BADM- Analysing Churn for a mobile service provider

Farah Sarfraz
Niharika Vempati
Pallavi Sabharwal
Shilpa Murthy

Data Preparation

- Substitute missing data: Several fields (some numerical) had 'NA' and 'Blanks'
  - Replaced 'NA' with Mode
  - Eliminated records with 'NA'
  - Left 'NA' as is

- Clean-up of data:
  - Similar records with similar entries
  - Eliminated Typos
  - Maintained uniformity

Data mining method

- Variable selection:
  - Classification tree to identify variables that contribute to label ‘Churn’
- Create a model:
  - Ran multiple models : classification tree, logistic regression etc
  - Best result with logistic regression
- Run on Test data:
  - Prepared test data along same line
  - Ran model on test data

Benchmarking + Critique

Benchmarking:

- Majority rule used initially
- Further, classification tree became benchmark for Logistic regression

Critique:

Positives	Negatives
No dummy variables	~40 variables- hard to implement
Best subset analysis can be used with only 2% error increase	Slightly higher error percentage when used without network duration
Increased data quality- data cleansing processes and categorical variables	Lot of data cleansing required as data needs to be converted to categorical variables

Recommendation

- We need to look at the levers under the management's control that have a high correlation with churn.
- There are many ways of engaging a potential 'churner'. We must refer to the 11 most important factors and optimize around each one depending on the cost. For example, we find that roaming charges and offers and promotions predict whether a consumer is a churner or not.
- Although we found network duration to be highly correlated with churn/non-churn, we found that it was a consequence/deriving factor, rather than an influence.