NOTTINGHAM UNIVERSITY BUSINESS SCHOOL

Academic Year: 2020 – 2021

MACHINE LEARNING & PREDICTIVE ANALYTICS

Coursework

CHURN PREDICTION FOR FOODCORP

Student ID: 20243144

May 2021

CONTENT

N LAB

A EXECUTIVE SUMMARY

This report provides methodology and implementation guideline in defining and predicting churn for FoodCorp. Based on findings from data exploratory and prediction model, some precautionary actions were also suggested to reduce customers' churning risk.

The overall structure includes three main sections, Churn analysis, Prediction model and Insight report.

The first part delivers a summary of ConsultingCorp's report, in which both the pros and cons of their churn definition were analysed. From that, a new way of defining churners was introduced, embracing a global period (38 days) derived from initial analysis as the strict churn delimitation while adopting a customised definition for each shopper based on their regular activity.

The prediction section first starts by presenting the approach in generating variables using tumbling windows. After that, features selection process is described, including a more detailed explanation on the churn definition stated above. By viewing the correlation between features and results from a simple model trial, two set of variables were created, a mixed group, and a purely temporal group contained only window aggerated features. The focus was then shifted to evaluation strategy and results from different models with different input set, in which a tuned version of Random Forest was selected as the final prediction estimator. It was also observed that the mixed set generally assisted better performance due to the involvement of more important features.

The last part of the report provides insights on differences between churners and non-churners, by investigating key inputs that separated them. Pen portraits were also included considering average values of important features in each group. Finally, a more in-depth look towards the churning group helped to identify two different types of potential lapsers and respective actions FoodCorp could take in retaining them.

Details on Data cleaning process, Implementation steps and Figure creation are included in 'Support Document' file. A separate Jupyter notebook with the codes to rerun weekly was also created.

B CHURN ANALYSIS

The purpose of churn prediction is to provide business information to have precautionary actions and improve customer loyalty. However, retention effort, in most cases, requires extra investment from the company that could become fruitless cost if the forecasting is inaccurate due to, not only model performance but also the churn definition. This report, therefore, will start off with an investigation consulting party's report to derive the final prediction target.

In terms of methodology, the consulting firm decided to apply a global churn definition for analysis, in which customers were assessed equally and would be considered as churners if being absent for a fixed, predefined period.

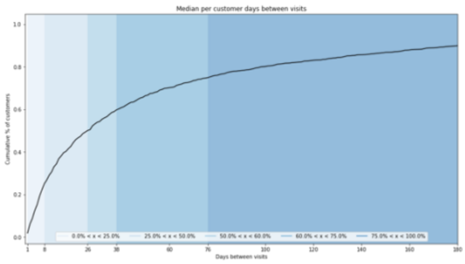


***Figure 1**. Cumulative % of customers based on avg. gap of visit by ConsultingCorp*

The report, then, provided two statistic results. The first one illustrated the cumulative histogram of customers based on their average (median) days between visits (Figure 1). Based on that, 60% of customer shopped at least once every 38 days, while the rest showed much more diverse values of visits' gap.

Note: The charts were reproduced using 'percentile_cont' function instead of 'quantile' to derive median value due to some issues with database

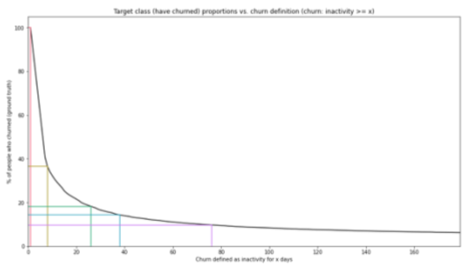


***Figure 2**. % of expected churn with different gap definition by ConsultingCorp*

The second graph (Figure 2) informed the percentage of people labelled as churners among active shoppers. Overall, the longer the defined churn period, the smaller the targeted population. Once again, the 38 days of gap appeared to be the most reasonable option as it stayed at the 'plateau' point of the curve, from which the churn proportion shows insignificant change with increases in inactive day.

Despite helping to generate insights as above mentioned, this approach still shows some limitations in defining churn.

First, while analysing the average gap of inactivity period, the report focused on the **quantity** of customers without considering their contribution to FoodCorp's revenue.

