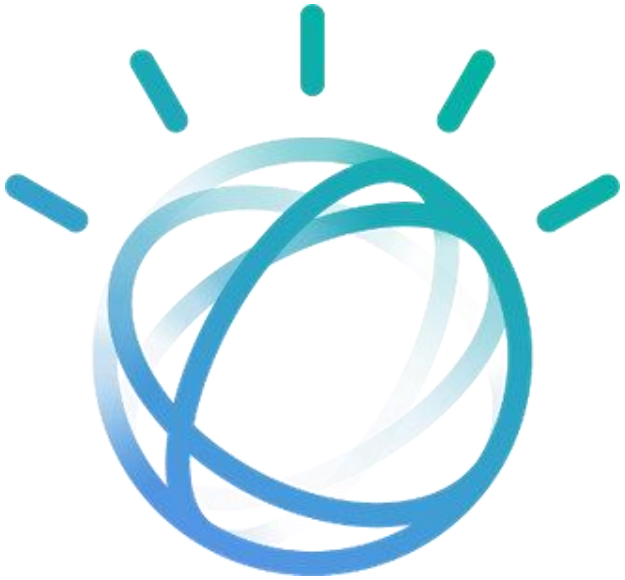
Sanofi CASA

Patient Enrolment Curve Prediction

TOC

- Why is this important and project Goals
- Result of the modelling vs Actual vs Baseline
- Model build
  - Data set description
  - EDA of set as is
  - Hypothesis and step by set methodology
    - Data normalization
    - Clusterization
    - Classification
    - Regression
- Next steps
- Tools



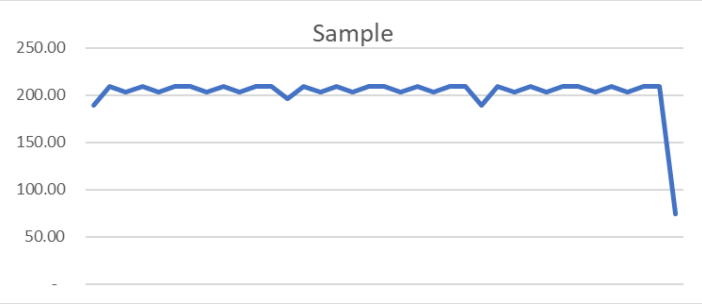
Why it is important

84 Unique Studies to forecast
547 – Unique country – studies combinations to forecast.
About 100 of them has patient target more than 50

2 weeks to complete the 1st cut of Baseline

Current process

As a result of short time team has no time to estimate Patient enrolment manually for each Country - Study.
The current approach is equally spread Patient Target across study-country duration and manually adjust Patient curve on total study level (not by country)

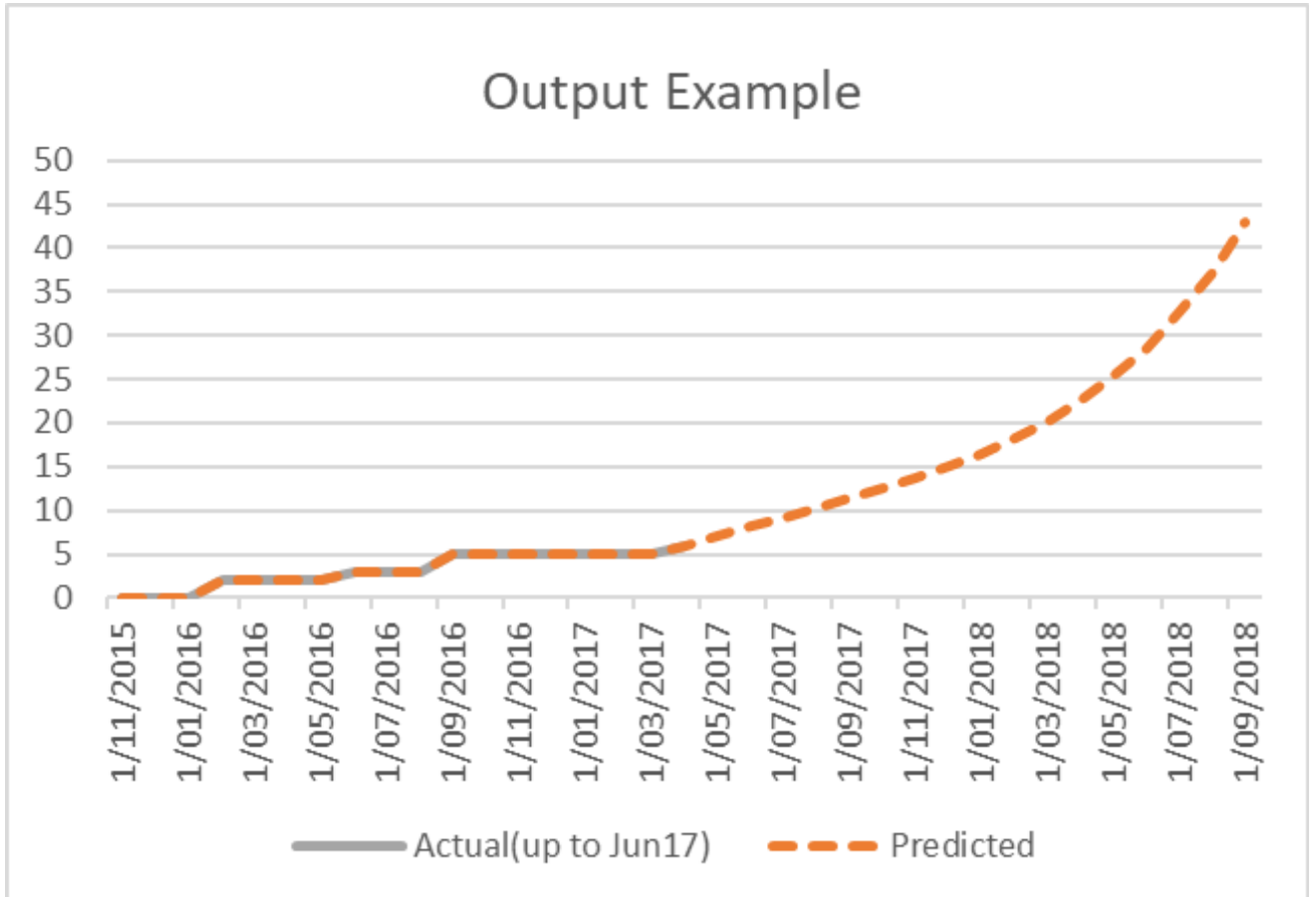


Why it is important

- Patient Curve is the main driver of Study expenses through the year.
- More precisely curve is predicted -- better the Cost Planning for the next year
- Adding automatically defined Patient curve can help on many aspects:
  - Improve precision of Cost prediction by country
  - Potentially, **Remove the step of manually TBD** (Dummy country) **adjustments on Study level** Which makes the overall calculation less intuitive and result less explainable.

