Identifying Unvaccinated Individuals in Canada: **A Predictive Model**

March 28th, 2016

DATA 5000 – Project Presentation

Team Antivirus!

Ardyn Nordstrom & Kevin Dick

Presentation Outline

- Introduction & Motivation
- Dataset
- Methodology: Data Preparation & Analysis
- Methodology: Model Development
- Results & Conclusions
- Implications

Introduction & Motivation

- 23% of Canadian children do not have their basic immunizations
- More than ⅔ of Canadians do not receive their flu vaccinations each year
- The Budget allocated $25 million dollars to vaccination efforts

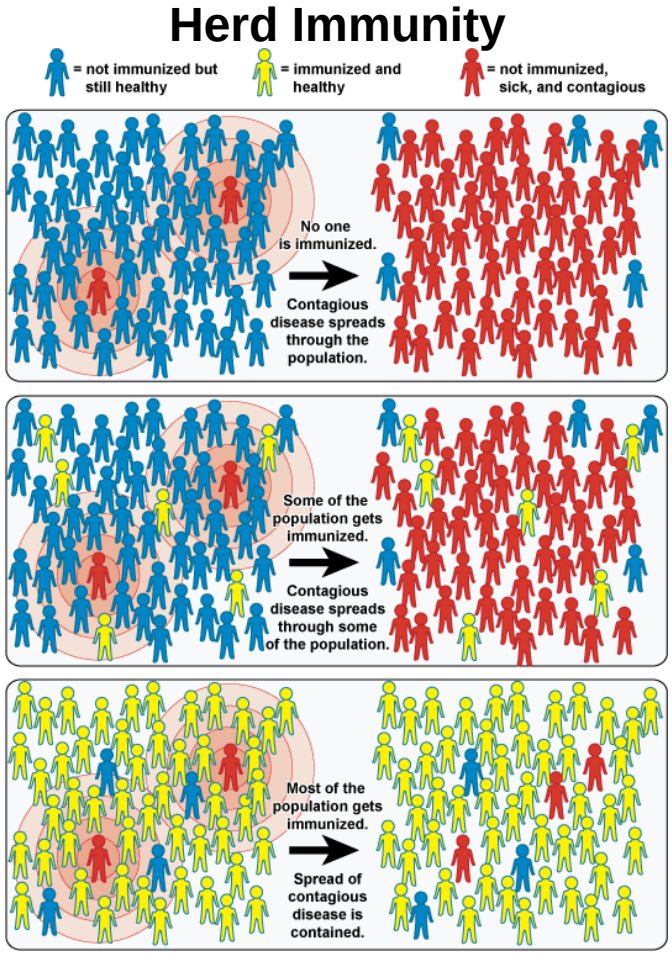


Image source: http://nrvs.info/faqs/herd-immunity-or-community-immunity/

Research Question

- What is the probability that Canadian households will receive their flu shot?
  - Canadian Community Health Survey (2009-2014)
  - Probit model for variable identification; random forest classifier
    - Training set: 2009-2013
    - Testing set: 2014

Result: This model can identify individuals at risk of not getting their flu shot.

Dataset

Canadian Community Health Survey
- Train: 2009-2013, Test: 2014

Includes details on health conditions, demographics, and life style
- 1,381 variables; 65,000 respondents sampled each year

Manually curated large groups of relevant variables

Turned categorical variables into binary variables

Excluded individuals with missing data (e.g. refused to declare or didn't know vaccination status)
- < 1% of observations

Methodology: Data Preparation and Analysis

- **Dependent Variable:** Flu shot in the last 12 months

- Focused on demographic and behavioural characteristics
  - Age
  - Smoking activity