Matching shoppers with their spending contribution, (Figure 3) we can see that 25% of visitors (visit gap < 8 days) were the company's core customers, contributing nearly 75% of total sales over the period. Similarly, nearly 95% of revenue was captured by 60% of customer base with average gap between visits less than 38 days.

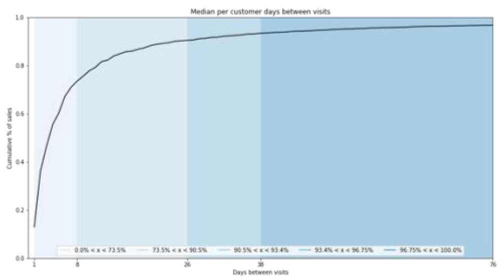


***Figure 3**. Cumulative % of sale based on avg. gap of visit*

Second, as the report had also stated, by using a global definition of churn, it ignored the customers' individual shopping pattern. As shoppers with different visit frequencies contributed differently to business, choosing any fixed churn period could lead to losing a certain number of important customers, as well as costing the company fruitless effort on irrelevant targets.

The recommendation, therefore, is to **keep a global churn period -** the strict point that embraces most of important customers to filter the active base for analysis (as mentioned 38 days), while using a **customised definition** based on each shopper visit frequency to define churn.

The detailed analysis on chosen decrease level versus regularity (prediction output) can be found in Section C – Feature generation approach. In summary, customers labelled as churners if they were predicted, in the next 38 days, come to the shop less than a half (49% specifically) of their average visit frequency in last 5 periods (190 days).

C PREDICTION MODEL

Before developing model, it is important to identify the prediction objective and therefore the evaluation parameters. Due to the unknown cost of both losing and retaining customers, the best estimator was defined as the one delivering most information (least trade-off between True positive and False positive rate). The final decision would then be made with threshold analysis to allow flexible adjustment on the churn rate. As such, Area under Receiver operating characteristic curve (AUC) was selected as the key focus in tuning and comparing models, with accuracy rate also involved to further explain the result.

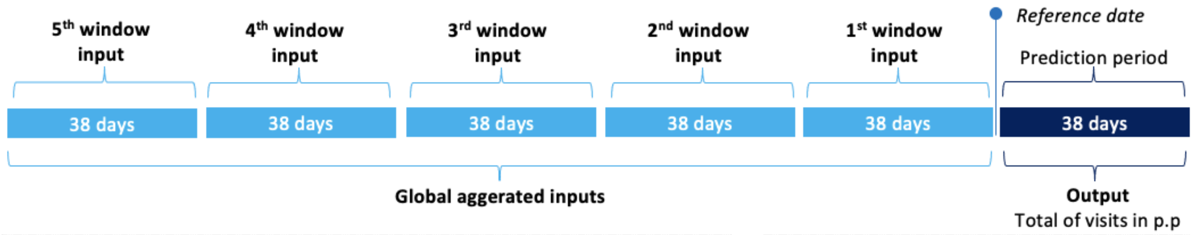
Feature generation approach

Since the prediction is expected to rerun every week, the approach was to set up tumbling windows that contained temporal variables. The size of each chunk equals to 38 days – the global churn delimitation, and the number of chunks considered before prediction is 5 periods– equivalent to customers' average time with the stores (205 days).
To be considered in the **active based**, customers must have **shopped during the 1st window period**.

The input

While the key factors that led to churn were unknown, the strategy was to create **all relevant behaviour variables**, investigating their relationship, selecting the best for modelling while reserving others in pen portrait analysis.

The groups of features, their time boundaries and acronyms are described in Figure 4. Overall, both 'global aggerates' and 'window aggerates' features were included.



