Contest Title: Data Mining Contest- MVAS Prediction

Submitted by: Team - BADM-Mohali-CtrlF

Group Member Name	PGID
Gagan Oberoi	61310627
Charanpreet Singh Arora	61310575
Nikesh Lamba	61310493
Anand Prasad	61310460
Dhruv Gandhi	61310862

Executive Summary

The Indian Telecom industry is amongst the most fiercely fought services market in India with more than 10 large-scale operators providing voice and data services at highly competitive prices. The Industry has witnessed significant reduction in profit margins in recent years, with the average revenues per user from voice telephony being amongst the lowest in the world. On the other hand, margins from value added services are largely seen as the next source of growth for mobile operators. The telecom companies therefore need to be innovative to come up with more relevant and value-adding services to gain a competitive edge in this cut-throat environment.

The client in this case is an advisory board to a consortium of Indian telecom operators consulting on the issue of declining revenues. The consortium wants to increase the subscription of MVAS services (services such as ring tones, sports updates, weather updates etc.) and identify potential customers within their existing bases to be targeted. They also want to understand key factors that would govern the adoption and usage of mobile value-added services.

Demand for MVAS in India has been on the rise in recent years, with growth in mobile penetration and customer purchasing power. According to a report published by Deloitte Consulting, the demand for MVAS is being driven by the following services:

- **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

These services determine consumer attitude towards MVAS and thus need to be considered for any predictive model developed to measure the likelihood of adoption. Other important factors that could be considered are **access to technology** (smartphone ownership, Internet enabled phone), **demographics** (age, gender, education etc.), **customer satisfaction levels** with mobile operators (call charges, network coverage etc.) and **current usage levels** (monthly expenditure on mobile services, Internet usage).

There are various classification techniques that could be employed to build such a predictive model, the choice of which would depend on the type of data available – categorical or continuous. We also need to keep in mind if the data will be available on future dates to help predict consumer behavior. It is typically considered useful to combine predictions from different techniques in order to reduce

inaccuracies and obtain more sound results. This 'combination-based' methodology called **ensembling** was employed by our team to reach final predictions. End results indicate valuable information on traits defining MVAS adoption with the younger age groups and the male gender dominating trends. Moreover, services such as social networking, news updates and GPS outscore other examples of MVAS as triggers of customer adoption.

Prediction Methodology – Classification (Supervised Learning)

The key to such predictions would be using mechanisms that could help 'classify' any customer into MVAS adopters and Non-adopters, based on previous available trends (called supervised learning). Such mechanisms would obtain valuable information/linkages from existing data and use the same to make future predictions (on finding similar traits in a future customer). We have further followed the principle of 'ensembling' and have combined the forces of different classification techniques to average individual prediction probabilities for each of the customers.

Phase 1 – Data Preparation & Exploration

The first step in the process is to have detailed understanding of the presented data and to explore it for obvious discrepancies. The training data set was analyzed and the following changes were made:

I. Checking for Missing Values – Removing records with high number of 'NA's in the dataset – These were records who had not completed the survey. Variables with lots of missing data were left out – e.g. Reasons to switch from previous provider

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

Phase II: Variable Selection

Even before the data set is examined, one needs to develop deep understanding of the business objective to be achieved, and study available data from the perspective of realizing the same. Having a sound understanding of all predictor variables and how they affect the final objective is imperative. This can be developed through secondary research sources and utilizing domain knowledge of industry insiders to better understand market trends and important business metrics.

