KDD Physics Task – Discussion of Modeling Approaches

Lalit Wangikar
Vice President
Inductis (India) Private Ltd
Gurgaon, India
lwangikar@inductis.com

Vineet Agrwal
Modeler
Inductis (India) Private Ltd
Gurgaon, India
vagrwal@inductis.com

Vivek Gupta
Senior Modeler
Inductis (India) Private Ltd
Gurgaon, India
vgupta@inductis.com

ABSTRACT

In this paper, we present the methodology followed by Inductis in developing the predictive models for Quantum Physics task in KDD Cup 2004. We discuss many challenges that we faced in approaching the task and how we overcame them. We explored the solution space with various classification approaches and finally used stochastic gradient boosting offered on the TreeNet platform. A set of TreeNet models was fit varying various parameters and its performance was measured.

Keywords

Stochastic Gradient Boosting, Logistic Regression, MARS, KDD Cup, TreeNet

1. INTRODUCTION

Inductis decided to attempt this problem as a part of its efforts to keep itself updated on the frontiers of predictive modeling and to take learnings from here to its business consulting space. This document presents the various approaches and analysis that we used for solving the KDD cup problem. Section 2 describes the challenges, section 3 discusses the pre-processing, section 4 presents the modeling summary including MARS, logistic regression with MARS and TreeNet models.

2. LACKING CONTEXT

Inductis primarily works by merging analytics with business consulting practice where problem context and managerial problem definition are fundamental to a successful solution. Defining the right problem is as important as finding the best analytical model. Our approach is guided by distinctive framework of Decideright which guides to define correct problem and search for the best answer.

In KDD Cup 2004, the problem's contextual knowledge had no significant bearing on the outcome. Hence the focus was on finding as good a predictive model as possible. We focused extensively on data understanding for finding the right meaning of variables. We utilized a variety of options from our traditional methodological framework for exploratory data analysis highlighting the characteristics of the problem space. We then developed a series of classification models and finally used stochastic gradient boosting based on its ability to extract information from a large number of variables.

3. EXPLORATORY DATA ANALYSIS

Our exploratory data analysis using Inductis proprietary tools included univariate and bivariate analyses of variables. We looked at the distributions of various independent variables and plotted their relationships with dependent variable. A series of bivariate plots between dependent variable and various covariates concluded that only 10 out of 78 variables have any observable relationship with the dependent variable. We also identified a set of multivariate patterns which had higher predictive relationships with dependent variables. Based on the first stage we imputed missing values with mean for various variables, understood multivariate outliers and capped the same. We classified the variables as categorical and continuous based on number of levels. We also created dummy indicator variables for outliers, missing values and important interactions effects.

4. Modeling Efforts

4.1 Sampling Plan

We took a 2/3 training and 1/3 validation sample on the training data and used it for modeling. Within 2/3 training set we used 20% sample for testing models in CART/ MARS and TreeNet.

4.2 MARS Results

We utilized MARS in two different ways: (1) To create variables approximating functional form specifications for use in logistic regression models. We identified various basis functions representing functional form approximations and interactions among variables. (2) To model the problem on MARS with principal components of variables. For developing the model using MARS we did a variable reduction exercise for overcoming multi-collinearity problems in the data. We varied a series of parameters in MARS for controlling complexity of resultant model. We however found that the MARS results were not superior to TreeNet preliminary results on this problem. A typical set of results from the exercise are presented in table 1.

Table 1 MARS Results

Prediction 1 **MARS**		
Parameters	Model 1	Model 2
Training Sample Size	30,000	35,000
Test Sample Size	20,000	15,000
Number of Basis Functions	100	150

Speed Of Search	3	3
Maximum Interactions	3	3
Observations Between Knots	50	50
Testing Degrees of Freedom	2	3
Accuracy	0.71302	0.71586

4.3 Logistic Regression Results

We approached the problem with a large set of logistic regression runs using a combination of imputations, approximating functional forms using MARS created variables etc. However we found that the results from this approach stuck at ceiling performance of around 71% on the holdout sample.

Table 2 Logistic Regression with MARS variables

Prediction 2 **Logistic + MARS Approach**	
Accuracy	71.07%
ROC	80.03%
SLQ	0.28
Cross Entropy	0.77

4.4 TreeNet Results

We decided to use stochastic gradient boosting algorithms. We used evaluation version of Salford Systems product TreeNet in this application. We made use of our learnings from exploratory data analysis and earlier modeling phase in choosing the parameter space as well as data treatment. We used logistic likelihood as the loss function and varied parameters like learning rate, number of trees, classification costs, nodes per tree and studied implications on holdout sample accuracy. Results from our best model are presented below.

Table 3 TreeNet Results

Prediction 3 – **TreeNet**	
Accuracy	74.95%
ROC	84.47%
SLQ	0.36
Cross Entropy	0.70

Compared to this our results on the KDD dataset dropped a bit but were still comparable to what we had been achieving.

5. ACKNOWLEDGEMENTS

We thank the team at Inductis comprising Arpita Chowdhary, Don Yan, Dinesh Bharule, Sandeep Tyagi, Titiksha Gautam who contributed in this effort.

About the authors:

Lalit Wangikar is the Vice President at Inductis (India) Private Ltd where he heads the India Consulting Operations. He has worked on various analytic consulting assignments in financial sector. In his recent engagement, he worked on predicting attrition with a large financial institution. He holds an MBA from Indian Institute of Management Ahmedabad.

Vineet Agrwal is a Modeler at Inductis (India) where he has been working on variety of modeling problems in financial industry. Vineet is B.Tech from Indian Institute of Technology Mumbai, India.

Vivek Gupta is Senior Modeler with Inductis (India) where he has worked on variety of modeling problems in Insurance and pharma. Vivek has worked on modeling problems in CPG, telecom sector in the past. His recent published work has been on choice modeling application in emerging markets. He holds a PhD from Indian Institute of Management Ahmedabad, India where his research was supported by Infosys Research Fellowship.