on the test will be asked SA
how is lift chart different from Confusion matrix and explain the use cases

Outline

- Explain and discuss predictive models and their business values to organizations.
- Identify data patterns using predictive models in (SSAS).

Illustrative Applications

- Customer Relationship Management
- Finance
- E-commerce and Internet

Customer Relationship Management

- Target Marketing
- Attrition Prediction/Churn Analysis
- Fraud Detection
- Credit Scoring

Target marketing

exam q

- Business problem: Use list of prospects for direct mailing campaign
- Solution: Use Data Mining to identify most promising respondents combining demographic and geographic data with data on past purchase behavior
- Benefit: Better response rate, savings in campaign cost

Example: Fleet Financial Group

IBM recently bought Red Hat for their customer base

- Redesign of customer service infrastructure, including $38 million investment in data warehouse and marketing automation
- Used logistic regression to predict response probabilities to home-equity product for sample of 20,000 customer profiles from 15 million customer base
- Used CART (Classification and Regression Trees) to predict profitable customers and customers who would be unprofitable even if they respond

Churn Analysis: Telcos

exam q

- Business Problem: Prevent loss of customers, avoid adding churn-prone customers
- Solution: Use neural nets, time series analysis to identify typical patterns of telephone usage of likely-to-defect and likely-to-churn customers
- Benefit: Retention of customers, more effective promotions

can use this information to call/contact customers at certain times to offer them benefits/packages to get them to stay

younger ppl tend to jump b/t companies -> likely to defect

Example: France Telecom

- CHURN/Customer Profiling System implemented as part of major custom data warehouse solution
- Preventive CPS based on customer characteristics and known cases of churning and non-churning customers identify significant characteristics for churn
- Early detection CPS based on usage pattern matching with known cases of churn customers.

neural networks is example of unsupervised learning, no labels

Fraud Detection

- Business problem: Fraud increases costs or reduces revenue
- Solution: Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously
- Benefit: Increased profits by reducing undesirable customers

Example: Automobile Insurance

- **Bureau of Massachusetts**
- Past reports on claims adjustors scrutinized by experts to identify cases of fraud
- Several characteristics (over 60) of claimant, type of accident, type of injury/treatment coded into database
- Dimension Reduction methods used to obtain weighted variables. Multiple Regression Step-wise Subset selection methods used to identify characteristics strong correlated with fraud

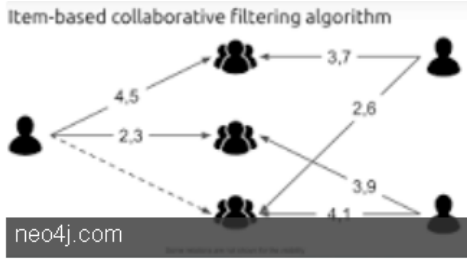
Risk Analysis

- Business problem: Reduce risk of loans to delinquent customers
- Solution: Use credit scoring models using discriminant analysis to create score functions that separate out risky customers
- Benefit: Decrease in cost of bad debts

E-commerce and Internet

In the newer, narrower sense, **collaborative filtering** is a method of making automatic predictions (**filtering**) about the interests of a user by collecting preferences or taste information from many users (collaborating).

Collaborative filtering - Wikipedia



- Collaborative Filtering
- From Clicks to Customers

Recommendation systems

- Business opportunity: Users rate items
- (Amazon.com, CDNOW.com, MovieFinder.com) on the web. How to use information from other users to infer ratings for a particular user?
- Solution: Use of a technique known as collaborative filtering
- Benefit: Increase revenues by cross selling, up selling

Clicks to Customers

- Business problem: 50% of Dell's clients order their computer through the web. However, the retention rate is 0.5%, i.e. of visitors of Dell's web page become customers.
- Solution Approach: Through the sequence of their clicks, cluster customers and design website, interventions to maximize the number of customers who eventually buy.
- Benefit: Increase revenues

Emerging Major Data Mining applications

- Spam
- Bioinformatics/Genomics
- Medical History Data – Insurance Claims
- Personalization of services in e-commerce
- RF Tags : Gillette (Collecting Data via RF Tags)
- Security :
  - Container Shipments which container contains drugs, etc.
- Network Intrusion Detection

