



BADM Data Mining Contest

CrowdANALYTIX Submission

Predictive Model – Mobile Value Added Services

Team – BADM-Mohali-CtrlF

November 2012



Problem Overview – Predicting Adoption of MVAS

Business & Data Mining Objective

Client

Advisory board to a consortium of Indian telecom operators consulting on declining industry revenues

Industry Context

The Industry is witnessing a rapid decline in voice tariffs amongst increasing pricing pressures and high competition. MVAS is widely believed to be the next source of revenue growth for telecom operators

Business Goal

To increase the subscription for MVAS services as a new revenue stream for telecom players

Analytics Goal

Predict the likelihood of acceptance of MVAS among the existing customer base and understand factors governing adoption and usage

Thought Starters (from Secondary Research)

MVAS consumers are interested in the following services and adoption of each needs to be studied

- Information based services such as news updates, health-related information, stock details etc.
- Application based services that need the user to play an active role such as checking the status of payments, GPS
- Enablement services which are a substitute to those provided by physical infrastructure such as a bank or a school e.g. person-to-person payments, travel reservations etc

Most important categories within the Non-SMS services include ring tones, music and entertainment, gaming and mobile browsing

Preparation Phase – Data Cleansing & Exploration

Data Preparation – Training & Test Set

1

Checking for Missing Values

- Removing records within very high no. of 'NA's in the dataset – These were records who had not completed the survey (Training only)
- Variables with lots of missing data were left out of the analysis – e.g. Reasons to switch from previous provider

2

Modifying Misrepresented Information

- Respondents who did not own a mobile phone and used landline only were listed as MVAS users. These were changed to Non-MVAS category
- Variable on age had erroneous entries e.g. listing 2011 as age – these were replaced with average figures

3

Variable Selection and Data Conversion

- Selected important variables based on secondary research and internal domain knowledge
- Multiple variables with similar intent were avoided e.g. ignored handset manufacturers for phone type (Smartphone/Internet enabled)
- Combined multiple options in variables into a binary form (Agree / Do Not Agree – Kept NAs as such)
- Created input variables as per the required format – e.g. binning continuous variables for Naïve Bayes

Data Mining Techniques – Supervised Learning

The Power of Ensembles !!

THE IDEA

- Combine results from multiple classification techniques in order to average out the prediction inaccuracies in data
- Technique selection was dependent on the type of data available – mostly categorical variables. Hence, decided to use
 - **Classification Tree**
 - **Logistic Regression**
 - **Naïve Bayes**
- Techniques such as Classification Tree and Logistic Regression would also provide information on key predictor variables
 - Predictors at the top of the classification tree and being repeated multiple times
 - Predictors that are part of the ‘Best Subset’ selection and follow Mallows rule
- Success Class – MVAS adoption

BUT XLMINER HAD OTHER PLANS !!

- Best Subset Selection refused to work – Reported “**Overflow Error**” – no information available on how to correct it
- Revised methodology to first use classification tree, derive important variables and thereafter run the logistic regression
- Started with around 20 predictors and obtained a list of 11 important variables from the classification model
- Scored test data using each of the three algorithms
- Averaged results using equal weights to derive the end probabilities

Learnings on Prediction Methodology

Validation Set Results

Classification

Error Report			
Class	# Cases	# Errors	% Error
MVAS	697	0	0.00
Non-MVAS	42	42	100.00
Overall	739	42	5.68

Misclassifications in top 10% - 6

Naïve Bayes

Error Report			
Class	# Cases	# Errors	% Error
MVAS	697	12	1.72
Non-MVAS	42	37	88.10
Overall	739	49	6.63

Misclassifications in top 10% - 4

Logistic Regression

Error Report			
Class	# Cases	# Errors	% Error
MVAS	697	0	0.00
Non-MVAS	42	42	100.00
Overall	739	42	5.68

Misclassifications in top 10% - 6

FINAL SCORE: 121 – Misclassifications - 3

Key Learnings !!

Keep Test Data

Important for techniques such as Classification trees where in the validation set is used for pruning, and performance is biased towards the data set

Logistic Regression Is Key

The accuracy could have improved significantly had the 'Best Subset' option been functional. Else keep time for testing multiple permutations of variables. We used all 11 in our model due to time constraints

Simple Averages Are Not Necessarily Good

Its useful to consider weighted averages with higher weights being given to models with greater lift ratio. Naïve Bayes could have been given a low weightage

Ensembles Seem to Work

The averaging resulted in changing the rankings of top 10% in the validation set. This can be seen in the reduced no. of misclassifications

Recommendations

Important Factors for Predicting MVAS Adoption

Consumers with the following characteristics are most likely to adopt MVAS:

DEMOGRAPHICS (WHOM TO TARGET)

- Important
 - Male Gender, Age Groups (15-45)
- Not As Important
 - Education (Professionals / Graduates/Rest)

MOBILE OWNERSHIP & USAGE (WHOM TO TARGET)

- Important
 - Smartphone Ownership, High expenditure (Monthly >700), High Internet Usage
 - Provider call charges (High ratings)
 - Provider of offers and promotions (High ratings)
 - Provider network coverage (High ratings)
- Not As Important
 - Handset brand, Mobile Service (GSM/CDMA), No. of SIM cards/mobile phones

INTEREST (WHAT TO OFFER)

- Important
 - Spend time on social networking websites – Internet enabled
 - Spend time reading newspapers – News Updates
 - Travel to unknown places – Need GPS
- Not As Important
 - Spend time watching sports
 - Like to track stocks

Thank You !!

Team CtrlF Members:

Anand Prasad (61310460)

Charanpreet Singh Arora (61310575)

Dhruv Gandhi (61310862)

Gagan Oberoi (61310627)

Nikesh Lamba (61310493)