MACHINE LEARNING AND ADVANCED ANALYTICS

COURSEWORK

SPRING SEMESTER 2019-2020

STUDENT CODE: 20184267

SECTION A: EXECUTIVE SUMMARY

FoodCorp is a medium sized retail chain operating in four stores across three cities. We received datasets comprising all transactional records and relevant customer metrics. The company's interest lies in identifying churners based on their purchasing behaviour over preceding time periods. So, a predictor model has to be constructed so as to target the customers likely to churn and re-engage them.

To deal with this temporal problem, after data exploration and cleaning, a set of features were engineered and selected so as to generate temporal hold out sets for training and testing of the three chosen predictor models - Linear Support Vector Machine, Random Forest classifiers and K-nearest Neighbours classifiers. The training and validation sets were constructed using the function defined for holdout set generation. Then over repeated training holdouts, the models were trained and tested on the corresponding validation holdout. Since the primary objective was to predict churners accurately and minimize misclassification of target class (churners) as active customers, the evaluation metrics of classification accuracy, sensitivity and precision were picked to assess the three models and pick the best performing model, which here, was Random Forest. This model was then trained on the previous holdout and was tested on an unseen test holdout set to gauge its prediction accuracy. The final model was then built according to the requirement of weekly runs. With a separate function to return input holdout set for each week, this model predicts the churn of a customer over the next week based on the previous holdout.

With tailored deals and offers, the average spends and frequency of store visits are likely to improve for customers who because of low engagement are likely to churn. The second recommendation would be to provide loyalty benefits to high and regular spenders to keep them engaged and ensure their retention. Depending on the accuracy of the predictions, if the marketing strategy is directed towards eventual churners efficiently, Foodcorp is likely to increase its revenues from a higher retention of active customers.

SECTION B: INITIAL CHURN RATE ANALYSIS

FoodCorp has done fair introductory analysis to help us understand the pattern of customers' store visits and the percentage of customers predicted as churners by the predictor model over the corresponding period. The two statistics explained are the distribution of time between consecutive visits of a customer and the percentage of customers predicted as churners. The first figure shows the variation in the percentage of customers corresponding to the median number of days between visits. As one can observe, the curve is parabolic and after increasing until a certain point, it becomes parallel to the X-axis. This suggests that after a certain point in time, a select portion of the customer base is likely to never return no matter how many days pass by. The second figure corresponds to the target class predictions made by the model. According to the figure presented, the proportion of the predicted target class initially drops significantly with increase in the duration of the inactivity period. Similar to the first figure, after a certain point, the predictor predicts the same proportion of churners even for an increase in the inactivity period. As per the figure, the analysis suggests that the company will have to target 88.22% of the customers who will churn if the inactivity period is 1 day and 11.66% for 57 days. It can be observed that the company will have to target a lesser proportion of the customer base if it expands its churn definition window up till a point after which the rate becomes stagnant. This however comes at a risky trade-off of losing the engagement of customers and possible revenue associated with them.

So, keeping in mind the importance of targeting customers who are likely to churn without incurring heavy expenditures from unnecessary targeting, a churn definition of 30 days was picked for the analysis since almost 60% of the customers return within this period and the predictor identifies around 17.58% of the active customers as likely churners and targets them. What drives this choice is the tendency of more than half the customers to shop for their supplies once every 30 days on an average. The statistics associated with this churn definition provides an ideal trade-off between correctly targeting identified churners and avoiding increased expenditure from unnecessary targeting of customers based on a short period of inactivity because they are likely to return sooner.

SECTION C: TECHNICAL IMPLEMENTATION

a) Data exploration and cleaning:

The five transactional datasets provided to identify trends in customers' purchasing behavior are customers, receipts, receipt_lines, products and stores containing information regarding the customer, records of visits, products, associated details and store related information respectively. The features that would most distinctively help pick a churn predictor model were picked from these datasets after data exploration.

Since most data cleaning was already done, exploration highlighted only one issue. of eliminating the rows in the receipt_lines table where values had been wrongly entered as 9999999. There were no missing values in the tables we used for our analysis.

b) Feature selection:

The approach required an ordering of the purchase dates in the given dataset. A new table dateranks was therefore created to account for the day number corresponding to each purchase date and was subsequently used in further analysis. Of all features in the provided datasets, the ones that seem relevant to predicting a customer’s possibility of churning are the quantities of products he is purchasing, the amount he is expending on the same, the frequency of his visits and the duration between two consecutive visits. So, using the ordered days from dateranks and the customer’s purchasing behaviour, a new table temp_features containing all the four features was constructed for feature analysis.