T/F question: Spam vs Ham

spam is fake ham, know that need to differentiate b/t fake vs real; from the content

^that was military slang term just like FUBAR

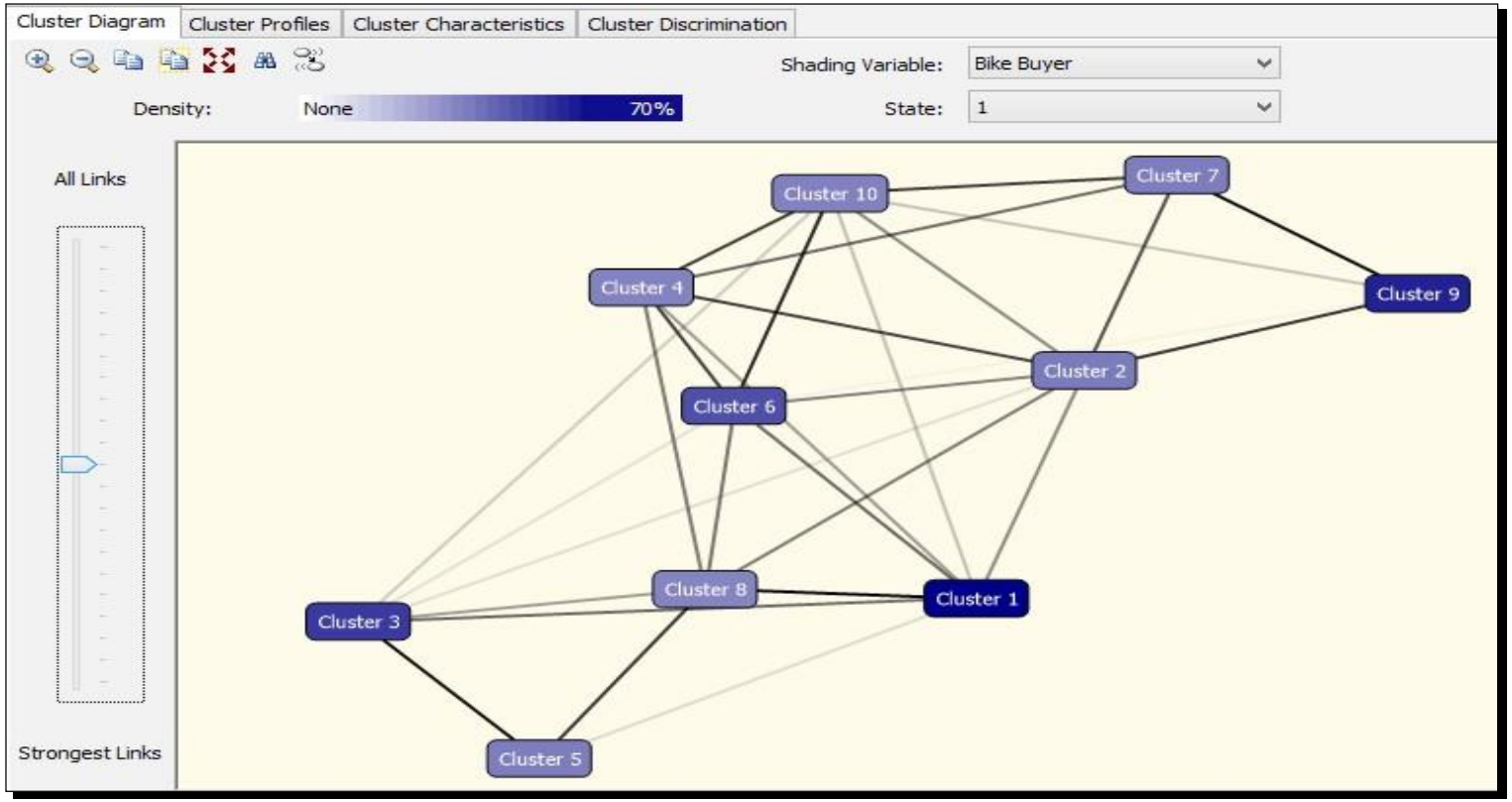
Radio-frequency identification uses electromagnetic fields to automatically identify and track tags attached to objects. The tags contain electronically-stored information. Passive tags collect energy from a nearby RFID reader's interrogating radio waves. Wikipedia

