Predicting Rural Migration in India using socio-demographic information

Varun, Akshad, Imran, Anjali, Richa

This document details out BIDM project done by our group on topic “Predicting Rural Migration in India using socio-demographic information”

Contents

Executive summary 2
Problem Statement 3
Data behind predictive models 3
Findings from the application of data mining models 4
Models 4
Findings 5
Results 5
Conclusion 6
Exhibits 7
Exhibit 1: Naïve Bayes Model 7
Exhibit 2: K-Nearest Neigbhors 8
Exhibit 3: CART 9
Exhibit 4: Tree Structure 10
Exhibit 5: Relative Column Importance 11
Exhibit 6: Visuals of few of the predictor variables that are eliminated based on visuals 11

Executive summary

India as a nation has seen a high migration rate in recent years. Over 98 million people migrated from one place to another in 1990s, the highest for any decade since independence according to the 2001 census details. However in 1970s migration was slowing down. The number of migrants during 1991-2001 increased by about 22% over the previous decade an increase since 1951. Apart from women migrating due to marriage, employment is the biggest reason for migration. The number of job seekers among all migrants has increased by 45% over the previous decade. Nearly 14 million people migrated from their place of birth in search of jobs. The overwhelming majority of these-12 million was men.

Migrants have created pressure on others who are in same job market. While freedom to migrate within the country is an enshrined right the uneven development, levels of desperation and other factors have created friction points. Most people migrate because of a combination of push and pull factors. Lack of rural employment, fragmentation of land holdings and declining public investment in agriculture create a crisis for rural Indians. Urban areas and some rural areas with industrial development or high agricultural production offer better prospects for jobs or self-employment.

The above issue led to our study, through which we wanted to understand if migration could be predicted depending on parameters associated with an individual. And if it can be predicted beforehand then it will provide the government and regulatory authorities with a very strong tool. The tool could then be used to identify the reasons of migration and take pro-active measures to ensure that rural Indians do not migrate. It will also enable state and central governments to provide region-wise incentives to rural Indians to check migration and hence prevent excessive pressure on urban cities.

1951-61	66
1961-71	68.2
1971-81	81
1981-91	80.9
1991-2001	98.3

Problem Statement

The core idea within our study is to identify the key drivers of migration in rural India. Through this project we hope given certain base information about an individual from rural India

1) The probability that he will migrate
2) The directionality of migration i.e. whether he will move to a rural or a urban locale

We thus hope to create a mechanism that will help address key questions for the following individuals/institutions

1) Planning agencies who want to forecast demand
2) The government for dis-incentivizing migration through schemes
3) Economists seeking to build on migration patterns
4) Corporations to help gauge impact on inventory levels and capacity in stores
5) The real-estate sector to gauge demand

Data behind predictive models

We use survey data commissioned by a government agency that was collected by administering this survey to over 200,000 respondents located all over India. The survey was administered manually through a 20 page document and the data was collated by agents. The data that was collated contained over 150 fields and had multiple pieces of information including demographic information, geographic details of location, certain direct question to gauge causality where the respondents were asked why they chose to migrate among others.

The data also contained records from both urban and rural India as it sampled the migration statistics of the country as a whole. The first task presented to us was therefore to whittle out that portion of the data which was of interest to us. We did this by identifying those records which had individuals from rural India. We then introduced a new variable in the system that could take on 3 values and indicated whether a person had migrated from rural India and if so whether he had moved to another rural location or to an urban zone.

The next task was to identify which of the 150 possible factors might be used in predicting the probability of migration. We therefore used the same variable decision and see how the remaining variables assume value for each of the values the decision variable took. We judged whether a variable brings value to our system by seeing whether the three buckets had different

"fingerprints" or distributions over the three categories[6]. We judged these fingerprints using histograms and saw if the data showed different patterns visually for each category and took factors having distinctly different patterns as predictors.

We thus arrived at 19 factors which helped define the model. Additionally we also dropped factors that gave duplicate redundancies. We thus ensured 19 highly distinct and diverse data points to predict migration. These included region, household type, household size, social group, land possession, age, education, self employment status, whether the person was self employed in the agriculture sector, whether the person was a regular wage employee in agriculture or a casual laborer, whether the person was a student or not and also self stated reasons for migration like employment, marriage and also geographic information like the region from where the person came etc.

