Nottingham University
Business School

UNITED KINGDOM · CHINA · MALAYSIA

Module title: Machine Learning & Predictive Analytics (BUSI4373)

Assignment title: Machine Learning Coursework- Churn Prediction for FoodCorp

Submitted By: S M Sazid Islam

Student ID: 20350823

Date of Submission: 12th May 2022

Executive Summary

The word “churn” is derived from change and turn which means discontinuation of a contract. In terms of a business, a churned customer is referred to a customer who has either undergoing a big pause with a company or have completely stopped using product or service of the company. Customer churn is one of the most important metrics for a growing business to evaluate because it defines the customer retention and satisfaction level with a company. Every company should aim to minimize the churn rate and keep it as close to 0%. Otherwise, companies experiencing high rates of customer churns might face severe consequences of extreme loss of revenue compelling the company to close the business. In order to predict customer churn, collection of historical consumer data is very important for which big companies like TESCO introduces membership cards (CLUBCARD) by which they can keep track of the customer’s behavior and analyze them to predict churning customers. Hence, collecting clean and accurate data is very much important to make proper analysis based on which further investment plans will be made to reduce the churn rate. This report is a detail analysis of churn prediction of customers based on the consumer behavioral data provided in a given time frame of the company named FoodCorp. The analysis is performed in Google Colab using Python programming language.

This report outlines the outcome of the analysis of customer churn prediction performed which includes, preparing the dataset and creation of a temporal dataset to perform the analysis, implementing classification model to predict the accuracy of class prediction. Based on the type of dataset, classification models like Random Forest Classifier and Support Vector Classifier (SVC) models has been used to assess the prediction. Moreover, further analysis like best model prediction and feature importance is also performed. Finally, the number of churned and non-churned customers has been predicted based on which FoodCorp can make further decisions.

This report is made on the basis of a very important factor called churn prediction for which the accuracy also needs to be maintained as a lot of decisions are related with this analysis. With accuracy comes challenges to get that accuracy for which necessary steps has been taken such as pre-processing the dataset by creating tables and sub dividing the customer information into temporal features so that the changes in features can be properly assessed. This involves the use of 3 important parameters, tumbling window size, output window size and reference date. The tumbling window size is taken as 7 days as FoodCorp want to predict the churners in 7 days period (a week). These parameters are essential to conduct the analysis with precision. In addition, when the models are used to assess the prediction, the business context was considered when finalizing the best fitted model. As well as, the type of the features used are crucial in helping FoodCorp grow in the long run and produce effective marketing campaigns to stop customers from churning and increase the loyal customer base.

The main aims of this report are to make the company, FoodCorp aware of the customers who are at churn risk by predicting as accurately as possible, helping the brand to grow by highlighting

the areas of importance and assisting in better budgeting of marketing campaigns which will allow the company to save cost as more specified target group can be influenced.

Current Levels of Churn:

To understand the current levels of churn, the individual report provided is considered and the findings are discussed based on the figures in the report. Figure.1 indicates that 75% of the customers makes repeat purchase or visits the store again when the days are more than 55 days. The graph can be observed to flatten after this point by which it can be stated that customers are returning.

