

Enhancing Business Decisions through Data Analytics and Use of GIS
A Business Application

Ms Uma V
CEO, Datafix Technologies Pvt. Ltd., Mumbai
DD: +91 22 2496 0026 | Cell: +91 98193 37150

And

Abhinandan Jain,
Indian Institute of Management, Ahmedabad
DD: 079 66324809

For Presentation at
1st IIMA International Conference on
Advanced Data Analysis, Business Analytics and Intelligence

June 6-7, 2009

Indian Institute of Management, Ahmedabad,
India - 380015

Enhancing Business Decisions through Data Analytics and Use of GIS

A Business Application

Key words: *data parsing, data enrichment, data linkages, spatial analysis*

Abstract

This paper presents the application of data analytics and spatial analysis (GIS) in one circle of a leading Indian telecom service organisation (*Bharti Airtel: BA*). BA was facing the problem of increasing bad debts and collection costs in one of the circles. BA turned to a data analytics organisation (Datafix Technologies Pvt Ltd¹) for resolving the issues by using a data based approach. The available data included filled up customer application forms and company's collection points. The customer data, like name, address, etc., was of poor quality. The methodology included (i) identifying relevant variables, (ii) splitting/ exploding the data fields and deriving new variables (iii) deriving linkages to identify unique customers and their relationships, (iv) using spatial analysis to study and link customers and collection centers. Paper uses, and shares the rationale of choosing the, specific tools in the Indian context.

The application helped BA in consolidating billing, reducing billing costs, identifying spatial pockets of higher defaults, identifying corporate clients for building relationships, and possibility of optimizing the location of collection centers. The paper shares efforts to generate emotional touch points from text data in Indian context.

Introduction and Issues

This paper reports on the use of data analytics and spatial analysis (GIS) for improving efficiency of collections² and reducing bad debts in a circle of a leading Indian telecom service organisation (*Bharti Airtel: BA*). The bills were sent to customers on a monthly basis. The dates of sending the bills and the due dates were staggered over the month

¹ Datafix Technologies Pvt Ltd is the current name for the entity formed from the earlier Spectrum Business Support Ltd (Spectrum). Datafix provides Data Quality Tools under the Ixsight family name. Ixsight finds mention in Gartners Magic Quadrant for Data Quality Tools 2008. The project mentioned above was done with Spectrum.

² The collection efficiency is described in terms of number of bills collected in specific time period and the cost incurred in collecting them.

across customers to smoothen the work load of billing and collections. The issues faced by BA were:

1. Non deliverability of bills to customer because of errors in the address (wrong postal code/ city name, incomplete or misspelt address, non availability of phone number, etc.).
2. Delayed payment by a significant proportion of customers although the bills were reaching the Customer in time. Some customers paid late because they didn't have a convenient collection centre or drop box located close by (say within 3 km.)
3. Frayed tempers at the Call Centre: The Call Centre kept calling customers for payments not realising what the problem was. Additionally a single customer may have more than one mobile subscription with BA and would be besieged by multiple calls. The call center could not identify him/her as a single customer.
4. Delinquent customers not identified early enough: In some cases, there was no payment because certain subscribers were habitual defaulters. They would take a subscription once, default and take another subscription under a similar sounding but different name.

BA had tried some actions like outsourcing the billing, employing collection agencies, sending bills on time, etc. These did not improve the situation significantly. BA then turned to a data analytics organisation (Datafix Technologies Pvt Ltd³) for help to resolve the issues by using a more data-based approach.

The key sub objectives of the project were jointly decided as: (i) Identifying geographical pockets of defaults, (ii) Reducing time taken for bill to reach customer, ensuring lower mail returns, and reducing billing costs, and (iii) Relocating/Increasing Collection Centers and Drop Boxes.

Data and Methodology:

³ Datafix Technologies Pvt Ltd is the current name for the entity formed from the earlier Spectrum Business Support Ltd (Spectrum). The project mentioned above was done with Spectrum.