Findings from the application of data mining models

Models

Based on the data and our objective we chose three classification models to predict the rural migration:

1) Naïve Bayes – A probabilistic classifier that can be used to predict the behavior of a dependent variable. Even though it makes simplistic assumptions, it is a very good tool to give decision under uncertainty.
2) *K*-Nearest Neighbors – The idea in this method is to identify *k* observations in the dataset that are similar to a new record that we wish to classify. We then use these similar records to classify the new record into a class, assigning the new record to the predominant class among there neighbors.

   Outcome: Final K value came out to be 11
3) CART (Classification and Regression Trees) – CART helps us predict the membership of a particular person in the classes – the person will migrate or the person will not migrate, based on one or more predictor variables. The advantage that CART has over the above two methods is that It also helps us determine the relative importance of the predictor variables. And this plays a very important role in understanding the impact of various variables.

   Outcome: Exhibit 4 shows the final tree structure.

Findings

The findings can be divided into two parts: First, we need to identify the importance of each of the above three methods and also their respective performances. We will then compare the performances of these methods to identify which one will be able to predict in the most accurate way. Second, we would like to identify which are the most important variables that drive rural migration.

Individual Performance: We used confusion matrix to measure performance of each of the three methods. The matrix compares the predicted values of the response variable with the actual values, for the given training and validation datasets. We observe that our first method, Naïve Bayes[1] predicts the response variable with 87% accuracy in case of training set and 86.4% in case of validation set. Whereas the second method, *K*-Nearest Neighbors[2] predicts with an accuracy of 90.85% in case of the training data and 89.22% in case of the validation set. And the third method, CART[3] predicts with an accuracy of 96% for the training set and with 89.8% for the validation set. We observe that the accuracy with which the classification is done is pretty high in all the three methods hence observe a strong dependence of the variables on migration.

Tree Structure: Tree was not pruned and hence final tree is very big. We here show a relatively condensed form of the tree which shows some initiation tree paths taken. The table in Exhibit 5 shows the relative importance of predictors using the classification tree. Most important predictors is "reas_marriage" followed by other_notlab, reas_emp, student, hhtype, state_region, mpce, hhsize, self_emp, causal_lab etc.

Results

On looking at the results generated by all the three methods together, we observe that CART gives the best result followed by the *K*-NN method. However, there is not

[1] Refer Exhibit 1
[2] Refer Exhibit 2
[3] Refer Exhibit 3
[4] Refer Exhibit 5
[5] Refer Exhibit 4
[6] Refer Exhibit 6

considerable different amongst the accuracies of the three methods and hence all the three are significant.

Metric	Naïve Bayes	*K*-NN	CART
% agreement	86.40%	89.22%	89.80%
Def Ratio 0	92.90%	95.15%	96.30%
Def Ratio 1	77.00%	85.92%	82.50%
Def Ratio 2	78.80%	74.19%	79.70%

Conclusion

Following are the conclusions from our analysis of all the three prediction models:

- Both classification tree and K-Nearest Neighbors came out to be more effective than Naïve Bayes
- Classification tree fared better on Agreement and Definitude ratios than K-Nearest Neighbors in terms of agreement ratios
- Final Predictions: Given the 19 predictors which can be easily known for an individual our models can strongly predict the migration and it's direction

Exhibits

Exhibit 1: Naïve Bayes Model

Training Data – Confusion Matrix

Input Node - Naive Bayes (3)

		Predicted			Totals
		2	1	0	
Observed	2	2838	624	130	3592
	1	817	3947	300	5064
	0	522	248	10941	11711
Totals		4177	4819	11371	20367

	Observed			Overall
	2	1	0	
% Agree	79.0%	77.9%	93.4%	87.0%

Positive Category - 2

Recall	Precision	F-Measure
79.0%	67.9%	73.1%

Validation Data – Confusion Matrix

Input Node - Predict (7)

		Predicted			Totals
		2	0	1	
Observed	2	1661	73	375	2109
	0	345	6526	151	7022
	1	526	186	2377	3089
Totals		2532	6785	2903	12220

	Observed			Overall
	2	0	1	
% Agree	78.8%	92.9%	77.0%	86.4%

Positive Category - 2

Recall	Precision	F-Measure
78.8%	65.6%	71.6%

Exhibit 2: K-Nearest Neigbhors

Validation error log for different K

Value of k	% Error Training	% Error Validation	
1	0.00	12.09	
2	5.70	13.73	
3	5.96	11.40	
4	7.34	11.46	
5	7.76	11.14	
6	8.18	11.35	
7	8.12	10.92	
8	8.50	11.02	
9	8.60	10.81	
10	9.12	10.91	
11	9.15	10.78	<--- Best k
12	9.44	11.11	
13	9.38	10.96	
14	9.66	10.93	
15	9.54	10.82	

