**Objective**

Forecast sales for a new product when launched based on other products. Try and make a model which shows geographic similarities or is normalised by country.

**Data Sources**

ProductMaster

* **Product** - product id/name
* **Manufacturer** - company manufacturing product
* **Molecule** - key chemical component of the product (molecule:product is 1:many)
* Class - mechanism of action (pill, injection, shot etc.)
* bioSimilar - product which is a knockoff. Here, "bioSimilar" holds <blank> for "not a biosimilar" and holds the base (copied) product if it is a biosimilar.
* Division - unknown column

One product has one molecule and one molecule should have one product + biosimilars

The market share is calculated based on the molecule

MySource (Sales Data) - Product + Indication should be unique

* **Country**
* Product
* Area - continent
* Sub\_Area - unknown column
* Date
* Value - quantity shipped
* **Template** - thereuptic area like immunology, oncology, HCV
* **Indication** - specific disease (leukemia) – indication belongs to one and only one template
  + product - indication has a many-many relationship
  + Molecule - indication has a many-many relationship

**Exploratory Data Analysis**

Phase = intro, growth, mature, maturity2, declining - could use K-means

#the first month look at market share -

#by indication, for each indication, qtysales of product/total sales of all products for that indication

#can have different lauches per indication

#can have difference lauch date by country

#molecule matters for efficacy only (with market size, look at molecule and molecule will correspond to product before LOE)

Compare competitors/biosimilars based on their phases - not biosimiolar might hurt space (if same indication at same time)

For each country, each product, you graph and then do clustering based on features

**Potential Features**

1) months since launch

2) indication or disease

3) competitors (biosimilars)

4) market size (group by aggregate sum of indication, date irrespective of products), market share = drugs sold/market size normalized by country

5) manufacturer

-

6) Popularity - using google trends

6) mechanism of action (pill, injection)

7) price

8) prevalence rate of disease for a country

9) Exogenous events like COVID-19 (data from John Hopkins) -

10) size of manufacturer

11) GDP

**ROUGH NOTES**

PLC Clustering -> lifecycle clustering to determine phases of product (3-4 PHASES)

Look into phases like growth phases and cluster countries which are similar to each other

#CLUSTERING 1

#For every indication, country, product pair :

# -- > MARKET SHARE OVER TIME

# -- > GROWTH = DIFF IN PATIENT VALUE (MONTH OVER MONTH)

--

# -- > MONTH SINCE DATA AVAILABILITY (Where Lauch Date is Date when it is >2%) (REMOVE) - this is a big assumption as first point was 89% for some as seen before

Indication \ Country \ Product \ Market

H \ Italy \ Prod15 \ Jan \ .2 \ 1%

H \ Italy \ Prod15 \ Feb \ .4 \2%

I \Germany\ Prod16\Feb \ .5 \2%

# Normalised using z-scores and did K-Means CLustering

# It separates out indication, country, product pair only and groups biosimlars together

# CLUSTERS : prod15, prod43, prod32,

#-----

#Heirarchical Clustering for each - (IN PROGRESS)

#Country, Indication, Product :

#- total Life length

#- max market share

#- avg market share

#- avg growth rate in plc growth phase till it reaches max market share

#- avg growth rate in decline is after that

\*ADD BIOSIMILAR INFO

**#Multiple products, multiple molecules, multiple indications, multiple countries**

**#OVERALL IDEA : New Product, we have certain (country, indication) - PLC**

1. **Per country+indication pair, which are similar regardless of product - LOOK FOR BROADER AGGREGATED FEATURES - biosimilar adoption rate, market size,competitive nature of the market, GDP of country (population normalised) - rather individual product - THINK OF FEATURE THAT FIND DIFFERENCES BETWEEN COUNTRIES**
2. **Per country + ind cluster, clustering of PLC into 4 phases**
3. **Forecasting**

#COUNTRY+INDICATION LEVEL FEATURES -

--

#2015-2020

----

5 year monthly data

COUNTRY/ INDICATION/PRODUCT/BIOSIMILAR/MARKET SHARE/VALUE/TIME

*40 8 14*

Italy A Prod1 Bio 5% 200 March/20

Italy A Prod2 Not 50% 100 ----->

Italy A Prod3 Bio 45% 200

Italy C Prod1

Germany A Prod1

Germany C Prod1

India D Prod1

India D Prod2

Aggregated over products at COUNTRY/MONTH level - save as CSV

COUNTRY/ INDICATION/BIO MARKET SHARE/MARKET SIZE (POP NORMALIZED)/GDP/POPULATION/GROWTH(TO DATE)

Italy A 50% 500 (can be normalized) March 20

Italy A 50% 500 (can be normalized) April 20

Italy C

Germany A

Germany C

India D

Country-Level features : (per month, per country) - dont cluster per month

ALIGNMENT AT TIME -> SMOOTHING AT TIME -> FEATURE CREATION -> ANALYSIS/VISULALISATION -> CLUSTERING

Visuals - Histogram per variable, Scatter plot per variable combination

-> aligning products over time

1. NUMBER OF BIOSIMILARS IN MARKET - visuals
2. PER CAPITA GDP - visuals
3. MARKET SIZE/POPULTATION - visuals
4. GROWTH RATE - visuals
5. MARKET SHARE OF TOP 4 PRODUCT - visuals
6. MARKET SHARE OF BIOSIMILARS - visuals
7. NUMBER OF PRODUCTS - visuals
8. HHI COMPETITION INDEX - visuals
9. CORRELATION BETWEEN COUNTRIES - from kartikay
10. WEB SCRAPING FEATURES
11. PHARMACEUTICAL PROFILING
12. RUTA’s SLIDES DATA - LOE STATUS etc.

Indication-Level features

Features :

1. BIOSIMILAR MARKET SHARE (PROXY FOR COMPETITON) - Arushi
2. NUMBER OF COMPETITORS - Arushi
3. MARKET CONCENTRATION RATIO BY INDUSTRY - Arushi
4. MAX MARKET SHARE - Arushi
5. AVG MARKET SHARE - Arushi
6. MARKET SIZE/POPULATION - Kartikay
7. GDP - Kartikay
8. AVERAGE GROWTH CAGR - Kartikay
9. AVERAGE DECLINE CAGR

Do we need another level aggregated over time ?

We may want to aggregate across indication as another level?

We might want to want aggregate at country level and use indication as a feature OR per indication we can create country clusters

ITALY A, INDIA B, INDIA C are same cluster 1

ITALY B, GERMANY A are same cluster 2

Train and Test divide here!

FOR EVERY COUNTRY/INDICATION CLUSTER :

COUNTRY/ INDICATION/PRODUCT/BIOSIMILAR/MARKET SHARE/VALUE

compare their PLC

Normalizing PLC’s after this - Exponential smoothing

Cluster it into phases

---

Kartikay Discussion - 24/9

1. General AIM : General New Product Launch PLC
2. Canada : Current PLC - growth, decline phase

Features - bayesian regression by Chandrakant

Flow - continuous

Clean code - reversibility of code

TO DO :

1. Test and train
2. @Tanmay - determine phases
3. Decide country clusters
4. Country based cluster - indication
5. NAN values - median
6. Fix normalisation in aggregation of plc