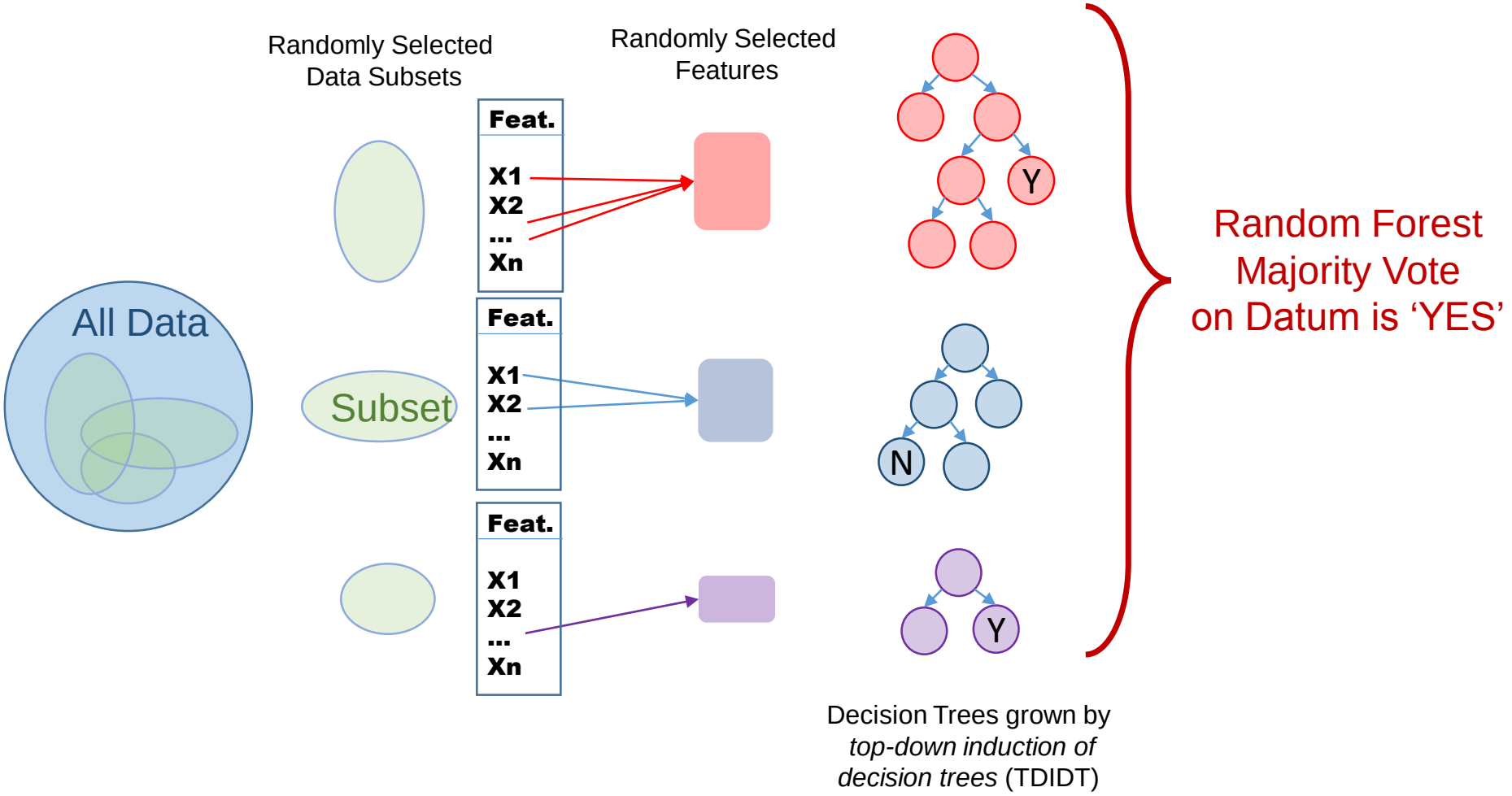
- Estimation of several probit models to identify variables and groups of variables with highest explanatory power
  - Final probit model (2012): Pseudo $R^2 = 0.1519$

- Identified 47 variables that were intrinsically and statistically significant in predicting vaccination outcomes

Variable Selection: Marginal Effects

Variables	Marginal Effect on Probability of Getting Flu Shot
Province of residence***	QC: 10.86% less likely – NS: 5.45% more likely
Diabetic***	11.89% more likely
Having a young child (age 0 – 5)***	11.47% more likely
Uses a cell phone while driving***	11.27% less likely
Getting a regular checkup***	9.37% more likely
Not wearing a seatbelt while driving**	8.9% less likely
Asthmatic***	8.89% more likely
Heart disease***	8.52% more likely
Having a family doctor***	7.35% more likely
Daily smoker***	7.29% less likely
Cancer***	7.15% more likely
Didn't attempt to find a doctor***	6.59% less likely
Chooses health for food content***	6.56% more likely
Female***	6.18% more likely
Arthritis***	5.57% more likely
Divorced, widowed, or separated***	4.21% less likely
Has strong social relationships***	4.15% more likely
Self-perceptions of good health**	Excellent health – 2.98% less likely
Married or common-law***	2.94% less likely
Frequently exercises***	2.57% more likely
University degree***	2.35% more likely

