Machine learning and Advanced Analytics (BUSI4373)

Churn Prediction for FoodCorp

Student Id: 20519860

08.05.2024

Executive Summary:

Foodcorp has identified customer churn as a significant concern and seeks to address it through data-informed strategies. Initial analysis provided a subtle understanding of churn, emphasising the trade-off between capturing potential churners early and minimising false positives. Based on this, a definition of churn was finalised, defining it as inactivity for 24 days. This definition strikes a balance, capturing a significant portion of potentially churned customers while maintaining a manageable false positive rate. Currently, the churn rate stands at 32%, with a retention rate of 68%.

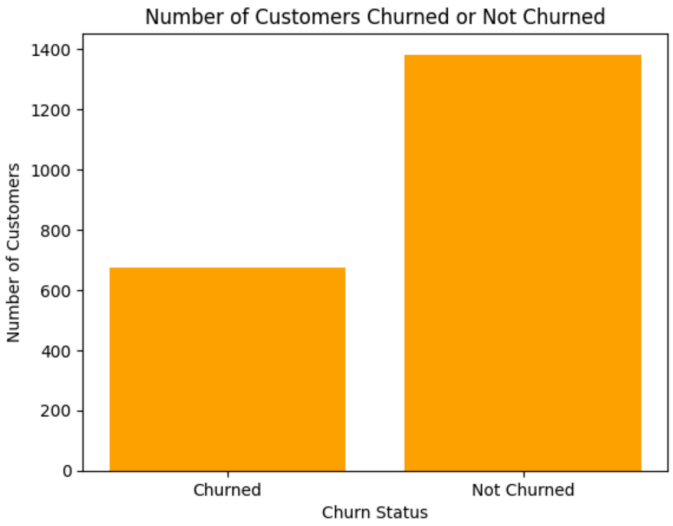
To address churn prediction systematically, a churn prediction system was developed. A temporal feature engineering approach was employed using customer transaction data, with the objective of capturing spending behaviour over time. Machine learning algorithms, including Decision Trees, Support Vector Classifier, and Random Forests, were evaluated, with Random Forests emerging as the optimal choice. Through parameter tuning and feature reduction, the model achieved an accuracy score of 0.74 on the test set, with a precision of 0.80, indicating its effectiveness in identifying potential churners while minimising false positives.

The identification of false positives and false negatives plays a pivotal role in the formulation of targeted interventions, with a particular focus on the retention of customers. The analysis of store-level dynamics provides insights into customer spending behaviour, offering additional granularity for the formulation of marketing strategies. By addressing false positives, false negatives, and true positives, marketing efforts can be optimised to achieve the greatest impact on customer retention and loyalty.

The technical report on insights derivation demonstrates a meticulous analytical approach, which employs key metrics and visualisations to evaluate model performance comprehensively. Performance evaluation metrics, confusion matrices, box plots, bar charts, and correlation analysis provide insights into spending behaviour, store-level dynamics, and feature importance. The insights derived offer actionable guidance for enhancing customer retention strategies and optimising marketing efforts to effectively mitigate churn.

Current levels of churn:

The choice of an appropriate churn definition necessitates a trade-off between the objective of capturing potentially churned customers at an early stage and the aim of minimising false positives. A longer churn definition, such as 44 days, may result in the identification of customers who are more likely to have truly churned, thereby reducing the number of false positives. However, this may also lead to the loss of opportunities for intervention with customers who churn earlier. Conversely, shorter churn definitions, such as 7 or 1 day, may capture at-risk customers early, enabling proactive intervention. Nevertheless, they may also result in a higher number of false positives, leading to inefficient resource allocation.



Churn Rate: 32.814778804083616
Retention Rate: 67.18522119591638

In this analysis, churn is defined as an active customer being inactive for a period of 24 days from the reference date. An active customer is defined as one who made a purchase during the 24-day period preceding the reference date. The selected definition strikes a balance between the need to identify potentially churned customers at an early stage and the minimisation of false positives. The median time between visits is 24 days, which allows this definition to capture a significant portion (60.15%) of customers who may be at risk of churn. At the same time, it maintains a moderate target rate of 17.87% of active customers for intervention. According to this definition, approximately 700 active customers are identified as churned, resulting in a churn rate of 32%, with a retention rate of 68%.