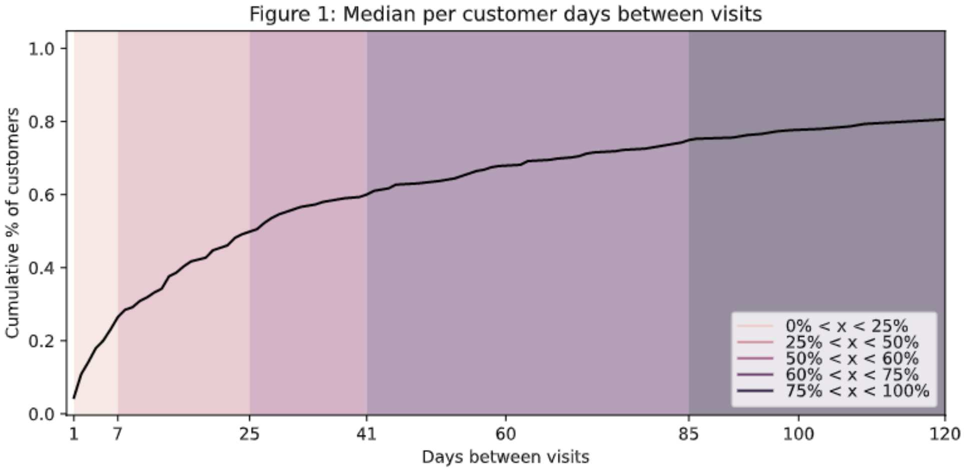


Figure 1: Median per customer days between visits

On the other hand, from figure.2 this can be noticed that from 80 to 100 days of inactivity, the number of returning customers has become stagnant and do not increases leading to a straight line.

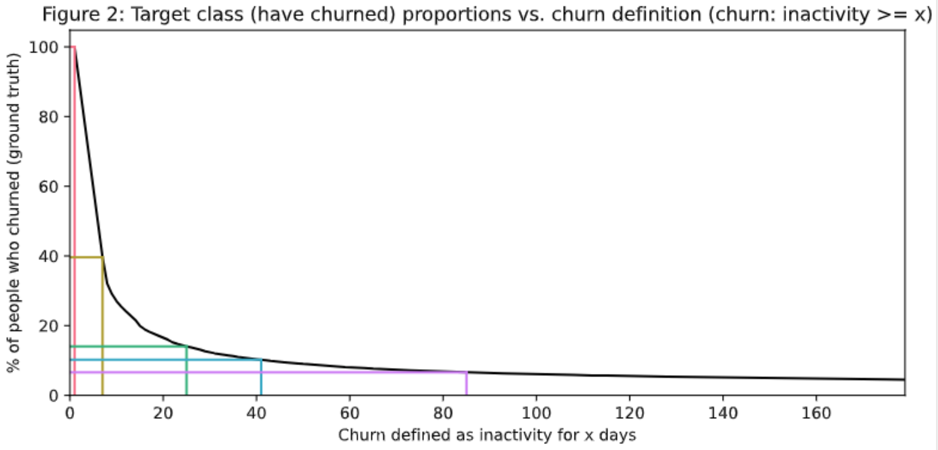


Figure 2: Target class (have churned) proportions vs churn definition

As a result, after analyzing the above figures this can be concluded that targeting mass people for the marketing purpose will be extremely expensive and might not be a good option for FoodCorp. Rather, in order to identify the churners and non-churners it is better to take fewer days as churn definition which will cover less people and more specified targeting can be done.

As per the report, 27 days is selected as churn definition which will cover less people, giving more accurate result and covering 58.7% of customers can give the company a better snapshot of the prediction of churn and non-churn customers.

Technical Approach:

This section includes the approach made to perform the analysis, creating the temporal dataset by producing suitable tables, implementing models, final prediction and feature importance.

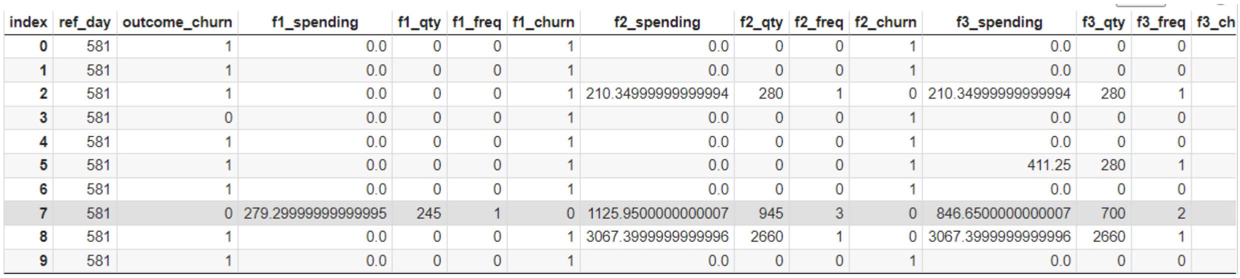
Creating temporal dataset:

The dataset provided consist of a number of features which were broken into temporal features to analyze the behavior of the customers against time frame. The whole analysis was conducted in Google Colab using sqlite3 as the file provided was in SQLite format. At first, the file named ml17.sqlite is uploaded in the Google colab notebook which is provided with this report. The tables present in the dataset is explored and then features are considered for the evaluation which are quantity of products purchased by each customer extracted using "qty", the number of times visited (frequency) of each customer using "freq" and total value contribution of each customer to the company by using "value". These are the factors which defines the churning behavior of a consumer which will help us to understand the changes in behavior and predict accurately. In order to evaluate the features, tables are created, (New_total_table) which gives the total quantity and total value of each customer. Next, a table named (Master_table) is created and all the other features taken are inserted into that table which is then converted into a data frame.