Project Goal and General Assumptions

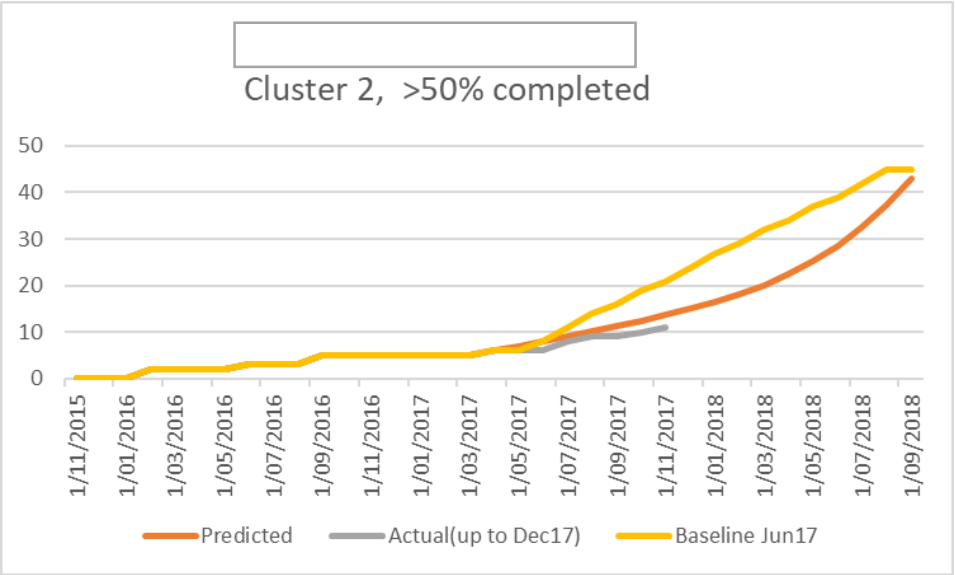
- The project Goal is to automatically predict monthly Patient Enrolment for each Study Country intersection.
- Each Country – Study has a given duration assumption
- Each Country Study has a target patient number which it should reach by the end of the Study Duration.
- There are on-going studies with some historical information
- There are not started Studies with no historical information

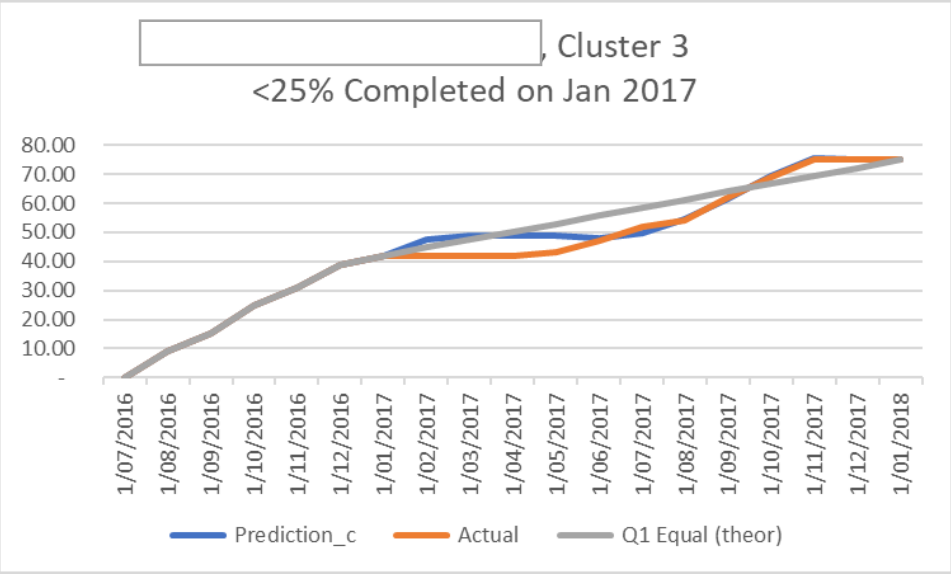


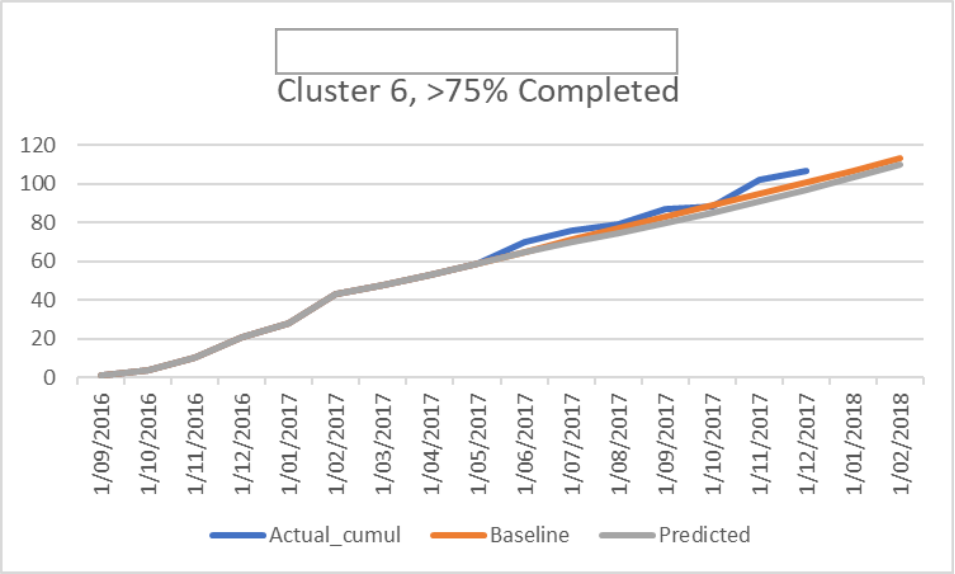
Modelling Result vs Actual vs Equal spread (currently)

The linear charts provide
Actual (by Dec 2017) vs Baseline vs Predicted data (train on actual up to Jul 2017) for different clusters.

To summarise, for studies with polynomial curve prediction model shows closer to real actual curve result







Data sets Description

Monthly Patient Enrolment

- Study
- Country
- **Year**
- **Month**
- Patient Enrolment amount

Study – Country Attributes

- Study
- Country
- **Start Date**
- **End Date**
- **Patient Target**

Study Attributes

- Study
- Therapeutic Area
- Study Priority
- Study Category
- Study Phase
- Ect..