Technical Report on the Developed Churn Prediction System

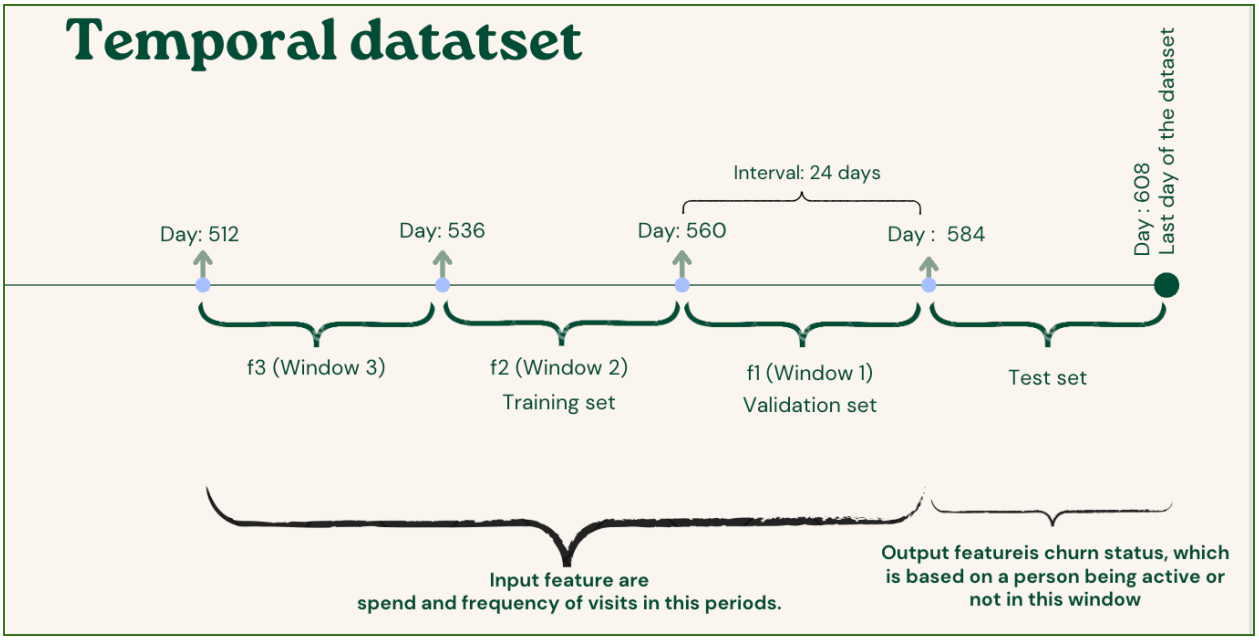
The dataset employed in this analysis comprises customer transaction data, encompassing variables such as purchase value, purchase date, product category, and store code. Each observation represents a unique customer, and the target variable is the churn status, defined as an active customer becoming inactive for a period of 24 days from the reference date.

1. Creation of the customer_summary table:

To facilitate temporal feature engineering, we created the customer_summary table, which aggregates customer transaction data. This table contains columns such as customer_id, total_value, no_of_days, frequency, and store_code. The no_of_days column represents the number of days since the first purchase date of the data set. The data was available for 608 days.

2. Feature Engineering:

Temporal features were engineered based on the customer_summary table in order to capture customer behaviour over time. These features include spending and frequency of purchases in rolling windows of 24 days. By analysing these temporal features, insights can be gained into customer activity leading up to the reference date, which is



crucial for predicting churn accurately. The image below shows how the temporal features were created.

In order to capture the temporal behaviour of customers leading up to the reference day, a windowing strategy was employed. Each window comprises a period of 24 days, with the first window spanning 24 days prior to the reference day (f1). Within each window, we calculate the spending and frequency of purchases for each customer. Furthermore, we create a second window that precedes the first window by another 24 days and another one that precedes the second window. This approach enables us to capture the behaviour of customers over three distinct time intervals leading up to the reference day. The features spend and frequency, and store_code from each window are the input features for the model. The churn status of an active customer in the window 24 days after the reference days is the output feature.

To determine the churn status, the spending behaviour of customers within the f1 window was analysed initially. Customers with zero spending during this period were categorised as inactive. Subsequently, inactive customers were filtered out from the dataset. Within the remaining active customer subset, spending in the output window was examined. Customers exhibiting zero spending in the output window were designated as churned and assigned a numerical value of 0, while those with nonzero spending were labelled as not churned, denoted by a numerical value of 1. This process effectively distinguishes between churned and non-churned customers based on their spending patterns across defined time windows.

