Health Outcomes in China for the Hukou Migrants: How Algorithms May Inform Public Policymakers

Marta Bengoa and Thierry Warin [Colin Powell School. City University of New York. University of Johannesburg and CIRANO; SKEMA Business School and CIRANO]

AEA 2020 - San Diego

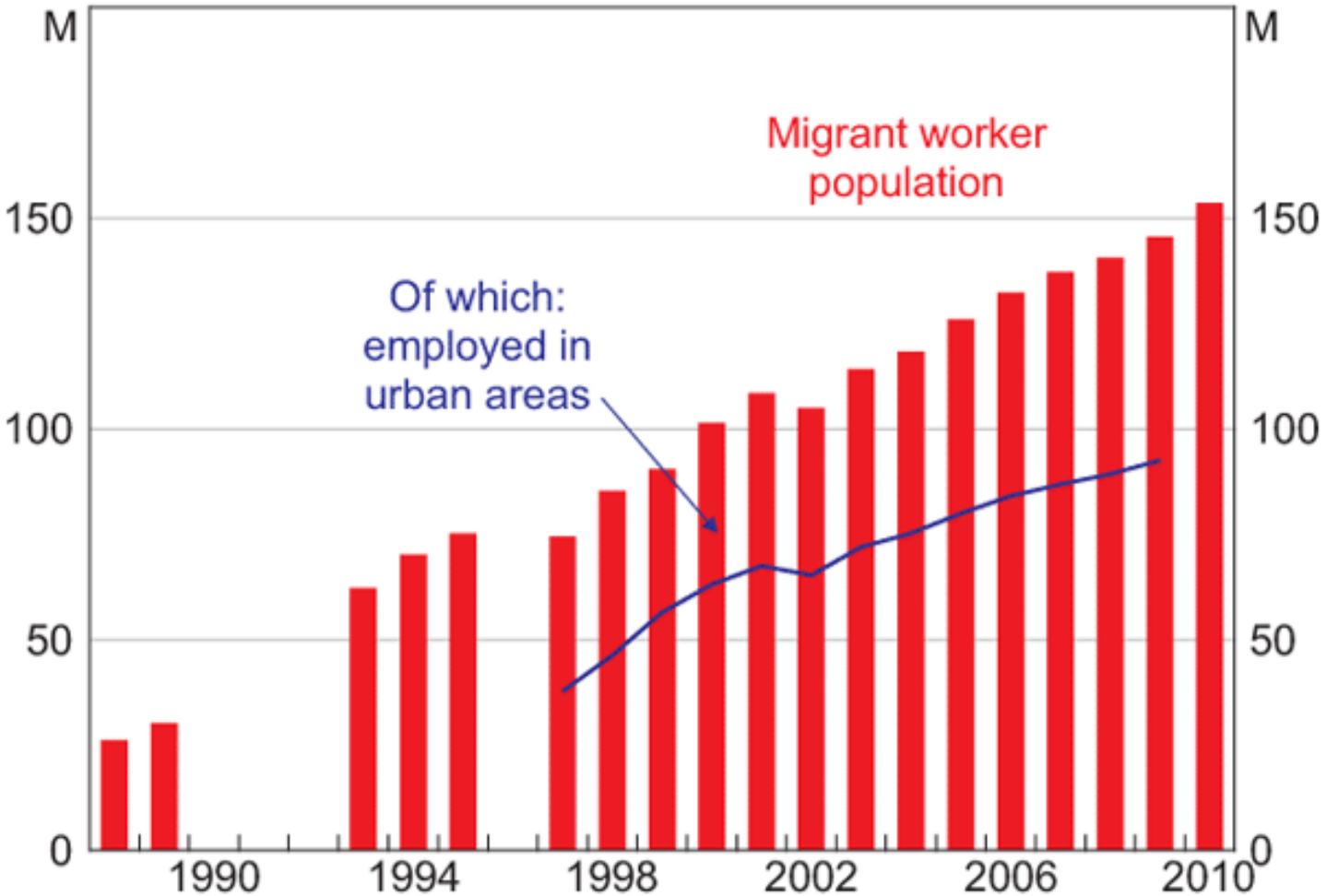
Section 1

Introduction

Introduction

- China's rapid development have spurred large migration from rural areas to urban areas
- Between 1990 and the end of 2015 the proportion of China's population living in urban areas jumped from 26% to 56%
- Currently estimated by census, there are more than 240 million rural-to-urban migrants and more than 160 million working in cities outside of their hukou. That accounts for approx. 30% of total rural labor force (China National Bureau of Statistics).
- The Hukou household registration system imposes restrictions and limits to where to live, which is determined mainly by birth.
- Hukou card is an internal passport that sets access to education and health services. It started in 1956-58, relaxed during the 60s and enforced again since 1978.

