POLY//ATICA

EMPOWERING

THE CITIZEN DATA SCIENTIST

13 SOHO SQUARE, LONDON, W1D 3QF
T +44 020 3468 1974 E INFO@POLYMATICA.COM

EMPOWERING THE CITIZEN DATA SCIENTIST IN YOUR ORGANISATION

Data is doubling every 18 months. Long gone are the days when a sophisticated business user could pull the whole dataset into their Excel spreadsheet for analysis. How have businesses handled this issue? Usually, by a combination of:

1. Looking at small subsets of data
2. Involving trained data scientists to analyse larger datasets
3. Accepting exponentially longer waiting times between a business question asked and an answer.

Polymatica is on a quest to create and empower the Citizen Data Scientist - a clever and business savvy user, who is able to ask the important business questions, get answers backed by data, and act without losing momentum.

Take an example of a marketing department needing to run a customer segmentation for a new targeted campaign. Traditionally, this would require getting your Data Scientist involved, who would take a subsample of your data to test out a few clustering algorithms, pick the best one for the subsample, and leave the winning algorithm to run overnight on all the data.

Now, meet the Citizen Data Scientist. She works in the marketing department and knows exactly what she needs to run her marketing campaign. Without writing any scripts, she can segment millions of customers based on their behaviour in minutes. The Data Scientist is left to do what he was hired to do - to uncover hidden trends in your business in new ways. The time from question to answer is decreased from a few days to a few minutes.

Empowering Citizen Data Scientists inside your organisation is easy with Polymatica. In addition to unrestricted data exploration and manipulation, they will get simple access to machine learning modules such as:

- Clustering

- Association rules
- Forecasting

They can explore and run analysis on the whole dataset - no size restrictions or slowdown as the business' data scales.

Data Science Modules

Empowering Citizen Data Scientists starts with giving them direct access to sophisticated analytical tools, without getting the Data Science team involved.

What normally takes weeks, can now be done in under a day. Below is an example of what a Citizen Data Scientist can accomplish between coming in to work and leaving in the evening:

1. Perform a quick explorative analysis of hundreds of millions of rows of granular data, get familiar with the shape and structure of the dataset in minutes. View and visualise the data by any combination of dimensions.
2. Come up with a new question – how can customers be segmented by their average spend, number of transactions and income?
3. Immediately find outliers with odd behavior – explore each outlier, remove them from the data if needed
4. Create a segmentation of customers in minutes – no need to worry about which algorithm to choose, or how many segments to go for
5. Look at typical profiling information for each segment – identify a customer segment that could be moved to a higher spending segment with the right marketing message.
6. For a chosen customer group, find what products are bought by the same customer by using association rules – prepare a targeted marketing campaign aimed to increase cross-purchasing

7. Forecast spend of each customer segment, and communicate with the wider business

We believe in letting Citizen Data Scientists do their own analysis and Polymatica modules were created to do just that. Bi-directional connection to the MultiSphere ensures all analysis is saved, so results obtained using one module can be used with another module.

Clustering

Polymatica's embedded clustering module allows you to quickly segment data based on any number of facts and parameters. Polymatica handles all the complexities - all the user needs to do is push a button.

Traditionally, a Data Scientist will face two major decisions when clustering data: picking a clustering algorithm and deciding on the number of clusters. Given the number of possible combinations and time to test each of those, this process is usually done on a small subset of data. Once the model and the number of clusters is selected, the clustering algorithm is run on the whole dataset. The danger in this approach comes from not seeing anomalies that are missing in the small dataset.

Let's take customer segmentation by spend, number of transactions and salary as an example. Using traditional methods, a Data Scientist runs tests to determine an ultimate algorithm and number of clusters on 100,000 transactions. He decides on a hierarchical clustering algorithm with 14 clusters. Just before leaving for the night, he sets the algorithm to run overnight on 1 billion transactions for 5 million customers. However, the Data Scientist is not aware that there are some outliers in the full dataset – people with a very high salary spending medium amounts on their card. They have been clustered with the medium earners, because we have pre-fixed the number of clusters to 14.

