Data Mining and Exploration
Interim Report

Thorvaldur Helgason (s1237131)
Daniel Stanoescu (s0838600)
Maria Alexandra Alecu (STUDENT NUMBER)

February 27, 2013

What we have done so far:

- Pre-processing:
  - Replaced missing values with both zeros and mean values.
  - Converted the dataset to a binary bag-of-features.
- Familiarized ourselves with Naive Bayes, SVMs, and Decision Trees.

The data we have chosen for this project is the Orange telecom customer behaviour dataset. We have been provided with a labelled training set for the data focusing on predicting the customer churn(switch providers), appetancy(tendancy to buy new products) or purcahse upgrades. Given the proposed dataset, we are looking at solving a classifcation problem. To respect customer privacy, both the order of the customers as well as the variables in the dataset have been shuffled.

After inspecting the data, we have identified that i contains both numerical and categorical values. We R to explore as well as preprocess the data. We began with preprocessing the numerical data. We replaced all the entries that were empty with zero. Secondly we replaced the zero entries with the mean of the values of each variable in the dataset. In the case of the categorical data, we changed the empty strings to Not Available. We then converted all the strings to numeric values so we can handle data operations easier.

We began experimenting with a series of algorithms in order to figure out which ones would perform better. We looked at employing SVMs but we quickly abandoned the idea as they would require a higher computational requirement than other methods.

What we plan on doing:

- See which pre-processing techniques and features work best for different classifiers.
- Familiarize ourselves with more classifiers.
- Compare the performance of the classifiers with the criteria described below.

We plan experiment with a linear classification method, more exactly, employing logistic regression methods.

What comparisons we want to run:

- Split data up randomly: 80% of instances will be the training set and 20% the test set.
- Perform 5-fold cross-validation on the training set and do final evaluation on the test set.
- For each classifier we store their accuracy, confusion matrix, ROC curve and AUC.