Introduction



* The red colums refer to all rural migrants without local 'hukou' where they are living; the blue line by Herd, Koen and Reuterswald (2010) calculates the level of rural–urban migrant employment, RBA estimate for 2009 number following their methodology

Sources: Chan (forthcoming) for 1988, 1989, 1993–1995 data; MOLSS

Introduction

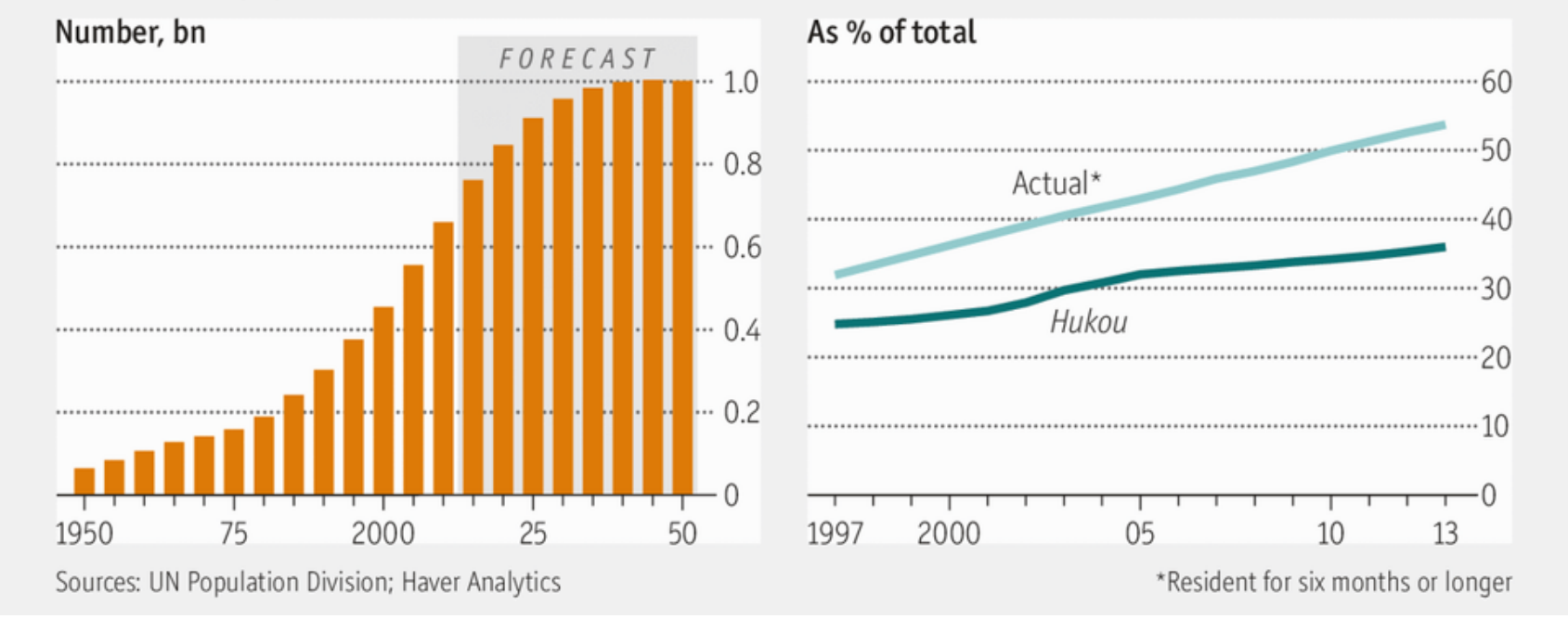


Figure 2: China's population

Introduction

- In 2009, China designed a healthcare reform that was intended to address inequalities in health access. Specific goals include:
  i) Provide wide basic health coverage for legal urban workers (mandatory) and to allow voluntary enrolment for urban residents without jobs (including students, elderly or disabled);
  ii) Voluntary cooperative medical system for rural citizens; and
  iii) Provide wide healthcare coverage to vulnerable and low-income population groups.
- The reform left the rural-to-urban migrant population outside of the provision of public healthcare.
- Several major cities have tried to introduce policy reforms to address the health challenges confronted by this phenomenon.
- Since 2006, Beijing is offering primary care health services and free healthcare to the children of migrant workers in community centres. There is also a pilot programs

Section 2

Literature Review

Literature Review

- There are not many studies that have addressed this link between migration with limited access to healthcare and health outcomes in developing economies. The most recent for China published by Sun (2015), uses self-reported outcomes (do I feel well or not, have I been sick..) which suffers from measurement errors.
- Other studies suggest that migrants are reasonably healthy at the point of migration but more likely to experience adverse effects than non-migrants. As they get injured and can't have access to health some return home while others remain in urban areas.
- Therefore, increases risk of workplace accidents, other contagious diseases (Chen, 2011; Lu and Quin, 2014; Wallace and Kulu, 2014).

Literature Review

- Studies mainly focus on the US, on the relationship of health of Mexican immigrants and lack of insurance. Goldman et al. (2014, Demography) find significant and strong correlation between non having insurance and decline in health of Mexican immigrants (compared to never migrants), specially immediately after migration.
- Wassink (2008, Demographic Research) analyses health insurance coverage for returned Mexican migrants to determine that lack of access is mainly driven by unemployment, revealing a negative association between lack of insurance and health

Section 3

Research Questions and Motivation

Research Questions and Motivation

- Motivation: Using Machine Learning techniques (+ ensemble methods) to study a very complex relationship in Social Sciences: health and migration
- Steps:
  - First, we use traditional econometric techniques for such a topic
  - Second, we use ML techniques to move from **correlations** (econometrics) to **predictive modelling** (ML)

Research Questions and Motivation

RQ1: This paper examines whether hukou has a predictive effect on migrants' self-reported health. We assess if there are observable differences in health outcomes migrant workers by hukou status. We use regression analysis, propensity score models and machine learning.

