Sainsbury's

Discontinued products: propensity model

Data Science team: Zakaria Alzilitni

Problem formulation

- The replenishment team want to predict, at any given point in time, which products will be discontinued after the next product range refresh.
- They estimated that currently around 30% of all product lines are replaced by new products.
- Throughout the year, stocks of products are replenished if the expected sales revenue is sufficient to justify the investment.
- However, some products sell less often than once per month in some locations.
- Thus, in order to confidently forecast the expected sales revenue of a product in a location, it is important to know the likelihood of it being discontinued at the next range refresh.
- To help the replenishment team identify products that will potentially be discontinued, we use past data to train classification algorithms.
- This with hope of identifying a relationship between known inputs and discrete output variables i.e. potential discontinuation or continuation of a product.
- As shown by **Figure 1**, these learned relationship pairs are then used to predict which products have a high propensity of being discontinued.
- Generally, propensity can be defined as a probability that a certain action or event will occur.

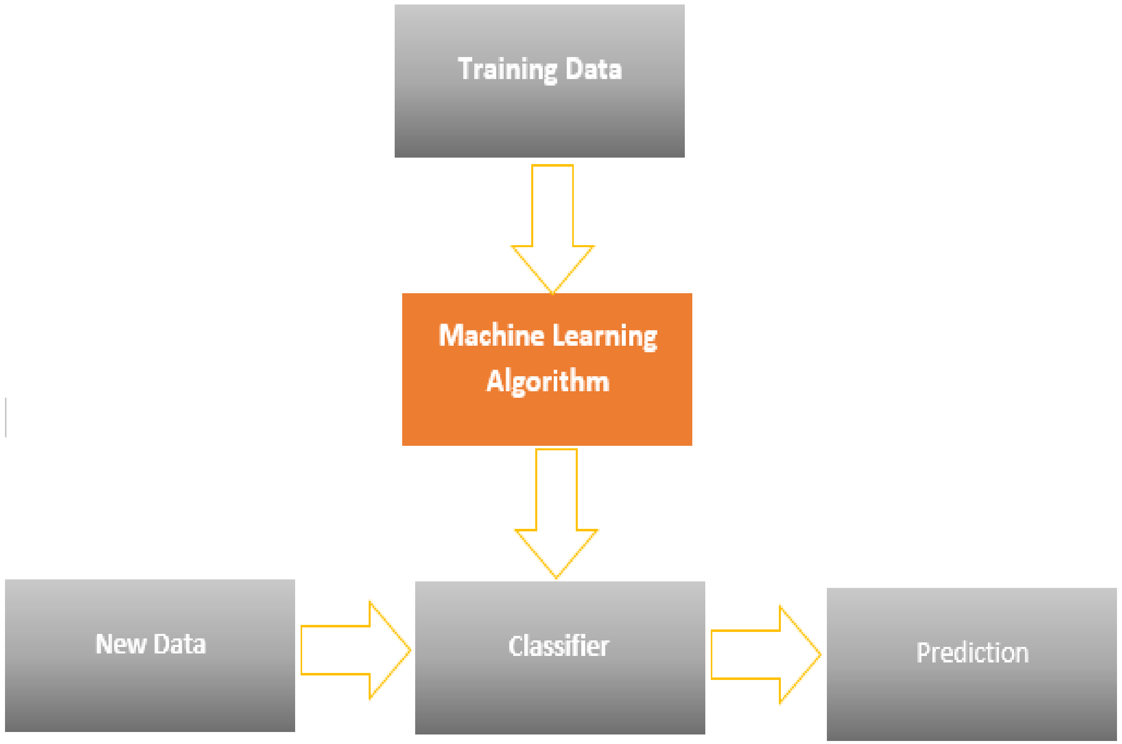

Figure 1:- Flow diagram of the machine learning process

Machine learning: Decision Tree

- Decision Tree is a predictive algorithm that identifies pathways that lead to a desired classification.
- More precisely, it utilise its branches to signify the relationship between features and the output variables.
- As shown in **Figure 2**, a decision tree is drawn upside down with its root at the top, before splitting into different branches.
- At each level of the tree, the algorithms uses the feature that allows it to split the observations at hand, in a way that leads the resulting sub groups to be as different from each other as possible.
- Paths then diverge at each node within the tree, whereby observations meeting the criteria go down the **YES** branch and ones that do not, go down the **NO** branch.
- Each row within the dataset will pass through the various branches of the tree before finally arriving at the leaf, where a final prediction is made.