A Citizen Data Scientist, doing the same analysis would just immediately run clustering on the whole dataset, with Polymatica suggesting 16 clusters – including the smaller cluster with the high earners producing medium spend. With the right marketing, they could be moved to the high earners- high spenders cluster.

Polymatica allows clustering to be done with a click of a button. In the background, a combination of 3 algorithms optimised for large data volumes, is run to ensure optimal clusters with 100% replicability. The results are evaluated against each other, and the user is then given the ability to choose the number of clusters for the best model. Polymatica displays the Quality for all numbers of clusters between 2 and 100. Quality is dependent upon how close to cluster means all the points are, and how far away the clusters are away from each other. Polymatica also recommends the optimal number of clusters.

Dynamic clustering analysis is also possible, whereby objects are grouped by their common behaviours in time – such as spending money on Friday nights, or specifically around the holidays.

Once clustering is run, a new dimension is created for each data row, indicating what cluster the row belongs to. This can be treated just as any other dimension, including filtering, visualisation and any further analysis.

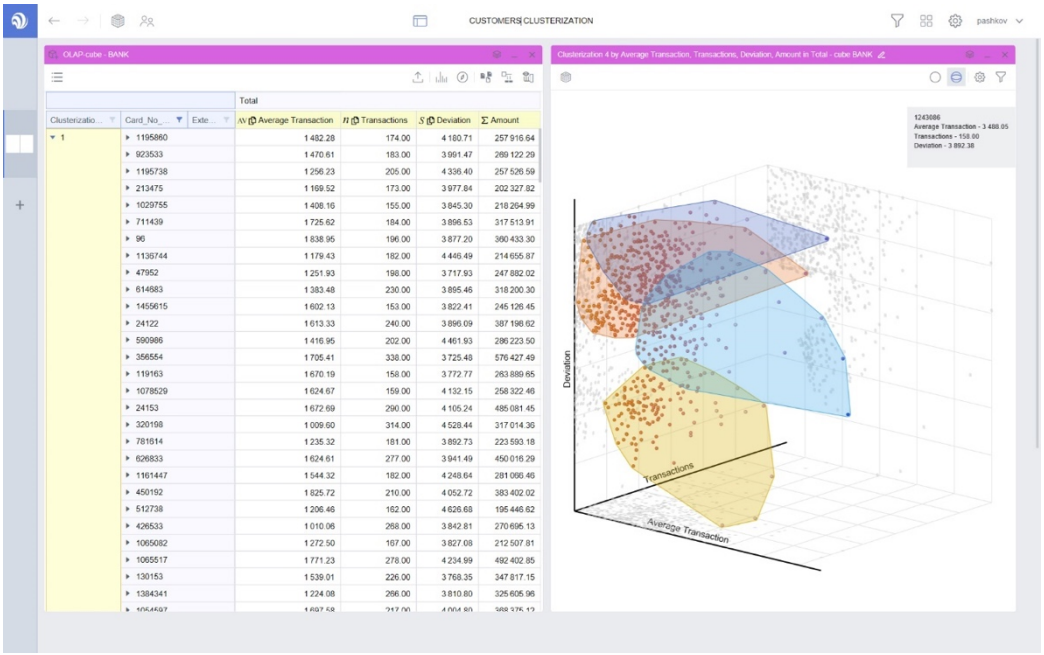

FIGURE 1: VISUALISATION OF CLUSTERING RUN ON 3 PARAMETERS

Association rules

Association rules allow the user to discover object features most frequently found together. For example, in retail, association rules analysis is used to find what products are bought by the same customer. The two important parameters that are obtained from association analysis are support and reliability. For an association rule X -> Y, support is the proportion

of all transaction that contain both X and Y, and reliability is the proportion of transactions that contain X, which also contain Y.