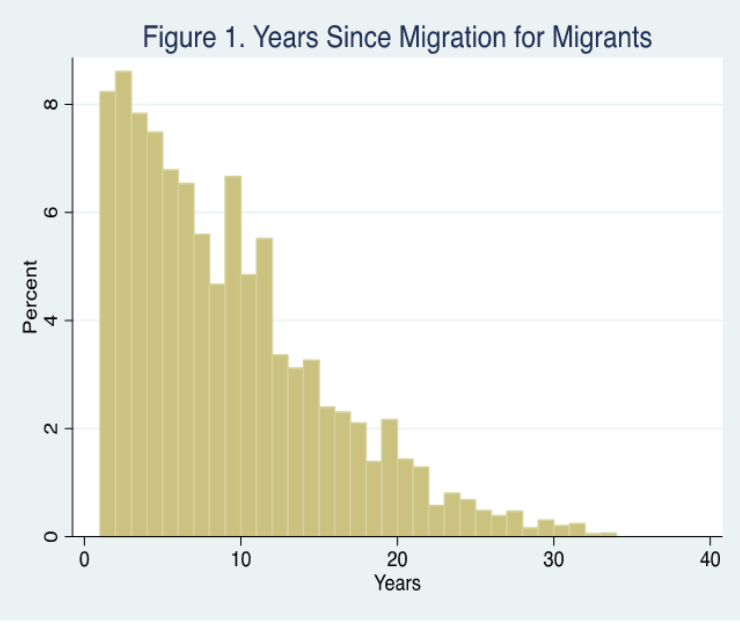
- We first estimate an ordered logit model based on self-reported health status to analyse if there exists a significant differential impact on health outcomes between rural-to-urban migrants.
- Secondly, we compared these results with those based on objectively measured health indicators:
  1) We try to assess part of the measurement error an check if there any differential effect when using objectively-measured health outcomes?
  2) Are health outcomes conditioned on years passed since migration?

Research Questions and Motivation

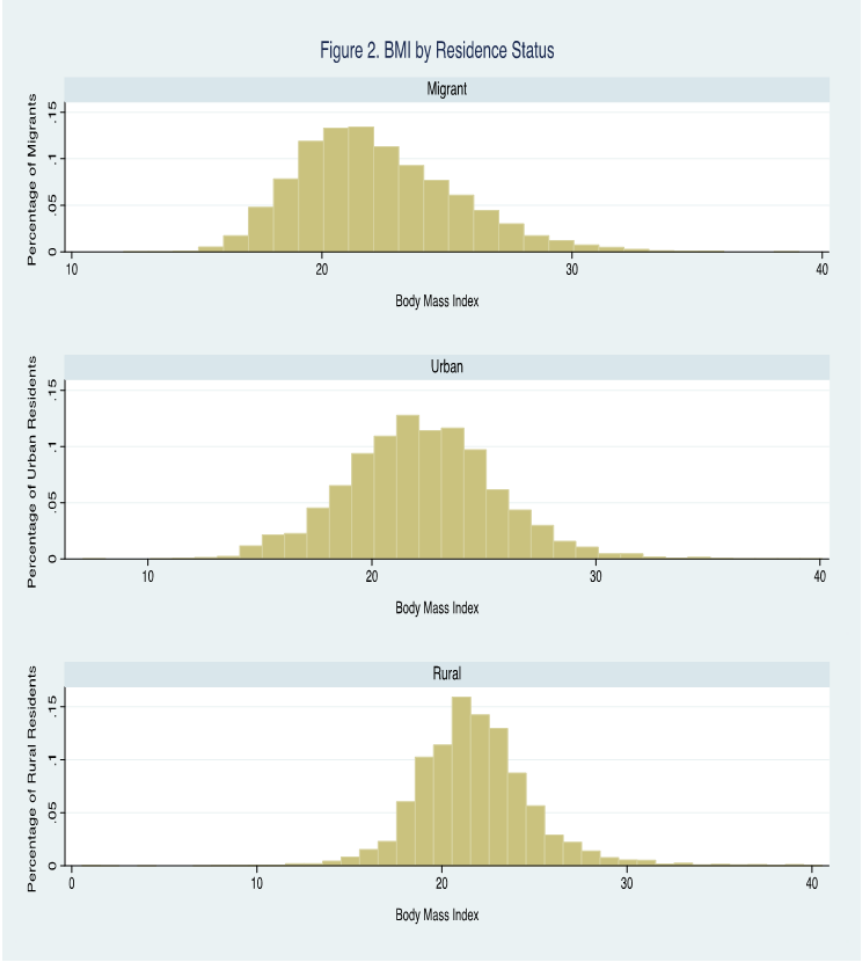
RQ: What we really wan to do here is use Machine Learning techniques to question whether hukou status plays a role in the health outcome of migrant workers.

