Customer Purchase Journey Prediction

Aditya Sumbaraju

Bellevue University

DSC680- Applied Data Science

Dr. Brett Werner

December 16, 2021

https://github.com/adityasumbaraju/DSC680

Business Problem

Customer purchasing predictions are common for many companies, including large and small-scale businesses. It is always an advantage to an organization that contains data wealth, and this will help them predict the possible purchasing behavior of a given customer. And it also allows clients to make informed decisions on maintaining inventory and providing customer recommendations on items, customer promotions, and predicting the customer's interests in buying a similar product from its competitor. In this case study, I would analyze and demonstrate some of these metrics on a publicly available dataset and develop an ML model to predict the future long-term purchasing in the form of customer lifetime values.

Background

In recent years analyzing shopping baskets has become quite appealing to retailers. Advanced technology made it possible to gather information on their customers and what they buy. Electronic point-in sales increased the use and application of transactional data in the next basket analysis. Analyzing purchase behavior patterns allows retailers to understand customer's purchase behavior, which helps to adjust promotions, store settings and serve customers better. Transactional data is used in mining information on co-purchases and adjusting promotion and advertising accordingly.

Accurate "next basket prediction" will enable next-generation e-commerce predictive shopping and logistics. I would be applying the deep learning technology behind the next basket prediction. Retail domain E-commerce online sales are predicted to double in the following three years, and it will be known for Covid-19 years. The pandemic accelerated this trend, and now even

groceries are going online, for example, amazon fresh. As per the experts, this process is considered to be irreversible. Groceries have many repetitive purchases occurring that generate buyer behavior patterns in time, and respective patterns can be discovered by machine learning models and subsequently be used to predict from the purchase transactions data.

Next basket prediction has been a primary ML topic for retailers for quite some time. Experts forecasted that machine learning would predict the next purchase to change e-commerce forever in the next year. This forecast did not materialize, and even seven years later, machine learning models are not good enough to make the next basket prediction which can be useful enough. Some estimate that even 60% prediction accuracy can enable predictive shopping, and Amazon is already experimenting with so-called anticipatory shipping.

Data Explanation

Data Source Urls:

- https://www.kaggle.com/chiranjivdas09/ta-feng-grocery-dataset
- https://stackoverflow.com/questions/25014904/download-link-for-ta-feng-grocery-dataset

About Data: The Ta Feng Grocery Dataset

Column definition: Transaction date and time, Customer ID, Age Group, PIN Code (Region), Product subclass, Product ID, Amount, Asset, Sales price

Fields of the Dataset are:

- Transaction date and time
- Customer ID
- Age: 10 possible values

- Residence Area: 8 possible values
- Product subclass
- Product ID
- Amount
- Asset
- Sales price

The Ta Feng Grocery Dataset is a supermarket dataset containing

- 817741 transactions from November 2000 until the end of Feb 2001.
- The data set contains information about 119578 shopping baskets that belong to 32266 unique users.
- In total, 1129939 items were purchased from available 23812 products.

Methods

As a first step, I will be predicting RMF values. Pass all the features that could help along with split values of customer_id and Recency (R), Monetary Value (M), Frequency (F). Finally, I will split the data into train and test sets to pass them to the Model. I would be using a simple recurrent neural network with 250 hidden units with Relu activation using L1 regularization and loss of mean square error. I would also be using xgboost to predict the Model, and based on the accuracy results; I will evaluate and provide the recommendations.