Applications

- Pattern Recognition
- Spatial Data Analysis:
- Image Processing
- Economic Science (especially market research)
- Crime analysis
- Bio informatics
- Medical Imaging
- Robotics
- Climatology

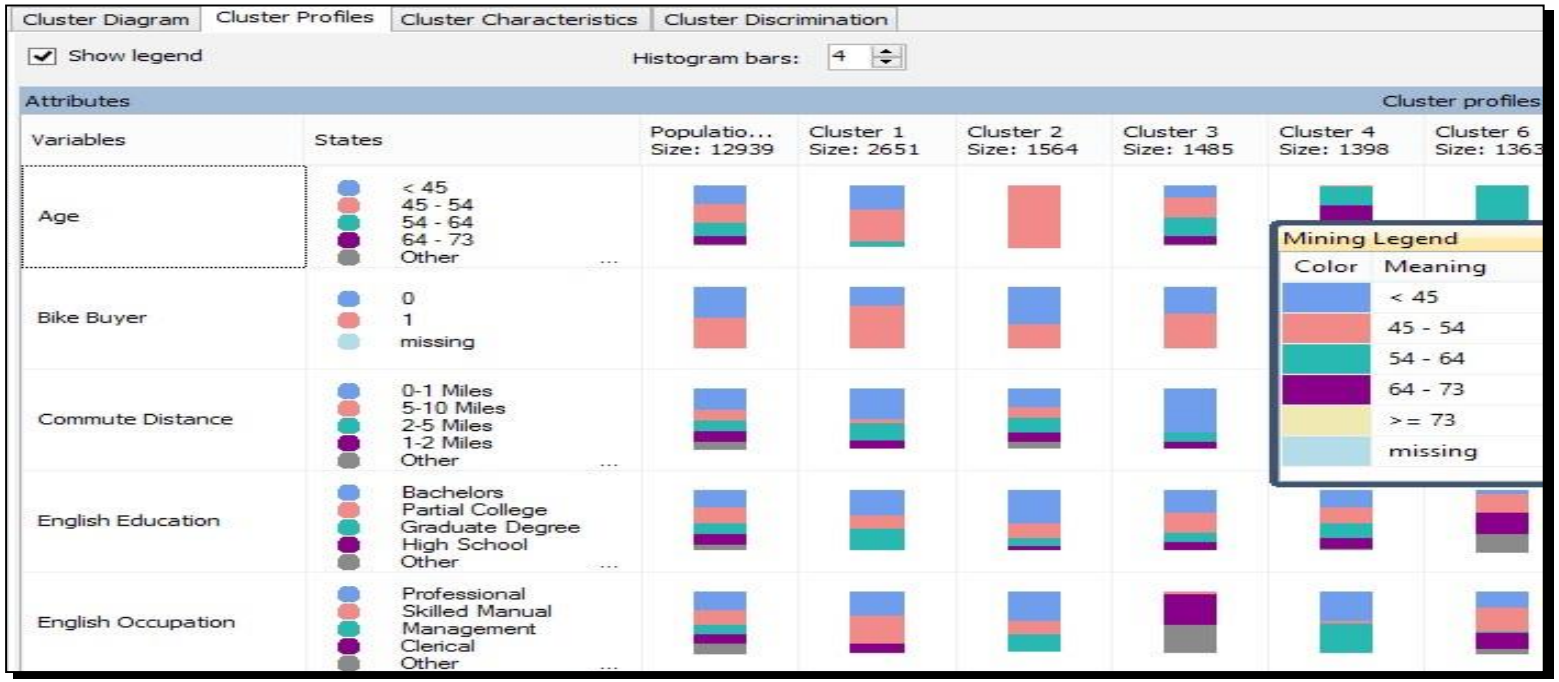
Cluster Diagram of Bike Buyers (State 1)

Clusters 3, 1 and 9 have the highest population



Cluster Profile – categorization of variables (States)