Colour	Underlined	Target
Orange	Y	1
Orange	Y	1
Black	Y	0
Black	N	0
Black	N	0
Black	N	0

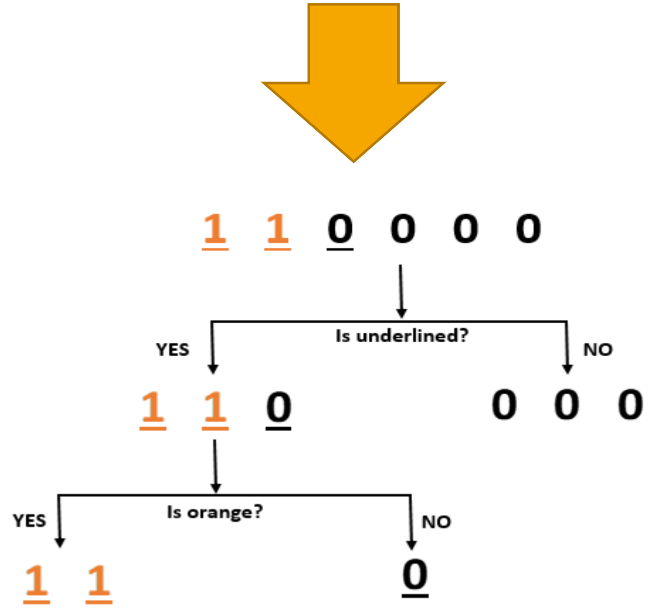

Figure 2 :- An intuitive example of the decision tree training process

Machine learning: Random Forest

- Random forest is a step up from decision tree in that it uses multiple simplistic and weakly correlated trees to form a strong classifier.
- As shown in **Figure 3**, these trees are trained in parallel, on different bootstrapped samples of the data.
- Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples.
- Additionally, each tree is only trained on a randomly selected subset of the available features.
- By doing so, the aforementioned characteristics, ensure that each individual decision tree within the "forest" is as unique as possible.
- Finally, the classifications provided by the trees are aggregated, and the majority decisions is taken as the final prediction.
- Thus, by aggregating the results of multiple trees trained on bootstrapped samples, a process formally known as bagging, we account for the disadvantage i.e. potential bias, of using a single tree.