Analysis

Illustrations

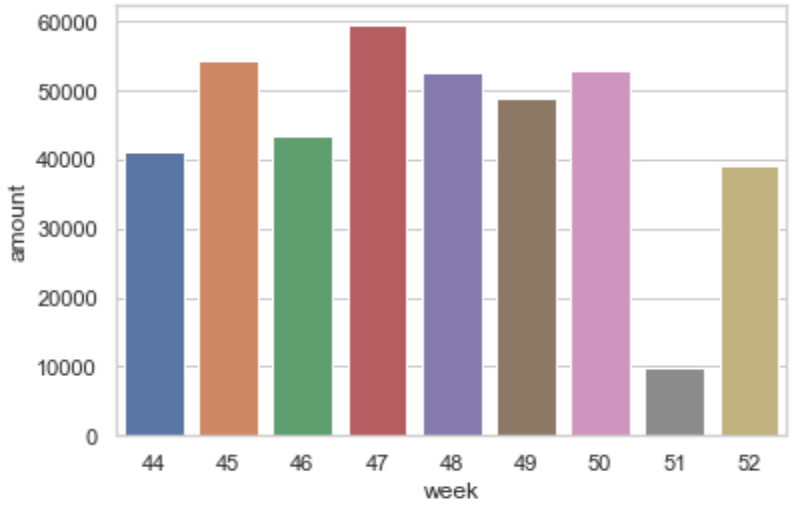

Illustration 1: Weekly Transactions Nov/Dec 2000

Observation: It is evident that Thanksgiving week accounted for more sales when compared to other weeks. We can see a dip in week 51, and sales look normal in week 52. From the limited dataset, it is tough to identify the root cause for the drop in week 51 sales, but we can have a strong assumption by comparing weeks 51 and 52 that inventory refresh could have caused a dip in week 51 sales.

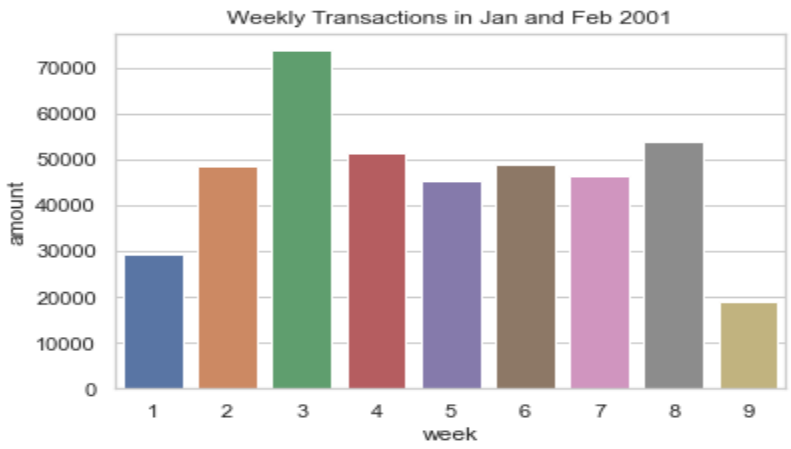

Illustration 2: Weekly Transactions Jan/Feb 2001

Observation: January 3rd week accounts for more sales than the first week. Ideally, Jan 1st week.2nd and 3rd week are considered peak seasons in the retail domain.

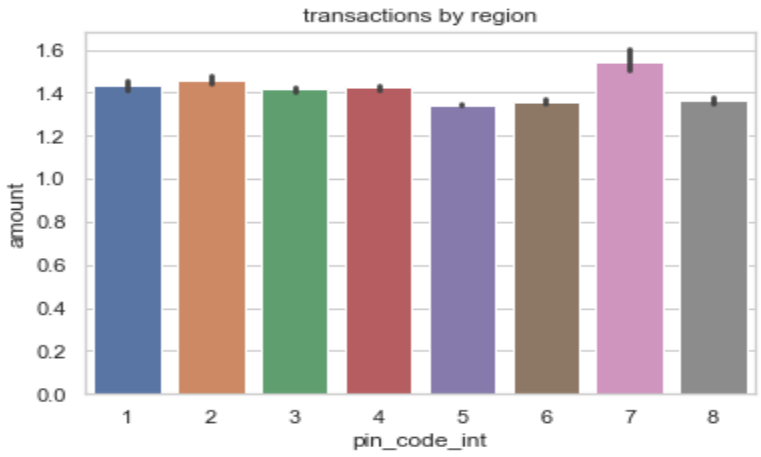

Illustration 3: Transactions by Regions

Observation: Region 7 is contributing more sales followed by 2 and 1. I think we need to apply more marketing strategies in region 5 to boost sales.