Random Forest Classification



Identifying a Subset of Features

Attribute Evaluation on 2012 Data												
Symmetrical Uncertainty			Chi-Squared			Gain Ratio			Information Gain		
0.054047	28	age	6772.461	28	age	0.047102	4	doc_dn_cntct	0.113254	28	age
0.037317	19	arthritis	1996.679	19	arthritis	0.044101	21	heart_disease	0.031996	19	arthritis
0.027628	15	family_doc	1250.035	6	reg_checkup	0.042009	20	diabetes	0.020896	15	family_doc
0.02573	20	diabetes	1139.733	15	family_doc	0.041693	19	arthritis	0.020655	6	reg_checkup
0.02517	21	heart_disease	1112.8	20	diabetes	0.036967	15	family_doc	0.017569	20	diabetes
0.022129	4	doc_dn_cntct	1057.695	21	heart_disease	0.034916	28	age	0.016683	21	heart_disease
0.021272	6	reg_checkup	918.4292	31	div_sep_wid	0.029709	22	cancer	0.014807	31	div_sep_wid
0.017102	31	div_sep_wid	693.4959	49	hh_inc	0.024135	3	phon_drive	0.013701	4	doc_dn_cntct
0.011796	9	daily_smoker	692.2072	4	doc_dn_cntct	0.020767	6	reg_checkup	0.011467	49	hh_inc
0.011625	22	cancer	581.4083	9	daily_smoker	0.01888	31	div_sep_wid	0.009985	9	daily_smoker
0.010803	33	child6_11	438.3868	33	child6_11	0.01596	33	child6_11	0.007735	33	child6_11
0.008161	32	child0_5	434.8571	22	cancer	0.013393	9	daily_smoker	0.006846	22	cancer
0.008018	49	hh_inc	348.421	29	female	0.012497	2	n_seatbelt	0.005916	32	child0_5
0.007502	34	health_fair	339.6505	32	child0_5	0.011897	45	NS	0.005782	29	female
0.00596	29	female	331.7553	34	health_fair	0.011775	32	child0_5	0.005502	7	hrs_sedentary
0.005454	40	QC	329.9432	7	hrs_sedentary	0.011363	34	health_fair	0.005305	34	health_fair
0.004726	45	NS	285.3821	1	HWTGBMI	0.010693	5	had_flu	0.00474	1	HWTGBMI
0.004569	7	hrs_sedentary	269.4548	40	QC	0.00628	40	QC	0.004566	40	QC
0.003832	37	health_excel	206.9048	48	pers_inc	0.005995	49	hh_inc	0.003492	48	pers_inc
0.003661	1	HWTGBMI	200.7167	10	social_rels	0.005823	29	female	0.003347	10	social_rels
0.003576	10	social_rels	186.1439	37	health_excel	0.004706	18	asthma	0.003144	37	health_excel
0.003359	27	psgrad	185.3323	27	psgrad	0.004534	37	health_excel	0.003029	27	psgrad
0.002837	18	asthma	175.9698	45	NS	0.004041	14	food_scarce	0.002794	45	NS
0.002753	14	food_scarce	124.3585	36	health_vgood	0.003766	7	hrs_sedentary	0.002064	36	health_vgood
0.002289	48	pers_inc	119.3797	18	asthma	0.003619	10	social_rels	0.001977	14	food_scarce
0.002163	36	health_vgood	116.3574	14	food_scarce	0.003539	27	psgrad	0.001924	18	asthma
0.002048	3	*phon_drive*	51.8026	3	*phon_drive*	0.002949	12	suicidal	0.001013	3	*phon_drive*
0.000852	26	postsec	43.3193	35	health_good	0.002887	1	HWTGBMI	0.000712	35	health_good

- Attribute evaluation found to agree with probit models
- Top 5-10 indicate highest potential for discriminability
- Compare subset of features to all features and compare performance