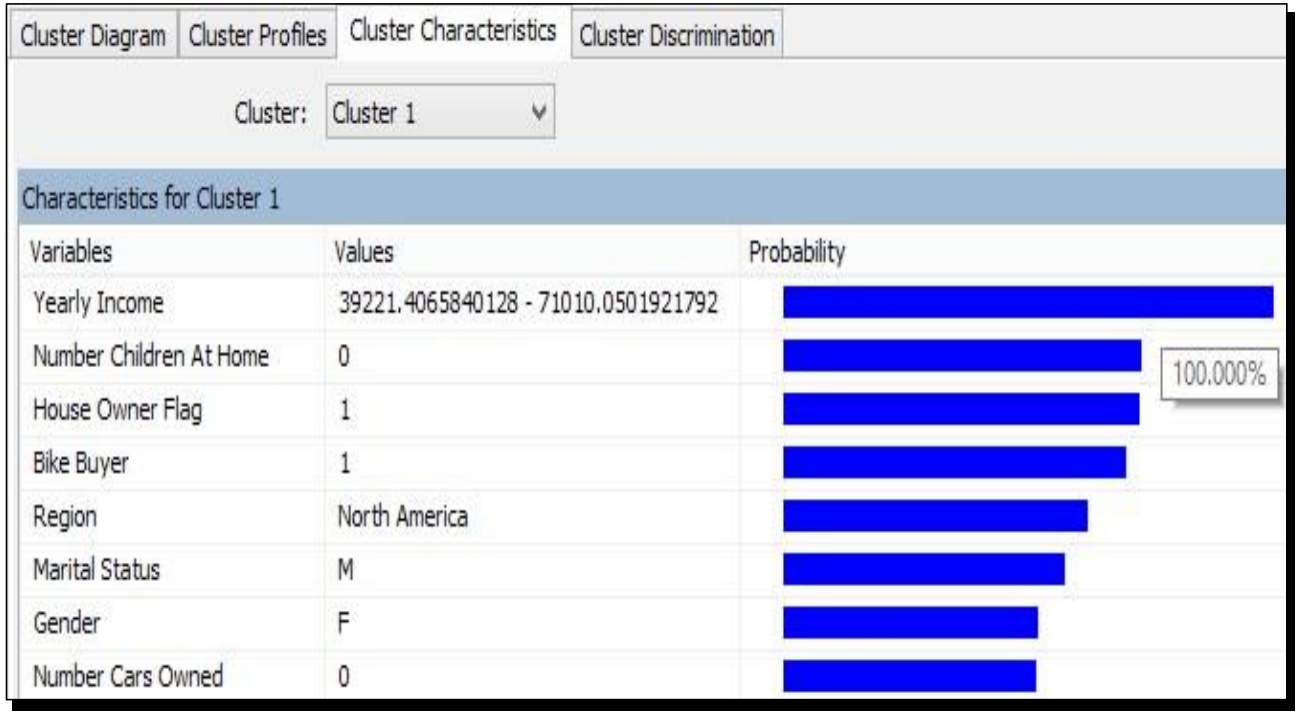
How each cluster is populated by the states



Cluster Characteristics

Probability Frequency ration of the variable having that value in the cluster.

