CDS513: GROUP PROJECT

Improving Sales Performance of the Ecommerce Website for an Electronics Store using Predictive Business Analytics Techniques

GROUP 3: LEE YONG MENG (P-COM0012/20) | LEE KAR CHOON (P-COM0130/19) | LIM HANG THING (P-COM0143/20)
School of Computer Sciences, USM | Semester II, 2020/2021 | CDS513: Predictive Business Analytics

Abstract

In recent years, the trend of purchasing items online through ecommerce websites have become a norm all around the world. In this project, a store owner selling electronics products is interested on customer buying behaviours in the ecommerce website for decision making such as product cross-selling and promotion. This is addressed via product affinity analysis by using market basket analysis (MBA) technique on customer transaction data. In large ecommerce websites, there may exist thousands of products of various brands sold within a single product category. An effective search algorithm is also needed for customers to search for the products that best match their preferences from the ecommerce website of the electronics store. Therefore, a recommender system (RS) is implemented to improve the customer experience when purchasing items on the website. In addition, the ability to identify the next sale trend of the products in the ecommerce website is also essential to the electronics store owner to plan for marketing spend to maximize the revenue. Therefore, time series (TS) forecasting is performed to help the store owner understand the factors contributing to the sales trend through sales forecasting.
Keywords: predictive business analytics, ecommerce websites, electronics store, market basket analysis, recommender system, time series forecasting

Project Background + Objectives

PROJECT BACKGROUND

Ecommerce, a global phenomenon, is facing great competition with multiple shopping items. Indeed, the contribution towards country economy growth is tremendous. Disappointedly, despite the growing big data yet it is under utilized, not well understood on its usefulness, & not fully helping in profitable decision-making process.

ISSUES & PROBLEM STATEMENT

Transactions under Ecommerce is huge, causing customer personality & buying power per transaction is hard to be understood. This resulted in the needs on product affinity analysis & likelihood of customer items purchase & enable cross-selling marketing.

Traditional methods are unable to recognize customers' preference and customers are not well treated. Study on the customer past event data would help to generate personalized list of recommended items to each customer.

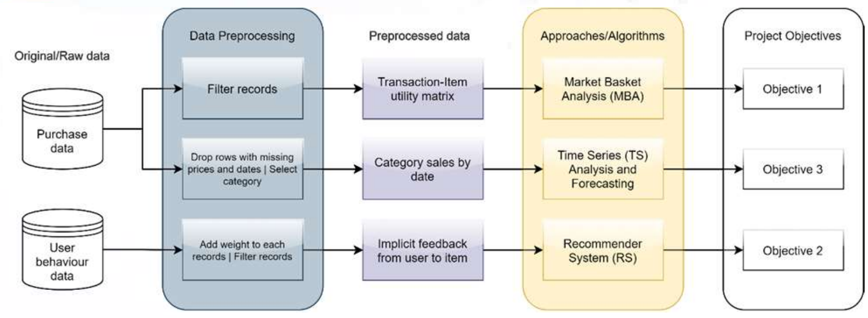
Sales trend prediction is missing but the forecast is beneficial, where the management can decide based on the predictive analysis in marketing budget plans & merchant inventory restock early notification.

OBJECTIVES

1. To discover the associations and relationships between electronic products in the transactions by **MBA** (i.e., support, confidence & lift measures), for items bundles sales to maximize revenues in ecommerce.
2. To generate personalized recommendations to every ecommerce customer by **RS**, targeted to improve the user experience, encourage user engagement & improve user loyalty towards the platform.
3. To discover sales trend & predict next seasonal sales using **TS forecasting**, to enhance store owner plan on next marketing actions & campaign.

Methodology

OVERALL PROJECT FRAMEWORK



Data Preprocessing steps are performed using Python (v3.8).

WHAT ARE THESE DATASETS ABOUT?

Purchase data (Source: https://www.kaggle.com/mkechinov/ecommerce-purchase-history-from-electronics-store)

>2.6M purchase records from the ecommerce platform of an electronics store For 11 months (Jan – Nov 2020) – each record corresponds to the purchase event of only one unit of item.

User behaviour data (Source: https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store)

>100M user behaviour history data in the ecommerce website of the electronics store in Oct & Nov 2019.

