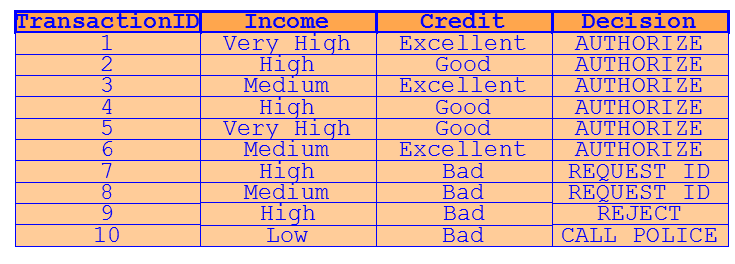
**Amazon Transaction Data Analysis**

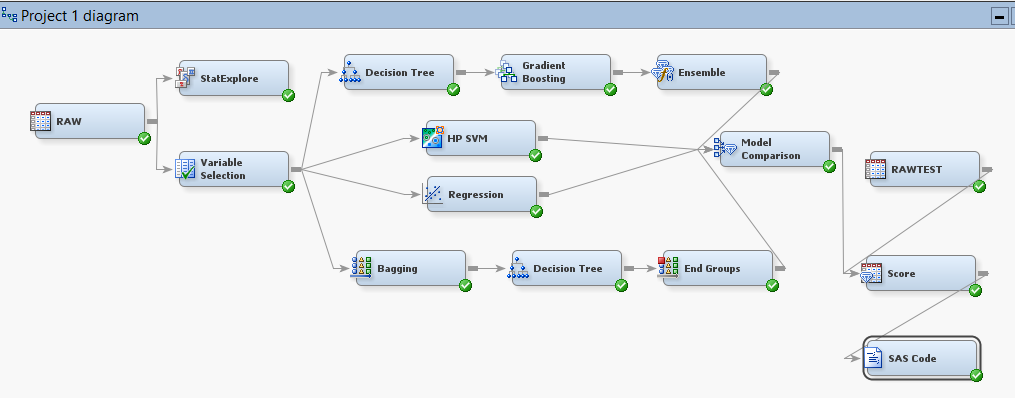
A pair of datasets are provided: a training dataset (amzntrain.arff) and a testing dataset (amzntest.arff). Data is taken from “An Empirical Analysis of the Value of Complete Information for eCRM Models”, Padmanabhan, B., Z. Zheng, and S. Kimbrough. MIS Quarterly, 30(2), 2006. Variables 1-15 are site-centric variables; 16-40 are additional user-centric variables and the last is the dependent variable.

|  |  |  |
| --- | --- | --- |
| No. | Variable | Description |
| 1 | gender | “1”—Male, “0” -- Female |
| 2 | age | Age of the user |
| 3 | income | Income of the user |
| 4 | edu | “0” – high school or less, “1”-- college, “2” – post college |
| 5 | hhsize | Size of house hold |
| 6 | child | “1” – have, “0” – not have |
| 7 | booklh | No. of bookings the user made at this site in the past |
| 8 | sesslh | No. of sessions to this site so far |
| 9 | minutelh | Time spent in this site so far in minutes |
| 10 | hpsesslh | Average hits per session to this site |
| 11 | mpsesslh | Average time spent per sessions to this site |
| 12 | booklc | Dummy variable, indicating if the user has booked at this site up to this point in the current session |
| 13 | httlc | No. of hits to this site up to this point in this session |
| 14 | minutelc | Time spent up to this point in this session |
| 15 | weekend | Indicating if this session occurs on weekend |
| 16 | bookgh | No. of past bookings of all sites so far |
| 17 | sespsite | Average sessions per site so far |
| 18 | sessgh | Total no. of sessions visited of all sites so far |
| 19 | minutegh | Total minutes of all sites |
| 20 | hpsessgh | Average hits per session |
| 21 | mpsessgh | Average minute per session |
| 22 | awareset | Total no. of unique shopping sites visited |
| 23 | basket | Average no. of shopping sites visited per session |
| 24 | single | Percentage of single-site sessions |
| 25 | booksh | Percentage of total bookings are to this site |
| 26 | hitsh | Percentage of total hits are to this site |
| 27 | sessh | Percentage of total sessions are to this site |
| 28 | minutesh | Percentage of total minutes are to this site |
| 29 | entrate | No. of sessions start with this site/total sessions of this site |
| 30 | peakrate | No. of sessions the user spend the most time within this site/total sessions of this site |
| 31 | exitrate | No. of sessions end with this site/total sessions of this site |
| 32 | SErate | No. of sessions start with this site/total sessions of this site |
| 33 | *bookgc* | Binary variable, indicating if this user has booked at any sites up to this point in the current session |
| 34 | *hitgc* | Total hits of all sites in the current session |
| 35 | *basketgc* | No. of shopping sites in this session |
| 36 | *minutegc* | Time spent of all sites in this session |
| 37 | *SEgc* | Indicating if this session uses search engines |
| 38 | *path* | Indicating if this site is an entry/peak |
| 39 | *hitshc* | Hits to this site/ hits to all sites in this session |
| 40 | *minutshc* | Minutes to this site/total minutes in this session |
| 41 | bookfut | Binary dependent variable, indicating if this user is going to book in the remainder of the session (after the clipping point) |

What decision would you make using different modelling techniques for a new transaction with Income = High and Credit=Bad?



Ran Ensemble method with gradient boosting, support vector machine (SVM), logistic regression model; and decision with bagging and found that Ensemble method is the champion model. Compared the results with respect to testing dataset; SAS programming on Score dataset for further results interpretation.



Prediction accuracy is **73.43%** at cut-off =0.5 as interpreted from score dataset.