- We use survey data reported in the Longitudinal Survey on Rural Urban Migration in China from the Institute for the Study of Labor (IZA). The survey collects data for 71,074 individuals (29,556 urban persons; 32,171 rural persons; and 9,347 migrants. Aprox 29% of rural persons) in two waves for the years 2008 and 2009.
- The survey contains data on socioeconomic indicators, such as education, income, ethnicity, and hukou registration.
- The RUMiC survey also includes data on many health indicators and outcomes. These include weight (kilograms), height (centimeters), dominant handedness, blood pressure, and grip strength

Econometric Model



Migrants BMI distribution is skewed to the left, 31% of migrants have a BMI below 18.0, which categorizes them as being underweight and subject to higher health risks



Econometric Model

The baseline specification is as follows with coefficients estimated by maximum likelihood (Williams, 2006 and Kleinbaum and Klein, 2010):

$$
\begin{aligned}
Pr(Y_j = i) = & \\
& Pr(\tau_{i-1} < \beta_1 MHukou_j \\
& + \beta_2 MHukou * YearsinceM_j \\
& + X_j\phi + Z_j\phi \\
& + \epsilon_j \leqslant \tau_i)
\end{aligned}
\tag{1}
$$

- Where Y_j states the self-reported health status for all j individuals living in urban areas, migrants and non-migrants.

Econometric Model

$$
\begin{aligned}
BloodPressure = & \\
& \beta_1 MHukou_i \\
& + \beta_2 Mhukou * YrsSinceM_i \\
& + K_i\psi + Z_i\phi \\
& + \epsilon_i
\end{aligned}
\tag{2}
$$

- Equation (2)s include a vector K of health indicators such as BMI, or being a smoker.

Econometric Model

- Health indicators could also be impacted by the length of time since a person migrated. of time since a person migrated. The direction of the relationship is unknown, but the longer the individual remains in the city, they can create more networks and gather more information about primary care clinics to access healthcare.
- We add health risk factor variables as well as socioeconomic variables to control for migration bias.

Preliminary Results

- Our preliminary results state that holding an urban hukou increases the chance for self-evaluated good health compared to dwellers with rural hukou in urban areas, although is only significant at the margin. Problems: mismeasurement, self-selection of migrants, do not observe the health of migrants before migrating.
- When using objective health measures the effect increases in magnitude and significance, but tends to disappear as migrants remain in the urban cities, suggesting a network effect of informal access to health providers, increase in incomes and access to private health.

Preliminary Results

- Migration will require adjustments in health provisions to accommodate the changing spatial demographics.
- Restricting migrants access to healthcare will clearly have an effect in the long run, including on migrant's health, productivity, and potential economic growth.

Section 4

Machine Learning Estimations

Why ML?

- The question of a parcimonious list in econometrics
- omitted variables
- model validity: time gaps, outliers, endogeneity, model uncertainty, parameter heterogeneity, measurement error, error correlation.
- In-sample model performance selection
- out-of-sample predictions

Data

- 27 variables
- 10,478 observations

Descriptive Statistics

	Rural Hukou	Urban Hukou
Age (mean)	32.1	39.2
Male %	56.6	48.8
Married %	73.37	80.3
Years of Education (mean)	9.2	10.8
Monthly Income (mean in RMB)	1677.6	1823.7
Wealth (mean in RMB)	4458.1	4667.9
Years Since Migration	8.5	9.8
Excellent or good health rating %	81.1	72.4
Average health rating %	16.8	27.3
Poor or very poor health rating %	2.1	3.1
Self-reported Health Rating (mean from 1-5 scale, 1 best)	2.0	2.1
Height (cm) (mean)	165.5	165.0
Weight (kg) (mean)	61.2	61.6
Smoker %	27.6	20.3
Cigarettes per day (mean)	3.7	2.7
Systolic Blood Pressure (mean)*	119.2	117.4
Diastolic Blood Pressure (mean)*	75.8	75.7
BMI< 18.5 (underweight) %*	6.5	4.5
BMI> 25 (overweight) %*	18.9	17.8
Health Insurance *(per centage)*		
NRCMS (Hukou based, targeted for rural pop, tax-funded and premiums)	58.4	10.8
URBMI (Hukou based, urban pop, tax-funded and premiums)	1.8	6.9
UEBMI (targeted to formal employees, contributions, offered by employer)	0.00	54.9
Commercial insurance	3.7	2.9
Other insurance	1.7	2.2
No insurance	30.5	20.2
Number of observations	7,598	2,706
Actions taken when ill or injured *(per centage)*		
Days missed due to illness or injury (mean)	3.7	6.3
Took No Action	20.1	10.6
Rested, but No Medicines	2.6	3.9
Obtained Medicine by themselves	45.8	38.2

Data Wrangling

- randomization
- Creation of a train dataset and a test dataset (70-30)
- Imputation of missing value values using KNN approximations
- Hot encoding for dummy variables
- range standardization fo our feature variables

Feature Selection in the overall dataset

- feature selection using recursive feature elimination (rfe):

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Variables	RMSE	Rsquared	MAE
Commercial Insurance	0.492	0.0166	0.483
Public Insurance	0.488	0.0333	0.479
Hukou	0.484	0.0558	0.474
Reason for Hukou Change	0.480	0.0734	0.469
City	0.477	0.0889	0.466

