Building Predictive Models for Election Results in India – An Application of Classification Trees and Neural Networks[1]

Vishnuprasad Nagadevara, Indian Institute of Management Bangalore

Abstract

The 2002 Judgment of the Supreme Court of India paved the way for compulsory disclosure of information with respect to the background of candidates in elections. This information includes the assets and liabilities as well as criminal antecedents, if any. The general elections held in 2004 were the first set of elections after the implementation of Supreme Court ruling. Thus, a fairly large amount of data on the candidate' background had become available for the first time. This data was used to build predictive models for forecasting the results of the Legislative Assembly elections of the state of Karnataka. Two different data mining techniques namely, classification trees and artificial neural networks were used to build the predictive models. The prediction accuracy ranged between 90 and 98 percent.

Keywords: Predictive Models, Classification Trees, Artificial Neural Networks, Elections, Data Mining

1. INTRODUCTION

The Indian general elections of 2004 were unique for more than one reason. The National Democratic Alliance (NDA) government led by the Bharatiya Janata Party (BJP) ruling at the center was so positive about its prospects of reelection they had advanced the elections by few months. Similarly, the state government in Andhra Pradesh, ruled by the Telugu Desam party advanced the elections to the state assembly in order to cash in on the positive feeling of the electorate. So was the case with the State of Karnataka where the ruling party was the Indian National Congress. The emergence of Janata Dal (S), which was led by the earlier prime minister of India, Mr. Deve Gowda and the failure of the ruling Indian National Congress was rather unexpected. The final results of elections had surprised many analysts. Even the exit polls turned out to be so close, some of the pollsters decided not to draw any conclusions. In any general elections, such results are not uncommon. But the results of the Karnataka Legislative Assembly elections of 2004 for the 12th Assembly were a major surprise. Table 1 shows the radical shift in the party positions in the 2004 assembly elections, for the 12th Assembly of Karnataka.

TABLE 1. PARTY-WISE POSITIONS IN THE 11TH AND 12TH LEGISLATIVE ASSEMBLY OF KARNATAKA

Party	11th Assembly	12th Assembly
Indian National Congress	133	66
Bharatiya Janata Party	43	79
Janata Dal (S)	10	58
Janata Dal (U)	18	5
Independents	19	12
Others	1	4
Total	224	224

The general elections of 2004 are generally considered as a landmark election for other reasons. It was the first time that the entire country voted using electronic voting machines. The Election Commission used more than a million electronic voting machines and the votes cast by more than 400 million voters were announced in less than 8 hours (Election Watch, 2004).

[1] Journal of Academy of Business and Economics, Vol. V, No. 3, 2005, pp 1-10

The general elections of 2004 could be considered as unique for a very important reason. In 2002, the Supreme Court of India delivered a judgment following a public interest litigation in the High Court of Delhi by the Association for Democratic Reforms asking for disclosure of candidates' background at the time of filing the nomination forms. The purpose is to make sure that the voters have sufficient information about the candidates in order to enable them to make an informed choice while casting their votes. The Delhi High Court delivered its judgment upholding the petition in 2000, giving directions to the Election Commission of India to collect the information about the candidates using the police and other such agencies of the government and assess their suitability for holding a public office and give wide publicity to such information. Consequently, the Government of India filed a Special Leave Petition in the Supreme Court of India in 2001 against the judgment of the Delhi High Court. Interestingly, several political parties have become intervenors to the case, opposing the Delhi High Court Judgment. The judgment pronounced by the Supreme Court in 2002 directed the Election Commission of India to ask for the following information from the candidates by way of affidavit to be filed by the candidates along with the nomination form:

- Whether the candidate is convicted/acquitted/discharged of any criminal offence in the past, if any, whether he/she is punished with imprisonment or fine or both?
- Whether the candidate, six months prior to filing the nomination, is accused of any pending case of any offence punishable with imprisonment of two years or more and in which charge is framed or cognizance is taken by the court of law.
- The assets (movable, immovable, bank balances etc.) of not only the candidate but also of his/her spouse and the dependents.
- Liabilities, if any, particularly to any public financial institutions or government
- Educational qualifications of the candidate