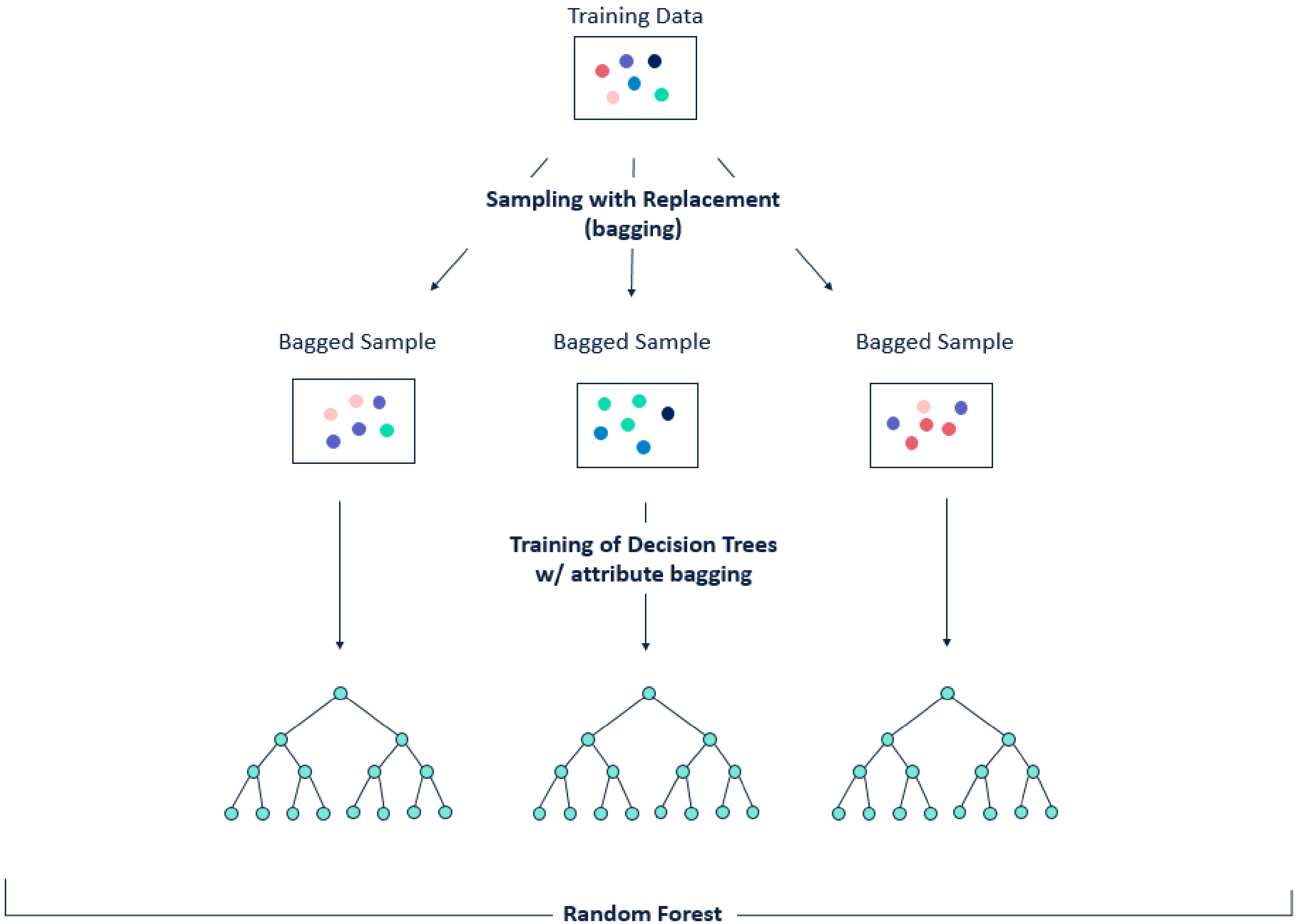

Figure 3 :- An intuitive example of the Random forest training process

Results

- In attempting to assess prediction power, we have used the figures derived from the confusion matrix in **Table 1**, to compute the classification accuracy of our chosen algorithms.

- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known

- As shown in **Table 2**, with a classification accuracy of **88%** and **86%** respectively, the random forest outperforms the decision tree as expected.

- Nonetheless, prediction power derived from classification accuracy can perhaps be misleading.

- The reason for this, is that the classification accuracy rate implicitly assumes that the distribution of class, i.e. the likelihood of discontinued and continued products is relatively balanced.

- However, this is seldom the case, as it is likely that discontinued products will be in the minority class for the majority of cases and thus, the validity of the accuracy measure will diminish as the skewness of class distribution increases.

- For this reason, we rely on Precision and recall. The former metric, is the ratio of correctly predicted positive observations divide by the total predicted positive observations (i.e. True positives + False positives).

- The latter, is the ratio of correctly predicted positive observations divided by all observations in actual class (i.e. True positives + False negatives).

- Thus, as both precision and recall scores for our decision tree and random forest models are above **0.8**, it suggests the models have strong predictivity power.

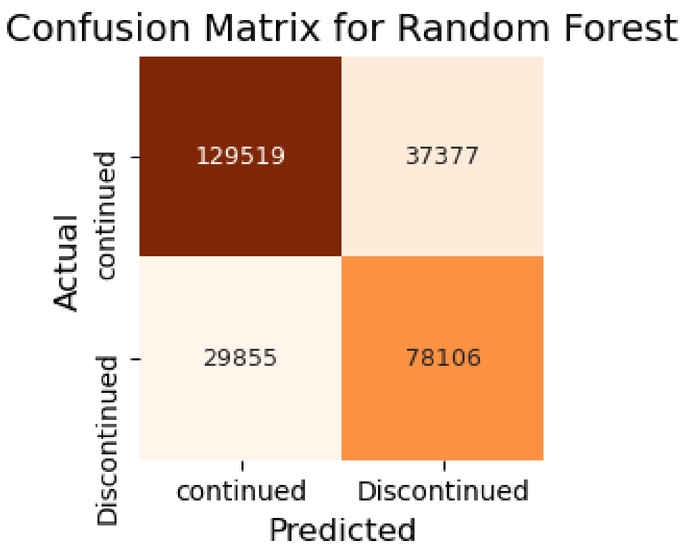

Table 1:- Confusion matrix illustrating the amount of correctly and incorrectly predicted observations

Models	Accuracy	Precision	Recall
Random Forest	0.87	0.85	0.84
Decision Tree	0.86	0.82	0.83