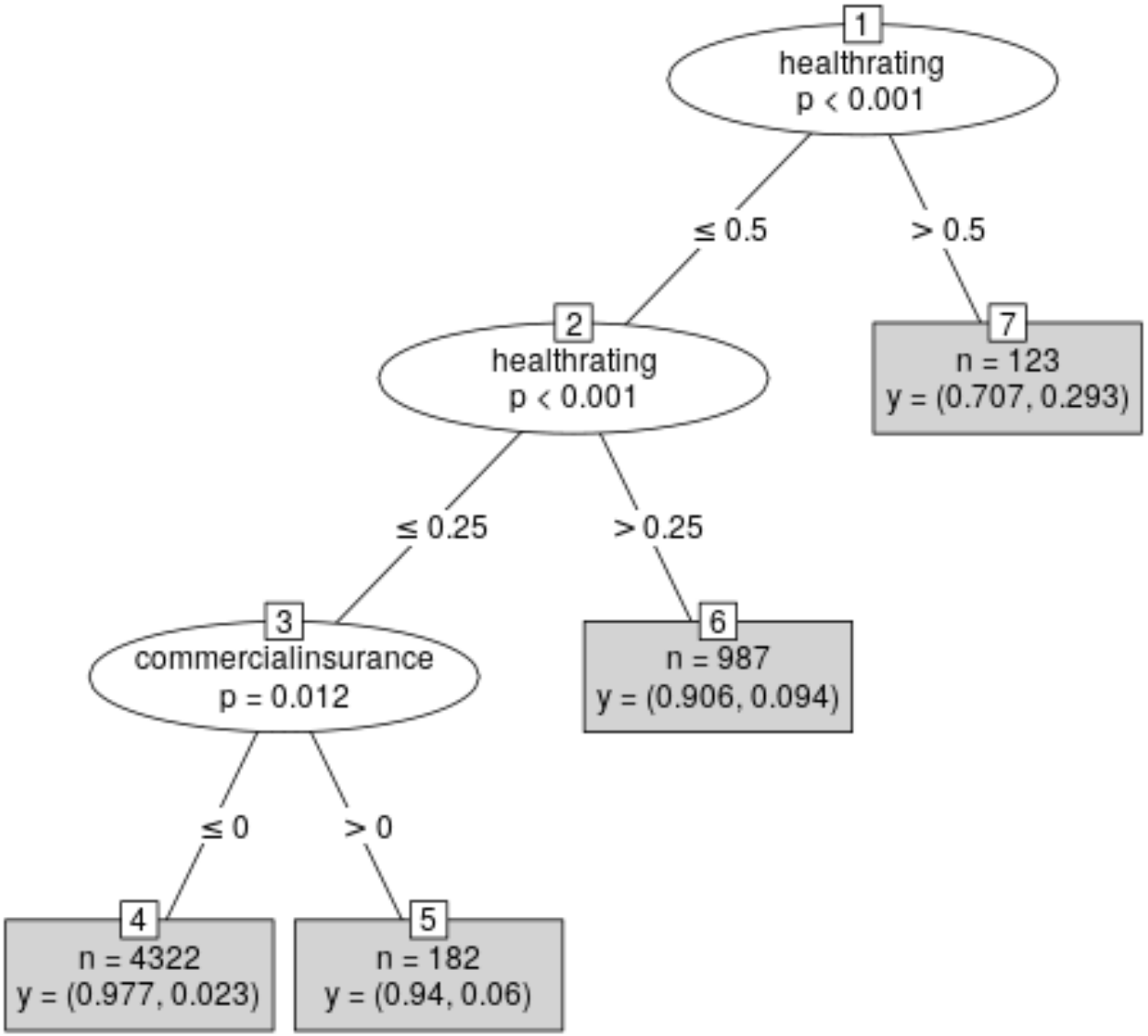
Scenario 2: Feature Selection

- 5 fold cross-validation for each model
- Models: a support vector machines with radial kernel, a random forest, a gradient boosting tree based model
- Automatic hyper-parameter tuning
- Model performance comparison over training and test sets, focusing on Accuracy and Kappa
- ensemble methods as well