When the Election Commission issued an order in June 2002 implementing the order of the Supreme Court, it created a flurry of activity among all the political parties. Twenty one political parties unanimously decided in an all party meeting that the order of the Election Commission could not be allowed to be implemented. An amendment to the Representation of People Act (which governs the electoral issues) was to be introduced in the Parliament in the Monsoon session of 2002. The bill retained the disclosure of pending cases but deleted the disclosure requirements of the assets and liabilities of the candidates. When the amendment could not be introduced in the parliament for various reasons, the government issued an ordinance which maintained that "no candidate shall be liable to disclose or furnish any information which is not required to be disclosed under the proposed bill, not withstanding anything contained in any judgment, decree or order of any court or any direction, order or any other instruction issued by the Election Commission". This ordinance led to a flurry of Writ Petitions in the Supreme Court. The Supreme Court delivered its judgment in March 2003, holding the amended act illegal, null and void. It restored the earlier judgment of 2002 and also declared that the judgment had attained finality.

Thus, the 2004 General Elections were the first elections to be held where the candidates were forced to make complete disclosures of their antecedents as well as their assets and liabilities. It was also decided to extend the provisions of the Supreme Court judgment not only to the parliament and state assembly elections, but also to local (village panchayat level) elections as well.

Thus, there was a lot of information about the candidates available to the voters, which could enable them to make an informed decision while casting their vote. It is also interesting to see if the information thus made available did make any difference in the outcome of the elections. Also, it is important to see if it would be possible to use the information to predict or forecast the results of the elections. Such forecast is very different from the forecasts based on "exit polls". Here, the attempt is not only to build predictive models in order to predict the possible outcomes, but also to identify various aspects of information, which could influence the final results of the elections.

Thus the objectives of this research paper are to

1. Develop predictive models which could be used for predicting the outcomes of the election
2. Identify various aspects of information made available consequent to the Supreme Court judgment and the subsequent orders of the Election Commission and
3. Evaluate the relative importance of these aspects of information in predicting the election outcomes.

2. METHODOLOGY

The elections for the lower house of the Parliament (Lok Sabha) and the legislative assembly of three states namely Andhra Pradesh, Karnataka and Orissa were held simultaneously in 2004. Subsequent to the order of the Election Commission, details of the antecedents of the candidates along with the assets and liabilities became available. The Association for Democratic Reforms had formed Election Watch Committees in the states to collate the information from the nomination papers and the appended affidavits. Data with respect to the candidates of the Karnataka State Legislative Assembly is obtained from the Election Watch Committee of Karnataka. In addition to the data collected from the Election Watch, other information available in the public domain is used to complete the information on each of the candidates. The election results of the candidates (win or loss) is used as the dependent variable for the predictive models. This is treated as a binary categorical variable. In addition, a number of variables on which information was available are used as independent variables. These variables included

- Age of the candidate (binned into 6 categories)
- Number of contestants in the specific constituency (binned into 4 categories)
- Movable assets (binned into 3 categories)
- Immovable assets (binned into 3 categories)
- Total Assets (binned into 3 categories)
- Liabilities (binned into 3 categories)
- Ownership of commercial buildings (binned into three categories including unknown)
- Ownership of residential buildings (binned into three categories including unknown)
- Whether the candidate belongs to the ruling party or not
- Revenue Division of the state (all the districts in the state are grouped into 4 revenue divisions)
- The areas to which the districts of Karnataka originally belonged (Karnataka state was formed by taking some of the districts of old Bombay and Madras Provinces, Princely States of Mysore and Nizam). This, along with the revenue divisions is expected to represent the demographic characteristics and the developmental differences across different districts of the state.
- Whether the constituency was reserved for the scheduled caste and scheduled tribe candidates
- Type of political party (binned into 6 categories including independents)
- Whether the candidate is an incumbent member of the legislative assembly
- Whether the candidate belongs to the incumbent party in the specific constituency
- Gender
- Educational level (binned into 6 categories)
- Whether the candidate had any criminal record
- Whether the candidates owns any agricultural land
- Whether the candidate has any liabilities to financial institutions
- Whether the candidate has any liabilities to government

Since all the variables are categorical in nature, usual predictive models that revolve around regression techniques could not be used for prediction of this specific case. On the other hand, other predictive models such as classification trees, neural networks, classification and regression trees etc. would be ideal for handling these types of variables. These techniques, which fall in the broad categorization of data mining techniques, are used for developing the predictive models.