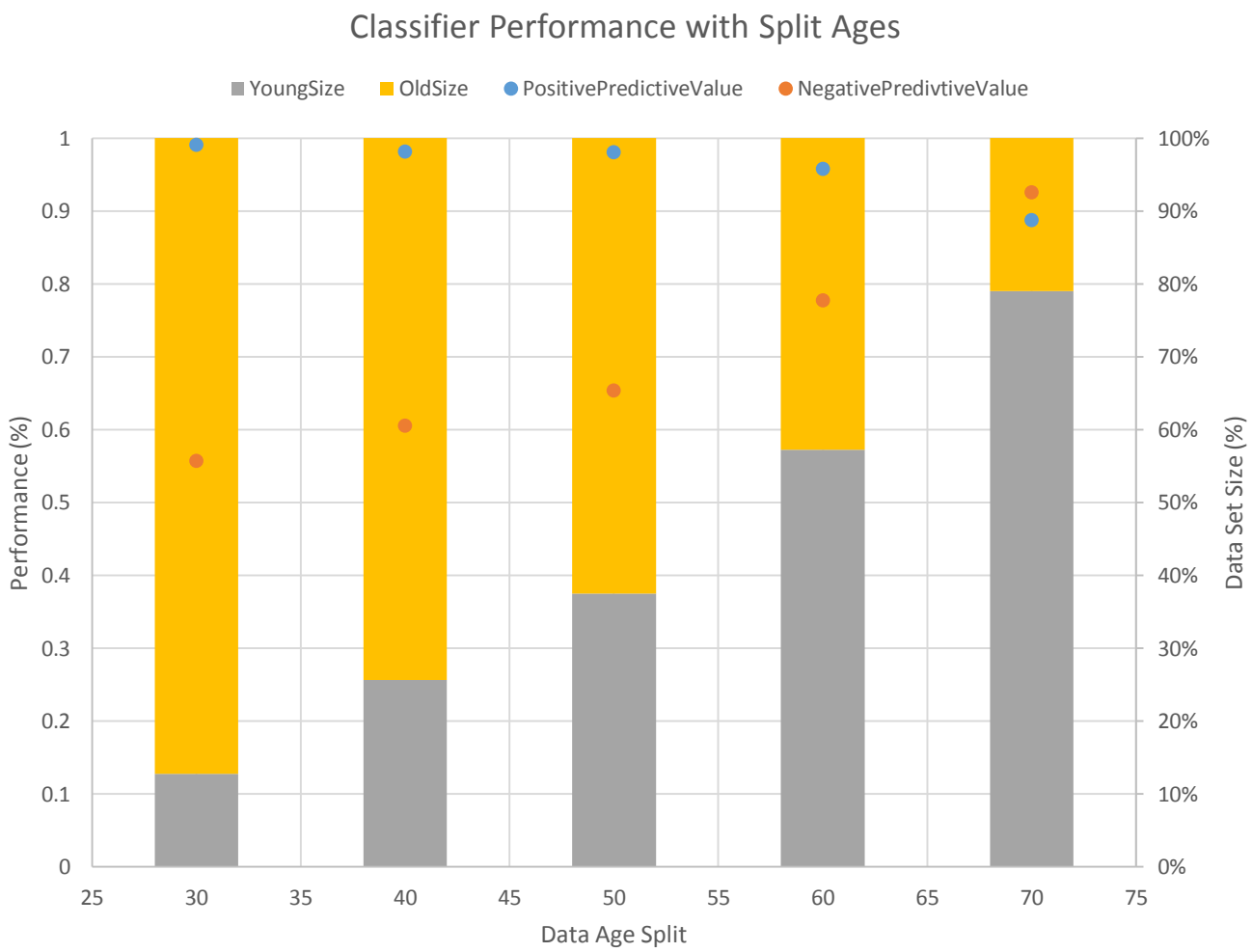
Baseline Results

Classifier Results - ALL 47 FEATURES Train: 2011, Test: 2012					
	Naive Bayes	J48 Decision Tree	Random Forest	Multilayer Perceptron	
Pos. Pred. V.	0.738	0.761	0.825	0.743	All 47
Neg. Pred. V.	0.616	0.584	0.635	0.656	
Accuracy	0.769	0.623	0.724	0.789	
ROC	0.729	0.701	0.733	0.768	

- Random Forest found to be stable with good performance.
- Subset of features were found to increase performance, but not across age groups.
- 'Young' individuals were difficult to classify.