3. Predictive approach:

Three machine learning algorithms, namely Decision Trees, Support Vector Classifier (SVC), and Random Forests, were employed to develop the churn prediction model. The performance of each model was evaluated based on accuracy scores, and 5-fold cross-validation was employed to ensure the robustness of the assessments.This technique involves partitioning the dataset into five equally sized folds, training the model on four folds, and evaluating it on the remaining fold. By repeating this process five times and averaging the results, a more reliable estimate of the model's performance is obtained compared to a single train-test split. This methodology helped to reduce the variability in model performance across different data subsets, thereby providing a more reliable evaluation of the models' effectiveness.

A Decision Tree classifier was trained as a baseline model and achieved an accuracy score of 0.69 on the validation set. Initial experiments indicated that the random forest classifier exhibited superior performance, achieving an accuracy score of 0.74 on the validation set, which means around 74% of the models prediction on a customer churning is right. Therefore, it was selected for further tuning. Subsequently, the finalised model's parameters were tuned to enhance its performance. Grid search CV was employed to perform a grid search of the random forest classifier's parameters. Subsequently, the optimal parameters for the model were identified, resulting in a validation accuracy of 0.77 for the tuned model. The parameters that were evaluated using the grid search CV for model tuning were as follows: 'n_estimators': [50, 100, 200], 'max_depth': [None, 10, 20, 30], 'min_samples_split': The optimal parameters were identified as follows: max_depth=10, max_features='auto', min_samples_leaf=2, min_samples_split=10, n_estimators=200. These parameters were selected to enhance the performance of the model.

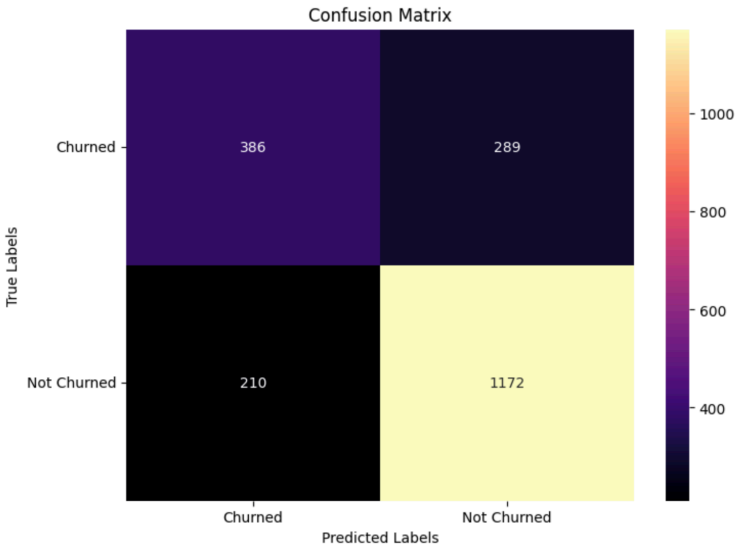
Once parameter tuning has been completed and a model has been selected, the next step is to attempt to reduce the number of features in order to remove those that are of least importance in order to optimise the performance of the model. In order to ascertain the importance of each feature and the impact it has on predicting the output feature, permutation importance was employed. Permutation feature importance is a model inspection technique that measures the contribution of each feature to a fitted model's statistical performance on a given tabular dataset. The most significant features identified were f2_spend, f3_spend, f2_frequency, f3_frequency, and f1_frequency. The least important features were store_code and f1_spend, These two features were kept as it is as it still showed to be

important for the model.

Following the parameter tuning, the Random Forest classifier model was applied to the test set, representing new data that was not seen during the model training phase. This step is crucial for evaluating the model's generalisation capability and its performance on real-world data scenarios. The obtained accuracy score of 0.74 indicates that the model correctly predicted the outcome approximately 74% of the time. This indicates a promising level of predictive accuracy, suggesting that the model is effective in making predictions on unseen data. However, further analysis may be warranted to assess other metrics such as precision, recall, and F1 score in order to gain a comprehensive understanding of the model's performance characteristics.