3. CANDIDATE PROFILE

The general profile of the candidates is presented in Table 2. There are a total of 224 constituencies for which the elections were held in 2004. Of these 224 constituencies, data on all the candidates was available for 195 constituencies.

TABLE 2. PROFILE OF THE CANDIDATES – ASSEMBLY ELECTIONS OF KARNATAKA

Category	Frequency	Percent
Education		
Unknown	272	16.58
Primary School	43	2.62
High School	353	21.51
Pre-University	231	14.08
Graduate	447	27.24
Post Graduate	294	17.92
Gender		
Female	86	5.24
Male	1555	94.76
Belongs to Incumbent Party In the constituency		
Does not belong	1469	89.52
Belongs	172	10.48
Incumbent		
Not incumbent	1490	90.80
Incumbent Candidate	151	9.20
Type of Party		
Unknown	387	23.58
BJP	177	10.79
Congress	196	11.94
Other National Party	210	12.80
JD (S)	193	11.76
Other Regional Party	295	17.98
Independent	183	11.15
Ownership of Houses		
Does Not Own	592	36.08
Owns	590	35.95
Unknown	459	27.97
Ownership of Commercial Buildings		
Does Not Own	600	36.56
Owns	573	34.92
Unknown	468	28.52
Belongs to Ruling Party		
Does not Belong	1445	88.06
Belongs	196	11.94
Age		
Less than 30	110	6.70
30 to 40	428	26.08
40 to 50	507	30.90

Category	Frequency	Percent
50 to 60	405	24.68
60 to 70	146	8.90
More than 70	22	1.34
Has Government Dues		
Unknown	437	26.63
No Dues to Government	1140	69.47
Owes dues to Government	64	3.90
Has Dues to Financial Institutions		
Unknown	526	32.05
No Dues to FIs	917	55.88
Owes Dues to FIs	198	12.07
Has Dues to Banks		
Unknown	396	24.13
No Dues to Banks	532	32.42
Owes Dues to Banks	713	43.45
Owns Agricultural Land		
Unknown	379	23.10
No Agricultural Land	368	22.43
Owns Agricultural Land	894	54.48
Criminal Record		
Does not have Criminal Record	1455	88.67
Has Criminal Record	186	11.33
No. of Candidates in the Constituency		
<=6	427	26.02
7 to 9	457	27.85
10 to 12	344	20.96
> 12	413	25.17
Reserved Constituency		
Does not belong	1391	84.77
Belongs	250	15.23

About 70 percent of the candidates have studied beyond high school level. Forty five percent of them are either graduates or post graduates. Only 86 of the candidates are female representing the domination of male members in the elections. The candidates appear to be younger with about 57 percent belonging to the age group of 30 to 50 years. Less than two percent of the candidates are more than 70 years old.

More than three-fourths of the members of the previous assembly (11th Karnataka Assembly) are again contesting in the 2004 elections. 151 out of the 195 constituencies have candidates who are incumbents. Similarly, 88 percent of the assembly constituencies had candidates from the incumbent party. All the national as well as regional parties had more or less equal distribution in terms of number of candidates contesting the elections.

More than 11 percent of the candidates have declared to have a criminal record. Very few candidates declared to have any dues to the government or financial institutions, where as those who have dues to banks accounted for 43 percent. About one-third of the candidates have declared ownership of commercial buildings and similar number declared ownership of residential buildings. About 55 percent of them own agricultural land. In general, the candidates are predominantly male with better educational qualifications and younger in age. The ownership of assets is mainly in agricultural land, residential and commercial buildings.

4. RESULTS

Two most commonly used classification techniques are classification trees and artificial neural networks. A brief description of the two techniques is given below.