Datafix executives first studied the data available from BA and arrived at critical data needed for resolving the issues. This was followed by developing a suitable methodology for addressing the issues.

The available *data* included the location of collection centers and drop boxes and individual subscriber wise data for the entire circle. The data on each subscriber in the circle was available from the source system of BA in the form of unstructured text file. It covered a variety of information pieces that were available in the filled up application form as well as those assigned to him/her by BA. The application form included key data on subscriber demographics and his location (address etc.). Those assigned by the company were Customer Type, Billing Cycle, Region Code etc. Each field was available separately.

Through joint discussion with BA Executives, the critical information pieces for resolving the issues were decided to be demographic and geographic information of subscribers. Therefore, the data fields that were chosen for analysis were as follows:

- Name
- Address, Out of the two addresses Billing Address was chosen
- Phone Numbers
- Date of Birth
- Region – This defined the region under which the subscriber was classified by the telecom provider
- Cycle – This defines the cycle of payment, sometimes for the same customer with multiple subscriptions each subscription would fall under a different billing cycle. This would further reflect on the customers bill payment behaviour
- Product Type – This defined the kind of product availed by subscriber which in turn reflected on their income or usage levels and therefore revenue potential may be.

The *methodology* employed by Datafix included three key steps:

- (i) **Data Preparation:** Through a process of data cleansing, parsing and enrichment, generate meaningful variables from the data available. This was achieved through application of a special data cleaning and parsing tool *Scrubbix*.
- (ii) **Identifying Data Linkages and Subscriber Relationship Size at Individual and Family Level:** The variables generated through (i) above to a unique set using a process of de-duplication to ensure single counting of a multiple subscriber. A specialized Deduplication tool *Deduplix* was used for this purpose.
- (iii) **Spatial Analysis:** Was done through a special tool *Spinfo's GIS* tool.

Data Preparation:

Data was imported from Text files of the source system of BA into a standard Database. Data integrity was checked by ensuring that no duplication was present, that the information was complete, and by validating content against metadata provided by BA.

Analysts explored the rules needed to be deployed for cleaning the data by addressing issues like (i) data points available on each variable (about 5 lakh points), (ii) extent of consistency in the information, (iii) quality of information⁴, (iv) duplicate/ multiple occurrences of same entities' names etc.

Exploding or Parsing Data:

Fields containing unstructured text were broken up into smaller elements on the basis of semantic rules and dictionaries such that each element was meaningful for analysis. Such

⁴ A typical definition of Data Quality (Refer item (1) of Bibliography)

- Data Accuracy – the degree to which Data reflects the real world object
- Data Completeness – the extent to which Data attributes are provided.
- Data Consistency – the absence of contradictions in the data
- Data Standardness – Semantic variations leading to inaccurate analysis
- Data Timeliness/currency – How recent is the data
- Data Auditability – Ability to track information to its original transaction

information extraction from data was a challenging task⁵. The static information like Name, Address through a process of Data Cleansing (segmenting or parsing, classifying and standardising) and Enrichment were converted into meaningful variables like Title, Gender, Religion, Language, Locality, City, Other Telephone Owner, Type of Corporate etc. For example: **address** was broken up into Street, Locality, City, etc. and **name** was split into name of corporate or individual.

Perfect Accuracy in extracting information even from postal addresses is not possible because they are replete with typographic, format and order errors. Indian Address Data in particular is free flowing as there is no national convention for Address laid down by the Postal Department. However using a parsing tool it is possible to extract relevant information for analysis. A good parser can extract the information from address to yield the following geographic information (i) Sub locality (surrogate for geographical coordinates), (ii) Locality (surrogate for geographical coordinates), (iii) Post Code (surrogate for geographical coordinates), (iv) City (geographic location), (vi) State, (vii) District (Rural or Urban), (viii) Taluka (Rural or Urban), (ix) Type of House (Chawl, Building, Bungalow, Kothi, Mohalla etc), and (x) Landmark. The locality, Urban/Rural and Type of House can further describe the *residential profile* of the person.