Scoring – Summary report (K=11)

Training set

Classification Confusion Matrix

	Predicted Class		
Actual Class	0	1	2
0	5484	142	83
1	33	2224	276
2	28	353	1377

Error Report

Class	# Cases	# Errors	% Error
0	5709	225	3.94
1	2533	309	12.20
2	1758	381	21.67
Overall	10000	915	9.15

Validation set

Classification Confusion Matrix

	Predicted Class		
Actual Class	0	1	2
0	10145	327	190
1	92	3936	553
2	49	776	2372

Error Report

Class	# Cases	# Errors	% Error
0	10662	517	4.85
1	4581	645	14.08
2	3197	825	25.81
Overall	18440	1987	10.78

Exhibit 3: CART

Training Data – Confusion Matrix

Input Node - Classification Tree (2)

		Predicted			Totals
		2	*1*	*0*	
Observed	*2*	3270	217	39	3526
	1	197	4783	84	5064
	0	104	166	11507	11777
Totals		3571	5166	11630	20367

	Observed			Overall
	2	*1*	*0*	
% Agree	*92.7%*	*94.5%*	*97.7%*	*96.0%*

Positive Category - 2

Recall	Precision	F-Measure
92.7%	*91.6%*	*92.2%*

Validation Data – Confusion Matrix

Input Node - Predict: Classification Tree (3)

		Predicted			Totals
		2	*1*	*0*	
Observed	*2*	1763	365	85	2213
	1	405	2557	137	3099
	0	93	160	6655	6908
Totals		2261	3082	6877	12220