GLOBAL AGGERATED INPUTS	
Total spending in last 5 periods	'total_spend'
Total days of shop in last 5 periods	'active_day'
Average basket value in last 5 periods	'basket_value'
Cost per item in last 5 periods	'unit_cost'
Total unique products purchased in last 5 periods	'num_pro'
Average number of visit each period	'average_visit'
Average number of spending each period	'average_spend
Most frequent store: store 0 (binary)	'store_code_0'
Most frequent store: store 1 (binary)	'store_code_1'
Most frequent store: store 2 (binary)	'store_code_2'
Most frequent store: store 3 (binary)	'store_code_3'

WINDOW AGGERATED INPUTS	
Total spending in 1st period	'sale_f1'
Total spending in 2nd period	'sale_f2'
Total spending in 3rd period	'sale_f3'
Total spending in 4th period	'sale_f4'
Total spending in 5th period	'sale_f5'
Total visits in 1st period	'visit_f1'
Total visits in 2nd period	'visit_f2'
Total visits in 3rd period	'visit_f3'
Total visits in 4th period	'visit_f4'
Total visits in 5th period	'visit_f5'

Figure 4*. Features generation approach*

The output

While defining any decrease in customer' average visit frequency as a signal of churn could help to prevent the slightest chance of lapsing, it could also lead to overestimate the churn rate as well as reducing the model accuracy.

To validate this assumption, the churn rate among active customers (output) and model performance were computed with different decrease levels of visit frequency ('average_visit'). 5 sets of repeated temporal hold-out were included to generalise the trend, with previous set acting as training material for following set.

For prediction records, a Random Forest was picked as the baseline model due to its robustness and stability in prediction, with minimum samples per leaf restricted to 30 to avoid overfitting.

The results illustrating in Figure 5 not only prove the above assumption, but also provide suggestion on the decrease level to be selected as churn signal, when the point of 0.5 (equal to half of customer's normal habit) observed a significant change in both the churn base and prediction performance. As the result, the closest index before this point (0.49) was selected as mentioned in Churn analysis.

With chosen approach, the average churn rate per period was 41.8%, considerably high for FoodCorp to provide one-size-fit-all offers for the whole group. Thus, this report will also provide a simple ranking after prediction to help the company decide on their prioritised target and customised offers.

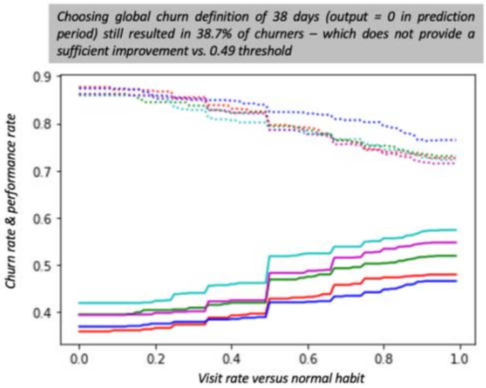


***Figure 5**. Churn rate and model AUC changes with % of normal visit habit*

It is worth noting that with this rate, there was no imbalance between output classes, therefore, no resampling step required in data processing.

Another approach using spending as the output instead of visits was also examined. However due to its widely spread in value even after coding, spending appeared to be less effective option. (Jupyter notebook: MLPA – Finding optimal output)

Feature selection

Feature correlation

A heatmap was plotted based on features' relationship in train set (ref date: '2020-09-09') which included 929 samples and 39% churn rate. (Figure 6)

Overall, while no single variable firmly decided the churn risk, most of extracted features showed medium impact to the output and helped to provide both business and modelling insights.

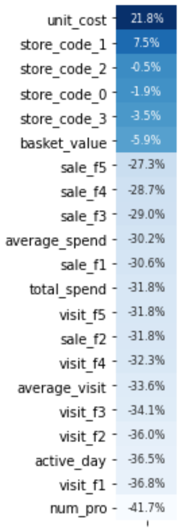


***Figure 6**. Correlation between input and output feature*

Cost per unit surprisingly showed a medium positive relationship to the output, and slight negative relation to customer value features, implies that customers purchased high-value products might belong to one-off visitor group and had higher churning possibility.

Static features regarding **store location** appeared independently to other features, including prediction target. Thus, they were removed from input pool.

Higher customer' value including record in **monetary** ('total_spend'), **frequency** (average and by-period visit and spending) and especially **loyalty** ('active_day', 'num_pro') indicated less chance of churning.

Inputs in **visit** and **spending** delivered similar effect in identifying target, with visit showed a slightly higher advantage, which is understandable as the output was defined by frequency of purchase. Since these two groups also greatly correlated with each other, spending features ('total_spend', 'average_spend', window sales) were removed to avoid the duplication of similar direction entities.