In order to find the reference date, maximum and minimum dates of the dataset provided is identified and using the Julianday python library the reference day is selected which is 608 according to the dates of the dataset. To figure out the change of the features with time, windows are created using python and parameters such as tumbling window size and output window size has been used. Since, the task is to identify the changes in behavior of the customers on a span of week, the tumbling window size is taken as 7, this will allow the company to know the customer's actions over a week's duration. The value of output window size is considered as 27, this is taken same as churn definition because it will let the company know whether the customer churned or not in the duration of churn definition.

To proceed with the analysis, a function is created (get_dataset) which creates a tumbling window of 7 days for 3 periods and the input and output features are returned. The dataset is then separated into x and y and the results of the temporal dataset created is visualized where the output feature is 0 or 1 showing that customers has churned or not and the frequency and quantity for 3 periods are also presented.

The Temporal Dataset created is shown below:



index	ref_day	outcome_churn	f1_spending	f1_qty	f1_freq	f1_churn	f2_spending	f2_qty	f2_freq	f2_churn	f3_spending	f3_qty	f3_freq	f3_ch
0	581	1	0.0	0	0	1	0.0	0	0	1	0.0	0	0	
1	581	1	0.0	0	0	1	0.0	0	0	1	0.0	0	0	
2	581	1	0.0	0	0	1	210.34999999999994	280	1	0	210.34999999999994	280	1	
3	581	0	0.0	0	0	1	0.0	0	0	1	0.0	0	0	
4	581	1	0.0	0	0	1	0.0	0	0	1	0.0	0	0	
5	581	1	0.0	0	0	1	0.0	0	0	1	411.25	280	1	
6	581	1	0.0	0	0	1	0.0	0	0	1	0.0	0	0	
7	581	0	279.29999999999995	245	1	0	1125.9500000000007	945	3	0	846.6500000000007	700	2	
8	581	1	0.0	0	0	1	3067.3999999999996	2660	1	0	3067.3999999999996	2660	1	
9	581	1	0.0	0	0	1	0.0	0	0	1	0.0	0	0	

Model Evaluation:

The task involves classification of the customers to churners and non-churners with the main goal to identify the churners and invest on them to prevent from being churned and improve relationship of FoodCorp with them and enhance customer satisfaction. Classification models, Random Forest Classifier and Support Vector Classifier (SVC) has been used for evaluation.

Random Forest Classifier is a powerful and easily adaptable supervised machine learning algorithm that grows and combines a number of decision trees and creates a forest. It can be used for both classification and regression purposes. It is used here because it is a classification analysis and it do not need any prescaling and standardization which makes it suitable for the dataset provided. As well as, it also gives an accurate accuracy score with a large dataset. Another reason for using Random Forest Classifier in this dataset is that the algorithm results in greater tree diversity, which trades a higher bias for a lower variance while producing a better evaluation model.

For Random Forest Classifier, train-test split technique is used to estimate the performance of the algorithm, using train-test split we split the dataset into two subsets, the first one used to fit the model called the training dataset and the second subset is the test dataset where the input element of the dataset is provided to the model, making predictions and compared to the expected values. The n_estimator value is taken to be 120 trees as this will provide the best accuracy and this the highest mark to improve the score. The dataset is tested and trained and the performance of the model is measured based on accuracy, precision and recall values.

The next model used is Support Vector Classifier (SVC) model, this is used to reduce the overfitting of dataset along with the kernel “rbf” can identify outliers properly. This is also helpful when working with a large dataset. Here, the value of gamma is taken as 1 because it might affect the decision boundary and resulting in overfitting issue. To perform SVC, classify_class_attrs library is imported and the accuracy, recall and f1-score values are assessed to determine the performance of the model. Accuracy is important because it describes the closeness of the prediction to the actual values. Precision defines the positive instances among the instances considered and Recall value indicates the relevant instances that were properly retrieved.