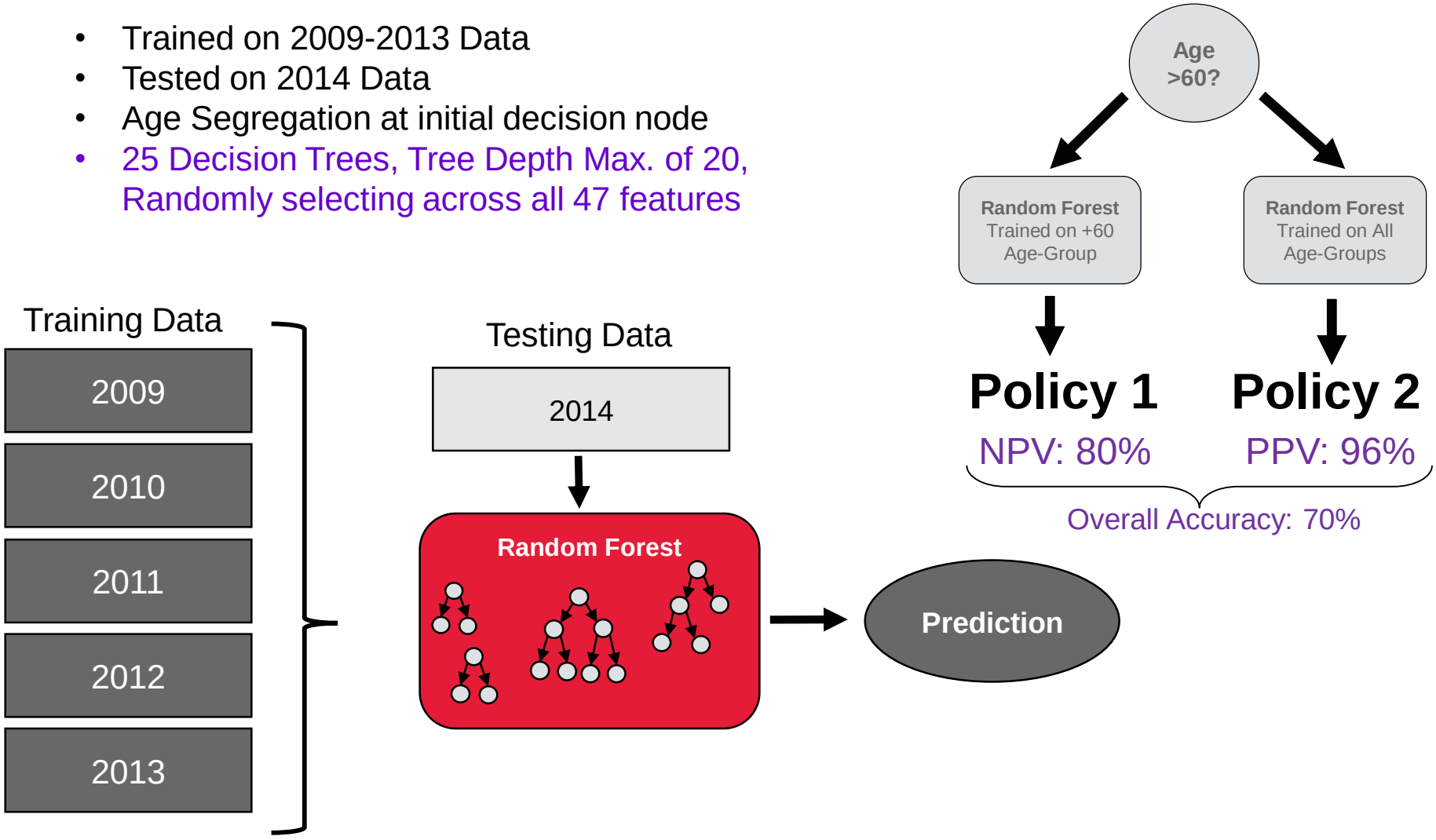
Classifier Results - 9 Features Selected Train: 2011, Test: 2012						
	Naive Bayes	J48 Decision Tree	Random Forest	Multilayer Perceptron	K*	Features
Pos. Pred. V.	0.784	0.861	0.851	0.862	0.874	heart_disease, cancer
Neg. Pred. V.	0.548	0.484	0.523	0.471	0.464	daily_smoker, age
Accuracy	0.698	0.723	0.731	0.719	0.724	family_doc, div_sep_wid
ROC	0.729	0.71	0.764	0.795	0.752	arthritis, hh_inc
						diabetes