- Distribution of variables within each Cluster



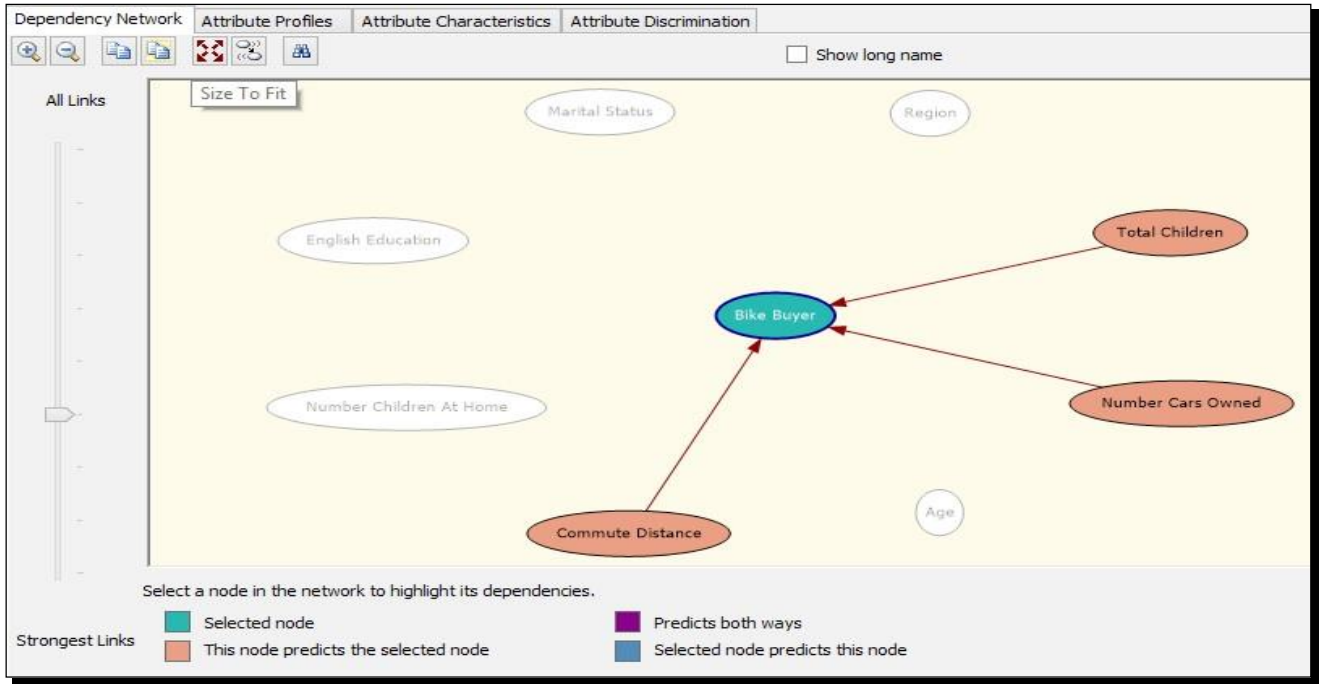
Cluster Discrimination – How the clusters differ

- In the cluster diagram cluster 1 is close to 8. This diagram shows contribution of each state of the variable differs in the cluster.

Naïve Bayes model

Varying the link shows the strength of the realtionship

- Shows the contribution of each node to the target node.



Lift Chart

- Compares the performance of the model with the Ideal and “Random Guess Model”.

Motivation ….Mailing Campaign

- Marketing dept. has a database of 10,000 customers
- Typically 10% of customers will respond to mail
- Has a limited budget , can only send out 5,000 mail

- How best to identify the 5,000 customers who are likely to respond to mail?

Two options to come up with mailing list

Randomly select 5,000 customers to target.

Use a mining model to target the 5,000 customers who are most likely to respond.