Table 2:- Different performance measures of machine learning algorithms

Results: ROC Curve

- ROC curve measures the predictive power of machine learning algorithms by plotting the sensitivity(Recall) against the output of the 1-specificity rate.
- The former of these measures the ability of our models to correctly identify products that are likely to be discontinued. The latter denotes products that are incorrectly identified as likely to be discontinued.
- In doing so, each combination of sensitivity/1-specificity values will correspond to a particular classification threshold.
- That is to say, as the threshold departs from its original default value of **0.5**, the cut-off points for positive and negative classifications (**0/1**) will become more stringent.
- This means a perfect predictive model would yield an area under the curve (**auc**) of **1**, while an area of **0.5** would indicate that a model is unable to differentiate between products belonging to the positive(discontinued) or negative class (continued).
- Consequently, as show in **Figure 4**, our models indicate a strong ability to distinguish between the two classes, as our random forest scores an auc of **0.85**, and by doing so, outperforming the decision tree.

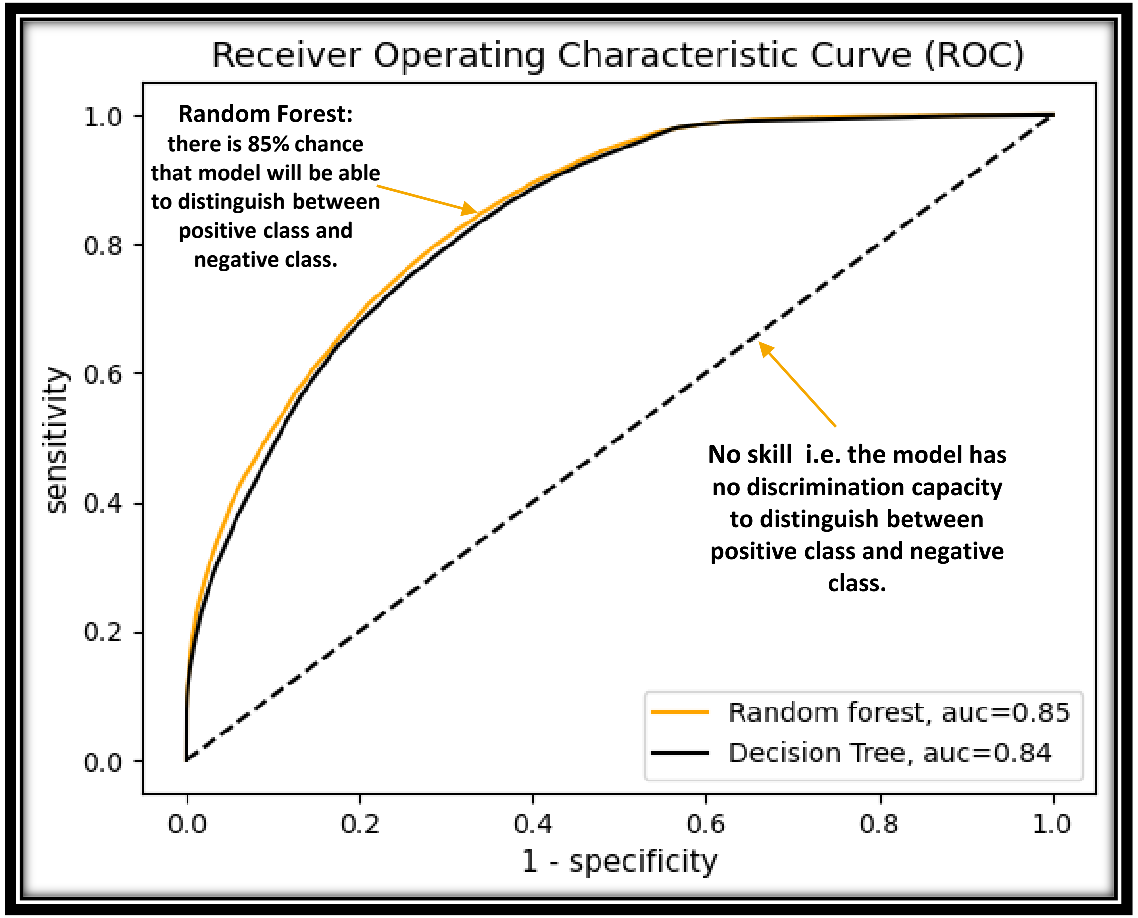

Figure 4 :- ROC curve illustrating the predictive power of our model

Results: Lift Curve

- The lift curve shown by **Figure 5**, is a widely used visualisation of model performance and empowers us to identify a number of products that are more likely to be discontinued.