APPROACHES AND ALGORITHMS

Market Basket Analysis (MBA) – RapidMiner Studio (v9.9)

- Purchase data → Transaction-Item utility matrix
- FP-Growth Algorithm: generate frequent itemsets
- Create association rules: support & confidence min
- Find interesting rules – discover hidden relationships and purchase patterns of electronics products sold.

Recommender System (RS) – RapidMiner Studio (v9.9)

- User behaviour data → User preference data
- Collaborative Filtering RS (CFRS): User behaviour (view, cart, purchase) as Implicit feedback
- User-to-User, Item-to-Item, BPRMF: k, learn rate
- Identify the best CFRS in recommending products to the customer on ecommerce platform.

Time Series (TS) Analysis and Forecasting - Python (v3.8)

- Purchase data → Daily category sales data
- ARIMA and Machine Learning (XGBoost Regressor)
- Discover sales trend, compare model performance in sales forecasting task for selected category.

Experiment and Analysis: Market Basket Analysis (MBA)

Experiment & Result

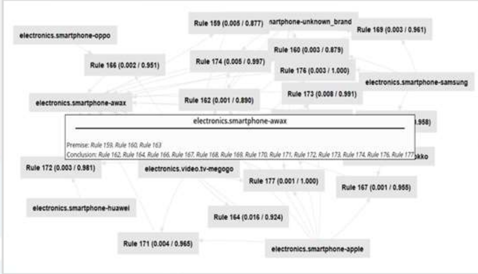
Support Min: 0.05; Confidence Min: 0.05

Support	item
16 %	Electronics-smartphone-Samsung
9.6 %	Electronics-smartphone-Apple
7.6 %	Electronics-smartphone-Huawei
5.4 %	Electronics-video.tv-Samsung

Support Min: 0.001; Confidence Min: 0.05

Support	Item 1	Item 2
1.6 %	Smartphone-Awax	Video.tv-Megogo
1.3 %	Smartphone-Samsung	Smartphone-Awax
0.8 %	Smartphone-Samsung	Video.tv-Megogo
0.7 %	Smartphone-Apple	Smartphone-Awax

Support	Item 1	Item 2	Item 2
0.8%	Smartphone-Samsung	Smartphone-Awax	Video.tv-Megogo
0.5 %	Smartphone-Awax	Video.tv-Megogo	Smartphone-N brand
0.4 %	Smartphone-Apple	Smartphone-Awax	Video.tv-Megogo
0.3 %	Smartphone-Samsung	Smartphone-Awax	Smartphone-N brand

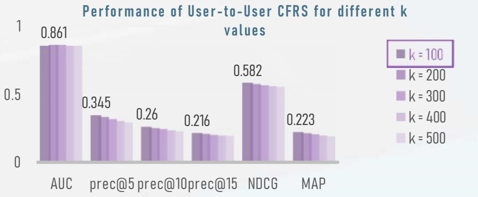


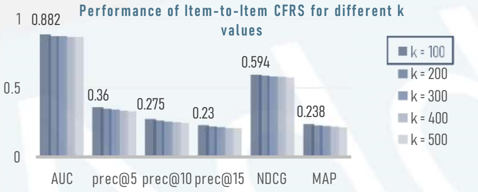
In short, higher level of support is preferred. Similarly, higher the confidence, greater the likelihood that the two itemset is accepted. Lift summarises the strength of association between the transaction which implies larger the lift, greater the link between the two.

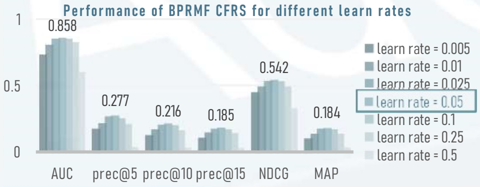
BIZ APPLICATION: Ecommerce Website, bricks-and-Mortar stores, Retail, Banks, Insurance, Medical, Telcos, Manufacturing

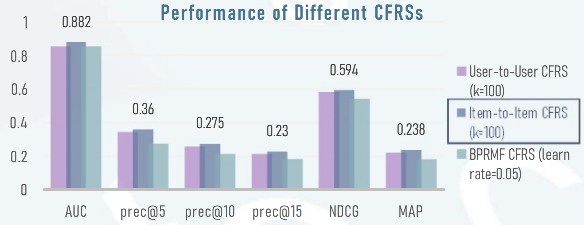
BENEFITS: Create bundle sales, increase sales & customer satisfaction, optimize product placement, special deals offers → improve shopping experience & brand loyalty & revenue

