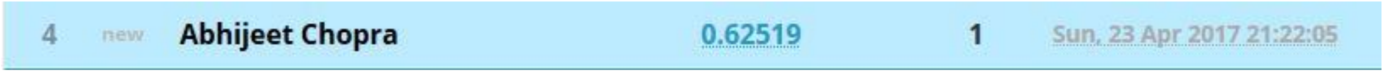
Kaggle Assignment 3

CSCI 527 Spring 2017 Competition 3

Due: 04/23/17, 11:59 PM CT

Metadata

1. Name : **Abhijeet Chopra**
2. CWID : *****
3. Kaggle Display Name : **Abhijeet Chopra**
4. Kaggle User Name : **abhijeetchopra**
5. Kaggle Email Address : *****@*****
6. Programming Language : **R**
7. Screenshot of best performing submission:



4 new **Abhijeet Chopra** 0.62519 1 Sun, 23 Apr 2017 21:22:05

Technique

1. Data preprocessing
    a. Converting numerical comma separated values into factors to ensure modelling functions treat them correctly.
    b. Shuffling rows to prevent bias due to too many consecutive same values.
    c. Item1id was not in numerial format so converted into numerical.
    d. Winner column was read as factors.
2. Data mining
    a. The **Decision Tree** algorithm using **C5.0** that was implemented in the R programming language in package "**C50**" was used.
    b. Model was obtained from all 379251 rows of training dataset and applied on 468120 test dataset rows.
3. Data post-processing
    a. Function **predict()** with argument **type="prob"** gives vector with probabilities for both **FALSE** and **TRUE** events. Hence, only the probabilities for the **TRUE** event were extracted from the given vector.

Innovations (Trials & Errors)

1. C.50, Rpart and custom Association Rules were applied to create predictive models from the given training set.
2. Predictive model was made from first taking all rows in to consideration which did not yield results. Then only 10,000 rows were selected.

References

1. Johnson, C. (2014, August 29). Decision Trees in R using the C50 Package | Retrieved from http://connor-johnson.com/2014/08/29/decision-trees-in-r-using-the-c50-package/.