Customer 1	1	0
Customer 2	0	1
Customer 3	0	1
Customer 4	1	1

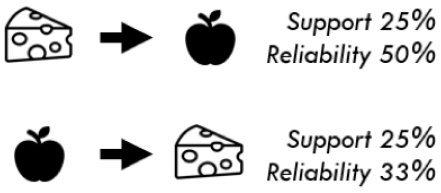

Figure 1: Support and reliability explained on an example of apples and cheese

Running association rules on a subset of data makes less sense than it does for clustering, as to find all the dependencies, the whole dataset is needed. Polymatica makes this easy, by running APRIORI algorithm optimised for large data sets at a click of a button. The user is able to choose the Support threshold they wish to look at – the minimal number of occurrences of events happening together. Support and reliability are presented for all the object combinations.

A Citizen Data Scientist is able to find common occurrences in 10 million transactions containing 10,000 purchase types in 2 minutes. By selecting a pair of products, the initial data can be filtered for further analysis.

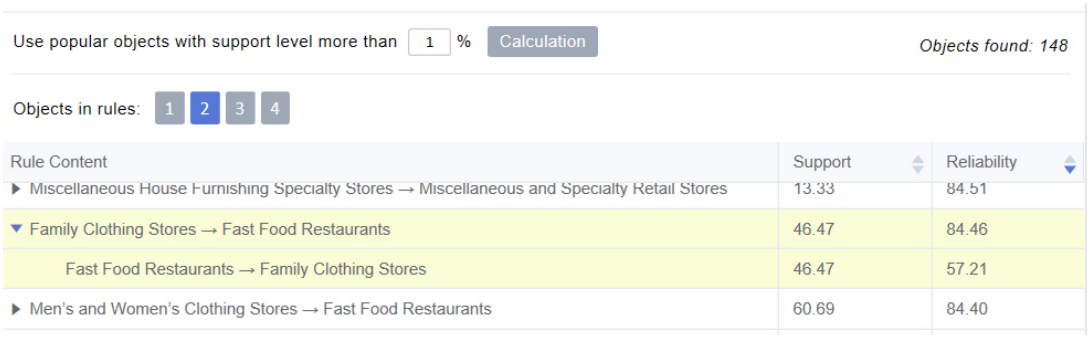

Figure 2 A snapshot of association rules module in Polymatica

Forecasting

Forecasting behaviour is another task usually left to Data Scientists, when selecting the best fit requires running tens of model variations, and smaller datasets are used for testing out the models first. Polymatica allows a Citizen Data Scientist to run forecasting algorithms on any facts and dimension elements.

For example, forecasting spend for the next month by customer will take 3 clicks of a button. Behind the scenes, Polymatica will evaluate 1,000 models including linear and polynomial regressions, ARIMA, ARIMA-T and Kalman filter. Each of the models will use 90% of the data points for training, and 10% for testing the model. The best fit will be presented to the user.

Running forecasting on large datasets is extremely important, as factors such as seasonality can be missed otherwise. Polymatica's ability to select an individual best fitting model for each dimension element also means that individual trends will be picked up.

While an overall trend can suggest that customers mainly shop during the weekend, some customers will have a different work schedule and will shop exclusively on Tuesdays. Polymatica can pick up this individual trend, and hence select the best fitting model to uncover hidden patterns.

Empowering the Citizen Data Scientist

By including built in data science modules, Polymatica gives Citizen Data Scientists full power to perform analysis traditionally reserved for individuals trained in Data Science.

About us

Polymatica is a data science company, pioneering new and accessible ways for global businesses to understand and use their data. The Polymatica platform uses embedded intelligence to deliver results and insights, at incredible scale and speed. Focused on business users throughout an organisation, Polymatica offers intuitive data analysis software, assisted decision-making solutions and visual storytelling tools, as well as consulting services and training designed to build the expertise of our clients.

For more information on Polymatica, please visit www.polymatica.com or call +44 (0) 20 3468 1974.