**EPSILON DX DATA SCIENCE CASE STUDY**

**PROBLEM - What are we trying to solve?**

Predict the top 3 product (amongst a pool of 21 products) that the customer is going to purchase

**DATA**

Training Data ~1 million rows with transactional data

TRANSACTIONAL DATA SAMPLE

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **transaction\_id** | **item\_id** | **item\_qty** | **InvoiceDate** | **amount** | **customer\_id** |
| **536365** | 85123A | 6 | 12/1/10 8:26 | 2.55 | 17850 |
| **536365** | 71053 | 6 | 12/1/10 8:26 | 3.39 | 17850 |
| **536365** | 84406B | 8 | 12/1/10 8:26 | 2.75 | 17850 |
| **536365** | 84029G | 6 | 12/1/10 8:26 | 3.39 | 17850 |
| **536365** | 84029E | 6 | 12/1/10 8:26 | 3.39 | 17850 |
| **536365** | 22752 | 2 | 12/1/10 8:26 | 7.65 | 17850 |
| **536365** | 21730 | 6 | 12/1/10 8:26 | 4.25 | 17850 |
| **536366** | 22633 | 6 | 12/1/10 8:28 | 1.85 | 17851 |
| **536366** | 22632 | 6 | 12/1/10 8:28 | 1.85 | 17851 |

**PERFORMANCE METRIC**

Match rate – Number of instances where the category purchased is part of the top 3 products predicted against the total predictions

E.g.: Let’s assume a total of 5 products available for the customer to purchase from (choice pool – A, B, C, D, E)

Number of transactions to be predicted – 5

Top 3 products predicted – ABC, Customer purchased A – Match

Top 3 products predicted – BCD, Customer purchased B – Match

Top 3 products predicted – BDE, Customer purchased A – Doesn’t Match

Top 3 products predicted – ABD, Customer purchased B – Match

Top 3 products predicted – CDE, Customer purchased A – Doesn’t Match

**Match Rate –**

Total instances where predictions match = 3, total predictions = 5

Match Rate = 3 / 5 = 60%

**DATASET**

Training Dataset: Total 22 months data

|  |  |
| --- | --- |
| 20 Month Data | 2 Month Data |
| Training Data for our model  ~ 1MM data points  5501 unique customers | Evaluation/Submission Data  Withheld Dataset for Performance Evaluation |

* While submitting, do ensure that the code is properly documented
* Please explain your approach in bullet points along with the code
* The case study will be provided on a Thursday and the submission (prediction file - submission.csv & code) is expected by Monday morning at 9 A.M
* This case study will be discussed in subsequent discussion