Build Propensity to Purchase Model in Python

Business Objective

Our Client is an early-stage e-commerce company selling various products from daily essentials (such as Dairy & vegetables) to high-end electronics and home appliances. It is a one-year-old company and they are witnessing many people coming to their platform and searching for products but only a few end up purchasing.

To increase the number of purchases, the business is planning to send discounts or coupons to users to motivate them to buy. But since it is an early-stage startup, they have only limited funds for this discount campaign. So, they have reached out to us seeking our help in building a model that would predict the purchase probability of each user in buying a product.

We will be making use of propensity modeling for this. Propensity modeling is a set of approaches to building predictive models to forecast the behavior of a target audience by analyzing their past behaviors. That is to say, propensity models help identify the likelihood of someone performing a certain action. We can then use this likelihood or probability score to create personalized targeting campaigns for the users thus reducing our total cost (targeting only a small set of users) and increasing our ROI.

To help us in predicting the propensity of purchase, we will also make use of RFM Modeling. RFM is a data-driven customer segmentation technique that allows marketers to make informed decisions. RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because the frequency and monetary value affect a customer's lifetime value, and recency affects retention, a measure of engagement.

It empowers marketers to quickly identify and segment users into homogeneous groups and target them with differentiated and personalized marketing strategies. This in turn improves user engagement and retention. With the probability scores from propensity modeling, the marketing team can filter only those users who need the actual push (in terms of discounts or coupons) to make a purchase rather than sending coupons to users who would have bought the product anyway.

Data Description

The dataset contains the purchase history of various users over a period of time for an e-commerce company.

Aim

1. To understand Propensity Modeling
2. To understand RFM Analysis
3. To build a model to predict the purchase probability of each user in buying a product for an e-commerce company with the help of the propensity model

Tech Stack

- Language - Python
- Libraries - pandas, sklearn, numpy, seaborn, datetime, matplotlib, missingno

Approach

1. Importing the required libraries and packages
2. Read the CSV file
3. Perform data preprocessing
4. Perform exploratory data analysis
    i. Univariate analysis
    ii. Multivariate analysis
5. Perform RFM Analysis
6. Feature engineering
7. Modeling Data Creation
8. Model building
9. Making predictions

Modular code overview

```
input
|_ config.yaml
|_ ecom_product_data.csv
|_ final_customer_data.xlsx
|_ final_customer_data_with_RFM_features.csv

lib
|_ PropensityData_EDA.ipynb"
|_ PropensityData_RFM_Features.ipynb"
|_ PropensityData_RFM_Features.ipynb"
|_ RFM_Modeling.ipynb"

src
|_ engine.py
|_ ml_pipeline
        |_ feature_eng.py
        |_ model.py
        |_ preprocessing.py
        |_ rfm.py
        |_ utils.py
|_ requirements.txt

output
|_ logreg.pkl
```

Once you unzip the modular_code.zip file you can find the following folders within it.

1. input
2. src
3. output
4. lib

1) input folder **-** It contains all the data that we will need for analysis.
    - A config file, with some basic configuration parameters which can be edited according to your dataset.

- A final_customer_data.xlsx file that has 2090 rows of customer transaction data with over nine features.
- A final_customer_data_with_RFM_features.csv file that is a merged dataset containing both customer data from final_customer_data.xlsx and the extracted RFM values
- An ecom_product_data.csv file which is a transnational data set that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. It is used for explaining how RFM modeling helps us in predicting the propensity of purchase.

2) src folder - This is the most important folder of the project. This folder contains all the modularized code for all the above steps in a modularized manner. This folder consists of:
    - engine.py
    - ml_pipeline

      The ml_pipeline is a folder that contains all the functions put into different python files which are appropriately named. These python functions are then called inside the engine.py file.

    - requirements.txt

      This file will help you install all the packages that are required to run the project successfully. You can install these packages using the command → pip install -r requirements.txt

3) output folder **–** The output folder contains the model we trained for this data. This model can be quickly loaded and used for future use and the user need not have to train all the models from the beginning.

4) lib folder **-** This is a reference folder. It contains the original ipython notebooks that we saw in the videos.

Project Takeaways

1. Understanding propensity modeling
2. How to perform univariate and multivariate analysis
3. Data Preprocessing
4. Understanding RFM modeling
5. Understanding the data used for RFM modeling
6. How to calculate RFM features?
7. How to calculate RFM rankings?
8. How to plot graphs using matplotlib and seaborn?
9. Understanding how to extract new features from existing data
10. How to encode categorical variables?
11. Data Scaling
12. Data Transformation
13. How to build a logistic regression model
14. Understanding how RFM features help in identifying the propensity to purchase
15. Understanding preferential treatments and high-value paths