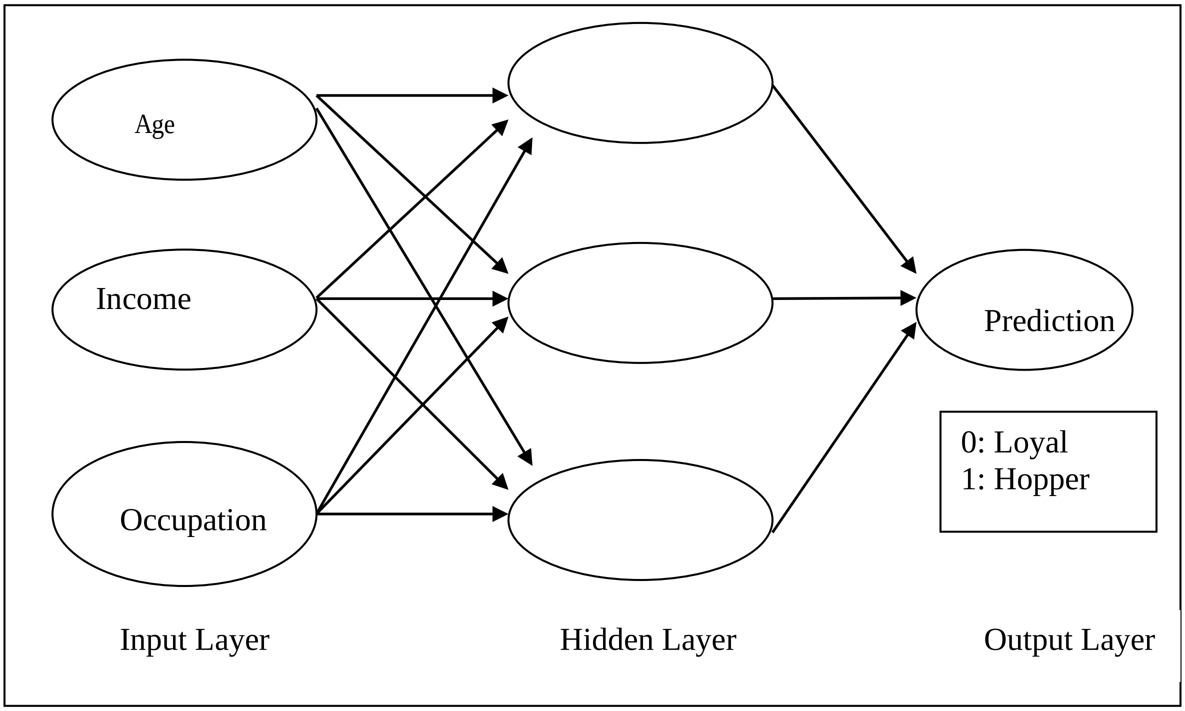
4.1 Classification Trees

A classification tree is a predictive model, which takes the form of a tree. Each branch of the tree is a classification question, and the leaves of the tree are partitions of the data set. The tree divides the data on each branch without losing any of the data. The technique picks predictors (independent variables) and the appropriate values for branching on the basis of the gain in information that the branching provides. The information gain can be defined as the difference between the amount of information that is needed to correctly predict the outcome before and after the split (branching) has been made. This difference is measured by the extent of entropy or Gini Coefficient or simple Chi-square analysis. The classification trees provide rules for prediction that are easy to understand and implement and hence are they used very frequently for building predictive models (Nagadevara and Tara, 2004).

4.2 Artificial Neural Networks (ANN)

The artificial neural networks (ANN) are generally based on the concepts of the human (or biological) neural network consisting of neurons, which are interconnected by the processing elements. The ANNs are composed of two main structures namely the nodes and the links. The nodes correspond to the neurons and the links correspond to the links between neurons. The ANN accepts the values of inputs into what are called input nodes. This set of nodes is also referred to as the input layer, as shown in Figure 1.

FIGURE 1. ARTIFICIAL NEURAL NETWORK



These input values are then multiplied by a set of numbers (also called as weights) that are stored in the links. These values, after multiplication, are added together to become inputs to the set of nodes that are to the right of the input nodes. This layer of nodes is usually referred to as the hidden layer. Many ANNs contain multiple hidden layers, each feeding into the next layer. Finally, the values from last hidden layer are fed into an output node, where a special mapping or thresholding function is applied and the resulting number is mapped to the prediction. The ANN is created by presenting the network with inputs from many records whose outcome is already known. For example, the data on age, income and occupation of the first customer (first record) are inputted into the input layer. These values are fed into the hidden layer and after processing (by combining these values using appropriate weights) the prediction is made at the output layer. If the prediction made by the ANN matches with the actual known status of the customer (either Loyal or Hopper), then the prediction is good and the ANN proceeds to the next record. If the prediction is wrong, then the extent of error (expressed in numerical values) is apportioned back into the links and the hidden nodes. In other words, the values of the weights at each link are modified based on the extent of error in prediction. This process is referred to as the backward propagation. The artificial neural networks are found to be effective in detecting unknown relationships. ANNs have been applied in many service industries such as health (to identify the length of stay and hospital expenses) (Nagadevara, 2004), hospitality (Nagadevara, 2005) air lines (Chatfield, 1998) etc.