4. Model Evaluation:

The evaluation phase entailed a comparison of the accuracy scores achieved by three distinct models. The utilisation of 5-fold cross-validation enhanced the rigour of this process, enabling the identification of the optimal model that exhibited the greatest generalisability to unseen data. Once the optimal model was identified, it was subjected to further testing on a separate dataset. The generation of a confusion matrix and the examination of additional performance metrics, such as precision, recall, and F1 score, enabled a more profound understanding of the model's efficacy across various aspects of classification. In this instance, false positives are to be focused on, as when the model predicts churn but in reality it did not happen, it means that the company could lose money on advertising to those customers. The final model has 210 false positive instances and a precision of 0.80. In this context, precision is the key metric to consider, as an increase in the precision score indicates a reduction in the number of false positives.




```
Accuracy: 0.757413709285367
Precision: 0.8021902806297057
Recall: 0.8480463096960926
F1 Score: 0.8244811818501583
```

5.Summary:

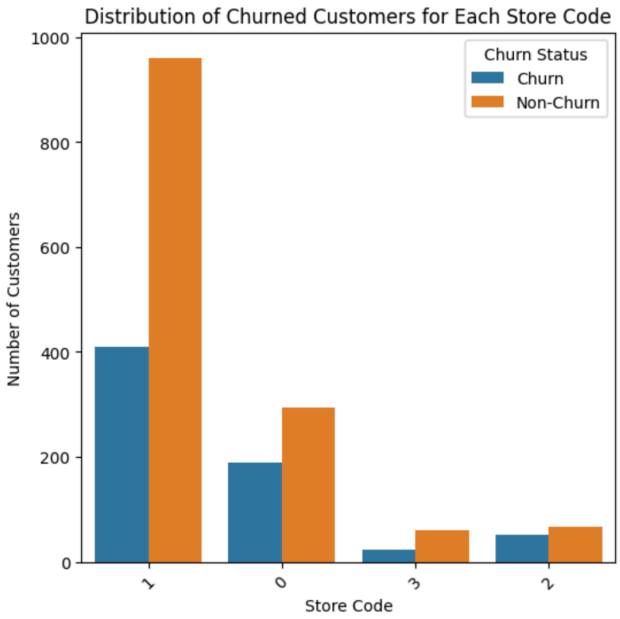
The analysis aimed to predict customer churn using transactional data, employing temporal feature engineering to capture spending behaviour over time. After evaluating multiple machine learning algorithms, Random Forests emerged as the optimal choice due to its superior performance. Through parameter tuning and feature reduction, the

model achieved an accuracy score of 0.74 on the test set, indicating a 74% prediction accuracy. Evaluation metrics highlighted the model's precision, recall, and F1 score, with a focus on minimising false positives to optimise marketing expenditures. The final model demonstrated a precision of 0.80, with 210 false positive instances, underscoring its ability to effectively identify potential churners while minimising the risk of misidentifying non-churners.

Insights

Section 1: Marketing Insights: Leveraging Churn Analysis

The occurrence of false positives and false negatives plays a pivotal role in the formulation of marketing strategies for customer retention and acquisition. Within our dataset, there were 221 instances of false positives, which signified cases where our model incorrectly predicted customer churn. In such instances, resources allocated towards advertising to these customers could potentially be misdirected, as these individuals did not actually churn. Conversely, false negatives occur when the model fails to predict customer churn accurately. Identifying and rectifying these instances is crucial, as it presents an opportunity to proactively engage with customers who are at risk of churning. By correctly predicting potential churners, the marketing team can intervene with tailored incentives and offers, effectively reducing churn rates and fostering customer loyalty. In order to be efficient, the marketing team should focus on reducing false positives and try advertising less. While this may not reach all customers, it is more cost-effective, If the marketing team is able to utilise the strategy effectively, or if they have a substantial budget and are keen on retaining customers, they can advertise offers to a large number of people, including those who may be false positives. However, this approach could result in greater reach, but at the same time, it would cost more. The decision between these two options is based on the marketing strategy followed by a company.