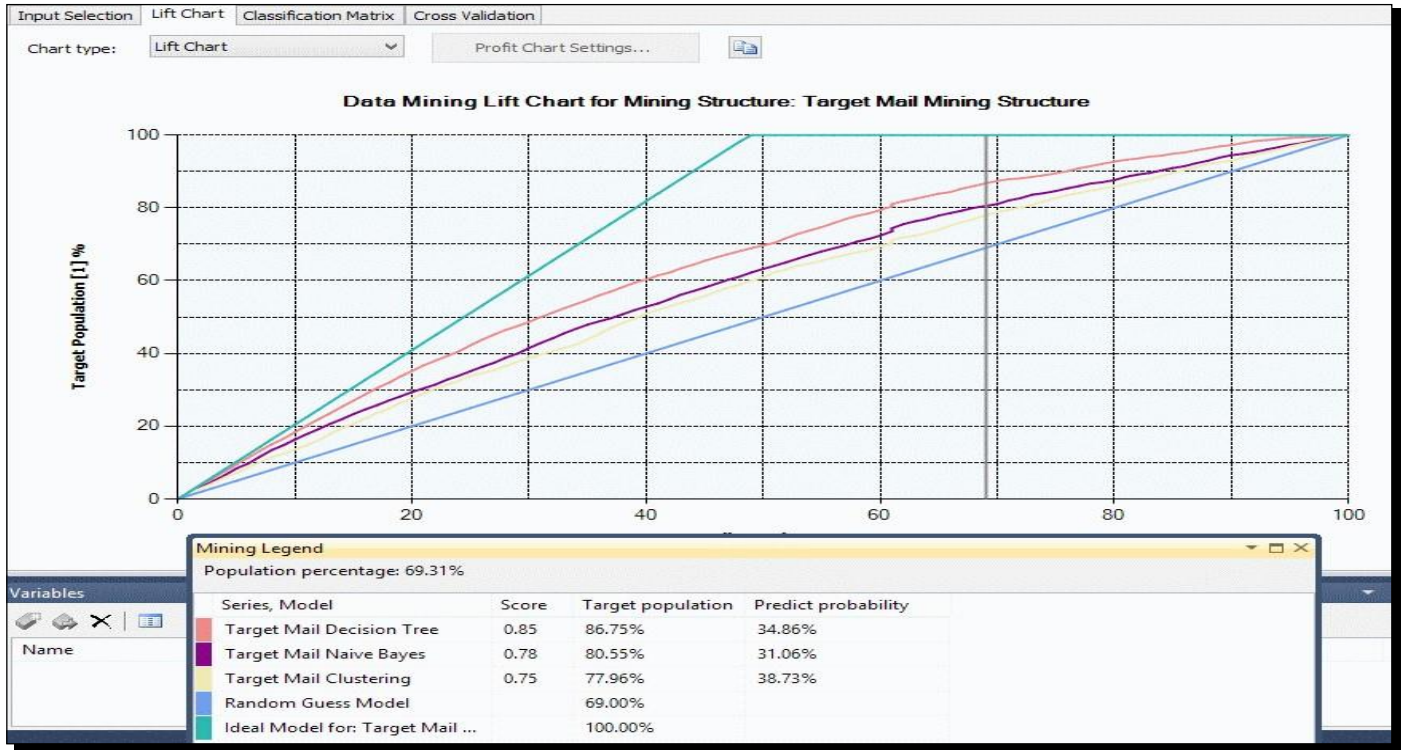
Interpreting the lift chart

- Random Line – typical response rate . For 5,000 customers, they expect to receive only 500 responses.
- Model response rate – you expect a better response rate because the model would identify those customers who are most likely to respond.
- Ideal line – perfect response, every target selected is correct.

Summary – Lift Chart

- A *lift chart* graphically represents the improvement that a mining model provides when compared against a random guess.
- The performance of the model falls between the ideal line and random guess.