4.3 Skewed distributions

The above two techniques, namely classification trees and artificial neural networks were applied to predict the results of the Karnataka Legislative Assembly elections of 2004. The entire data with respect to the 1641 candidates was used to train the models as well as for testing the effectiveness of the prediction. Both the techniques resulted in interesting predictions. The predictions of these two models are given in Table 3.

TABLE 3. PREDICTIONS BASED ON THE TWO MODELS BEFORE ELIMINATING THE IMBALANCE

		Prediction					
		Neural Network			Classification Tree		
		Lose	Win	Total	Lose	Win	Total
Actual	Lose	1427	19	1446	1422	24	1446
		98.69%	1.31%		98.34%	1.66%	
	Win	44	151	195	153	42	195
		22.56%	77.44%		78.46%	21.54%	
	Total	1471	170	1641	1575	66	1641
Error		3.84%			10.79%		

The overall misclassification with respect to the neural network was only 3.84 percent. At the same time, the misclassification for the classification tree was about 11 percent. Both the models are very effective in predicting the "loser" category (the misclassification is less than two percent). On the other hand, both the models are rather ineffective in predicting the "winner" category. For this category, the accuracy level of the classification tree was only 21.54 percent where as that of the neural network is 77.44 percent. In neither of the cases, the accuracy of this category is nowhere near that of the "loser" category. This type of problem in training the model, with both the classification tree as well as the neural network is not uncommon with skewed data sets. The data set consisted of 195 constituencies and consequently the total number of winners is only 195. On the other hand the total number of candidates was 1641. The proportion of winners among the total candidates was less than 12 percent. Thus, if the model predicts that "every one is a loser", the model would be misclassifying a maximum of 12 percent of the cases, resulting in an accuracy level of 88 percent! The behavior of the model would be such that it tends to predict more cases as losers since the data set is heavily skewed in favor of losers. In such cases standard classifiers tend to be overwhelmed by the large class and ignore the small or minority class (Chawla, 2002, Chawla 2003). This problem of skewed data sets with "minority classes" can be handled with different approaches. At the algorithmic level, solutions include adjusting inflating the costs to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf, in the case of classification trees etc. At the data level, the solutions include random over sampling and under sampling as well as directed over sampling.

The basic idea of over sampling or under sampling is to eliminate or minimize the imbalance or rarity by altering the distributions of training examples (Weiss, 2004). Typically the class distribution is altered to reduce the problems associated with rare classes. Under sampling eliminates majority class examples. This is achieved by taking a smaller sample of the large class and combining it with the entire training set of the smaller class. In order to avoid any sampling bias, a number of such random samples are drawn from the large class and the models are trained using these different samples. Ultimately, the classification rules obtained from different samples are combined or bagged to evolve a single set of classification rules. The over sampling duplicates minority class examples. Over sampling can lead to over-fitting because it involves more numbers of exact copies (Chawla, 2003 and Drummond and Holte, 2003). Over sampling does not actually make new data available and consequently, could become ineffective in improving the predictability of the minority class (Drummond and Holte, 2003). At the same time the over sampling had given better classification in other studies such as (Japkowicz and Stephen, 2002).

In the case of election data of Karnataka, it is felt that under sampling will not be appropriate because the elimination of some of the losing candidates from the database will result in loss of important information required for classification. Consequently, it was decided to use over sampling techniques by replicating the minority class there by increasing the total number of records to 2226. The minority class in the over-sampled data set constituted about 35 percent of the total there by avoiding the pitfalls involved in the classifying skewed data sets.

4.4 Final Results

Two techniques of classification, namely classification trees and neural networks are used for classifying and predicting the winners and losers of the election using the data set with over sampling. The misclassification matrix for the two techniques is presented in Table 4.