Business analytics greatly values 'parsimony' i.e. a model with lesser number of predictor variables but with similar accuracy levels is far better than a model with a large number of variables. Thus it is

important to select an initial set of parameters that would be most helpful in predicting the likelihood of MVAS, and modify them accordingly as per the algorithm to be employed. In this case the following steps were undertaken after selecting the important predictors based on industry 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)
- Created inputs as per the required format e.g. binning continuous variables for Naïve Bayes

Phase III: Training, Validation & Scoring on Test Data

After the Data Preparation and Variable Selection phase we proceeded with running multiple classification techniques on the training data, and with assessing the prediction performance. Thereafter, each of the algorithms was used to calculate a probability of adoption for the 'test' customer data, which were averaged to calculate the end result. We have used the following three classification/prediction methods:

- **Classification tree** – This also provides us with a list of important factors governing adoption. Those at the top of the classification tree and being repeated multiple times are key. We started with around 20 predictors and obtained a list of 11 important variables from the model.
- **Logistic Regression** – Using important predictor variables derived from the classification tree. It also has a 'Best Subset' option that helps judge the important predictors
- **Naive Bayes** – An important technique for cases requiring a ranking of records (customers). This is also preferred when we have categorical variables for predictions (similar to the case here)

Prediction Errors on Validation Set (Kept Separate from Training Data Used to Develop the Model)

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 AFTER COMBINING – 121/124

Misclassifications in top 10% - 3

Key Dos and Don'ts While Devising Predictions

I. Keep Separate Test Data – More important for techniques such as Classification trees where in the validation set is used for pruning, and the model performance is biased towards the data set. A separate test data will help check the predictive accuracy better, before deploying the model in the actual world

II. Logistic Regression Is Key – Best subset option in XLMiner may have issues and report an 'overflow error' which is difficult to navigate. The accuracy could have improved significantly had the 'Best Subset' option been functional. In such a case, one needs to test multiple permutations, taking different collection of predictors at a time, and select the best option (with minimum prediction error)

III. Simple Averages are Not Necessarily Useful – While using ensembling, it is useful to consider weighted averages with higher weights given to models with greater lift ratio. Probabilities from Naïve Bayes could have been given a low weight as it is used mostly to rank and not predict exact probabilities.

Key Takeaways for the Client

Besides predicting the likelihood of MVAS adoption, the other underlying objective of the data mining exercise was to identify key variables which contribute to why a customer would opt for Mobile VAS. After various parameters and characteristics are identified, telecom operators can then decide how to segment the market better and develop specific solutions to market offerings as per customer needs. This will help gain significant competitive advantage and increase the top line by diversifying into new service categories. From the analysis conducted, the following points are important to market MVAS to customers

I. Understanding WHOM to Target – Consumer Demographics

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

II. Understanding WHAT customers prefer – Mobile Ownership & Usage

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

III. Understanding WHAT services to offer – MVAS Preferences

- 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

Mobile Handset & Company/Network Quality: Amongst important characteristics for a customer to use MVAS are "Having a smartphone" and "the call quality, network coverage and call charges" offered by the telecom operator. Thus it is imperative to identify customers using smartphones or handsets which support VAS services and then specifically target them rather than targeting the entire customer base. This will significantly reduce the customer acquisition costs. It is further important to perform well in "Points of Parity" parameters such as having a good network coverage and competitive pricing. This will help companies expand customer base as well as retain existing customers using MVAS

Customer Spending & Promotional Offers: Another important variable is to understand the spending habits and purchasing power of customers. The predictive model discovered that customers with substantial monthly expenditure (> INR 700 monthly) were using MVAS. Hence, targeting customers with high monthly expenditure on mobile services would have higher likelihood of acceptance. Another parameter having significant effect on MVAS adoption is the kind and number of promotional offers provided by telecom operators. Thus, if telecom operators can come up with relevant and innovative solutions it would help them increase their market share. A telecom operator may also decide to target customers with lower monthly spending, but needs to offer solutions which cater to specific service needs at competitive prices.

Customer Usage Habits: It is easier to market MVAS to customers who are aware of and feel the need to use different kind of services such as news updates, GPS – or are actively using the Internet on their cell phones for social networking, for gaining access to information, to make payments etc. One of the most important parameters which can be leveraged upon is to identify customers based on behavioral parameters and their likes/dislikes – i.e. target those who spend a substantial time on mobile web, who usually read newspapers online or those who travel to new places by road, for various MVAS services.