Within visit frequency group, the closer the observation period, the higher negative correlation it had towards the output, showing that temporal features could help in predicting churn when being considered as chain of events. To examine this assumption, remained variables were arranged in 3 groups for further examined as show in Table 7.

***Table 7**. Feature sets*

'All' set	'Global aggerates' set	'active_day', 'basket_value', 'num_pro', 'unit_cost', 'average_visit'
	'Window aggerates' set	'visit_f1', 'visit_f2', 'visit_f3', 'visit_f4', 'visit_f5'

Feature sub-group in prediction

To observe their prediction power, 3 groups of features were fed into modelling. The results including their average accuracy and AUC were derived through 5 repeated hold-out sets, up to the validation date ('2020-09-09'). Once again, Random Forest (min samples leaf=30) was used as the baseline model.

	Accuracy	AUC	Methodology
All	75.8%	82.1%	RandomForestClassifier(random_state=42, min_samples_leaf=30)
Global aggerates	75.6%	81.6%	
Window aggerates	75.3%	81.6%	
Benchmark	52.4%	51.1%	DummyClassifier(random_state=42, strategy='stratified')

***Table 8**. Accuracy and AUC on different feature sets*

Table 8 illustrates the outcomes of this process. Overall, while both the 'global aggerates' and 'window aggerates' groups provided similar and stronger results versus a benchmark of dummy classifier, their combination showed a slightly higher performance.

Based on above results, the rest of this section will consider model optimisation for 2 input sets, the 'all' and the 'window aggerates'.

Model evaluation strategy

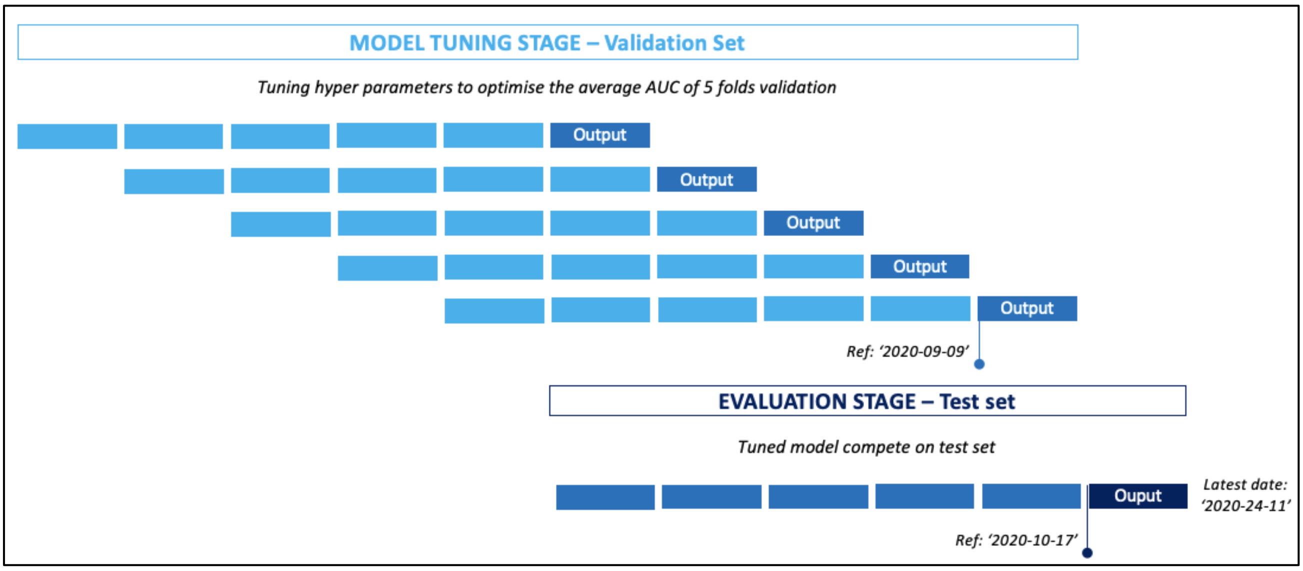


***Figure 9**. Model evaluation strategy*

Figure 9 describes the model evaluation strategy for this project. Overall, this included 2 main stages: hyper parameter tuning and model comparison.

