ASSIGNMENT SUBMISSION FORM

Treat this as the first page of your assignment

Course Name: Business Analytics using Data Mining

Assignment Title: **Crowdanalytix - Predicting Churn/Non-Churn Status of a Consumer**

Submitted by: **Team FoodiesBADMMohali**

Group Member Name	PG ID
Shilpa Murthy	61310803
Niharika Vempati	61310452
Pallavi Sabharwal	61310622
Farah Sarfraz	61310296

(Let us not waste paper, please continue writing your assignment from below)

Executive Summary:

One of the most critical factors in Customer Relationship Management that can make or break a company's long-term profitability is **churn**. If a company can predict whether a customer is likely to churn, it can take a more targeted approach to running promotions to reduce churn. This is a sophisticated evolution from the traditional approach to incentivize all customers equally to reduce churn as it allows companies to spend their marketing budget more effectively. It is this managerial usefulness of being able to predict churn that attracted us to this assignment. In this project, we have mainly used classical data prediction techniques of classification tree and logistic regression to obtain accuracy with error rate of 42%. Taking into consideration the fact that misclassifying a churner as a non-churner, we lowered the cutoff of probability of churn for the classification to 0.4. Although this has resulted in a higher rate of error, it has reduced the overall cost of misclassification, which is the objective of the assignment. Below we outline the steps involved in deriving the prediction for the test data:

1) Data Preparation:

Although this step was the most mechanical, it was unfortunately also the most cumbersome and it took up 60-70% of our time. It was crucial however for us to spend this time as we wouldn't have been able to convert categorical values to numerical ones and get sensible results otherwise. We thoroughly and patiently combed through 40 columns of data to remove any anomalies before proceeding on running the model. Below are the steps we followed for both the training and test data sets separately:

- Handling missing data: There were many fields which had values of NA or were missing. While it was ok to have NA values for some factors, it did not make sense to include records with other NA values. Hence we used different approaches to handle missing data including replacing with NA, elimination and mode. For example, we applied mode for Age. Also, we deleted records with NA as service provider as those records were junk with most fields empty.

- Cleaning up data: There were some typos and spaces in the fields which were throwing off the analysis. We eliminated those & adjusted formatting in other fields where it wasn't uniform. For example, Average sms-es/day had some strange formats of date, which interpreted and replace with numbers. Initially we were unable to proceed with any data analysis as there were columns with over 30 distinct values. This was unacceptable by the tool, so we had to deep dive into the data and fix spelling errors for the tool to work. For example, service provider mails.

- Data partitioning: We split the Training data into 60% Training Set and 40% Validation set.

- Converting variables from categorical to numerical: As many of the variables were often categorical, we need to convert them to numerical category values. This was an essential step before being able to use these columns in the classification tree and logistic regression. For example, we converted Provider.Network.Coverage which contains Somewhat agree/Somewhat disagree to numerical values of 0, 1 etc.

2) Business Objective:

The main business objective was to minimize the cost of misclassification, as defined as:

{ (# churn misclassifications) x (X - cost of churn misclassification) + (# non-churn misclassifications) x (5X - cost of non-churn misclassification) } / n

This is dependent on not only reducing overall error rate, but also reducing the error rate for classifying a churner as a non-churner by 5-times as much as vice-versa keeping in mind the high cost of misclassification of the former.

3) Data Mining Method:

- Removing factors that were less relevant: The first step we took was to pare down the number of variables/columns to a manageable number. As there were too many columns in the original file, data processing was becoming cumbersome, so we went through the columns and took a judgment call about whether there was any sense in including the field in the analysis. We then deleted the factors we believed would have little or no prediction power. For example, we deleted all the TV viewing habit related columns.

- Using Classification Tree to Shortlist Variables: To determine which factors to keep and which to drop, we decide to employ the classification tree method to shortlist the factors. Using this tool, we were able to narrow down to the 12 most important factors that predict whether a customer churns or not. We found that 'Network Duration' was the single factor that was highly correlated with churn and was able to predict all the variability. Initially, we were happy to be able to get such high accuracy rates, though very surprised at the rawness of the solution. However, upon further thought we became suspicious about this variable and upon consultation with the professor realized that it was the antecedent to predicting the churn and hence could not be used in the model. This left us with the below 11 most important variables for predicting churn, in reducing order of importance:

1. Age
2. Service.Provider..Primary._ord
3. Provider.Roaming.charges_ord
4. Provider.Offers.and.promotions_ord
5. Usual.top.up.size.Pre.Paid.User._ord
6. Average.use.of.Value.Added.Services..VAS...music.video.downloads_ord
7. Average.use.of.Value.Added.Services..VAS...Document.Reader..pdf..word.etc.._ord
8. Budget.conscious_ord
9. Travelled.to.foreign.countries_ord
10. I.have.a.large.circle.of.friends._ord
11. Total.No.of.people.in.your.house

- **Building the prediction model:** After running the classification tree above, we re-ran the tool to give us to give us the prediction for churn vs. non-churn. In classification tree, although error rate for training data was extremely low, however the error rate for the validation data set was extremely high. This indicated overfitting and the model wasn't very robust. Hence we decided to explore another data mining model – logistic regression. The results from logistic regression were more promising for both the training and the validation data, with sensible error rates.