Source: TM1 , 1158 **Country-Study** historical data

	Study	Country	Year	Year_Period	Month	Exercise	Measure	Amount
0	[redacted]8	France Corporate	2017	Y-05	M08	Actual	Monthly Inclusion	29
1	[redacted]8	France Corporate	2017	Y-05	M09	Actual	Monthly Inclusion	58
2	[redacted]8	France Corporate	2017	Y-05	M10	Actual	Monthly Inclusion	35

Source TM1, 1808 **Country Study**, historical/on-going and new

Study	Country_dim	Number of Centers	Patient Target	Number of Visits	FPI/FSI	LPI/LSI	FPI/FSI Date	LPI/LSI Date
[redacted]	Russia CSU	12	31	NaN	2012-07-17	2013-03-18	41107	41351
[redacted]	Ukraine	4	17	NaN	2012-08-02	2013-03-18	41123	41351

Source TM1, 795 **Study**, historical/on-going and new

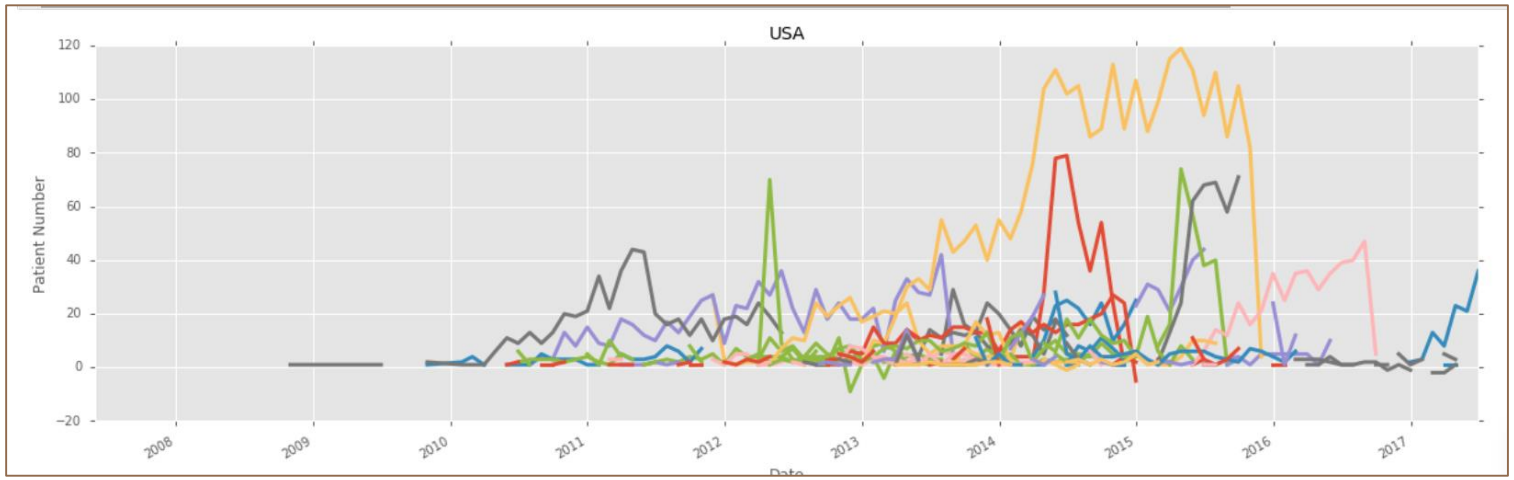
Out[457]:

ınamed:	Study	Country_dim	Patient Target	Number of Centers	Number of Visits	FPI/FSI Date	LPI/LSI Date	Priority value	Project code	Study Phase	Category
	[redacted]	Corp	8	1	5	42137	42137	Priority 1	GZ385660	I	G
2		Corp	32	14	8	41291	41541	Priority 2	SAR100842	IIa	G
5	TAC12340	Corp	60	15	22	41541	41724	Priority 1	SAR231893	IIa	G

First look at the data set

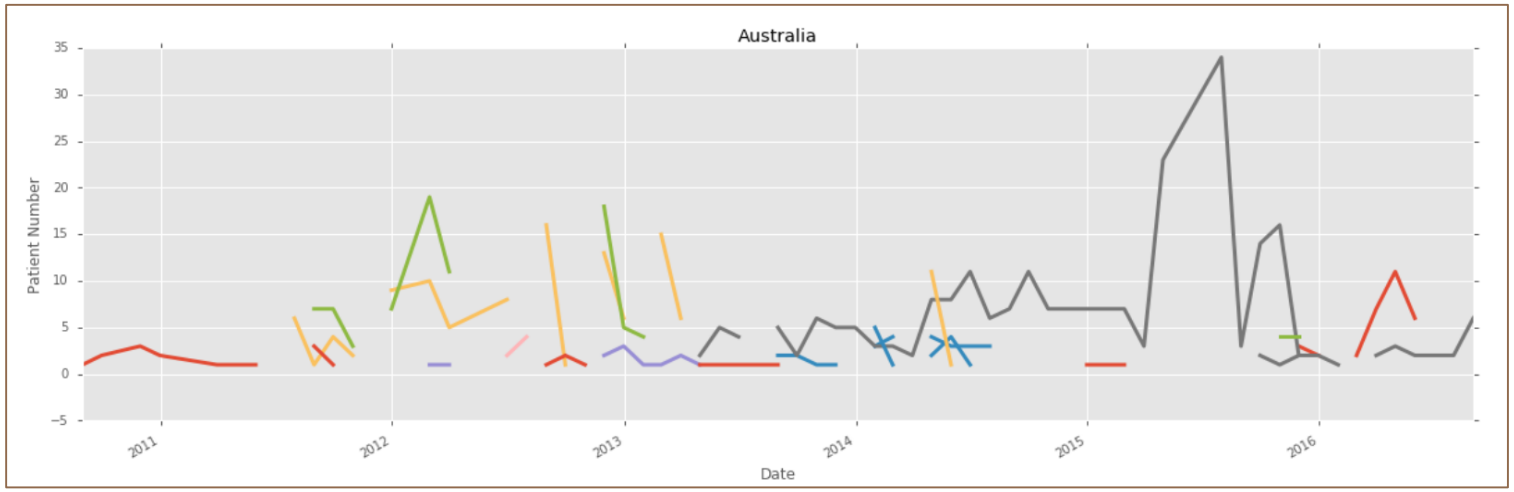
The line graph provides information on historical data of Monthly Patient Enrolment in USA for all completed studies, based on TM1 Data.

Overall, all lines hardly can be described by 1 function. The assumptions of the studies are quite different, so we can not directly compare them.



