iCARE: A framework for Big Data based Customer Analytics

Authors: N. Sun, J. G. Morris, J. Xu, X. Zhu, M. Xie
DATA: AN EVER-INCREASING QUANTITY

- Systems, sensors and mobile devices transmit data
- Volume, Velocity and Variety

A new style of IT emerging

Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11 million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users
RISE OF UNSTRUCTURED DATA
BIG DATA AND BANKS - SOURCES

- Nature of data has become complex
- Structured and unstructured
- Internal and external sources
- Challenges in handling and generating valuable information for banks
BIG DATA POSSIBILITIES

Big Data Opportunities in the Financial Services Sector

- Increase credit score accuracy based on new sources of data (ex: social media data) for improved risk control and more competitive pricing
- Improve Risks and Pricing Management
- Enhance Customer Experience
- Personalize offerings, identify new cross-sell opportunities and control churn based on deep insights on customer behavior, drawn from multiple sources of data
- Increase Operational Efficiency
- Identify New Business Models
- Optimize data management costs using Big Data technologies, improve operational efficiency through straight-through processing and real-time, performance management
- Monetize raw anonymized customer data or behavioral insights that enable improved market analysis
TYPES OF DATA

▪ Hard Data
  – Numbers easy to store and transmit
  – Bank transaction records, phones, etc.

▪ Soft Data
  – Mostly exists in the form of text
  – Social media, feedbacks, etc.
  – Have been a focus for developing important insights
TRADITIONAL CUSTOMER ANALYTICS

- Mainly focused on hard information
- Focused on 4 dimensions with standard data mining techniques:
  - Customer Identification: Classification, Clustering
  - Customer attraction: Regression
  - Customer retention: Classification, clustering
  - Customer Development: Association
- “More than 70% of banking executives worldwide say customer centricity is important to them”
- “Research indicates that only 37% of customers believe that banks understand their needs and preferences adequately “
BIG DATA CHALLENGES

- Handle massive complex data in a cost effective manner
  - Increase in both hard and soft information
  - Analyzing large data using traditional analytics tools takes a long time

- Effectively generate business value and obtain competitive advantages
  - Understanding of the problem should be combined with problem-solving techniques to improve decision making and bring real business value to a bank
  - Trillions of dollars and Euros can be generated in value worldwide on the basis of big data
iCARE ARCHITECTURE

Data Mining
- Cross-selling
- Customer Retention
- Personalized Recommendation
- New Customer Acquisition

Data Storage
- Structured Data
- Internal & External Unstructured Data

Data Preprocessing
- IBM InfoSphere BigInsights

iCARE Analytical Models
- IBM SPSS Modeler & Analytic Server
PHASES IN iCARE

- Data Acquisition
  - Structured data: Internal and external sources
  - Data converted to a standard input format before putting it in IBM BigInsights
  - Unstructured data: web log files, call records, social media

- Data Preparation
  - Unstructured data: work to regularize or schematize data before modelling
  - Structured Data: handle incorrect, incomplete and irrelevant data
  - Once all the data has been prepared and cleansed, a data integration process is conducted

- Data Modelling

- Business Applications
iCARE ANALYTICAL MODELS

▪ Statistic and machine learning algorithms customized to suit business scenarios
  – Interactive decision tree for customer retention
  – Business strategies automatically generated

▪ Parallelized models for better computing models
  – Algorithms designed to follow MapReduce model
  – Save time and lower costs
CUSTOMIZED PARALLELIZED K MEANS CLUSTERING

- Select K data points as the cluster centers
- Assign each data point to the closest cluster using the standard K-means algorithm and Manhattan distance

\[ f(x, y) = \sum_{i=1}^{D} (|x^i - y^i|) \]

- Update the cluster centers using:

\[ C_{k, new} = \sum_{j=1}^{J_k} w_{k,j} x_{k,j} \]

where \( (w_{k,j}) = \frac{1}{d(x_k, C_{k,old})} / \sum_{j=1}^{h} \left( \frac{1}{d(x_k, C_{k,old})} \right) \)
CONTINUED...

- Redistribute the data points to their closest cluster and drop any data point \( x_j \), which is far away from any cluster center
- Repeat until convergence
BUSINESS APPLICATIONS

- Customer segmentation and preference analysis
- Potential customer identification
- Customer Network Analysis
- Market Potential Analysis
- Channel allocation and operation optimization
CASE STUDY: A BANK IN SOUTHEAST CHINA

- 20TB data analyzed to generate insights for retaining active online banking customers
- Large amount of structured data with ambiguous definitions in their various systems
- Online/mobile banking log files were also provided for analysis
- More than two hundred of attributes were generated from different sources including personal information, account information and transactions
FRAMEWORK IN ACTION

- Data from multiple database sources were integrated
- Customer ID: Online Banking and ECIF system
- Card Number: Online Banking and ECIF system
PERFORMANCE EVALUATION

Performance of customized decision tree in a real case of iCARE
Comparison of computing time: big data platform versus a single host
CONCLUSION

▪ Beyond limitation of traditional banking systems: use of unstructured data

▪ Use of parallelized version of k means and other algorithms gives the flexibility to scale

▪ Results can be interpreted as business rules which help in decision making
REFERENCES


Discussion...
STRENGTHS AND WEAKNESSES

▪ Strengths
  - Well organized with sufficient background details
  - They explain the problem with traditional techniques and how big data can help in giving deep insights about customers

▪ Weaknesses
  - Not much technical depth. Just an overview
  - Focused on distributed model but showed results favoring single host
  - Use of social media not explained in their model
  - Performance for large number of records is questionable
FUTURE WORK

- Can be extended to applications other than customer relationship management.
- Since this model is scalable, therefore it lays ground work for even bigger projects using more data.
 RELATED READINGS


DISCUSSION QUESTIONS

▪ Use of Manhattan Distance instead of commonly used Euclidean Distance?
▪ Why do you think the big data platform performed badly as compared to single host?
▪ What soft features could be considered for better customer understanding?
▪ How can social media help in getting deeper insights about the customer in banks keeping in mind the privacy of the customer?