	Observed			Overall
	2	*1*	*0*	
% Agree	*79.7%*	*82.5%*	*96.3%*	*89.8%*

Positive Category - 2

Recall	Precision	F-Measure
79.7%	*78.0%*	*78.8%*

Exhibit 4: Tree Structure

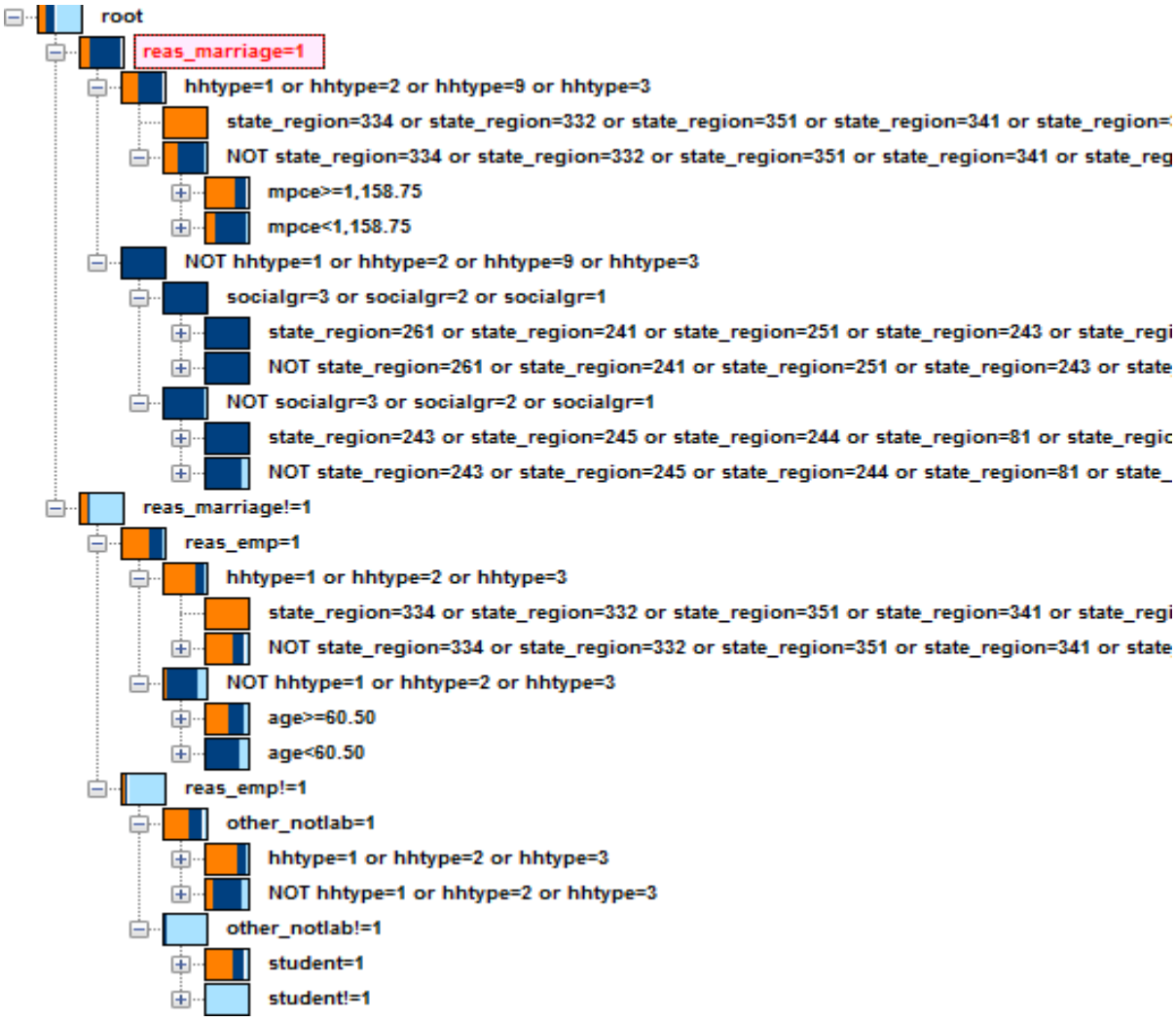

Exhibit 5: Relative Column Importance

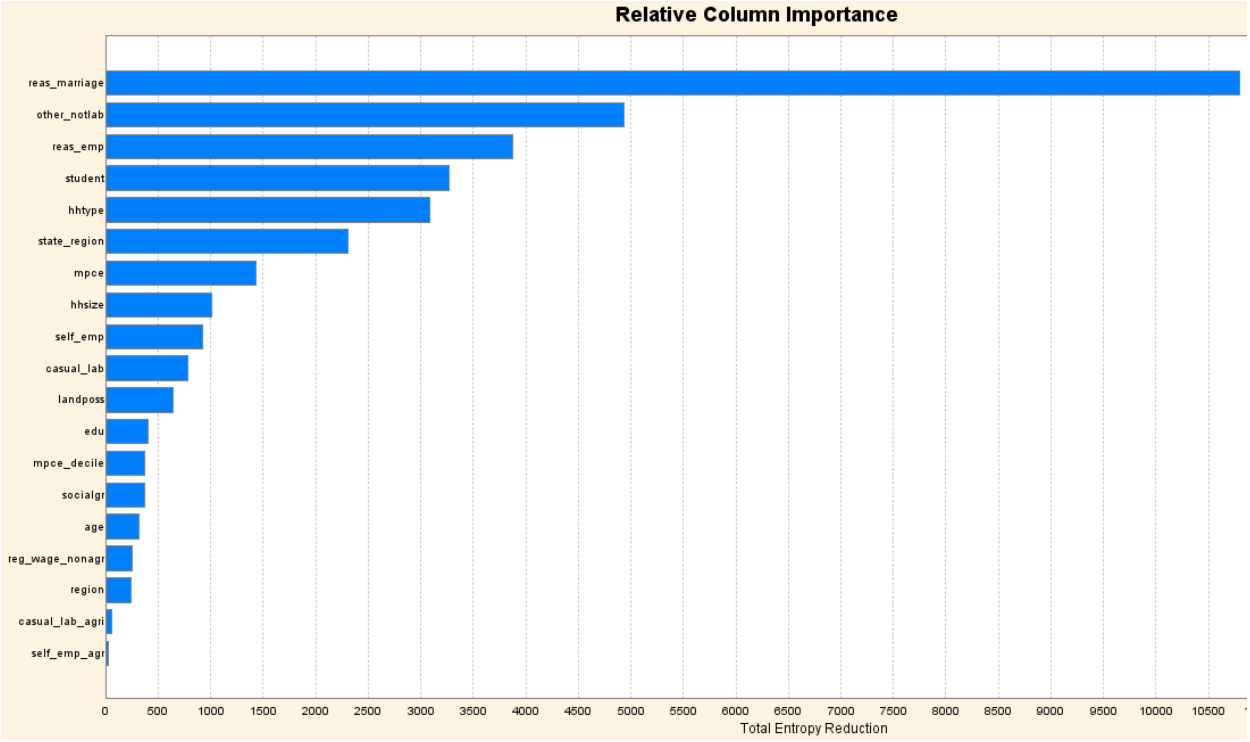

Exhibit 6: Visuals of few of the predictor variables that are eliminated based on visuals