The line graph provides information on historical data of Monthly Patient Enrolment in Australia for all completed studies, based on TM1 Data.

In addition to USA analysis, we visually see that some studies have some significant gaps in data. As a result, applying Timeseries prediction functions can be challenging.



Hypothesis and Action Plan

The **hypothesis** is that although studies are different and have different assumptions, some of them have similar features and trends. We can group Studies with similar graph lines and identify features they share within group. Allocate new studies to one of the group so we can use historical data of them to predict Enrolment for a new study.

To check the hypothesis I have break down project to steps to do:

1 Normalization

- Use Cumulative Amount instead of monthly
- Normalize Number of Patients for each Country/Study to make them comparable
- Index Time periods
- Normalize time periods

2 Clusterization

- Identify for each study 3 point of compilation: 25%, 50%, 75%. Number of points are chosen arbitrary
- Apply hierarchical clustering algorithm. (one of the option)
- Limit number of clusters to less than 10 (selected arbitrary)

3 Classification

- Train the classification model on historical data
- Use the Model for Classification Of new and on Going studies
- Explore Feature importance of given model

4 Regression

- Prediction model for each cluster
- Predict Patient enrolment for each country study based on assigned class

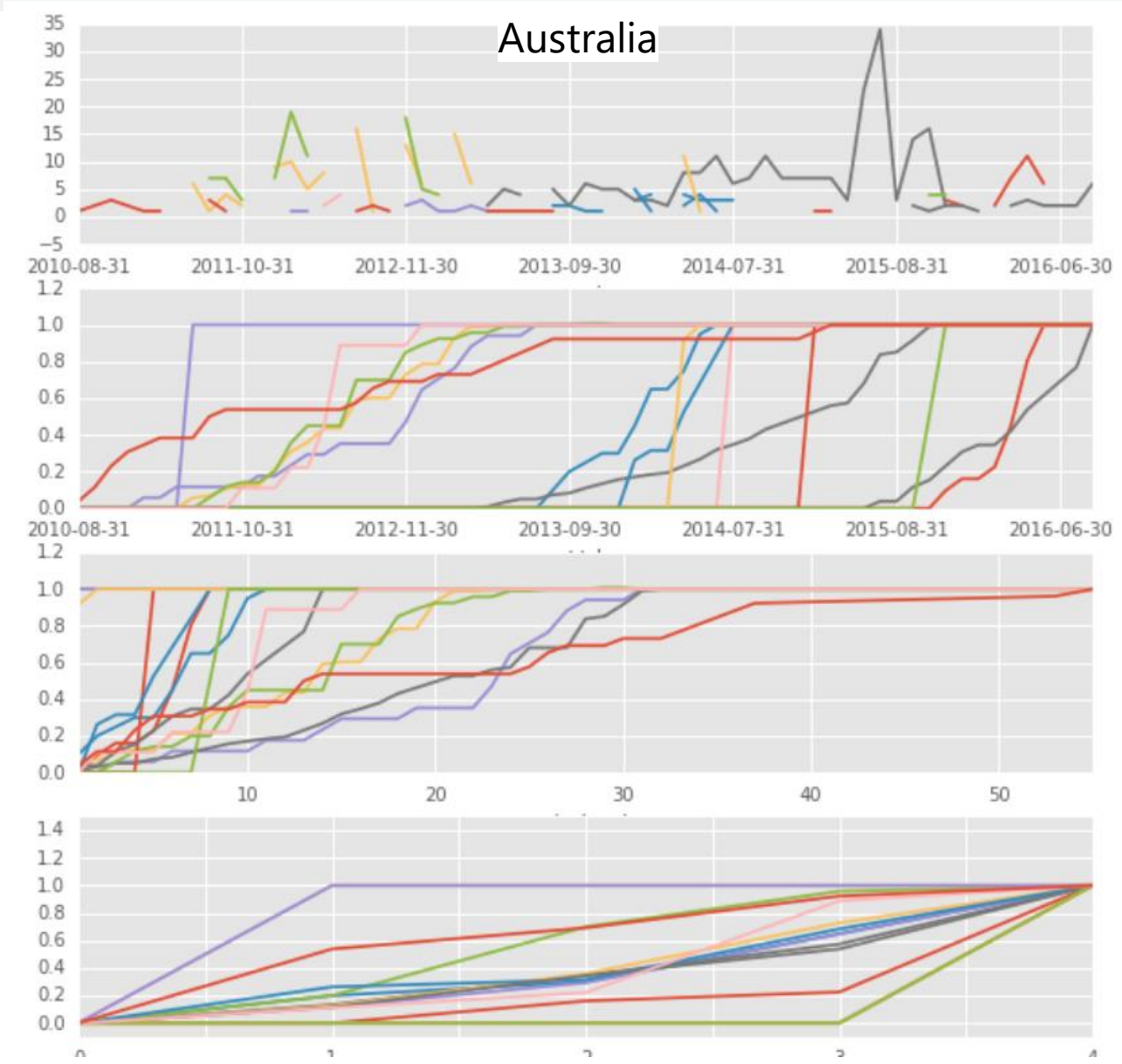
1 Normalization. Step by step transformation.

The chart provides monthly patient enrolment by study in Australia for historical studies only (by Jul 2017)

- Scale Patient Amount(0-100%) and Use Life to Date(cumulative) Amount instead of monthly amount.

- Replace datetime by Period indexes. The start point is the same for all studies.

  - Scale Period Index for all studies. Reduce number of Index point to 5. (0%, 25%, 50%, 75%, 100%).
  - Number of points is chosen arbitrary.