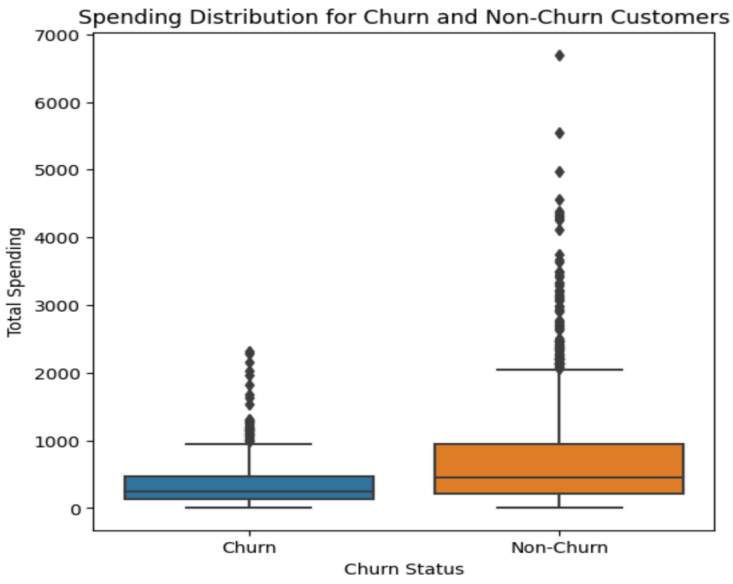


Furthermore, marketing strategies should also encompass true positives, representing customers correctly predicted to churn. These individuals warrant special attention, as they

are most likely to discontinue their patronage. By focusing efforts on retaining these customers through personalised outreach and retention initiatives, the marketing team can effectively mitigate churn risks and preserve valuable customer relationships.

Additionally, the analysis reveals valuable insights into store-level dynamics that can inform strategic marketing decisions. Store 1 emerges as the store with the highest number of active customers, indicating its significance in the customer base. However, with approximately 400 churned customers, which is nearly double that of Store 0, targeted marketing efforts towards Store 1 could yield substantial reductions in churn rates.

Store 2, while experiencing a lower overall churn count, exhibits a concerning trend with more than one-third of its customers churning in the last window. Despite its smaller customer base compared to Stores 0 and 1, it is imperative to focus on improving retention strategies for Store 2 in order to maintain a healthy customer base and optimise overall performance.Furthermore, it has been found that customers can be identified as being on the verge of churning or not based on their spending in the previous window. In most cases, customers who are about to churn spend less than £1,000 in the 24 days preceding the reference day.



In summary, insights derived from churn analysis enable the marketing team to tailor strategies effectively, targeting both individual customer segments and specific store locations in order to enhance customer retention and drive business growth. By addressing false positives, false negatives, and true positives, marketing efforts can be optimised to achieve maximum impact on customer retention and loyalty.

Section 2 : Technical Report on Insights Derivation

The process of deriving insights from the churn prediction model involved a meticulous and analytical approach, focusing on key metrics and visualisations to comprehensively evaluate

model performance and extract actionable insights. To objectively assess the model's classification effectiveness, a range of performance evaluation metrics including precision, recall, F1 score, and accuracy were meticulously calculated.

Additionally, a confusion matrix was constructed to provide a detailed understanding of the predictive model's outcomes, allowing for a granular analysis of true positives, true negatives, false positives, and false negatives. Subsequently, box plots were utilised to compare spending patterns between churned and non-churned customers. These visualisations effectively highlighted discernible differences in spending behaviour, with churned customers consistently exhibiting lower spending levels compared to their non-churned counterparts. This observation serves to highlight the significance of spending behaviour as a predictive feature for churn propensity.

Further analysis was conducted using bar charts to investigate the distribution of churned customers across different store codes. This exploration revealed significant disparities in churn rates among various store locations, indicating potential opportunities for targeted marketing interventions aimed at curbing churn. Moreover, a correlation analysis was performed to delve into the relationship between customer spending and churn status. The correlation matrix yielded intriguing insights, particularly those pertaining to the correlation between the frequency of visits (f1_frequency, f2_frequency, and f3_frequency). Although these features exhibited correlation, they were retained in the dataset as they provide valuable insights into customer behaviour across different time windows. This suggests that customers tend to exhibit a consistent purchasing pattern across each window, thereby reinforcing the importance of temporal features in churn prediction

Additionally, permutation importance analysis was conducted to rank the features and evaluate their significance in predicting churn. This analysis provided valuable insights into feature importance, informing feature selection and model optimisation strategies.The random forest classifier model utilised in this analysis was further fine-tuned using gridsearch CV to optimise model performance. This iterative tuning process resulted in a maximised accuracy score of 0.77.

In conclusion, the insights derived from this comprehensive analysis offer valuable guidance for businesses seeking to enhance customer retention strategies and optimise marketing efforts to mitigate churn effectively.