In the first step, chosen models were iterated with different parameters over 5 temporal hold-out sets, up to the validation reference date of '2020-09-09'. The average performances were then used to select the best indicators for each model.

In the comparison part, optimised models were used to predict the test set. Results from both the training set and test set were taken into consideration. The selected model was expected to show not only best performance in the test set, but also stability versus the training result.

2 groups of features were separately examined through different algorithms to select the best fit ones. This included Random Forest (RF), k Nearest Neighbors (kNN), Decision Tree (DT) and Logistic Regression (LR) for the 'all' set, and RF, kNN (with extra parameter for dynamic time warping) and Time Series Support Vector Classifier (SVC) for 'window aggerates' set. For distance-based kNN model that applied for 'all' set, features normalisation (MinMaxScaler) was conducted in advance to balance variables' unit (Table 10).

Model evaluation result

	Tuning stage		Comparison stage (AUC)	
	Model	**Best parameters**	**Train set (5 folds)**	**Test set**
'All'	RF	max_depth = 2 n_estimators = 50	82.2%	83.6%
	kNN <with scaling>	metric = 'l1' n_neighbors = 18	80.4%	80.2%
	DT	max_depth = 5 min_samples_leaf = 70	81.1%	82.2%
	LR	C = 0.001	81.5%	81.7%
'Window aggerates'	RF	max_depth = 4 n_estimators = 95	81.7%	83.6%
	kNN	metric = 'l2' n_neighbors = 19	80.4%	80.1%
	SVC	No tuning due to high computation expense	None	82.2%

***Table 10**. Model evaluation results*

Evaluation results can be found in Table 10. Overall, Random Forest slightly overperformed other models in both features set and consistently with different periods.

Another observation was that, for kNN, more simple metrics ('l1','l2') showed better results versus Dynamic Time Warping method, which is understandable as the number of windows was small and consistent between samples.

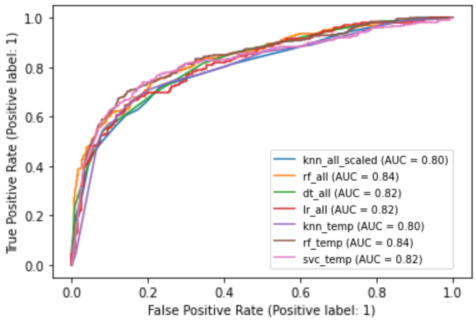


***Figure 11**. Receiver operating characteristic curve (ROC) Between models in test set*

Plotting the ROC of all models on test set (Figure 11) revealed more insights on each algorithm performance.
RF won in both input set thanks to its capability to achieve better results in precision when recall rate stayed in upper range from 78% to 95%. Outside of mentioned ranges, all models produced similar outcomes in predicting the output, with kNN and SVC suffered the most in low threshold (higher recall) dimension.

To ensure a fair competition between RF in 'all' and RF in 'window aggerates' set, their tuned versions were fitted to 10 repeated temporal hold-out data up to latest date. The results (Figure 12) showed that overall, the 'all' set delivered better and more consistent results in both accuracy and AUC, hinting that some global aggerated features added extra information in predicting churn versus the solely temporal set.

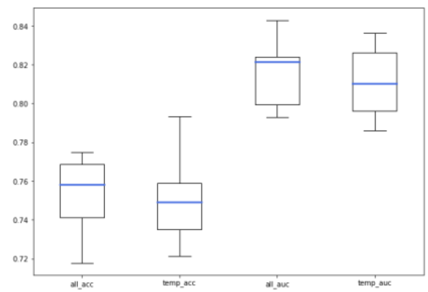


***Figure 12**. Accuracy and AUC by tuned RF in 'all' and 'window aggerates' sets*

In conclusion, Random Forest (max depth = 3, estimators = 50) and 'all' set were selected as the final combination for prediction. At the threshold of 0.5, the model achieved 77.1% in accuracy, 73.3% in precision and 66.6% in recall for the test set. Applying the model to most update data (reference date = '2020-11-24') resulted in 34% of active customers (300 people) labelled as churner. As discussed, such proportion would be too high and costly for FoodCorp to deploy retention program. Not to mention, the list might contain different types of churners that should not receive same offer.
Therefore, a customer grouping task based on churn prediction probability was made to not only short-list the key target for business, but also provide information to understand churner and non-churner profile.