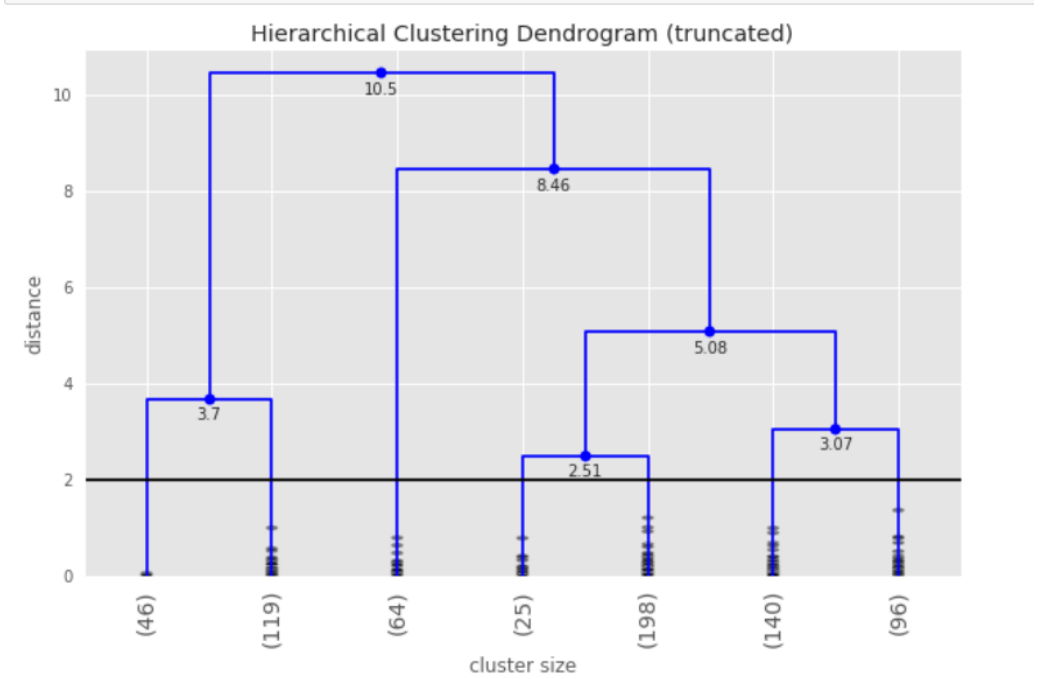
2 Clusterization. Hierarchical Clusterization.

Clustering techniques are used to group data/observations in a few segments so that data within any segment are similar while data across segments are different.

This **hierarchical clusters** are represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the country-study, the leaves being the clusters with only one country-study

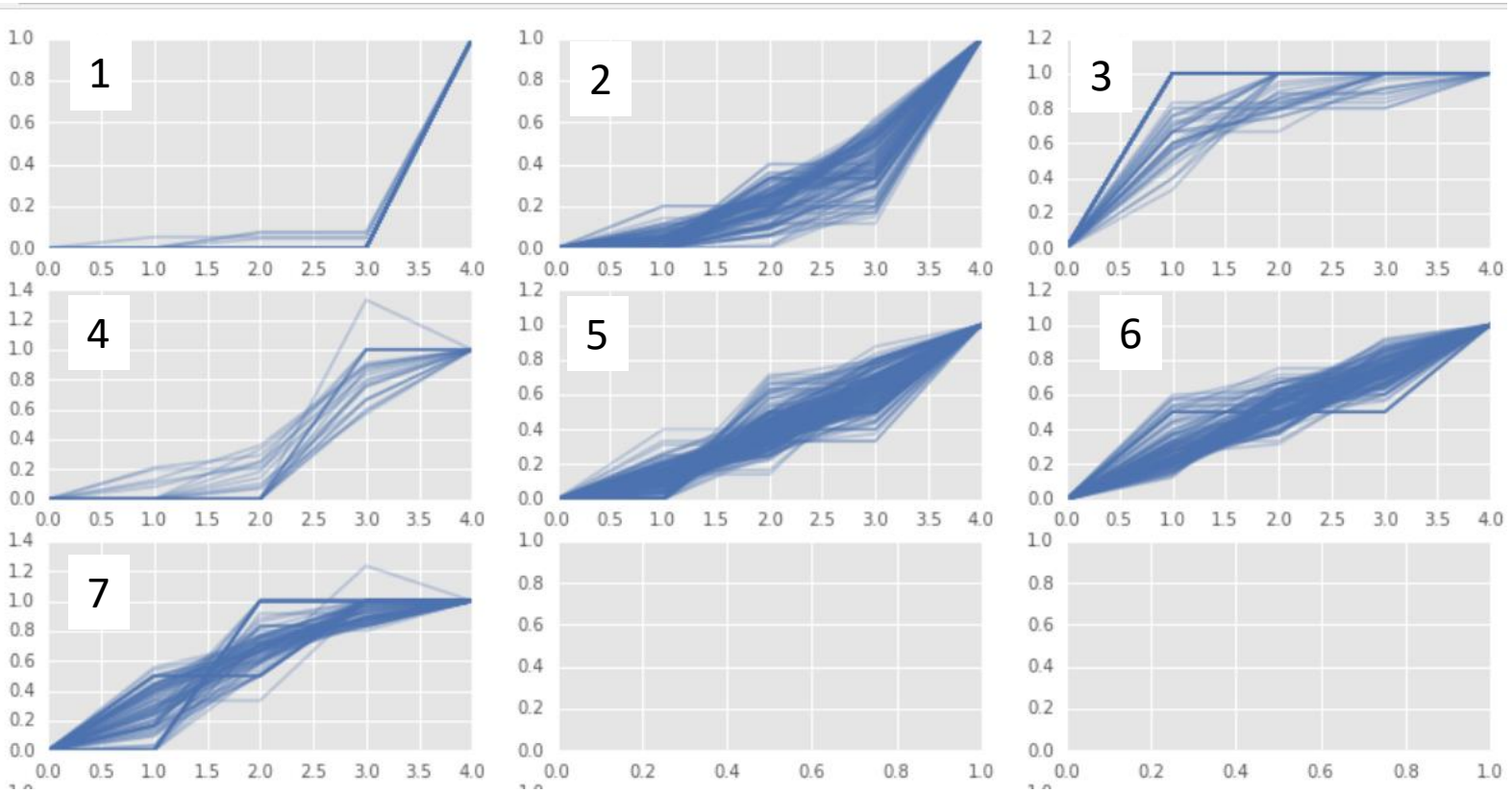
It uses common mathematical measures of distance to group samples.

The method is used in this concept called **Ward**.
It minimizes the sum of squared differences within all clusters.



- On the x axis you see labels. They are the indices of Country-Studies.
- On the y axis you see the distances (of the 'ward' method in our case).
- Horizontal black line is selected arbitrary to cut 7 Clusters

2 Clusterization. 7 clusters Result.



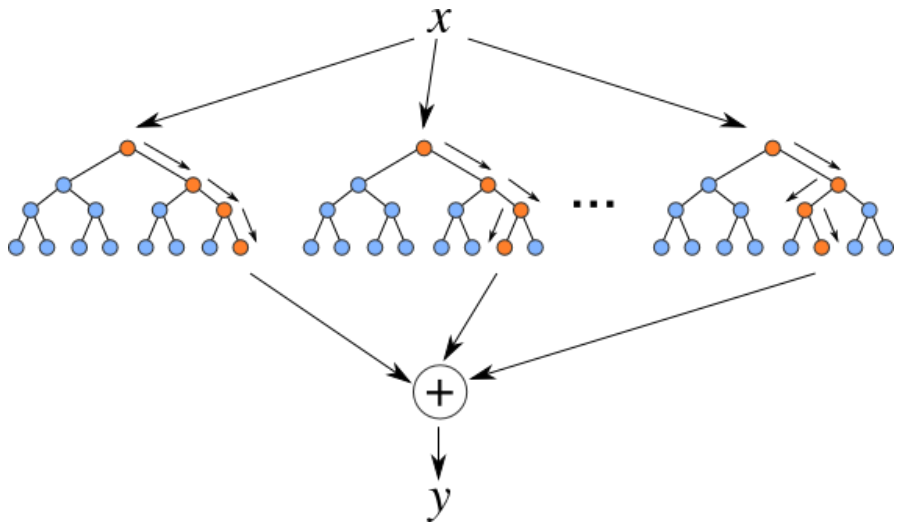
The multiple line graphs provide Cumulative Normalized Patient Enrolment Information of Country-Studies which were completed by Jul 2017, segmented to 7 clusters by Hierarchical Clusterization Algorithm.

