Question

- E-commerce companies are not only able to collect customers' profile information, they are also able to observe the online behaviours of customers. How may we use advanced analytics in marketing to identify, understand and target our customers more effectively?

Introduction

- We may further refine our problem statement to as such: How can analytics be used to identify suitable up-selling/cross-selling opportunities for different profile customers?
- The basis of cross selling and upselling lies in enticement. To correctly do that, a deeper understanding of customer behaviours and purchasing patterns are required. Thus, this is where data science is particularly useful to provide meaningful insights that can serve as drivers of sales for e-commerce firms by recommending the right product to the right customer at the right time.
- To provide a dynamic customer experience and increase more value to customers, we aim to build an e-commerce recommendation model based on data mining techniques.

Proposed approach

Data processing

- Data can be sourced from a myriad of places, like customer account information, past purchases, and even scientific experiments for market segmentation, market basket analysis, and uplift modelling respectively.
- As a prerequisite step and before any analysis can take place, it is important to ensure data integrity as garbage data only produces poor analysis. Thus, data cleaning must be carried out to remove any duplicates, rejected transactions, orders that are not processed successfully, irrelevant observations, or observations with missing data. Furthermore, structural errors should be fixed and outliers filtered away as well.

Market segmentation

- To ensure the right customer group is targeted, datasets should be first segmented according to certain demographics of the population. This enables marketing teams to better understand customers and therefore tailor the best marketing techniques to approach different customer segments.
- We can rely on clustering techniques to identify different user groups or clusters according to the common features of the dataset. Such features range from age, gender, location, product type, amount spent, etc. How customers are allocated to the segments depend on the business goals of the company.
- One way of doing so is to use density based clustering techniques, particularly the DBSCAN (Density-based spatial clustering of applications with noise) algorithm, which groups points of high density based on how close they are and a minimum number of points. This algorithm is chosen as it can identify patterns and associations in the data that are difficult to detect on the onset, and is also resistant to outliers.
- The algorithm is easy to implement and takes only 2 parameters, epsilon (eps) and miniPoints. The former specifies the distance between points to be considered as a cluster while the latter is the minimum points that will count as a cluster.
- Additionally, parameter optimisation is also important to ensure we get accurate results or clusters. To get the optimal number of eps, we can use the K-distance graph. Where the maximum curvature is on the K-distance graph, that is where the optimal eps will be. For number of miniPoints, this number varies and depends on domain knowledge. Nonetheless, the latter can be derived by doing repeated DBSCANs over the dataset until an optimal or desired number is achieved.

Market Basket Analysis

- Historical purchases are good indicators of what customers will likely buy in the future. Hence, this is where market basket analysis (MBA) comes in, which is a technique that analyses customers purchasing habits and uncovers any interesting associations or patterns between products. This permits companies to better serve their customers by recommending suitable items which may entice consumers.
- Here, we will use association rule mining algorithms, namely the Frequent Pattern Growth algorithm. This algorithm is selected due to its computational efficiency and speed compared to other methods like the Apriopri algorithm. FP-Growth does not generate candidate itemsets, uses less memory, and adopts a compact data structure i.e. tree, making it more suitable and scalable for larger databases.
- The first step comprise of counting the frequency of items in the dataset of transactions and storing the results in a table. Items in each transaction are sorted in a descending order and those that fail to meet the minimum support threshold will be discarded. The minimum support threshold acts as a filter to ensure only frequent items will be used to define significant associations. The FP-tree structure is then constructed with a root "null" transaction by transaction, with each node representing a frequent item and its corresponding occurrence that increases as more transactions are inserted into the tree. This process ends once all transactions are processed and all frequent itemsets are generated.
- The next step comprise of generating strong association rules for ascertaining the correlation and relationship between items in the dataset. We will use the following association rules:
    - **Support**: Frequency of the itemset in the dataset
    - **Confidence**: Proportion of transactions with item A, in which item B also appears in
    - **Lift**: Ratio of the confidence of items A and B over the support of item B. A value > 1 indicates a strong association between the items while a value < 1 indicates a weak association between the items
- Once these rules are generated, we can observe a number of associations between the purchases, for instance, knowing which items are commonly purchased together and which items are the most popular.

Uplift modelling

- Once customers are segmented and MBA done, the next thing to predict is how well they would respond to up-selling or cross selling marketing actions. This makes sense since a company should not waste its resources on non-persuadable customers, but focus its resources on customers that can bring in revenue.
- Thus, we would use an uplifting model to measure the effectiveness of marketing actions and build predictive models to determine response. This approach can be condensed into the following: *Impact of treatment* (t) *on target* (y) *given features* (x).
- To implement this, we would first need to conduct a randomised controlled experiment i.e. A/B testing to gather data on the treated and control (untreated) groups. The treatment and outcome depends on the business goals, and an example of the former is whether a customer has been sent a marketing email while the latter is whether the customer has patronised the particular e-commerce website after receiving the email. Both the control and treated groups will used for the modelling.
- To model uplift, we can use the Class Transformation (CT) method, which is simple and performs better than the two model approach. This method aims to predict a new target called Z_i that is made up of the *groups* (treated and control groups) and *outcome* i.e. whether a purchase was made at the online shop. With the new target, a customer is either in the treated group and have a positive outcome or in the control group and have a negative outcome. An assumption is also made where both the control and treated groups are of the same proportion.
- The uplift model will then be evaluated, using methods like bins of uplifts to estimate the treatment effects.