Chapter 7:
Data Mining
Overview

Topics discussed:

- The Need for Data Mining and Business Value
- The Data Mining Process:
  - Define Business Objectives
  - Get Raw Data
  - Identify Relevant Predictive Variables
  - Gain Customer Insight
  - Act
The Need for Data Mining

- Companies developed a need to understand their customers better
  - They must be proactive and anticipate customer desires

- Many companies have full access to detailed customer data
  - Data mining provides businesses with the ability to make knowledge-driven strategic business decisions

- The increasing data deluge instead of lack of data generates a problem for many companies which leads into an obstacle (to use data)
- Companies need to implement a standardized data mining procedure in order to extract customer intelligence and value from that data
The Business Value of Data Mining

- Data mining can assist in selecting the right target customers or in identifying customer segments with similar behavior and needs.

- Applications of data mining include the following:
  - Identifying customers that are likely to stop business with the company with the help of predictive AU1 models.
  - Increasing customer profitability by identifying customers with a high growth potential.
  - Reducing marketing costs by more selective targeting.

Source: Kumar and Reinartz (2012)
The Data Mining Process

(Re)Define Business Objectives
- Define objectives and expectations
- Define measurement of success

Get Raw Data
- Extract descriptive and transactional data
- Check quality (technical and business)

Identify Relevant Variables
- Rollup Data
- Create analytical data
- Enhance analytical data
- Select relevant variables

Gain Customer Insight
- Train predictive models
- Compare models
- Select models

Act
- Deploy models
- Monitor performance
- Enhance models
Today: Most time is spent on data extraction, transformation and data quality

60 – 70 % of process time

(Re)Define Business Objectives
Get Raw Data
Identify Relevant Variables
Gain Customer Insight
Act

< 30 % of process time

Tomorrow: Most time spent on business objectives and customer
Extent of Involvement of The Three Main Groups Participating in a Data Mining Project

**Groups**

1. Business
2. Data Mining
3. IT

(Re)Define Business Objectives → Get Raw Data → Identify Relevant Variables → Gain Customer Insight → Act
Extent of Involvement of The Three Main Groups Participating in a Data Mining Project

- Data mining group
  - Understand business objectives and support the business group to refine and sometimes correct the scope and expectations
  - Most active during the variable selection and modeling phase
  - Share obtained customer insights with the business group

- IT resources
  - Source and extract required data used for modeling

- Business group
  - **Check plausibility and soundness** of the solution in business terms
  - Take lead in deploying new insights into corporate action (e.g., a call center or direct mail campaign)
Data Manipulation

- In a simple, two-dimensional data table **columns** represent descriptive variables, whereas **rows** represent single observations.

- Manipulation on **columns**:
  - Transformation
  - Derivation
  - Elimination

- Manipulation on **rows**:
  - Aggregation
  - Change detection
  - Missing value detection
  - Outlier detection
Data Preparation

- For modeling, incoming data is sampled and split into various streams as:
  - Train set: Used to build models
  - Test set: Used for out-of-sample tests of the model quality and to select the final model candidate
  - Scoring data: Used for model-based prediction

The data sets must be carefully examined and designed to assure statistical significance of the results obtained
Define Business Objectives

- Modeling of expected customer potential in order to target the acquisition of customers who will be profitable over the lifetime of the business relationship
- Mathematically define the target variable of customer behavior that has to be predicted
- Distinguish between customers by setting a target variable for specific customer groups, e.g., 1 for one group and 0 for another
- Establish likely threshold levels indicating which customer should be targeted in the marketing campaign
Define Business Objectives (2)

- Define a set of business or selection rules for the campaign (e.g., identifying customers that should be excluded / included in the target groups)
- Define the details of project execution specifying the start and delivery dates of the data mining process, and the responsible resources for each task
- Define the chosen experimental setup for the campaign
- Define a cost/revenue matrix describing how the business mechanics will work in the supported campaign and how it will impact the data mining process
- Establish criteria for evaluating the success of the campaign
- Find a benchmark to compare results obtained in the past for similar campaign setups using traditional targeting methods instead of predictive models
Cost/Revenue Matrix

- A Cost / Revenue matrix describes how the business mechanics will work in the supported campaign and give business users an immediately interpretable table

  Example: Call Center Campaign

- Assuming average cost per call is $5, each positive responder (purchaser) will generate additional cost due to:
  - Administration work required to register him as a new customer
  - Cost of the delivered phone handset ($100)
  - Customers who respond positively will generate average revenue of $1,000 per year
# Cost/Revenue Matrix – Call Center Campaign

<table>
<thead>
<tr>
<th>Cost/Revenue Matrix</th>
<th>In reality prospect did not purchase</th>
<th>In reality prospect did purchase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model predicts prospect will not purchase (not contacted)</td>
<td>Cost: $0  1\text{st} \text{ year revenue}: $0  Total: $0</td>
<td>Lost business opportunity of +$895</td>
</tr>
<tr>
<td>Model predicts prospect will purchase (contacted)</td>
<td>Cost: -$5  1\text{st} \text{ year revenue}: $0  Total: -$5</td>
<td>Cost: -$5; -$100  1\text{st} \text{ year revenue}: +$1000  Total: +$895</td>
</tr>
</tbody>
</table>
Get Raw Data

- Business objectives need to be translated into data requirements
- Identified data has to be extracted and consolidated in a database
- The quality of the analytical data has to be checked (business context)

- Three steps of getting raw data:
  - Step 1: Looking for Data Sources
  - Step 2: Loading the Data
  - Step 3: Checking Data Quality
Step 1: Looking for Data Sources