For every first visit of a customer at a store, the gap feature, partitioned over customers and stores, was modified to store the duration between two consecutive visits as zero. Using python, temporal holdout generation considering all features was then done. The four features considered for each time-period were the frequency of visits, sum of the gaps between two consecutive purchases, sum of the quantities of items and the sum of the values associated with the purchases over the period. The output feature has the target class as churners(0) and non target class as active customers(1). The output feature for a customer in period i was evaluated as 0 or 1 depending on the frequency of visits of the customer in the next period i+1. Using a generated holdout for a period of 30 days preceding the reference day of 608(last recorded day of purchase), the feature relevance was then studied by applying permutation importance to the feature set. As per the feature importance values, duration between consecutive visits(gap22-gap28) has zero importance associated with it. The other features, amt, qty and freq have varying degrees of positive importance suggesting high relevance to the output target class(churners). Therefore, considering the same, after excluding gap, the final feature set contains store visit frequency, quantities of items purchased and the value expended.

c) Prediction approach:

After importing the relevant libraries and datasets and conducting data cleaning on the same, the importance of the features was studied and the final feature set was selected. The absence of missing values in the tables rules out the need for imputation. Encoding is another preprocessing step which is not required. The dataset was split into training and test sets.

Since, this is a temporal problem, a function to return input and output holdout sets was defined prior to splitting. The function returned holdout sets containing the values in the defined feature set for each period. The two aggregate window definitions to be considered are tumbling windows and output windows. The tumbling window defines the period over which the purchasing behaviour of the customer has been considered and aggregated. The output window governs the size of the exclusive test holdout set for measuring predictive accuracy of the final model on unseen data. The size of the tumbling window here while training and testing was set to 30 days. The tendency of 60% of the customers to shop once every 30 days on an average makes this churn definition an ideal choice. So, a tumbling window size of 30 days (our selected churn definition) will help us study the purchasing pattern of a customer on a monthly basis. With 7 such tumbling windows, the models can be adequately trained to make predictions for the validation holdout set. The prediction of possible churn for active customers is based on their attributes over the past periods.

The selection of the models has been justified below. The training of the models for predictions has been done across repeated holdout sets to average out the results over the iterations. By moving the reference date back by an interval equal to the tumbling window size, this becomes a systematic method to ensure that the training/test sets do not overlap over time and that predictions are performed only on data unseen relative to the reference day. Using the evaluation metrics as outlined below, the scores of each model was assessed. A final predictive model was picked and its performance on unseen data was evaluated. For the final churn prediction model to be re-run on weekly basis, a separate function was defined to generate the test input holdout. After changing the parameters(reference day being the date of execution, tumbling window size being 30 days and output window size being 7 days-churn definition), the final predictive model can be executed every week.

d) Model evaluation:

The three classifier models chosen for the modelling of the training dataset to test the effectiveness of their classification are Linear SVC, Random Forest and K-Nearest Neighbours (KNN) classifiers.

- K-NN is a swift yet efficient classification model that is extremely easy to implement. Moreover, the absence of outliers and missing data will make the performance of a K-Nearest Neighbours classifier a strong baseline for the other models being assessed.
- Linear SVC are efficient classification models performing well in high dimensions. Since, we are considering 21 features while training, in addition to the memory efficiency of such models and the non-overlapping nature of the target classes, the linear SVC is likely to produce accurate predictions.
- Random Forests are flexible and robust ensemble classifiers whose performance is extremely unlikely to be affected by overfitting. Its high accuracy even without hyper-parameter tuning makes it an obvious pick for historical training of the dataset.

The models were optimized using certain parameters and by training the dataset with the input features selected in the previous section.

The parameters considered while tuning the selected models are:

- K-Nearest Neighbours classifier: *n_neighbors=3, weights= 'uniform'*

The n_neighbours value signifies the number of neighbors to be considered while casting a vote. An uniform weight assigns equal weight to all points.

- Random Forests classifier: *n_estimators=100, max_depth=20*

The n_estimators parameter=100 will ensure weighted accuracy. A max_depth of 20 means the maximum depth of a constituent tree is 20.