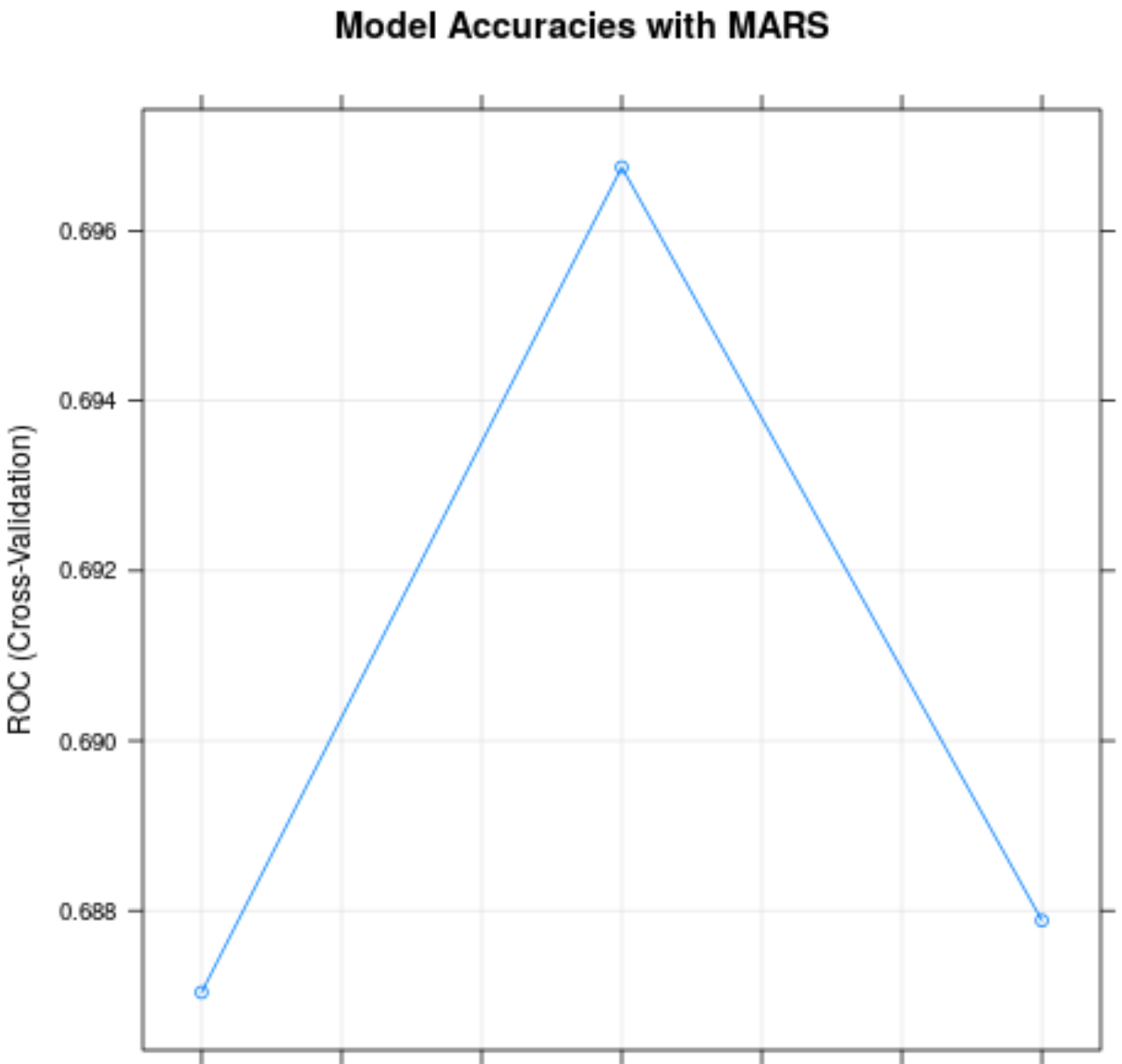
Variables

Target Variable	Feature Variables
daysmissedforsick_city	city, hukou, reasonforhukouchange, healthrating, commercialinsurance, publicinsurance, yearsofeduc, yrssincemigration