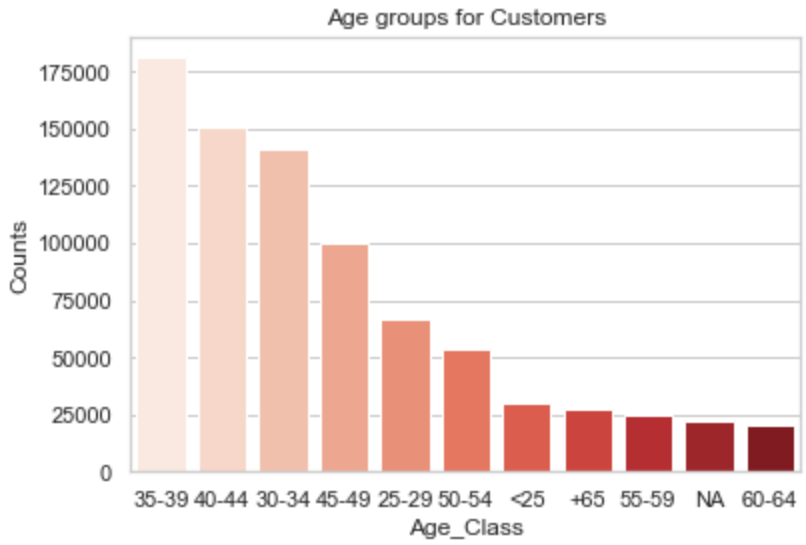

Illustration 3: frequent shoppers w.r.t age groups

Observation: The frequent shoppers come from the age ranges of 30 to 44; this would seem consistent for a grocery store if you are a young adult shopper usually young adults tend to shop more compared to other age groups. Presumably, older adults only need to buy for themselves and their partners.

Exploratory Data Analysis

Impact of RFM scores:

customer_id	Recency	Frequency	Monetary	R	F	M	RFM_Score	RMF_Segment
00001069	19	11	1944.0	6	2	3	11	First segment customers
00001113	54	18	2230.0	3	3	3	9	Second segment customers
00001250	19	14	1583.0	6	2	2	10	First segment customers
00001359	87	3	364.0	2	1	1	4	Third segment customers
00001823	36	14	2607.0	5	2	3	10	First segment customers
00002189	57	62	14056.0	3	4	4	11	First segment customers
00003667	21	13	11509.0	6	2	4	12	First segment customers
00004282	47	9	967.0	4	2	2	8	Second segment customers
00004381	103	11	701.0	1	2	1	4	Third segment customers
00004947	81	36	3363.0	2	4	3	9	Second segment customers

EDA 1: RFM Detailed Scores

RMF_Segment	Recency	Frequency	Monetary	
	mean	mean	mean	count
First segment customers	15.7	37.4	4906.3	18796
Second segment customers	59.9	9.8	1364.5	10647
Third segment customers	97.6	3.6	387.6	2823

EDA 2: RFM aggregated Scores

Observation: EDA 2 represents RFM scores that are easier to read table summarized by the relative RFM score as shown in EDA 1. The customers with high RFM scores are more important customers to the business. Group customers based on RFM score to classify customers into First Segment, Second Segment, and Third segments for more straightforward interpretation.

1. First Segment customers= RFM Score equal or greater than 9
2. Second Segment customers = RFM Score between 4 and 9
3. Third Segment customers = Anything else

Assumptions and Limitations

Although next basket analysis is computationally efficient and straightforward, several significant limitations exist. Since there is no available data on households and individual consumers, interdependencies across purchases of individual consumers or households are neglected. In other words, due to data restrictions, homogeneity across purchases is assumed. Personal household-level data would be very beneficial for the analysis. A combination of insights on product dependencies and household level data could help retailers better pricing and promotion decisions for different customer segments. My area of interest is investigating sequences of purchases and events concerning the customer.

Assuming the sequential time series analysis would be an appropriate technique to use while dealing with anonymous transactions. Please note that anonymous transactions do not unveil information on consumer behavior. Availability of household data would be very beneficial for the second research question. A model-based approach can be used for prediction and forecasting. Running a multinomial logistic regression with more independent variables would effectively predict product choice. Having non-anonymous household data also allows applying techniques like clustering, decision trees, or artificial neural networks to provide more insightful information on the consumers and their preferences.

Challenges

- Data contains Chinese characters; I have imported the data using ISO-8859-1.
- Anticipating whitespaces in data and need to work on data alignment.
- Incorrect variable types.
- Python package-related issues.
- There is no detailed data description; it doesn't contain currency details for the sales price and the units used for the variables "amounts" and "Assets."
- I am relying on the Kaggle dataset as this is usually clean, and I anticipate no missing or insufficient data that needs to be substituted with dummy data.
- Since the data is not current, Modeling and forecasting may not apply to the present-day scenarios.

Recommendations

Future Uses:

Retailers can use the insights gained from the next basket analysis in several ways, including:

1. **Cross-Sell**: Group products that customers frequently purchase together in the store's product placement.
2. **Campaigns & Promotions (Marketing)**: Target marketing campaigns to customers and entice them to purchase related products for recently purchased items.
3. **Online shopping websites**: Provide recommendations that are associated with frequently purchased products. For example, Amazon and BestBuy display the message "Customers who purchased this product also viewed this product…" right after adding the items to the basket.