- This is because the x-axis presents all products that will potentially be discontinued according to their propensity scores in descending order from highest to lowest.
- The y-axis plots the cumulative lifts that provides an estimation of how much more likely a product is to be discontinued.
- The baseline represents a random model where the percentage of targets for each group is more or less equal, thus, it is uniformly set at 1.
- For example, if the replenishment team used our current model to target the top **30%** of products they are **2** times more likely to find discontinued products than in any equally sized randomly selected sample of products.
- As these evaluations are always based on previous data we can use such lift curves to evaluate how better off we are using a propensity model than without one.

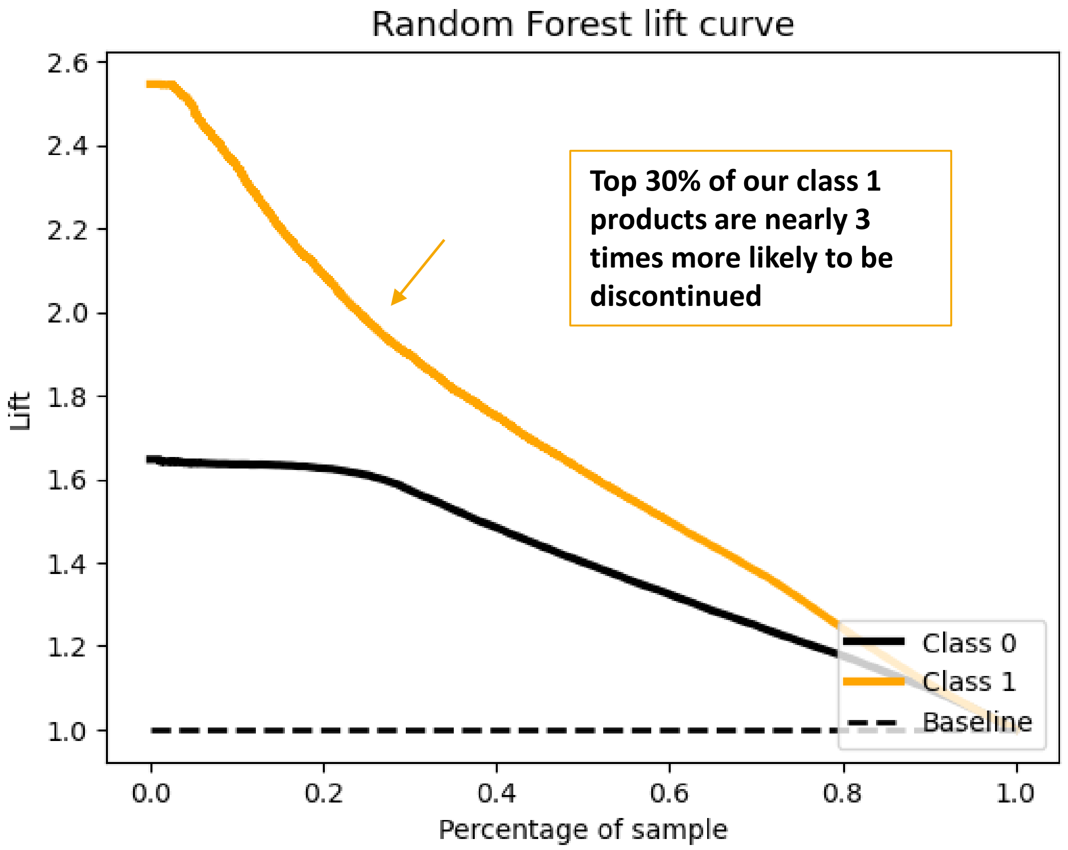

Figure 5 :- Lift curve illustrating the predictive power of our model

Future suggestions

Predicative power could be improved by:

- Enriching data with new fields such as location, economic statistics and trends

Future work to complement current analysis:

- Discrete event simulation of the replenishment process to reduce potential inefficiencies
- Topic modelling to identify the topics discussed in user comments on products and why they have or have not been purchased
- Sentiment analysis of comments on products. The output of this can be also used as a way to predict whether products will be discontinued
- Unsupervised analysis to understand the segment or group of products the are discontinued

Any questions ?