Churn profile & business recommendation

While looking at the profiles of 'real' churners or non-churners would help in differentiate the two groups, further analysis on the model mechanism and 'predicted' labels at the latest data allows more in-depth understandings on key features that define them. Analysis and implementation steps that derived below findings were included in the Technical report section.

Based on that, churners were identified by their **time with the stores**, as the most important signals in defining churn were the **number of visits in last period** ('visit_f1'), total of **shopping days** ('active_day') and **number of unique products** purchased ('num_pro') (Figure 14 - SHAP calculation). Specifically, they were **new customers** that mostly had just started from last period, with an average of 2 shopping days and around 11 purchased items during that time.

Another highlight is that despite having inferior customer values, the cost per item ('unit_cost') among this group was significantly higher than the non-churners (£3.8 versus £1.7). This indicates that rather than shopping daily essentials, churners were more interested in higher value items.

To help business utilise their resources in targeting, priority ranking based on predicted churn probability was added to customer profile. By that, two smaller groups of potential lapsers were detected – the 'First time' (Group 1,2), and the 'Rising interest' (Group 3,4,5) customers. Overall, while the 'First time' shoppers had only visited the store once, purchasing high value but possibly non-essential items, the second group had repeated purchase with the store and a highly resembled behaviours ('unit_cost', 'basket_value') with the non-churned groups. Hence, it is suggested that FoodCorp take customised approach towards them.

For 'first-time' customers with highest risk and limited exposure to the store, it is essential to attract them come back on second purchase. High investment can be put into these customers as their current proportion is relatively small (17% customer base). Special offers, promotions or product sampling could be useful, especially for more essential products to drive them on the track of loyal customers.

For 'rising interest' group that had passed the one-off barrier, it is more critical to provide them convenience and satisfaction when shopping. As their number of purchased products was quite significant, a recommender system could help in suggesting them relevant products. Bundling or cross-selling based on their current basket is another option to further increase the engagement.

On the final note, beside recommended activities on retention, FoodCorp should also improve its customer's records. The current database is lacking 88% of shoppers' birthdate, as well as having no information on customers' contact and occupation. This prevents several strategies to be deployed, such as providing reminders for customers to comeback or informing of promotion and offers. As different segments of people often possess different shopping behaviours, added information on background could also help improve the prediction model.

	**NON-CHURNERS** GROUP SIZE: 66% VALUE SIZE: 96%		**CHURNERS** GROUP SIZE: 34% VALUE SIZE: 4%
Period (38 days) mean stats		**Period (38 days) mean stats**	
ACTIVE DAY: 22 days	NUM PRO: 84 products	ACTIVE DAY: 2 days	NUM PRO: 11 products
AVERAGE SPEND: £72	BASKET VALUE: £17	AVERAGE SPEND: £16	BASKET VALUE: £18
AVERAGE VISIT: 4 times	UNIT COST: £1.7	AVERAGE VISIT: 1 time	UNIT COST: £3.8

***Table 13**. Churner and Non-churners pen portraits*

Technical report

The first step in identifying key differences between churners and non-churners was to understand the key features that supported prediction. SHAP values calculation (SHapley Additive exPlanations) was selected for this purpose due to its capability in delivering not only feature's importance but also the relationship direction (negative or positive) between inputs and output.

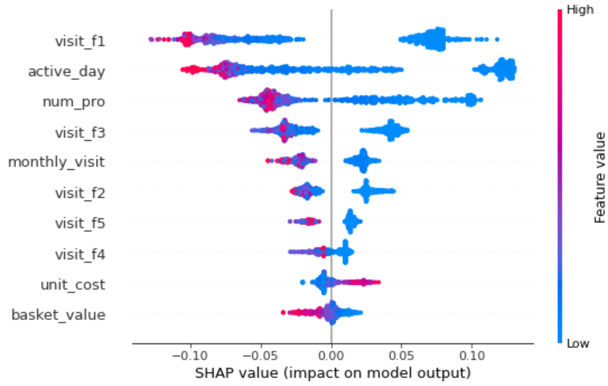


***Figure 14**. Feature importance in test set by SHAP*

Applying SHAP on the final version of Random Forest, overall, the model focused on detecting new, one-off customers as this group possessed highest churn risk. This showed in the top ranking of 'visit_f1' and 'active_day' variables.