To fit the most recent date in the models, ‘now’ entity with the max date is entered based on which the behavior is measured. For model evaluation, x and y are tested and trained by subtracting the output_window_size from the now date. In addition, the model is fitted using fit command.

When evaluating the results of a classification model, making decision based on the accuracy only will not be proper. Other parameters like precision is also important because it allows the company to understand the cost of false positive, it can identify whether the investment made is made on the proper customers or not. This will stop the company from spending money on wrong customers and make more cost saving and effective marketing plans. While, making an important decision of investment by a company it is very important to identify the number of actual positives among the positives captured by the model. Because this can have the negative effect which might lead the company to lose a potential loyal customer as the company will then consider a non-churner as churner.

Using Random Forest Classification model, the accuracy is found to be 83% which implies that the model is a well-fitted model, while the Precision and Recall values are also 83% as per the weighted average. This indicates that this model is fitting properly with the dataset. On the other hand, using SVC model, the accuracy is 71% which is also a good fit but since Random Forest has higher accuracy, this is selected for the churning.

After selecting the best model, pen portraits of the churners and non-churners are produced to identify the mean, maximum, minimum and majority behavior of the two types of people.

Churn Prediction:

The max date which is the most recent date, 2022-03-22 is considered as “now” and the churning customers for the next 27 days are evaluated. We considered 1 as churner and 0 as non-churner. The number of churners (1) is found to be 392 and the number of non-churners (0) is found to be 975. This shows that the company still has time to work on reducing the churn rate as churners are less than non-churners, however this is a big group of churners, about 30% of the population considered for the prediction is churners, this should be minimized to reduce the loss of revenue through proper customer engagement.

Feature importance:

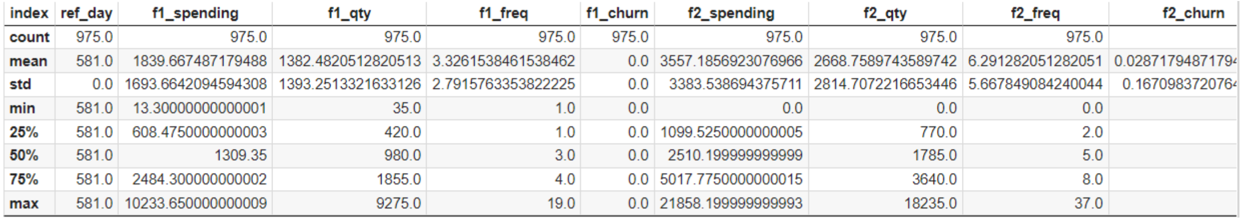
The features considered for the analysis are further assessed by checking the top 10 temporal features among the features used. Prioritizing the features will allow us understand the main factors causing the churn of customers and the company will be able to identify the areas of improvement. Features that impacts the revenue of the company directly are considered in this analysis such as value, frequency of visits and quantity, these are vital but differs from one another for which it is important to identify which are the main issues of the churning as huge marketing and management cost is involved with each improvement. Based on the analysis performed, spending has the highest rank which shows that revenue (value) is the most important feature among the others

Once the features have been identified, to funnel out and keep the best features only, the features with low variance with variance threshold set to 0.0 are removed. This is done to specify the main features and so that it is clearer for the business to focus and reduce the number of churners.

Insights from the Analysis:

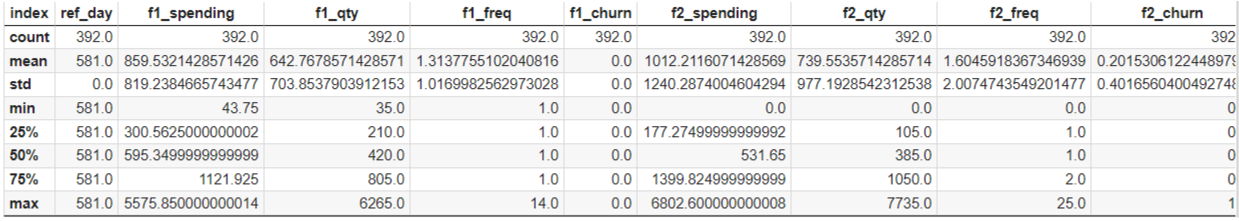
This section involves the insights from the analysis performed as well as how the business scenario is related with the outcome. This will also allow the company to save money by not focusing on huge customer base while creating marketing campaigns. Hence, this is can noticed that out of 1367 active customers, 392 customers are churners and the other 975 people are non-churners and the vital features which caused the outcome are frequency and revenue value. To summarize the analysis, the analysis made the prediction on 1367 people which are detected as active customers. Initially, the dataset provided was processed to set the dataset for performing the analysis, then machine learning algorithms are applied for churn prediction and classification models are used to understand the accuracy of the prediction. Based on which, pen profiles are created to get a clear vision of the scenario and feature importance helped to identify the importance features related to the analysis. Finally, the churners are predicted, these are the people who might end relationship with FoodCorp. One way of stopping the predicted churners from churning can be the use of email and SMS marketing, sending promotional information of bundle offers designed specifically for each churn individuals based on their likes and interest derived from the shopping history data of these predicted churn customers.

Using pen profiling, we have analyzed the majority behavior of the churners and non-churners. The output of the pen profiling of predicted non-churners is visualized below:



index	ref_day	f1_spending	f1_qty	f1_freq	f1_churn	f2_spending	f2_qty	f2_freq	f2_churn
count	975.0	975.0	975.0	975.0	975.0	975.0	975.0	975.0	
mean	581.0	1839.667487179488	1382.4820512820513	3.3261538461538462	0.0	3557.1856923076966	2668.7589743589742	6.291282051282051	0.02871794871794
std	0.0	1693.6642094594308	1393.2513321633126	2.7915763353822225	0.0	3383.538694375711	2814.7072216653446	5.667849084240044	0.1670983720764
min	581.0	13.30000000000001	35.0	1.0	0.0	0.0	0.0	0.0	
25%	581.0	608.4750000000003	420.0	1.0	0.0	1099.5250000000005	770.0	2.0	
50%	581.0	1309.35	980.0	3.0	0.0	2510.199999999999	1785.0	5.0	
75%	581.0	2484.300000000002	1855.0	4.0	0.0	5017.7750000000015	3640.0	8.0	
max	581.0	10233.650000000009	9275.0	19.0	0.0	21858.199999999993	18235.0	37.0	

The output of pen profiling of predicted churners is shown below:



index	ref_day	f1_spending	f1_qty	f1_freq	f1_churn	f2_spending	f2_qty	f2_freq	f2_churn
count	392.0	392.0	392.0	392.0	392.0	392.0	392.0	392.0	392
mean	581.0	859.5321428571426	642.7678571428571	1.3137755102040816	0.0	1012.2116071428569	739.5535714285714	1.6045918367346939	0.2015306122448979
std	0.0	819.2384665743477	703.8537903912153	1.0169982562973028	0.0	1240.2874004604294	977.1928542312538	2.0074743549201477	0.4016560400492748
min	581.0	43.75	35.0	1.0	0.0	0.0	0.0	0.0	0
25%	581.0	300.5625000000002	210.0	1.0	0.0	177.27499999999992	105.0	1.0	0
50%	581.0	595.3499999999999	420.0	1.0	0.0	531.65	385.0	1.0	0
75%	581.0	1121.925	805.0	1.0	0.0	1399.824999999999	1050.0	2.0	0
max	581.0	5575.850000000014	6265.0	14.0	0.0	6802.600000000008	7735.0	25.0	1

Revenue(‘spending’):

Based on the outputs of the report, the first feature compared is revenue generation by both the groups. It can be stated that the churners spend on an average (£859 to £1083) in the stores of FoodCorp in the time frame while the non-churners spend on average (£1839 to £5170) on their regular visits to FoodCorp, this shows that there is a very significant difference between non-churners and churners in terms of revenue contribution of FoodCorp. It is visible from the average spent by the two group of customers that the non-churners have a better relationship with the brand than the churners, they find FoodCorp more reliable and more trusted place to shop from. While, the churners do not tend to spend more. While comparing the maximum spent by each group, the maximum money spend by churners in a week time is (£5575 to £10852), while for the non-churners the maximum spend is (£10233 to £31854). This also indicates that the there is a churn risk associated with the customers spending less. Revenue is very much important aspect of a business as it ensures long term growth and existence of a company.