We can see on the chart that each cluster has some visually recognizable and unique shape.
More importantly, some clusters provide us information's on typical country-studies *outliers.*
For instance, **Clusters 1 and 4** grouped all study profiles where actual patient enrolment started after 50% of study duration had already passed. Where **Cluster 3**, clearly shows us that a Patient Target has reached before end of duration.

3 Classification. Random Forest Algorithm

Why Random Forecast Algorithm?

- Random forest models are one of the most widespread classifiers used.
- Random Forest classifier can be modeled for categorical values.
- Random Forests are an *ensemble* or collection of individual decision trees.



Based on a model built on historical data we need to classify New and on-going studies.
On-going studies already have some historical data, so we can use it as additional variable to increase our accuracy.
The table shows that accuracy score increases for studies on a different completion stage

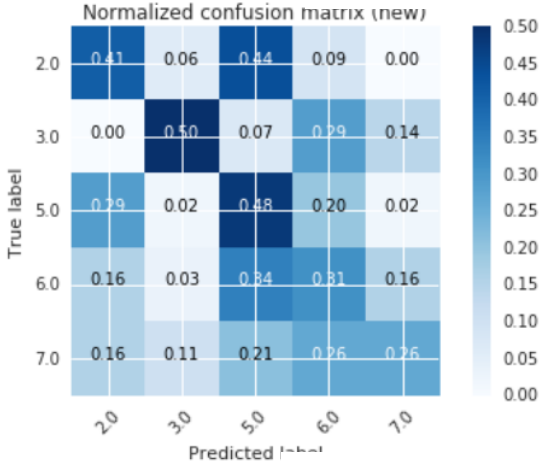
Country Study Subset to Classify	Classification Average Accuracy
Not Started Studies	33%
On-Going Studies <50% completion rate	57%
On-Going Studies <75% completion rate	81%
On-Going Studies >75% completion rate	**93%**

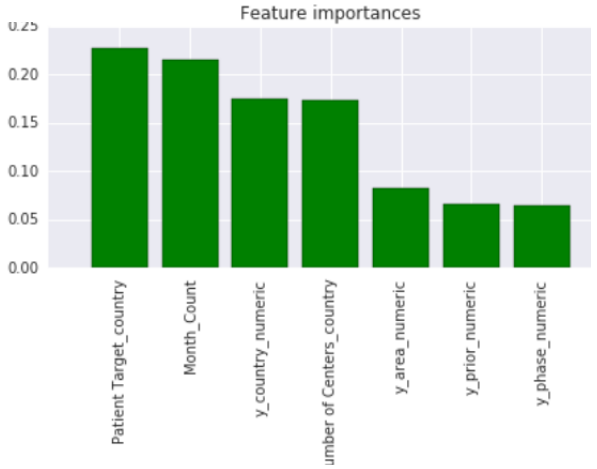
3 Classification. Confusion Matrix & Feature Importance

The matrix shows True Classes vs Predicted on a historical data set. For not started studies we can correctly predict only **33%.** And the chart next to it shows model **feature importance** – what are the most impactful variables. Patient Target and Study Duration has the biggest impact.

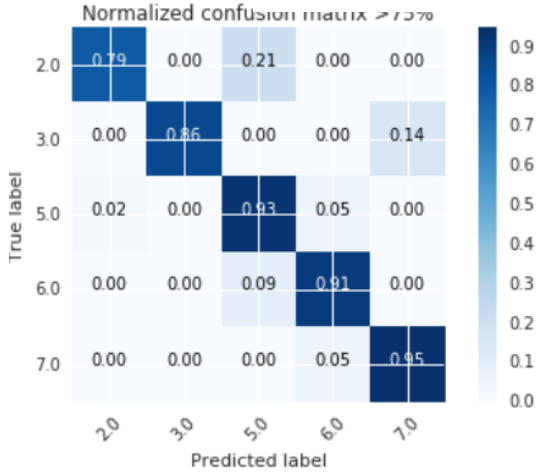
For studies which have more than 75% completion rate we can correctly predict only **96%.** Patient target and Duration is not that important as values in 1,2,3 (25%, 50%, 75%) of Patient Enrolment.

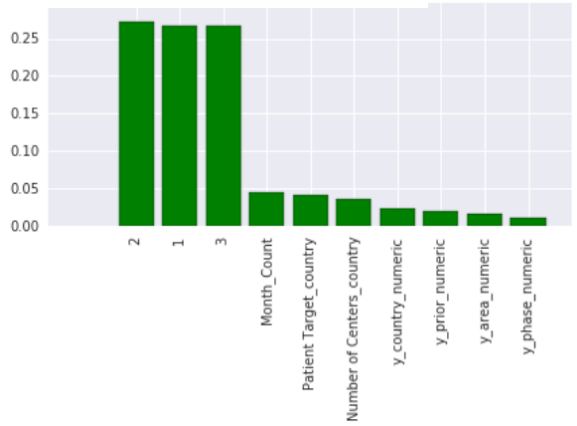
New studies prediction





>75% completion rate studies prediction





4 Regression Analysis

After classifications of all new and on-going country-study combinations we can predict patient enrolment ratio by identify the best – fit function for each cluster.