- *Linear SVC: Default parameters were used. No hyper parameter tuning was required.*

The focus while formulating the models is to identify customers who are likely to churn in the immediate future. Based on the insights gained from the prediction model, targeting possible churners translates to engaging them better and ultimately an improved customer retention rate. Regularity and loyalty among customers is a driving factor of revenue generation. This underpins the need to efficiently predict churners for targeted marketing and engagement. Since our target class is churners, our primary motive is to correctly predict the set of active customers that are likely to churn. Failing to predict the target class equates to the lost opportunity of retaining a customer. Since, our objective is to reduce the levels of churns, minimizing the number of false negatives is an important objective. Similarly, falsely predicting the target class(churners) comes with the added expense of targeting associated the customer.

Therefore, while formulating the models, minimizing the false churner predictions has also been considered, i.e., the number of false positives.

The instances of the target class are fairly balanced. Therefore, considering the objectives of correct prediction and minimizing customers lost to churn from a wrong prediction, classification accuracy, sensitivity and precision are the primary evaluation metrics chosen. The metrics are in line with the focus of efficiently targeting all churners without compromising heavily with the expenses related to incorrect targeting.

e) Final model:

On comparing the models based on the selected evaluation metrics, the classifier that performs best considering classification accuracy, sensitivity and precision is a Random Forest classifier. With a classification accuracy of 77.6%, sensitivity of 85.1% and precision of 80.3%, the model does the best job when it comes to making correct predictions about customers that would eventually churn.

A high classification accuracy implies that the model is well trained to make correct decisions on most occasions. From a business point of view, a high classification accuracy means that the model identifies the likelihood of churning of customers correctly on a majority of occasions. A high precision value indicates that if the target customer has been predicted to churn, the likelihood of the customer not churning is relatively low. A high sensitivity implies that churners are accurately predicted and only on rare occasions, the opportunity of targeting an eventual churner is missed due to a wrong prediction. Although not an evaluation metric, this model also has a decent specificity of 65.3% indicating that the non-target class (active customers) is identified correctly with low chances of false prediction, minimizing the costs associated with inessential targeting.

Predicting churners (True positives) has been the primary focus while picking the winning classifier. So, apart from minimizing the number of actual churners misclassified as active customers, our second motive has been to minimize the expenditure from non-essential targeting. Keeping both motives in mind and after trading off between the performance measures for each classifier, random forest is chosen as the winning classification model with the best performance measures.

f) Predictive model implementation:

The final random forest predictor model was then used to make predictions for the unseen holdout test set. This model when used for predictions on the test holdout, had a classification accuracy of 78.6%, sensitivity of 86.5% and precision of 80.5%.

Using the same hyper-parameters as tuned with during training and model assessment, the final predictor model was then constructed with an output window size of 7 days and tumbling window size of 30 days for re-runs on a weekly basis. A separate function was defined to generate the test input holdout. After training the winning classifier with the holdout preceding the test input holdout, predictions were made for the input holdout.

SECTION D: INSIGHTS

The churners and active customers as predicted using the test holdout set by the final model display very distinct characteristics across the features that have been identified as important through the permutation feature importance scores. The differences across the attributes in the two classes (target class being churners) was studied to gain insights from the data and incorporate into marketing strategies designed to engage the customers better and keep them from churning.

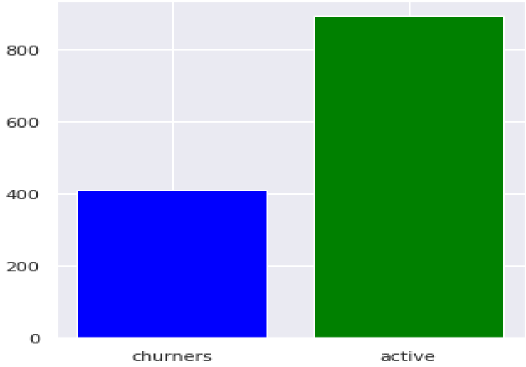

Fig 1: Churners vs Active customers count

The accuracy of the prediction model makes the structure of the prediction dataset very similar to that of the actual dataset. Of 1304 active customers, it is predicted that 411 customers will churn over the next window. The actual number of customers that churn over the period are 492. The rest of the customers are active customers. The active customers and the ones predicted to churn display extreme disparity when it comes to the frequency of store visits, the amount they spend and the quantities of items purchased.