Similarly Name can be exploded further into: (i) First Name, (ii) Last Name, (iii) Title, (iv) Gender.

From the above, the following can be extrapolated if not already available with in the name: (i) Gender, (ii) Title, (iii) Religion, (iv) Language, (v) Individual/ Corporate, (vi) Corporate Type.

⁵ Information extraction populates a database from (Refer item (2) of bibliography) unstructured or loosely structured text; data mining then discovers patterns in that database. Information extraction involves five major subtasks

- ♦ Segmentation – finds the starting and end boundaries of the text field that will fill a database field
- ♦ Classification – determines which database field for each text segment
- ♦ Association – determine which fields belong together in the same record
- ♦ Normalization – puts information in a standard format so that it can reliably be compared
- ♦ Deduplication – collapses redundant information so that you don't get duplicate information in the database

Data Enrichment:

The extracted data was enriched through deriving new attributes from existing attributes based on correlation with original information or third party data. Once the Information has been extracted from data, new pieces of information can be derived using (i) existing information based on derivative logic or (ii) third party information. For example: (i) using the deduplicated Information, the total revenue per subscriber and per subscriber family can be estimated and (ii) using the Name Information – Religion, Language, Type of Customer (Corporate or Individual) can be identified. Similarly, using Address Information – Residential Type, urban/Rural, geocode can be populated

Scrubbox, Datafix Technologies’ cleansing tool to cleanse, parse and enrich the data uses Parsers and Rule based heuristics to clean and parse the data. Based on the criteria⁶ for selecting a tool for such purposes, DQAS, Scrubbox and Deduplix – the Audit Solution, Cleansing Solution and Matching Solutions of Datafix Technologies Pvt Ltd were chosen

Identifying Data Linkages and Subscriber Relationship Size:

For the purpose of linking data, the first task was to create a set of records that identify unique parties /customers. For example Party A may have 10 relationships with the organisation but should be considered only once for the purpose of certain types of

⁶ Criteria for selection of Tools:

These criteria have evolved based on actual experience of Datafix and are vital for getting maximum efficiency. Needless to say architecture, performance and ability to work with a variety of databases are important as would be with any generic software solution.

- Tools should be able to address the various issues pertaining to Data Quality Assessment, Data Cleansing and Data Deduplication (the whole cycle of Data Assessment and Information Extraction)
- Tool should be able to work with 80 % plus efficiency on Indian Data
- Tool should be able to achieve enrichment using third party Data
- Tool Supplier should also provide services which are dependent on linguistic expertise and Indian Name & address semantics
- Tools should be supported by updated dictionaries
- Tool Supplier should be able to supplement the tools with manual expertise in terms of rule additions for Information Extraction and dictionary enhancement
- Tool should have been proven in a similar environment on large scale Indian Data

analyses. Subsequently, data linkages were established between unique customers. For example, several customers may be members of the same household. Such linkage could also provide a useful measure of loyalty and relationship value. In the example below, Ravish Ahuja has three mobile subscriptions but should be considered only once for purpose of profiling. The Unique individual ID generates a Match ID number (Unique Individual) based on Name, Address Combinations. These can also include Name, Address, and Telephone combinations and other variables that may be available. Similarly Members of the same family as indicated by Last Name and Address were given a Family ID.

Such Linkages also provide a useful measure of loyalty. For example: Loyalty for an individual could be described as a weighted index of (i) years of relationship (current date less start date), (ii) number of family members using services (got through deduplication), (iii) default level, (iv) usage level, and (v) recency of latest subscription. Once a loyalty status has been arrived at, it can be further profiled on the basis of age, locality, religion, marital status or any other profile variables.