Identifying Appropriate Age Split Threshold



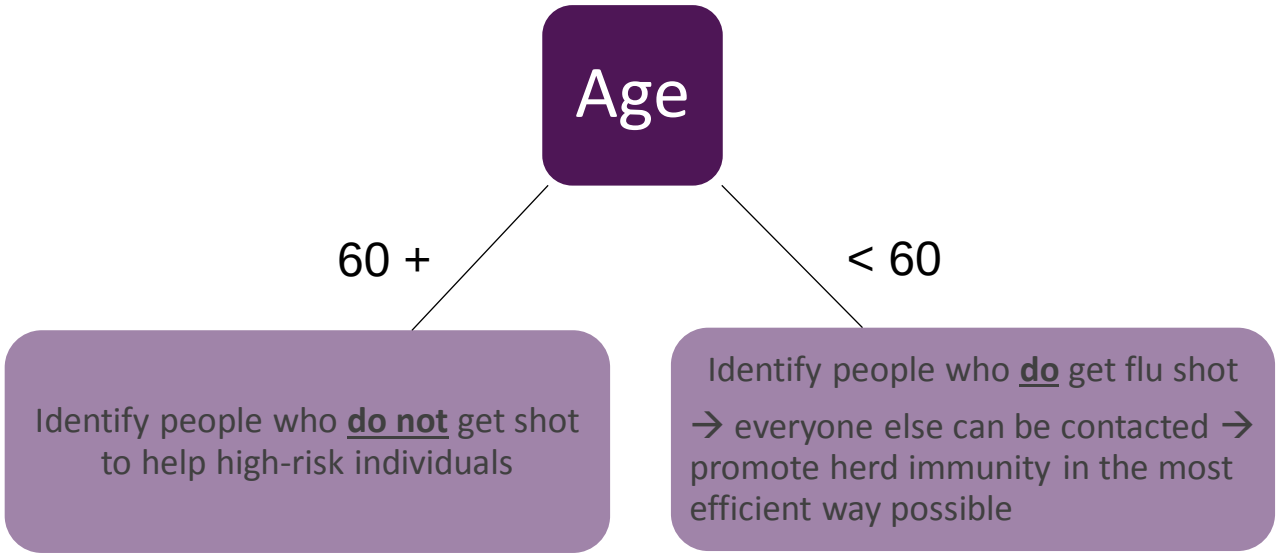
Final Classification Model

- Trained on 2009-2013 Data
- Tested on 2014 Data
- Age Segregation at initial decision node
- 25 Decision Trees, Tree Depth Max. of 20, Randomly selecting across all 47 features



Conclusion & Future Directions

- Survey data can be used to accurately identify people that can benefit from vaccination interventions:
  - Promote herd immunity
  - Identify high-risk individuals
- Different identification strategies for people above / below 60 can allow vaccination resources to be efficiently



Thank you!

Questions?

Come check out our poster at DATA DAY!

Identifying Unvaccinated Individuals in Canada: A Predictive Model

Kevin Dick
Systems and Computer Engineering
kevin.dick@carleton.ca

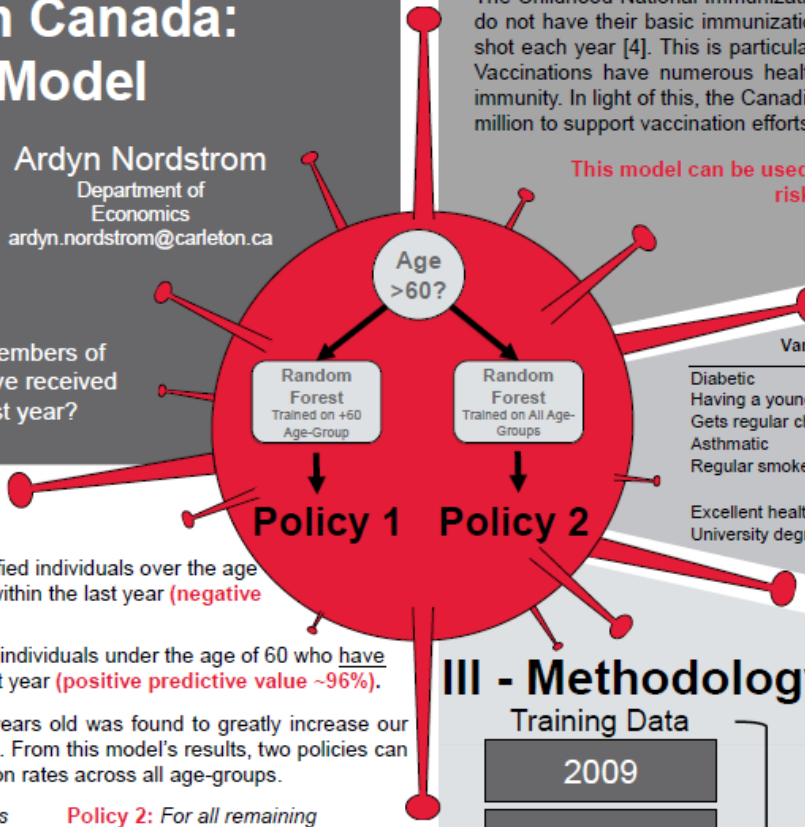
Ardyn Nordstrom
Department of Economics
ardyn.nordstrom@carleton.ca

