Using Databricks for data engineering and analytics

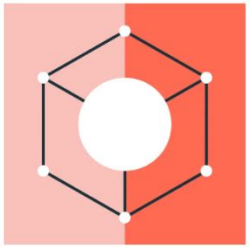
Customer segmentation with RFM

Muzychuk Andrii, InterLogic

Agenda

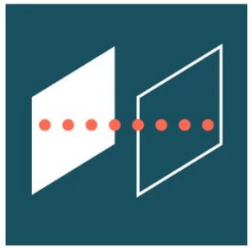
- Databricks Intro
- Use case motivation and model assumptions
- How to calculate RFM values
- Simple customer segmentation based on quantiles
- Customer segmentation with K-Means
- What to do next

What is Databricks?



Simple

Unify your data warehousing and AI use cases on a single platform



Open

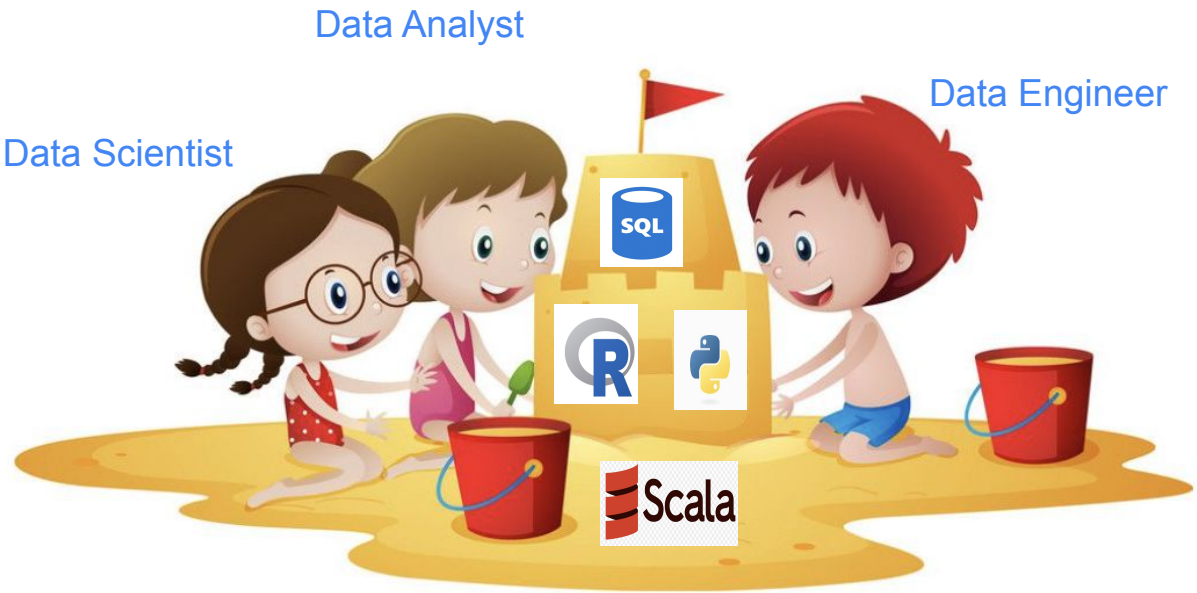
Built on open source and open standards

Multicloud

One consistent data platform across clouds

https://www.databricks.com/

So, why Databricks?



Visit https://docs.databricks.com/introduction/index.html to find out more …

Use case motivation

From a marketing perspective, it is valuable to understand the characteristics and preferences of your best customers for at least two reasons:

- to keep them as customers
- to target marketing efforts toward prospects who are most likely to respond.

A bit of history

First employed by direct marketers sending catalogs via direct mail in the 1940s

Main **objective** was to avoid sending costly print catalogs to customers who were unlikely to convert.

Catalogers would maintain and update a 3×5 index card for every customer in their file. Each index card was ranked by:

- when the customer made their last purchase
- how often they purchased, and
- how much the customer had spent in their lifetime

… and it worked !

RFM Model assumptions

- Customers who have purchased more recently are more likely to purchase again when compared to customers who have purchased less recently
- Customers who purchase more frequently are more likely to purchase again when compared to customers who have purchased only once, or less frequently
- Customers who have higher total monetary spend are more likely to purchase again in the future when compared to customers who have spent less monetarily

R - recency **F** - frequency **M** - monetary value

CDNOW Dataset

- Contains the entire purchase history up to the end of June 1998 of the cohort of 23,570 individuals who made their first-ever purchase at CDNOW in the first quarter of 1997.
- Contains 1/10th systematic sample

Link to download data: https://www.brucehardie.com/datasets

```
→  rmf head -3 data/CDNOW_master.txt
00001 19970101  1    11.77
00002 19970112  1    12.00
00002 19970112  5    77.00
→  rmf head -3 data/CDNOW_sample.txt
00004 0001 19970101  2    29.33
00004 0001 19970118  2    29.73
00004 0001 19970802  1    14.96
```

Note the difference in structure of full data set and sample

Demo time

To reproduce:

1. Sign up to Databricks Community Edition
2. Download the CDNOW dataset
3. Upload notebook to Databricks
4. Upload CDNOW dataset sample to Databricks
5. Set path variable in the notebook.
6. Have fun :)

Summary

RFM concept is a relatively simple and easy to interpite, but yet, quite a powerful

RFM values are used as input for Customer Lifetime Value model (CLV)

repo : https://github.com/anmuzychuk/rfm-des23

Where to next?

- Databricks has a lot to offer, explore documentation and solution sections

- Learn / improve data transformation with pyspark

- Improve segmentation model by choosing optimal? number of segments using Silhouette score

- Explore RFM modifications

- Check this lecture for CLV motivation