- Data sourcing
  - Mixed **top-down** and **bottom-up** process driven by business requirements (top) and technical restrictions (bottom)

- Data warehouse infrastructures with advanced data cleansing processes can help ensure working with high-quality data

- All metadata available has to be collected to fully understand data types, value ranges and the primary/foreign key structures

- Build a (simple) relational data model onto which the source data will be mapped
Step 2: Loading the Data

- Defining further query restrictions in order to model subsets of the full data

- After data management is requested to deliver the specified data, IT teams prepare the necessary data queries

- Deliver extracted data to the data mining environment in a pre-defined format

- Further processing and using data to fill previously defined database environment as part of the ETL process (Extract-Transform-Load)
Step 3: Checking Data Quality

- Importance of data quality
  - According to Olson (2003) the costs of poor data quality are estimated at 15-25% of operating profit

- Assess and understand limitations of data resulting from its inherent quality (good or bad) aspects

- Create an analytical database as the basis for subsequent analyses

- Relevant aspects of data quality
  - Accuracy (consistency and validity)
  - Relevance
  - Completeness
  - Reliability
Step 3: Checking Data Quality (2)

- Carry out preliminary data quality assessment
  - To assure an acceptable level of quality of the delivered data
  - To ensure that the data mining team has a clear understanding of how to interpret the data in business terms

- Data miners have to carry out some basic data interpretation and aggregation exercises

- The data available for the mining project must be analyzed to answer the following questions:
  - Does the data correspond to the original sourcing requirements?
  - Is the quality sufficient?
  - Do we understand the data?
Identify Relevant Predictive Variables

- Step 1: Create Analytical Customer View – Flattening the Data
- Step 2: Create Analytical Variables
- Step 3: Select Predictive Variables
Step 1: Create Analytical Customer View – Flattening the Data

- Individual customer constitutes an observational unit for data analysis and predictive modeling
- All data pertaining to an individual customer is contained in one observation (row, record)
- Individual columns (variables, fields) represent the conditions at specific points in time or a summary over a whole period – this requires denormalizing the original relational data structures (flattening)
- Definition of the target or dependent variable- values should be generated for all customers and added to the existing data tables
Step 2: Create Analytical Variables

- Introduce additional variables derived from the original ones
- When needed, transform variables to get new and more predictive variables
  - Example: Transform customer birth date into age
- Increase normality of variable distributions to help the predictive model training process
- Missing value management is key for enhancing the quality of the analytical data set
Step 3: Select Predictive Variables

- Inspect the descriptive statistics of all univariate distributions associated to all available variables

- Variables that can be excluded
  - Taking on only one value (i.e. the variable is a constant)
  - With mostly missing values
  - Directly or indirectly identifying an individual customer
  - Showing collinearities
  - Showing very little correlation with the target variable
  - Containing personal identifiers

- Check if all variables have been mapped to the appropriate data types
Gain Customer Insight

- Step 1: Preparing Data Sample
- Step 2: Predictive Modeling
- Step 3: Select Model
Step 1: Preparing Data Samples

- Analyze if sufficient data is available to obtain statistically significant results

- If enough data is available, split the data into two samples
  - The train set to fit the models
  - The test set to check the model’s performance on observations that have not been used to build it
Step 2: Predictive Modeling

Two steps:

- The rules (or linear / non-linear analytical models) are built based on a training set
- These rules are then applied to a new dataset for generating the answers needed for the campaign

Guidelines:

- Distinguish between different types of predictive models obtained through different modeling paradigms: supervised and unsupervised modeling
- Find the right relationships between variables describing the customers to predict their respective group membership likelihood: purchaser or non-purchaser, referred to as scoring (e.g., between 0 and 1)
- Apply unsupervised modeling where group membership is not known beforehand
Step 3: Select Model

- Compare relative quality of prediction by comparing respective misclassification rates obtained on the test set.
- Economic implications of a model by applying the previously defined cost / revenue matrix will be included.
- Predictive models, for instance, deliver a score value, or likelihood, for each customer to show the modeled target behavior (e.g., purchase of a credit card).
Act

- Step 1: Deliver Results to Operational Systems
- Step 2: Archive Results
- Step 3: Learn
Step 1: Deliver Results to Operational Systems

- Apply the selected model to the entire customer base
- Prepare score data set containing the most recent information for each customer with the variables required by the model
- The obtained score value for each customer and the defined threshold value will determine whether the corresponding customer qualifies to participate in the campaign
- When delivering results to the operational systems, provide necessary customer identifiers to unambiguously link the model’s score information to the correct customer
Step 2: Archive Results

- Each data mining project will produce a huge amount of information including:
  - Raw data used
  - Transformations for each variable
  - Formulas for creating derived variables
  - Train, test and score data sets
  - Target variable calculation
  - Models and their parameterizations
  - Score threshold levels
  - Final customer target selections

- Useful to preserve especially if the same model is used to score different data sets obtained at different times
Step 3: Learn

- Referred to as “closing the loop”
- Obtain the facts describing performance of data mining project and business impact
- These facts are attained by monitoring campaign performance while it is running and from final campaign performance analysis after the campaign has ended
- Detect when a model has to be re-trained
Summary

- Data Mining can assist in selecting the right target customers or in identifying previously unknown customers with similar behavior and needs.
- A good target list is likely to increase purchase rates and has a positive impact on revenue.
- In the context of CRM, the individual customer is often the central object analyzed by means of data mining methods.
- A complete data mining process comprises assessing and specifying the business objectives, data sourcing, transformation and creation of analytical variables and building analytical models using techniques such as logistic regression and neural networks, scoring customers and obtaining feedback from the field.
- Learning and refining the data mining process is the key to success.