Subscriber Number	Name	Address	Same Person (Unique Individual)	Same Family (Members of Same Family)
9868668912	Ravish Ahuja	33/4 Galaxy C G Road Ahd	1	1
9452446921	Ahuja Raweesh	Flat 33 Bldg 4 Ahmedabad CGRd	1	1
9564717412	Ravish K Ahoj	CGRd Galaxy 33-4 Next to Vijay Supermarket Ahd -2	1	1
9451367932	Siddharth Ahuja	Galaxy 33 CG Road Ahmedabad	2	1
9456678312	Leena Ahuja	Apt 4 CG Marg Blk 3 Near Amul Shop Ahmedabad	3	1

A Deduplication tool, **Deduplix** (Datafix Technologies' Deduplication tool), was used to find record linkages and identify unique records (customers) through a process of fuzzy matching. Deduplication enables identification of data that is inherently same but semantically different and possibly having formatting and content inconsistencies.

For example, let us consider a Name like Meenakshi Iyengar which may have three types of variations.

1. **Variation 1:** The Name could be spelt in the following various ways: (i) Minaxi Iyengar, (ii) Meenaxi Aiyengar, (iii) Aiangan Minaksi, (iv) Meenakshee Iengar, (v) Meenakshi Iye ngar, (vi) Iyanger Minakshi. The above describes a *phonetic* variation.
2. **Variation 2:** This could be further compounded with variations through the introduction of a middle name. For example, (i) Minakshi R Aiyengar, (ii) Ramchandra Minakshi Iyengar, (iii) Minakshi Ram Chandra Aiyengar.
3. **Variation 3:** A further noise in the data could be present as follows: (i) Mrs Minakshi Aiyengar, (ii) Shrimati M Aiyengar, (iii) Minakshi Ramchandra Aiyengar (Ref: 4567)

The above are variations of a single variable: "Name". Similarly Address, Telephone, Email, PAN numbers are all subject to inconsistencies of format, semantic, typographic, content, etc.

A good Deduplication⁷ program will enable identification of records as similar based on a weighted combination of several fields with thresholds defined from each field. The

⁷ Characteristics of a good dedupe tool are:

- Ability to match in spite of semantic inconsistencies
- Low False Positives and False Negatives
 - False Positives – Identifying two records which are intrinsically different as same
 - False Negatives – identifying two records which are same as different
- Ability to match using several rules
- Ability to provide weights and thresholds
- Support for various types of phonetics as used in Asia, Middle East or Europe
- Ability to Match on a large scale and high speed
- Ability to Provide various kinds of groupings say family ID, Individual ID, Corporate ID etc

Dedupe program should have good phonetic logic that can deal with misspellings and variations.

Deduplix from Datafix Technologies was chosen because it had proven capabilities to provide all the above apart from extremely good matching capability on Names and Addresses. Deduplix was also proven in the Indian context where dirty data poses a real challenge

Spatial Analysis of Data:

Exploded data - sub-locality, locality and street – was used to geocode onto digital city maps. The process was iterated several times over till at least 80 % of the points were plotted. *Spinfo's GIS tool* was used for this purpose.

Spinfos GIS tool is a homegrown mapping tool where the solution encompasses the data, map as well as mapping software. Spinfo geocodes various address points (based on the parsed and cleansed addresses). Most importantly full support was available for geocoding atleast 75 % of the data. This is particularly relevant since India does not have a latitude/ longitude attached to street addresses. Spinfo has detailed city maps of Bangalore, Chennai, Hyderabad, etc along with a robust Geocoder application that takes in to account the Indian addresses. As the addressing logic changes from city to city the Spinfo Geocoder is tuned to city specificity. Spinfo Mapping tools provide the backend data at a street level which is continuously updated. The mapping tool also provides a convenient way to update new information on to the map and also gives various thematic views. More than one layer can be simultaneously viewed.

Results and Learnings:

Benefits for BA: The project was able to help in

- (i) Consolidation of billing Cycles on the basis of record linkages: BA had various billing cycles allocated for various subscribers. A subscriber having more than one connection (multiple subscriptions) could be across different billing cycles. Thus a subscriber would get bills at varying times for various connections and may not be able to manage his/her payments on time. The

process of cleansing and deduplication helped identify all the connections of one subscriber and link them to one billing cycle.