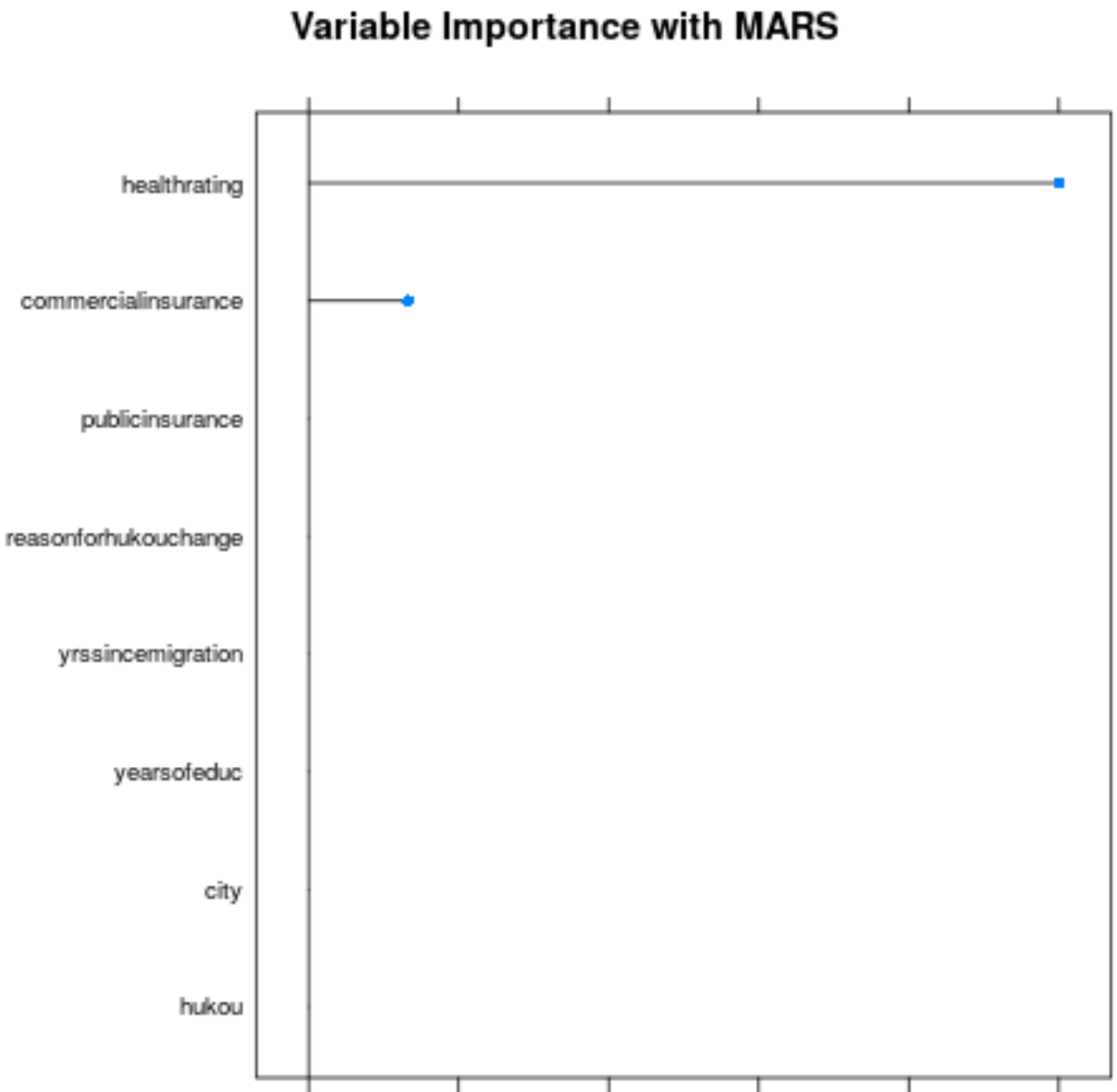
Conditional Inference Tree



Fitting a Multivariate Adaptive Regression Spline



Performance



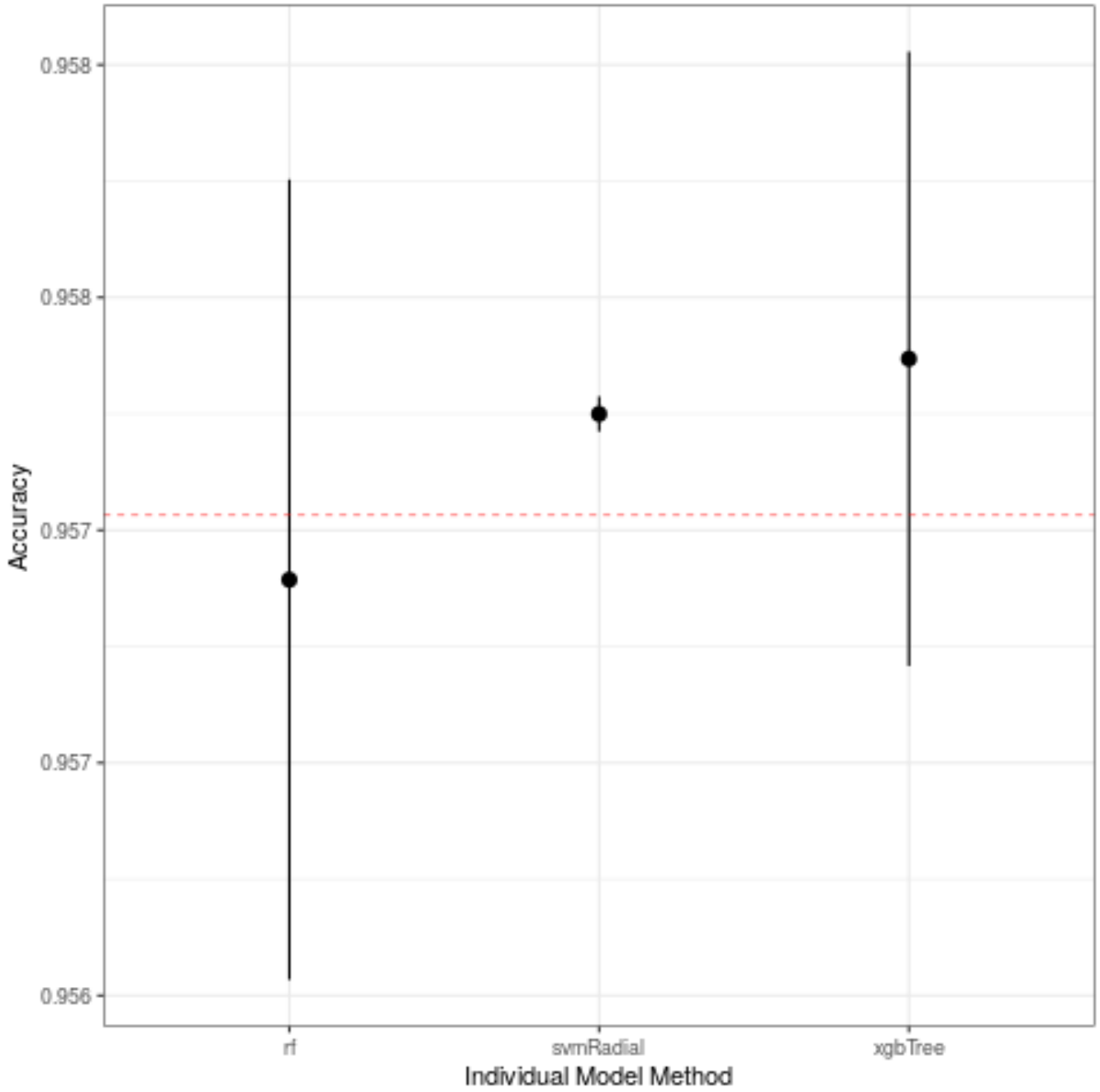
Prepare the test dataset and predict

Confusion Matrix and Statistics

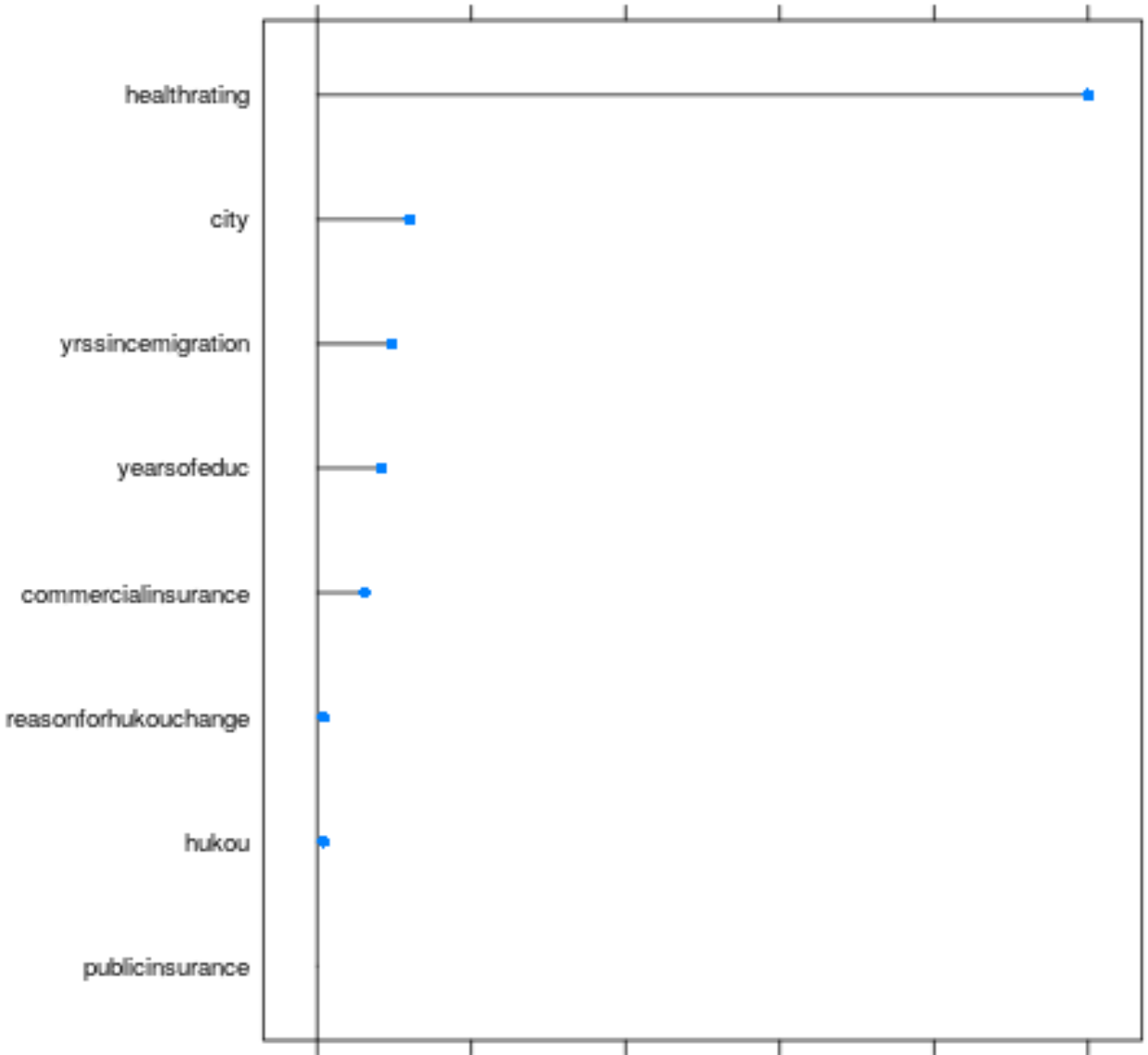
```
               Accuracy : 0.949
                 95% CI : (0.939, 0.957)
    No Information Rate : 0.948
    P-Value [Acc > NIR] : 0.487

                  Kappa : 0.029

            Sensitivity : 0.9996
            Specificity : 0.0161
              Precision : 0.9492
                 Recall : 0.9996
                     F1 : 0.9737
```

Ensembling the models

Performance	ens_1	ens_2	SVM	RF	xgbTree
Accuracy	0.62464	0.9193	0.948	0.948	0.957
Kappa	0.00835	-0.0191	0.000	0.000	0.052


0.958
0.958
0.957
0.957
0.956
Accuracy
rf
svmRadial
xgbTree
Individual Model Method

Prediction on the test dataset using



Section 5

Conclusions

Conclusions

- China's Hukou reform is a move in the right direction. Pilot programs in 29 provinces are helping to raise awareness about the necessity to eliminate barriers to health access which are now linked to geography. In 2015, Shanghai had approximately 9.8 million migrant workers holding a rural hukou.

Conclusions

- ML techniques are SO! interesting techniques for **predictive modelling**