Despite having the highest negative correlation with the output, the number of unique products purchased ('num_pro') lost its first position, mostly due to its high correlation with the active days, and therefore did not provide the model significantly added information in defining churners.

Another observation from the plot was that 'active_day' delivered the clearest breakpoint between churners and non-churners, indicating that it helped to optimise model's 'precision' and 'accuracy'. Similarly, 'basket_value' and 'unit_cost' appeared to further increase them. This explained why 'all' features assisted better performance compared to the 'window aggerates' group as seen in the Model evaluation section.

The second step in deriving pen portrait insights was comparing the profiles of two groups. To provide an intuitive comparison, the **average values** (mean) of key features were **standardised** and grouped to respective labels. Polar charts were then generated illustrating these values. The detailed and unstandardised average stats of each group can be found in Jupyter notebook – MLPA Final model evaluation.

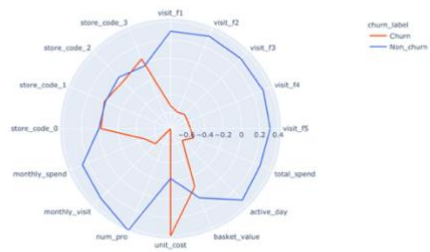


***Figure 15.** Features' average value (standardised) in churned and non-churned groups*

In the final step, customers were labelled with ranking in target priority, with 1 as the most likely and 9 as the less likely churned group. Figure 16, 17, 18, 19, 20 respectively illustrate key statistics of the 9 groups, including their proportion,'total_spending', 'active_day', 'unit_cost', 'basket_value' which helped to fuel the findings in Insight section.

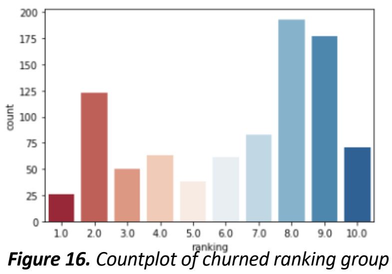


***Figure 16.** Countplot of churned ranking group*

Group	Churn prob.	% Size	% Revenue	Group	Churn prob.	% Size	% Revenue
1	>= 90%	2.9	0.05	6	40-49%	6.9	1.6
2	80-89%	13.9	1.0	7	30-39%	9.4	3.0
3	70-79%	5.6	0.6	8	20-29%	21.8	15.3
4	60-69%	7.1	1.1	9	10-19%	20.0	31.6
5	50-59%	4.2	1.0	10	<10%	8.0	44.7

***Table 17.** Key stats of churned ranking group*

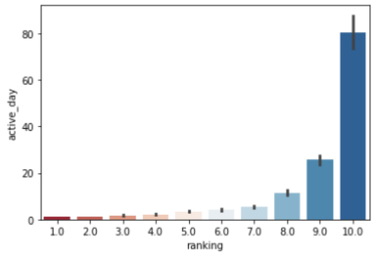


***Figure 18.** Average active days of churned ranking group*

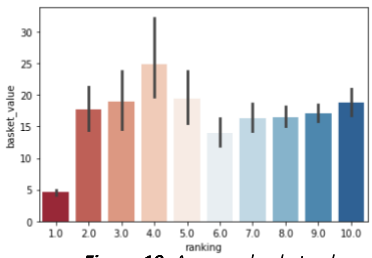


***Figure 19.** Average basket value of churned ranking group*

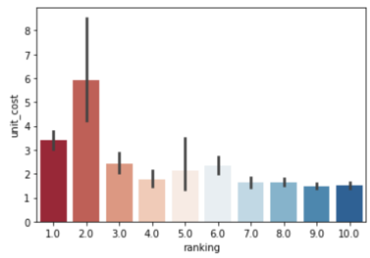


***Figure 20.** Average unit cost of churned ranking group*