify as many churners as possible.

For Logistic Regression, a tailored standardization strategy was applied to ensure feature comparability while avoiding data leakage. Specifically, non-temporal features were standardized using the mean and standard deviation computed solely from the training set, and these values were then consistently applied to both the training and test data to maintain evaluation integrity. In contrast, temporal (lagged) features were standardized at the row level to preserve individual behavioral patterns over time. Rather than comparing

each customer to the overall population, this approach allowed the model to capture how recent activity deviates from that customer's own short-term history.

Model evaluation was conducted using an out-of-time validation strategy, which uses five reference dates, each spanning one month to the next. This method was selected because it respects the temporal structure of the data and avoids data leakage that may occur with conventional cross-validation, which is unsuitable for temporal prediction problems.

The models were evaluated using the area under the ROC curve (ROC AUC), with the positive class defined as churn. The Random Forest Classifier achieved the highest average score (0.7958), followed closely by XGBoost (0.7920). The Logistic Regression model performed slightly lower, with an AUC of 0.7503. These results confirm that tree-based models perform better in capturing the non-linear relationships present in customer behavior, while Logistic Regression still provides a competitive baseline.

After selecting the Random Forest Classifier as the final model, a feature selection step was performed to validate the removal of low-impact variables. As shown in the evaluation results, excluding recency alone led to a slight decrease in performance (ROC AUC from 0.7958 to 0.7936). Interestingly, when all frequency-related features (freq_lag1 to freq_lag4) were also removed along with recency, the performance slightly increased to 0.7947. This indicates that neither recency nor frequency-based features contributed meaningfully to model performance.

These results confirmed that the exclusion of both feature types did not negatively impact the model and even led to a marginal improvement. Consequently, both recency and the lagged frequency variables were removed from the final feature set, supporting a simpler model architecture without loss in predictive accuracy.

Before evaluating the model, a parameter tuning step was performed to optimize the performance of the Random Forest Classifier, considering the same five reference dates. A manual grid search was conducted in combination with the out-of-time evaluation strategy to ensure that parameter selection was aligned with the temporal structure of the prediction task. The grid explored variations in **n_estimators**, **max_depth**, and **min_samples_split**, as these parameters are critical for controlling model stability, preventing overfitting, and balancing complexity with generalization.

Among all combinations tested, the configuration with **n_estimators** = 200, **max_depth** = 5, and **min_samples_split** = 5 achieved the highest performance, reaching a ROC AUC of 0.8014 and a recall of 0.4487. This combination was therefore selected as the final setting for the model. The tuning process confirmed that performance gains were consistent and not sensitive to minor parameter changes, further reinforcing the model's robustness for deployment.

Finally, the model was evaluated on a hold-out test set using the best parameters. The ROC curve was constructed to determine the optimal classification threshold with the goal of maximizing recall. A threshold of 0.3 was selected, as the impact of misclassifying non-churners is not as significant as the cost of failing to detect a churner. Losing a customer typically incurs much higher costs to recover them. At this threshold, the model achieved a recall of 0.7404. This recall value shows that the model successfully identified over 74% of

all true churners, making it effective at capturing a significant portion of the target group, although some churn cases may remain undetected.

3.4. Summary

A comprehensive analysis was conducted to prepare and model churn prediction effectively. Initial data cleaning addressed anomalies such as outliers and negative values in transactions, ensuring minimal distortion of customer behavior. Exploratory analysis revealed that most customers shop once or twice per month, with loyalty reflected in both frequency and regularity of visits. Based on these insights, a set of temporal features (spend, frequency, diversity over four weekly lags) and non-temporal features (recency and average days between purchases) were engineered using a tumbling window structure. Feature importance analysis indicated that frequency and recency were of limited predictive value and were excluded from the final feature set. Multiple models were evaluated using an out-of-time strategy, with the Random Forest Classifier outperforming others (ROC AUC = 0.79). When applied using "01/02/2022" as the reference date and the best parameters identified through hyperparameter tuning, the model achieved a recall of 0.80. This demonstrates its strong ability to identify churners and highlights its potential for supporting proactive retention strategies.

4. Insight Report Part 1: Summary for Marketing

The behavioral analysis conducted on customer transactions reveals meaningful differences between customers who churn and those who remain engaged. These findings can support the development of targeted engagement strategies by highlighting segments at risk and providing justified indications of behavioral change.

Churners tend to display unstable shopping habits. Their spending decreases over time, particularly in the weeks leading to inactivity. They also interact with a narrower set of product categories, and while some show a slight increase in product variety just before churn, this is not sustained. Importantly, they return to the store less frequently, with long and irregular gaps between visits. These patterns suggest that churners may be disengaging due to limited interest or satisfaction with the store's offerings.

The data shows that churners reduce both spending and diversity prior to churn. This indicates that collecting feedback from customers with narrowing product baskets could help uncover early drivers of disengagement. Understanding whether this reflects dissatisfaction, changing needs, or lack of awareness may support more personalized marketing interventions.

Non-churners, on the other hand, show signs of loyalty. Their visits are more regular and frequent, with shorter and more consistent intervals between purchases. They also purchase a wider and increasingly diverse range of products over time. These characteristics are consistent with a pattern of increasing engagement.

The data shows that non-churners tend to diversify their purchases over time. This suggests that encouraging product exploration may be associated with stronger retention. Monitoring drops in product variety could help identify loyal customers at risk of disengagement.

The combination of declining spends, low diversity, and increasing time between visits is associated with customer churn. These behaviors can be monitored using the developed prediction model. Customers showing such trends may benefit from proactive engagement. For example, when the system flags a likely churner with recent drops in spending and category diversity, further qualitative information (a short feedback prompt or preference survey) could reveal potential areas for re-engagement, such as missing promotions or unmet expectations.

5. Insight Report Part 2

The behavioral insights presented in the marketing summary were derived through systematic analysis of transactional data and predictive modelling. The purpose was to identify patterns between customers who churn and those who remain active, based on the churn definition adopted (35 day inactivity threshold).

The process began with exploratory data analysis. Customer visit frequency and inter-visit intervals were analyzed, revealing significant behavioral segmentation (Figure 1). Additional distribution analysis (Figure 2) confirmed that customers who visit less frequently also display greater variability in their time between visits, which may signal instability in shopping habits. This observation supported the creation of features such as average days between purchases and its standard deviation, both of which proved informative in characterizing churn behavior.

To capture evolving customer patterns, temporal features were engineered using a tumbling window approach (Figure 3). Aggregates such as weekly spend, frequency of visits, and product diversity were computed across four-time windows, enabling the model to detect behavioral trends. For instance, churners were found to decrease their weekly spending in the weeks leading up to churn, while non-churners showed the opposite tendency, steady or increasing expenditure. Similarly, churners tended to maintain low product diversity, suggesting reduced engagement with the store's range of offerings.

In the modelling phase, the Random Forest Classifier was used to identify the most influential features. Feature importance techniques (SHAP values and permutation analysis) consistently highlighted that lagged spending and diversity were the strongest predictors of churn. This confirmed that changes in purchasing patterns are crucial indicators of disengagement. In contrast, frequency and recency features were shown to contribute minimally and were excluded from the final model to enhance interpretability.

The insights summarized in Section 1, such as the declining spending and narrow product engagement of churners, versus the consistent and varied activity of non-churners, were derived directly from the behavior of these key features across both customer segments. These insights are grounded in observed trends over time and are validated by their predictive relevance in the final model.