The pen portraits of the active customers and the churners were developed by analysing the summary statistics of the test_X dataset based on the output value (1 for active and 0 for churn) and their corresponding indices (customer ids) on the preds dataset. The summary statistics was studied to identify differences across the features. Since, the pattern is similar for all features across the temporal periods, the comparison plots have been produced only for the period preceding the one being predicted for. The plotting has been done on python using Matplotlib. The pen portraits of the churners and active customers can further provide insights that will assist in optimizing the efficiency of the marketing strategy adopted for targeting.

The active customers tend to spend more. This accounts for their higher mean spend as compared to the customers that were predicted to churn. Active customers have a mean spend of £40.86 over the past period as compared to £3.2 for churners. Similarly, they tend to buy a higher number of products during each purchase. Active customers on an average buy 34 products during a month as compared to the average of just 2 products bought by a churner. Lastly, the frequency of store visits of active customers is an indicator of their regularity at the store. Active customers visit the stores thrice during a 30 day period on an average as contrasted to a single visit of an eventual churner.

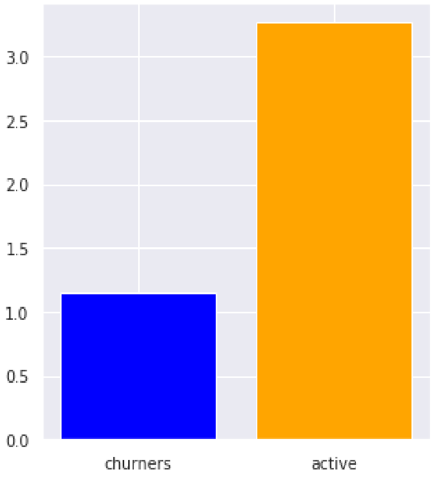

Fig 2: Comparison by number of store visits

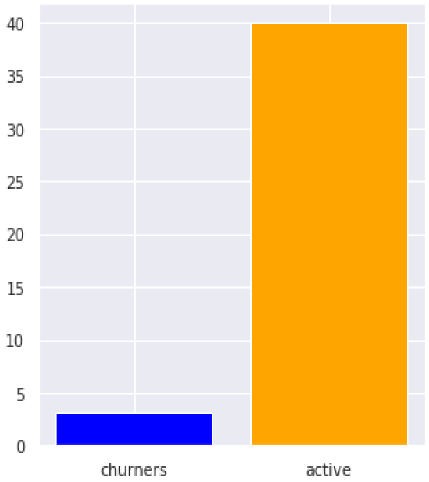

Fig 3: Comparison by amount spent

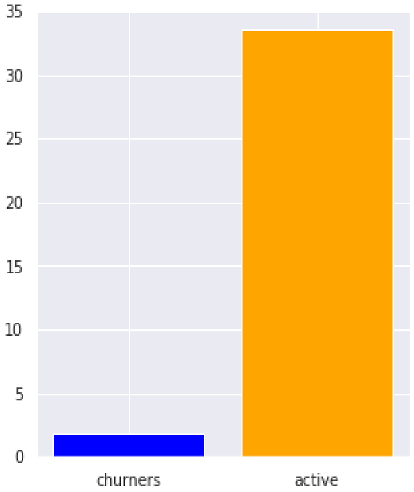

Fig 4: Comparison by item quantities purchased

These insights were derived from an analysis of the summary statistics of the input features corresponding to each customer in the prediction list. Filtering the customers based on the prediction of churn will help us separate the dataset into churners and active customers. The variation of the input features corresponding to each customer in test_X can be then be studied by plotting the summary statistics of the features identified as most relevant as per permutation importance.

In summary, for prediction of customers likely to churn in the near future, classification using temporal holdouts was opted for and a Random Forest classifier was used as the learning algorithm to predict customers likely to churn based on their purchasing pattern. Firstly, customer segmentation can be conducted so as to offer tailored deals and incentives, with which the average spends and frequency of store visits is most likely to improve for customers who are currently lowly engaged and are more likely to churn. Advertising loyalty benefits to high and regular spenders identified from the pen portraits might ensure their retention and unwavering engagement. Loyalty rewards are an effective way of boosting engagement and reducing the likelihood of churn. In conclusion, with efficient targeted marketing of customers likely to churn, the store chain is sure to maximize its revenue collection over time.