TABLE 4. PREDICTIONS BASED ON THE TWO MODELS AFTER ELIMINATING THE IMBALANCE

		Prediction					
		Neural Network			Classification Tree		
		Lose	Win	Total	Lose	Win	Total
Actual	Lose	1425	21	1446	1257	189	1446
		98.55%	1.45%		86.93%	13.07%	
	Win	24	756	780	56	724	780
		3.08%	96.92%		7.18%	92.82%	
	Total	1449	777	2226	1313	913	2226
Error		2.02%			11.01%		

The overall misclassification for the neural network is about 2.02 percent as compared to 11.01 percent for the classification tree. There was no significant improvement in the overall misclassification after over-sampling the data set. At the same time, there is a significant increase in the accuracy levels of the predictions with respect to the winners. The accuracy levels have improved to 92.82 percent in the case of classification tree and to 96.92 percent in the case of neural network. This improvement in the accuracy levels is significant when compared to the results obtained earlier without over sampling (21.54 percent for the classification tree and 77.44 percent for the neural network). The improvement in the case of the classification tree is through a trade off in the accuracy levels of the predictions with respect to the losers. The prediction accuracy of this category had come down to 86.93 percent as compared to 98.34 percent earlier. The classification tree is presented in Figure 2. It can be seen that the variable “Party

Type" is at the top of the tree indicating that it is the first variable over which the branching is carried out. The criminal record and the movable assets appear in the next level.

The neural networks do not provide similar levels to the variables. Nevertheless, the neural networks do attach a numerical value to the variables indicating how sensitive the prediction is with respect to a change in the value of these variables. Table 5 presents the sensitivity levels of these variables grouped into 5 categories namely, demographic characteristics, ownership details, extent of liabilities, political factors and others.

FIGURE 2. CLASSIFICATION TREE

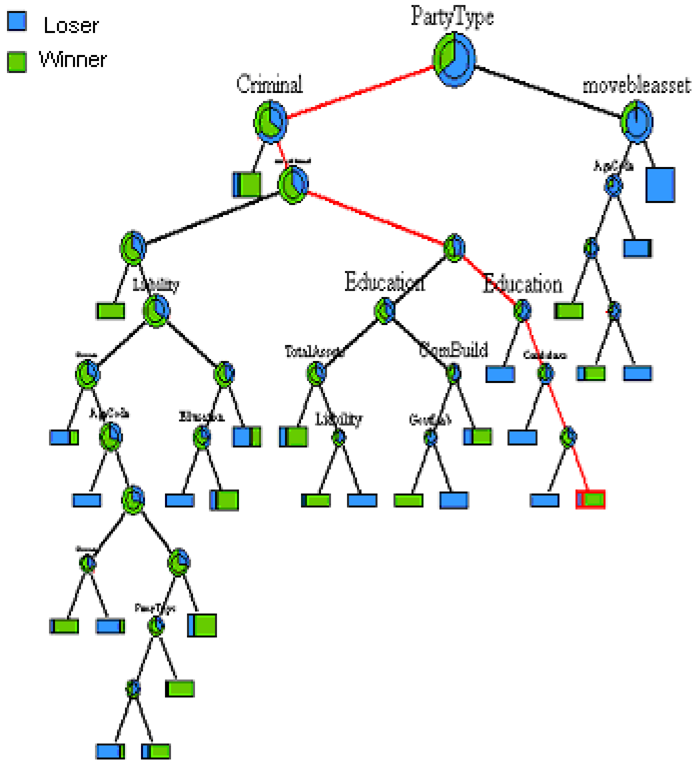


TABLE 5. ARTIFICIAL NEURAL NETWORK – SENSITIVITY VALUES

Demographic Characteristics		Ownership		Liabilities		Political Factors		Others	
Variable	Sensitivity	Variable	Sensitivity	Variable	Sensitivity	Variable	Sensitivity	Variable	Sensitivity
Age	7.4	Agricultural Land	4.1	Bank Loan	3.9	No. of Candidates	4.7	Criminal Record	2.4
Education	7.2	Commercial Buildings	4.1	Loans from FIs	4.3	Incumbent Party	3	Original Divisions	3.8
Gender	1.3	Residential property	2.9	Liabilities to Govt.	5.1	Incumbent	1.7	Reserved	1.2
		Immovable Assets	2.5	Total Liabilities	5.3	Party Type	19	Revenue Divisions	4.9