Experiment and Analysis: Recommender System (RS)









List of products in the list of top 10 recommended products by 3 CFRSs and their corresponding average rankings (Customer ID: 551211823)

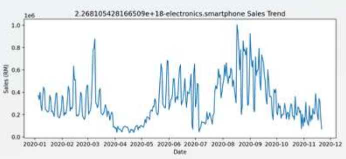
No	Product ID	Product Name	Average Ranking
1	2681	electronics.smartphone-xiaomi	1.33
2	2641	electronics.smartphone-apple	1.67
3	2655	electronics.smartphone-huawei	4.33
4	2571	electronics.clocks-apple	4.67
5	2668	electronics.smartphone-oppo	5.33
6	2425	electronics.audio.headphone-sony	5.67
7	2375	electronics.audio.headphone-jbl	6.33
8	2325	electronics.audio.headphone-apple	9.00

User-to-User CFRS: Optimal k = 100
Item-to-Item CFRS: Optimal k = 100
BPRMF CFRS: Optimal learn rate = 0.05
(Bayesian Personalized Ranking Matrix Factorization)

OVERALL: The best CFRS in recommending products to the customer on ecommerce platform is the Item-to-Item CFRS with weighted k-NN (k=100). It has the highest values of AUC, precision metrics and NDCG. The store owner can implement an item-to-item CFRS which is capable of recommending relevant products that might have not known, but similar to those customers have interacted before (e.g., recommending Xiaomi smartphone to customer interacted with other electronic products of Xiaomi brand). The list of recommendations is generated by comparing the item similarities in terms of the interactions between customers and products in the system.

Experiment and Analysis: Time Series (TS) Analysis and Forecasting

Time Series Plot



The line chart shows the smart phone category sales from January to November 2020.

Differencing

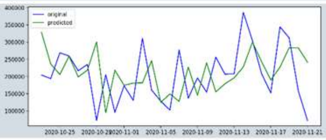
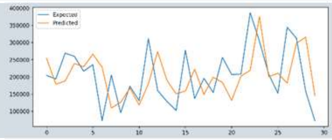
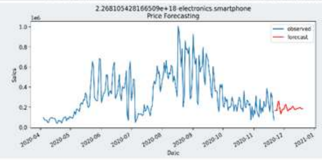
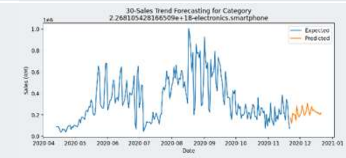
```
ADF TEST:

The test statistic: -2.164866
p-value: 0.219243
Critical Values:
1%: -3.451
5%: -2.871
10%: -2.572
```

The ADF Test is conducted on the category sales. It shows the p-value is 0.22 which is more than standard significant value of 0.05, so the null hypothesis is not rejected, and it is non-stationary. Differencing is needed.

ARIMA Residual Diagnosis

The ACF and PACF plots of residuals shows almost 100% of sample Auto-correlations are meets the criteria for a good forecast and predict.

ARIMA & XGBoost Experiment and Analysis

	ARIMA(7,1,0)	XGBoost Regressor (Best Model)
Parameters	p = 7, d = 1, q = 0	n_estimators: 100, learning_rate: 0.1, max_depth: 10
Performance metrics	MSE: 8954281994.78 RMSE: 94627.07 MAE: 74608.42 MRE: 0.50	MSE: 7061066569.06 RMSE: 84030.15 MAE: 67641.38 MRE: 0.40
In-sample 30-days Prediction		
Out-of-sample 30-days Forecast		

XGBoost regressor model has lower MSE, RMSE, MAE, and MRE than ARIMA model, which means XGBoost is good forecasting model for this category sales dataset.

Conclusion: MBA helps the business owner to make better decisions on marketing strategies and planning, such as cross-selling and bundle sales. RS can be implemented to provide personalized recommendations to customers. TS forecasting helps to predict future events and business owner can take early actions such as restock and market retreat planning.

Future Research: RS can be implemented using more advanced techniques such as Deep Learning and Reinforcement Learning. More advanced TS technique involving multivariate TS data can be implemented to study the sales data for multiple categories and the relationships among them.