- (ii) Consolidation of bills on the basis of Address: Again as described in (i), the process of deduplication helped consolidate all the bills of a family so that they could be bunched together in a common envelope (though individually in separate envelopes) and enable savings on delivery costs.
- (iii) Optimization of location of Drop Box and Collection Centers: The process of cleansing enabled geocoding whereas the process of Deduplication reduced the redundancy and enabled customer views rather than subscriber views. Superimposing the drop box and collection centre data on customer data yielded useful insights as to which customers were well facilitated by collection points and which were not.
- (iv) Identification of pockets of default at sub-locality level: As BA had knowledge as to which customers were defaulters and which were not, superimposing defaulter information on customer data provided a spatial analysis of default patterns. These were then further analysed by the BA team for identifying other variables which could have possibly lead to the default.
- (v) Identification of variables like religion and language (the “Emotional Touch Points”): Based on Religion and Language – customers were sent emails and communication during important festivals.
- (vi) Identification of corporate customers: The process of cleansing clearly identified corporate and individual customers. Corporate Customers had their names abbreviated or spelt badly. Through the process of segmenting, classifying and standardising – top corporate customers were easily identified.
- (vii) Use of digital Maps to guide customers to nearest drop box or collection centre at call centre: The project threw up a further interesting possibility. If the map was made available at the call centre, the call centre executive could guide irate customers to the nearest call centre or drop box by looking at the map which gave both customer location as well as the nearest collection points.
- (viii) Enhancement of customer address to ensure lower mail returns: The enhancement of the customer address further meant more accurate and

consistent information and therefore lower mail returns. Cities, pincodes were corrected. Landmark information was separated and highlighted.

- (ix) Consolidating 80:20 analysis on the basis of revenue and relationships: The process of consolidation and mapping billing revenues to customers rather than subscribers yielded the top 20 % customers who contributed to 80 % of the revenue.

Learnings:

While tabular analysis could yield some of the above information and results, spatial analysis was superior because

- customer concentrations could be aggregated for geographically contiguous areas though they appeared far apart on a tabular view
- Several different pieces of information could be overlaid so that relationships between the variables could be spatially viewed.

Examples: (i) Correlation between proximity of collection points to localities of customer concentration, (ii) Correlation of customer concentrations to service centres of showrooms, (iii) Defaulter locations to Age profile, etc.

Unfinished Tasks:

The issues with respect to collections were at the initial stage where the solutions could be addressed by using the above analyses. With the analysis already done, it was possible to find the gaps that existed in the data. After filling the data gaps, analysis of efficiency of Billing Collection could be on the basis of identifying further analysis: (i) correlation between variables and default rate and (ii) location of optimal drop box points on the basis of solving linear programming algorithms.

- These tasks can be accomplished through a Quarterly Collection Efficiency Meter on lines of Brand Monitor
- Customer Satisfaction Index linked to Customer Collection efficiency can also be studied
- The sample should include Corporates, Individuals, those who pay late, don't pay, pay after several calls and those who pay on time could be the sample cells

- The sample for those who are bad payers should also include customers who are affected positively by: (i) Complete Address, (ii) Drop Box Convenience, (iii) Multiple Bills in one envelope, etc.
- It is also important to note whether those already paying on time deteriorate or not – and these reasons may form the part of next study
- An advantage of such a study is that one need not visit the customer, such information is readily available online month on month.

Suggestions for Use of Data Analytics and Spatial Analysis:

The nature of customer data available in India can facilitate creation of variables impinging on emotional touch points for creating better relationship building strategies and plans. The application demonstrates the need for (i) identifying appropriate variables for managerial decision making, (ii) collecting complete data from customers even in a situation of fast growing customer base, (iii) identification/ use of suitable tools for data parsing, enrichment, establishing data linkages, and plotting of data on digital maps, and (iv) use of higher order statistical techniques for decision making.

XX

References:

1) Thomas C Redman, 1992 Data quality, Management & Technology. New York: Bantam Books

2) Andrew McCallum, Distilling Structured Data from Unstructured Text, University of Massachusetts, Amherst, November 2005