	Movable Assets	3.6		Ruling Party	1.9	
	Total Assets	4.7				

Among the demographic characteristics, age and education appear to be predominant in predicting the outcomes. It may be recalled that most of the candidates in this election belong to an younger group with fairly good educational background. From the assets and liabilities side, the total assets and total liabilities appear to be important, followed by the dues to the government and ownership of agricultural land and commercial buildings. Among the political factors, the results are most sensitive to the Party Type. Incumbency does not appear to be of high importance while the number of candidates in the fray appears to be more important. Finally, the results are not very sensitive to the criminal record of the candidate. The revenue divisions of the state, which are incidentally also indicators of the level of development within the state is more important in the results. The reason for criminal record not being important could be that only 11 percent of the candidates have criminal record. In the southern state of Karnataka, criminal antecedents of the candidates is not as predominant an issue as in some of the other states. Data with respect to other state assemblies is not readily available for comparison purposes. Nevertheless, an analysis of the candidates from different states for the 2004 Lok Sabha elections reveals that only 9.8 percent of the candidates from Karnataka were known to have criminal antecedents, where as the corresponding percentages were 23.3 percent in West Bengal, 20.1 percent in Bihar and 19.6 in Uttar Pradesh (Election Watch).

5. CONCLUSIONS

The Supreme Court judgment of 2002 with respect to the disclosures of the background of candidates for elections in India resulted in providing voters with sufficient information. While this information was primarily meant to enable the voters to make a well-informed choice, the availability of such information made it possible to build effective predictive models for forecasting the election results. Two techniques namely classification trees and artificial neural networks were used to build the predictive models for the Karnataka Assembly elections. Over-sampling technique was used to eliminate the predictive biases introduced by the skewness of the data set. The overall accuracy of the predictive models varied from 90 to 98 percent. The important variables in predicting the election outcomes were age, education, ownership of assets, liabilities, type of (political) party, as well as the number of candidates in the fray. The extent of economic development as indicated by the revenue divisions of the state was also found to be an important predictor. The real test of the predictive model would be to apply the model to the data on candidates in the next general elections and validate the results.

6. REFERENCES

Chatfield, C. (1998) ‘Time Series Forecasting with Neural Networks: A Comparative Study using the Airlines Data’, *Applied Statistics,* 47, Part 2, pp. 231-250

Chawla N, Bower K, Hall L, Kegelmeyer W. “SMOTE: Synthetic Minority Over-sampling Technique”**,** *Journal of Artificial Intelligence Research,* Morgan Kaufman Publishers, pp321-357, 2002

Chawla N. “C4.5 and Imbalanced Data Sets: Investigating the Effects of Sampling Method, Probability Estimate, and Decision Tree Structure”, *in Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC, USA, 2003*

Drummond C and Holte. R. C. “C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling”, *In Workshop on Learning from Imbalanced Data Sets II, International Conference on Machine Learning, Washington DC, USA, 2003.*

Election Watch, Association for Democratic Reforms, Ahmedabad, India, 2004

Japkowicz N., and Stephen S. “The class imbalance problem: a systematic study”. *Intelligent Data Analysis*, 6(5): 429-450, 2002.

Nagadevara, V and Tara S. N., “Improving the Effectiveness of Post Literacy Programme Through Data Mining Techniques”, Towards E-Government, Management Challenges, Ed. MP Gupta, Tata McGraw Hill Publishing Company, New Delhi, 2004

Nagadevara, V “Application of Neural Prediction Models in Healthcare”, Proceedings of the 2nd International Conference on e-Governance, Nov 29 – Dec 1, 2004, Colombo, Sri Lanka, pp 139-148.

Nagadevara V., “Improving the Effectiveness of Hotel Loyalty Programmes through Data Mining”, Proceedings of the “International Conference on Services Management, Mar 11-12, 2005, New Delhi, India

Weiss G. M. **“**Mining with rarity: a unifying framework”, ACM SIGKDD Explorations Newsletter Special issue on learning from imbalanced datasets Volume 6, Issue 1 June 2004

Author Profile:

Dr. Vishnuprasad Nagadevara obtained his Ph D from Iowa State University, Ames Iowa. He is currently Professor in the Quantitative Methods and Information Systems Area at the Indian Institute of Management Bangalore. His current research interests are Data Mining, Application of Management Techniques to Education and Entrepreneurship.