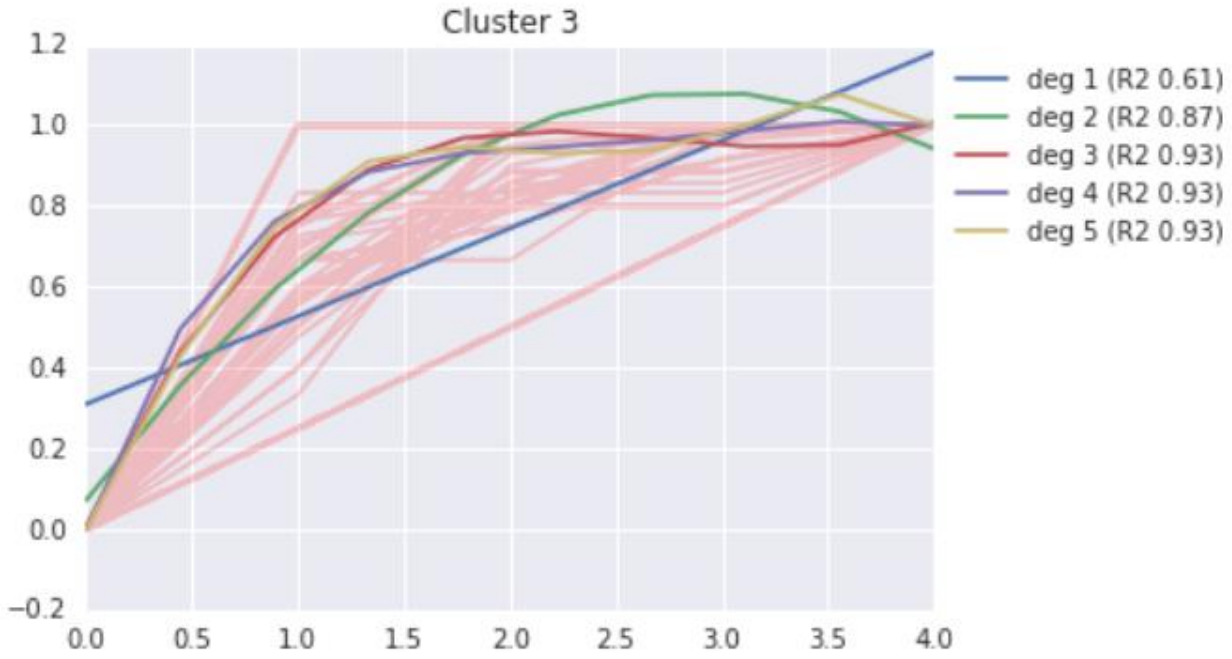
Most of the clusters have polynomial function, so in order to find the best fit we preform several attempts to find the best possible fit.

For each cluster we try to fit 5 different degree polynomial functions. The closer metric R2 to 1 than better fit we have.

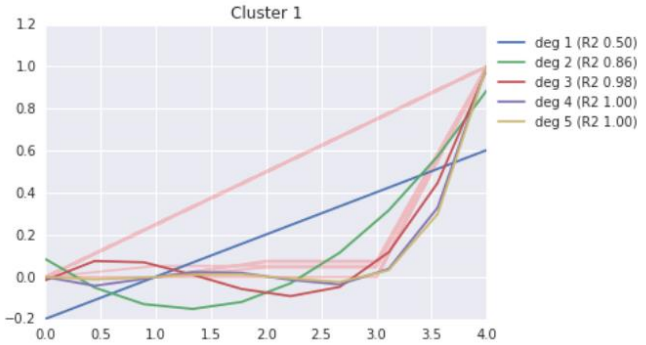
The chart provides several information of all historical studies Patient enrolment curves (in pink color).

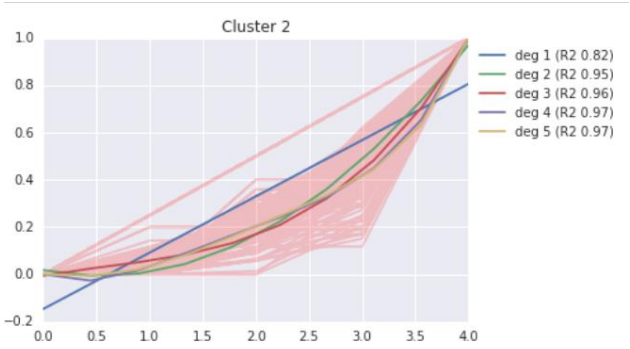
On top of it we can find 5 attempts of fitting the function. We see that starting from degree 3 we have a best fit.

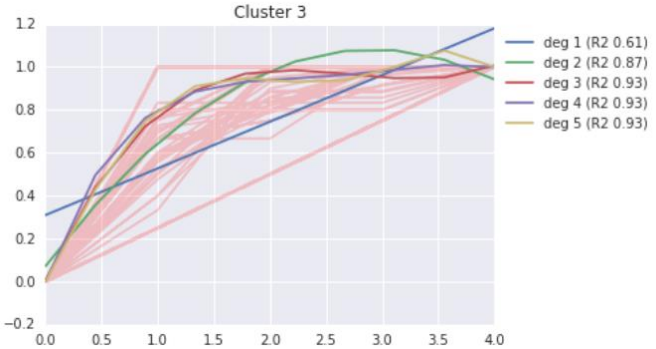
So we can use the result to predict New and on-going studies of Class 3