Lift chart

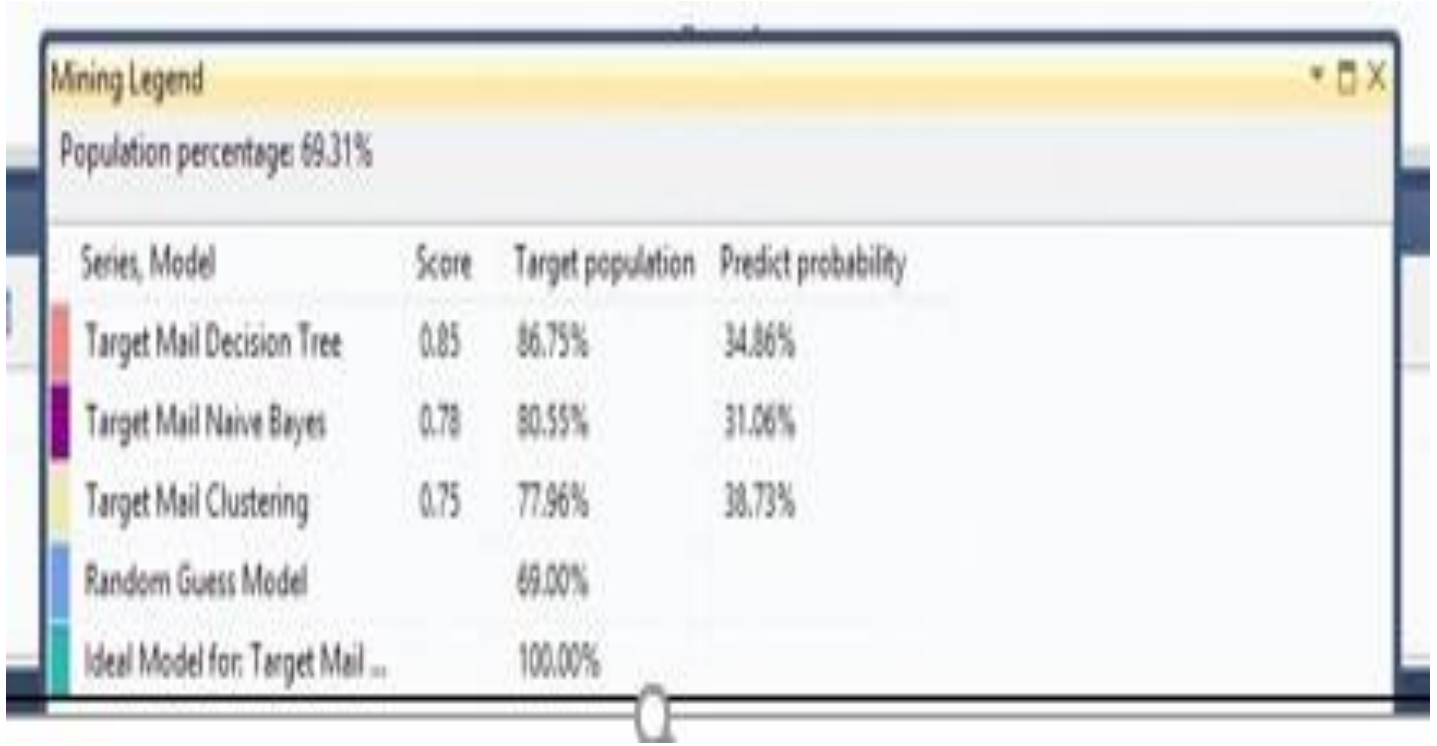

Input Selection
Lift Chart
Classification Matrix
Cross Validation
Chart type:
Lift Chart
Profit Chart Settings...
Data Mining Lift Chart for Mining Structure: Target Mail Mining Structure
Target Population [1] %
100
80
60
40
20
0
0
20
40
60
80
100
Mining Legend
Population percentage: 69.31%
Series, Model
Score
Target population
Predict probability
Target Mail Decision Tree
0.85
86.75%
34.86%
Target Mail Naive Bayes
0.78
80.55%
31.06%
Target Mail Clustering
0.75
77.96%
38.73%
Random Guess Model
69.00%
Ideal Model for: Target Mail ...
100.00%
Variables
Name

Lift chart cont’d

- Ideal case must be a 45 degree line , every selection hits a target
- Random guess must reflect the proportion of the “target” in the dataset. We have to go thru the whole dataset to find all the targets.
- Plotting the performance of each model must be between these 2 extremes.

Assessing the performance of the model on reduced population sizes

- By moving the “gray” line one could adjust the population of the dataset and read out the performance of the mining models.



Profit chart

- Used to compare the profitability of using various model.
- Uses a simple cost structure to compute this.

Simple Cost structure

Setting	Value	Comments
Population	20,000	Set the value for the total target population Your database might contain many customers, but to save on mailing expenses you might choose to target only the 20,000 customers who are most likely to respond. You can get this list by running a prediction query and sorting by the probability output by the predictive model.
Fixed cost	500	Enter the one-time cost of setting up a targeted mailing campaign for 20,000 people. This might include printing, or the cost of setting up an e-mail campaign.
Individual cost	3	Enter the per-unit cost for the targeted mailing campaign. This amount will be multiplied by a number equal to or less than 20000, depending on how many customers the model predicts are good prospects.
Revenue per individual	400	Enter a value that represents the amount of profit or income that can be expected from a successful result. In this case, we'll assume that mailing a catalog results in purchase of accessories or bikes averaging $400. This amount will be used to project the total profit associated with high probability cases.

Comparison of profitability of models