4. **Bioinformatics:** discovering co-occurrence relationships among patients' diagnosis and active pharmaceutical ingredients prescribed to different patient groups.
5. **Manufacturing**: predictive analysis of equipment failure
6. **Customer purchase behavior:** Tying up customer purchase data with demographic and household income.

In the current trend, organizations are discovering ways of using purchase prediction analysis to gain valuable insights into pattern associations and hidden relationships. The market leaders such as Amazon, BestBuy, Walmart continue to explore the technique's value. A predictive version of next basket analysis makes in-roads across many sectors to identify sequential purchases.

Implementation Plan

The implementation plan contains the below steps to accomplish the project

- Business understanding
- Data Preparation
- Exploratory data analysis
- Data prep for Modeling
- Modeling
- Model Evaluation

Ethical Assessment

The project reminds me of a rational maxim, "good ethics is good business"; Does it mean that unethical conduct is penalized in the Retail-Ecomm domain and ethical conduct rewarded? It is proved that, in many circumstances, firms may be immune to marketplace sanctions and more inclined towards customer retention. It is worth examining the conditions under which the organizations are called to task "the indefinite campaigns" or reward the existing consumers, i.e., as reflected in consumer attitudes toward the brand, the organization, and purchase intentions. This includes investigating the relationship between shopping experience and consumer purchasing behavior. Different algorithms may give other output for a given dataset, provided their limitation is associated with the individual machine learning models. This seems to be a potential problem in choosing a model for my use-case, and I would pick the best-suited algorithm based on its characteristics.

Importantly, I would seek customers' consent before using the customer data in predictive modeling applications that drive the marketing strategies.

Conclusion

This work aims to automate Marketing campaigns using customer segmentation data as an input. The purpose is to simplify the Marketing team's job by avoiding analyzing thousands of rules associating customers with their next item. Recurrent Neural Networks make it easy to find the sequential patterns that return the most probable item sequence. We can retain the customers based on the model-based recommendations by sending appropriate promotional emails.

References:

Smartbridge, . (2021, January 12). *Market basket analysis 101: Anticipating customer behavior*. Smartbridge. Retrieved December 19, 2021, from https://smartbridge.com/market-basket-analysis-101/

Kordik, P. (2020, November 14). *Deep learning for recommender systems: Next Basket Prediction and sequential product recommendation*. Medium. Retrieved December 19, 2021, from https://medium.com/recombee-blog/deep-learning-for-recommender-systems-next-basket-prediction-and-sequential-product-recommendation-796228b34dee

Cavique, L. (2006). *(PDF) next-item discovery in the Market Basket Analysis*. Next-Item Discovery in the Market Basket Analysis. Retrieved December 19, 2021, from https://www.researchgate.net/publication/224693768_Next-Item_Discovery_in_the_Market_Basket_Analysis

Singh, S. (2021, May 26). *Predicting purchases with Market Basket analysis*. Medium. Retrieved December 19, 2021, from https://medium.com/geekculture/predicting-purchases-with-market-basket-analysis-d6ad2152bf6e

Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., & Pedreschi, D. (1970, January 1). *[PDF] next basket prediction using recurring sequential patterns: Semantic scholar*. Next Basket Prediction using Recurring Sequential Patterns. Retrieved December 19, 2021, from https://www.semanticscholar.org/paper/Next-Basket-Prediction-using-Recurring-Sequential-Guidotti-Rossetti/b59d6ee45a50a5f5dc7be654cb706897e4ff147c

Appendix

The data set contains information about 119578 shopping baskets belonging to 32266 unique users.

In total, 1129939 items were purchased from available 23812 products.

Features:

- transaction_dt
- customer_id
- age_group
- pin_code
- product_subclass
- product_id
- amount
- asset
- sales_price
- frequency
- recency
- monetary

Additional features:

- Week_number
- Total_sum
- Age_group
- Unit_price
- Log_unit_price

Hyperparameters

1. SimpleRNN

2. Relu activation

3. 250 hidden units

4. L1 regularization at 0.0001

5. MSE loss

6. Batch size 120

7. Shuffle=True

8. 200 epochs

10 Questions:

1. How this Model does help in marketing strategies?
2. How do end-users access the results?
3. From EDA, What are the best-selling products?
4. Why use a Recurrent Neural Network?
5. What is the sales trend for my best-selling products?
6. What evidence or rationale supports your findings?
7. How would you go about designing an operational solution?
8. How would you go about estimating the time/cost/skills needed to build something?
9. What modeling techniques are applied?
10. List the existing applications that use this Model, if any?