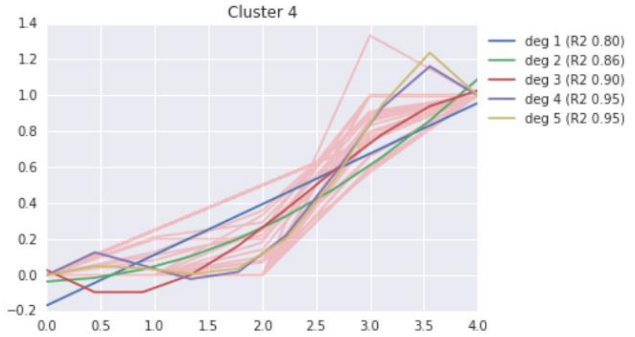


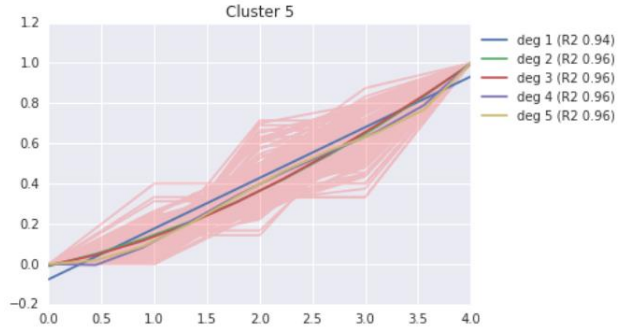
4 Regression analysis of all clusters

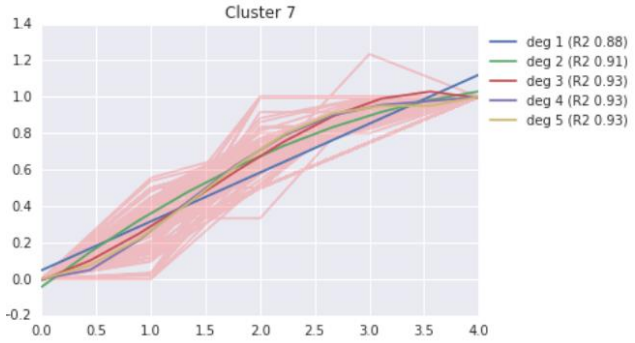








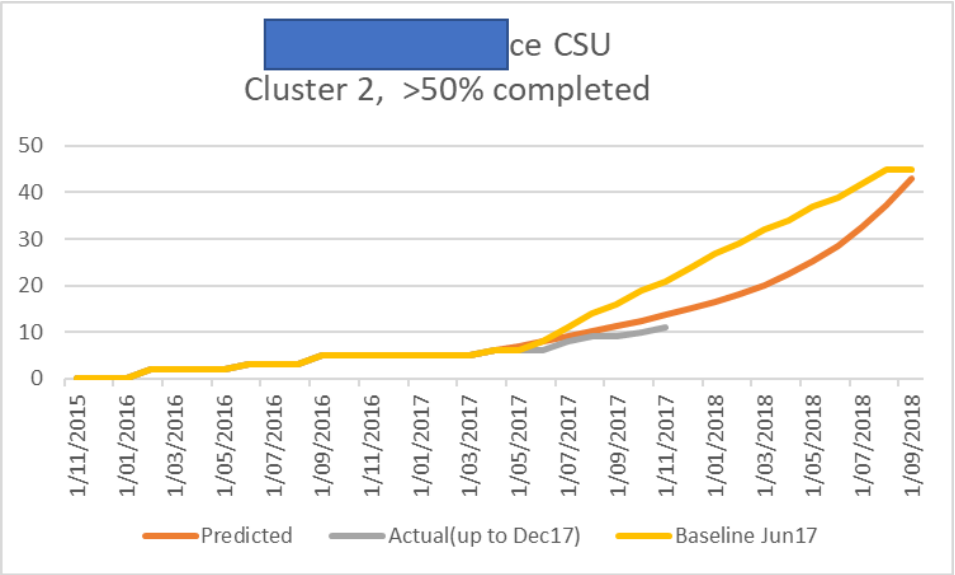


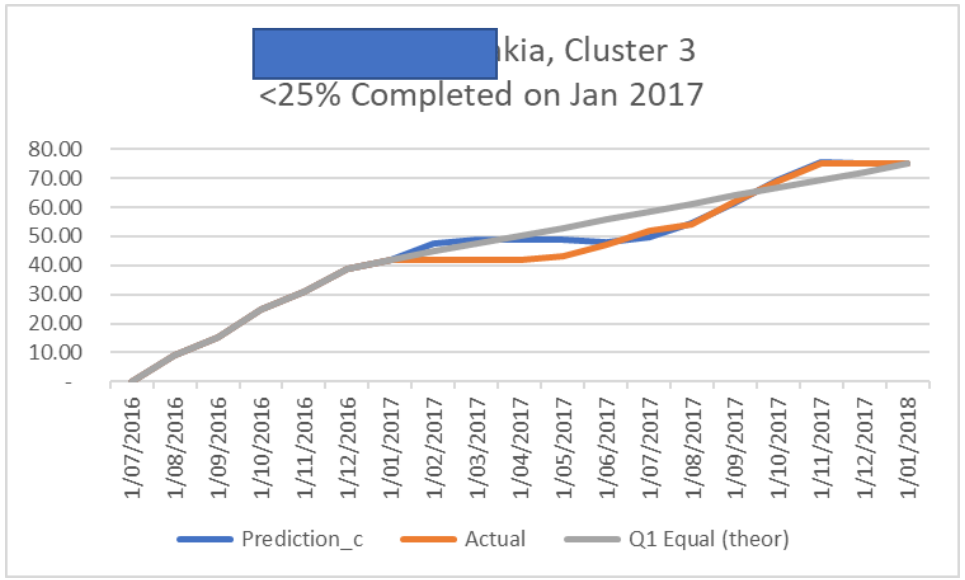


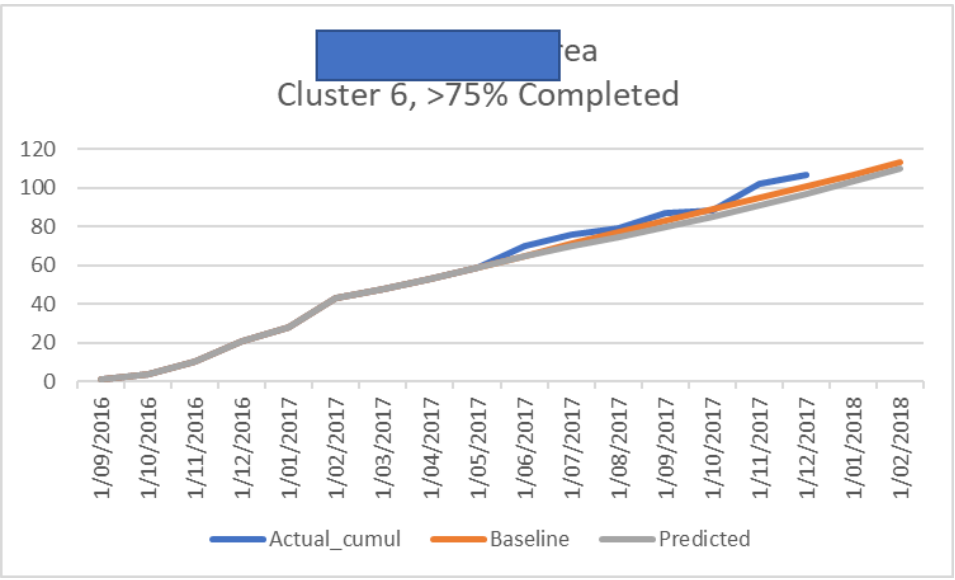
Modelling Result vs Actual vs Equal spread (currently)

The linear charts provide
Actual (by Dec 2017) vs Baseline vs Predicted data (train on actual up to Jul 2017) for different clusters.

To summarise, for studies with polynomial curve prediction model shows closer to real actual curve result







Next steps

- For both classification and regression task, the same random forest algorithm can be used

- Data exploration and clustering highlighted some that dome study duration can be challenged
- Timeseries for specific cases
- Tune every step algorithm
- Check if Machine learning algorithm performs better on that cases

Tools used

- IBM DSX environment with Python 2.0 notebook and Spark
- The code is platform independent and can be added as a standalone or as a part of IBM Cloud solution