4) Results:

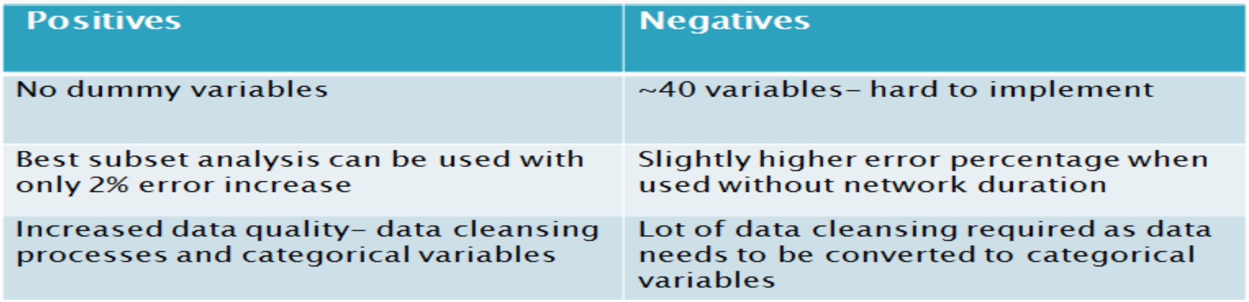
- **Error rate for validation data:** Using the logistic tree model, we were able to achieve the lowest error rate of 38% for the validation dataset.
- **Adjusting for the cost of misclassification:** As the cost of misclassifying a churner as a non-churner was 5 times that of classifying a non-churner as a churner, we decided to compromise on the error rate to minimize the total cost of misclassification. We reduced the cutoff value for the probability of classifying a record as churn to 0.4, which resulted in more records being labeled as churn and a higher error rate of 42%. However, it minimized the overall cost of misclassification as fewer churners were being incorrectly labeled as non-churners.
- **Picking between parsimony and precision:** The above model was run using 40 variables. To make the model more managerially useful, we could retain the top 11 variables only and drop the rest, however this would result in a drop in accuracy of prediction. When we re-ran this model with 11 variables, the error rate increased by 2%. Depending on how costly these errors are, the manager must decide to compromise one option for the other.
- **Applying to test data:** We ran the logistic regression model of 40 variables on the test data and please find attached the final excel sheet with the results.

5) Benchmarking and Critiquing:

While we got similar results as other teams, there were certain characteristics that we really liked in other teams' projects which we hadn't considered in our model as well as aspects which we had considered but chose to leave out of our model. Here they are:

- **Clubbing of highly correlated variables:** Some groups combined highly correlated variables to give a ‘super variable’. This increased prediction power and parsimony. The drawback of this method though is that it would reduce managerial usefulness. For instance, if we found that age and talk time were highly correlated, how would the new hybrid variable X of the two help managers take decisions?

- **Partitioning of data into three parts**: We only partitioned data to training and validation. Had we had a test data set, we could have used three levels to check prediction power of model.
- **Ensembles:** One of the teams combined results from multiple models to get higher accuracy. We thought this was a very nice touch!
- **Trends:** One of the teams suggested that having information about trends such as usage patterns would be a good predictor of whether a person is likely to churn. Great idea!
- **‘Black box’ models:** We took a decision early on not to employ black box models such as Naïve Bayes as the focus of our assignment was to offer managerial insight. We believed that not being able to visualize how each variable affects the outcome of churn/non-churn was a non-negotiable drawback and we would rather compromise on accuracy than managerial usefulness.



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

5) Recommendation:

- **Demographic related factors:** The results of this model highlight some factors which can predict churn, which are either related to the demographic of the customer (eg. Age, gender etc.) or the actions of the service provider (eg. Call plans, roaming charges). Decisions managers can take vary for both these types. For demographic related factors, mangers can preempt churn from vulnerable demographic types and offset their propensity to switch providers by offering superior quality in the other service related variables that matter to customers (as from the classification tree). This can help in segmentation and targeted promotions and marketing. I this case, age, travel and circle of friends and family were the most crucial.

- **Service related factors:** The level of service a telecom provider decides to offer depends on how much value an average customer places on these factors. This model allows the company to formulate a general strategy of service based on precise quantification of the cost of improving service vs. the cost of losing a customer. Decision making on numbers rather than intuition!
- **The company can employ two approaches: Spend on & retain non-churners:** Direct efforts towards strengthening relations with those customers who are likely to stay on, and cut out customers who are likely to churn and are hence less profitable in the long-run. **(OR) Spend on & reduce churners:** Some companies find that it is actually cost-effective to direct marketing efforts towards vulnerable customers who would've otherwise churned. This reduction in churn would actually increase long-term profitability of the firm due to retention of the customer's lifetime value. The company would typically, in this case, not expend budget on non-churners as they believe that these customers would not switch irrespective of the promotional spend.
- **Network Duration:** Although network duration was excluded in this model as it was an antecedent of churn/non-churn status, however in practical usage, we believe that the time a person has spent with a service provider would be a good indicator of whether a person is likely to churn or not. If I've stuck on to a particular provider for 5 years, the chances of me switching to another are much lower than that of a brand new customer.