Quantity(‘qty’):

Quantity is the total number of items bought buy a customer in the time frame provided. According to the analysis, shoppers who are tend to churn buys 642 to 800 items on an average while the non-churners purchases 1382 to 3873 quantity from the stores. This decline in items purchased indicates that churners are considering a second brand to meet their demand and needs while the non-churners prioritize this brand as their regular shopping brand. The maximum quantity purchased by the churners also represents that they are not satisfied with the brand and might be in a stage of either looking for a new brand or they buy from FoodCorp as a secondary option.

Frequency(‘freq’):

Frequency is the number of visits made by a customer in a week time for this analysis. Here, analyzing the frequency of visits among churners and non-churners can make us understand how frequently the customer makes repeat purchase. The average frequency of the churners is 1 to 2 times while the mean frequency of non-churners is 3 to 9 times. This shows that shoppers how are tend to churn visits FoodCorp occasionally and is not a loyal customer of the business. If the maximum number of visits is viewed, churners visited 14 to 33 times during the time frame and the non-churners visited 19 to 57 times. This big difference clearly indicates how the shoppers who are tend to churn are at churn risk and might end relationship with the brand in short period of time.

The graph shows the distribution of customers who are predicted to churn (churners) and who are predicted to retain (non-churners) is shown using bar and pie charts:

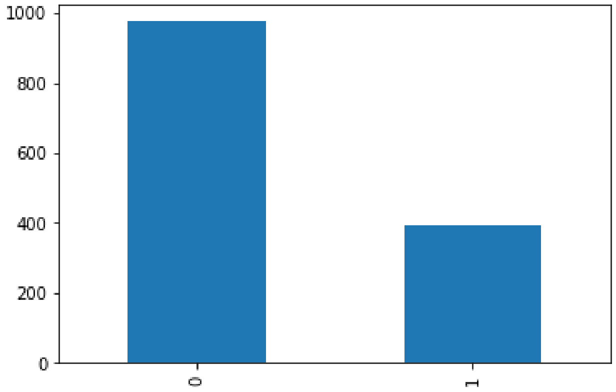


Figure 4: Bar chart to show the distribution of Churn and Non-churners

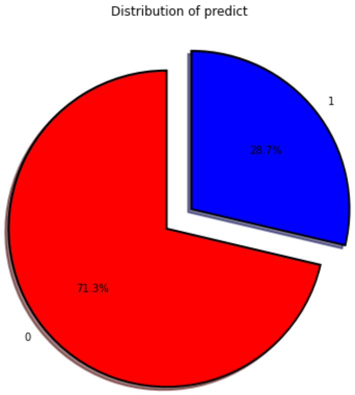


Figure 3: Pie chart to show the distribution of Predictions

The above figures show that the non-churners who are retaining customers is higher in percentage than customers who are high at churn risk. The percentage of possible churners (1) is 28.7% and the percentage of loyal or retaining customers (0) is 71.3%. From this distribution this can be perceived that the company is doing well but should focus on minimize the tendency of the high-risk churn customers as otherwise, they will experience loss of customers.

The above insights were derived from the analysis performed using Google colab and Python. The variables used are at first identified from the dataset provided, then using machine learning models and pen profiling, these insights were discovered and displayed in this report.

Conclusion:

In conclusion, this report was made based on the churn prediction analysis performed in the Google colab notebook provided with the report. The best efforts have been made to highlight the customers who are at high churn risk. However, better prediction could have been achieved using bigger dataset and with more machine learning approaches. The present scenario of FoodCorp according to this report indicates that they are doing good in the business as 71% of the population considered are non-churners. However, they should be aware of the customers who are at churn risk because if steps have not been taken to reduce the number of predicted churners, this group might influence the non-churners causing FoodCorp to lose potential customers and more importantly revenue will be impacted.