Research Question:
What is the probability that members of Canadian households will have received their flu vaccination in the past year?

I - Introduction

The Childhood National Immunization Coverage Survey in 2013 found that as many as 23% of Canadian children do not have their basic immunizations (e.g. whooping cough) [3]. Fewer than one third of Canadians get their flu shot each year [4]. This is particularly dangerous for seniors as the flu can lead to more serious health problems. Vaccinations have numerous health benefits for immunized individuals as well as for the public through herd immunity. In light of this, the Canadian government's latest budget allocated an additional $25 million to support vaccination efforts [1].

This model can be used to identify individuals at risk of not getting a flu shot.



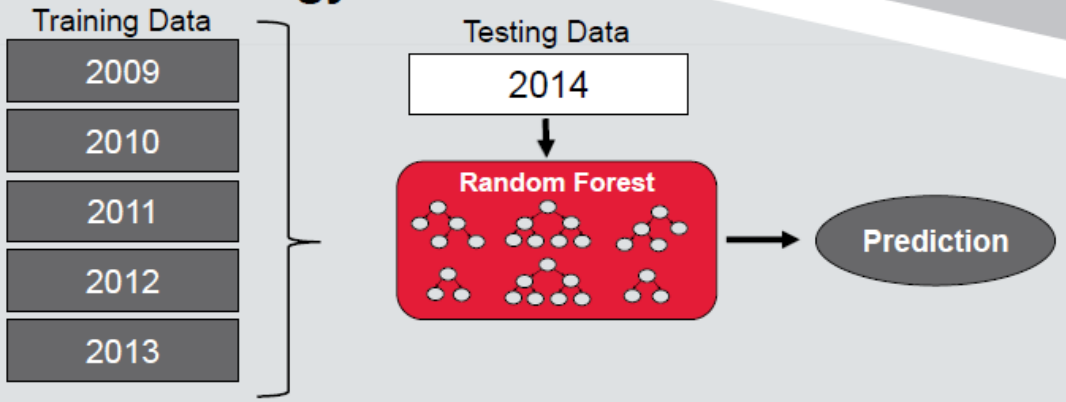
II - Data

Variable	Marginal Effect on Vaccination
Diabetic	12% more likely
Having a young child (age 0-5)	11% more likely
Gets regular checkup	9% more likely
Asthmatic	9% more likely
Regular smoker	7% less likely
⋮	⋮
Excellent health (perceived)	3% less likely
University degree	2% more likely

Using the Canadian Community Health Surveys (2009 – 2014), 47 variables out of 1,381 were identified as being intrinsically and statistically significant in predicting vaccination outcomes. The marginal effect from a probit model were used to infer the significance of each variable.

Discrete characteristics (e.g. province of residence) were turned into binary variables for intuitive use in the model.

III - Methodology - Model



IV - Results

Our method has successfully identified individuals over the age 60 who have not been vaccinated within the last year (negative predictive value ~80%).

Our method can accurately identify individuals under the age of 60 who have received their vaccination in the last year (positive predictive value ~96%).

The division of age groups at 60 years old was found to greatly increase our accuracy (overall accuracy ~70%). From this model's results, two policies can be developed to increase vaccination rates across all age-groups.

Policy 1: *For high risk individuals (60+), this model can be used to target specific persons who likely have not received their flu shot.*

Policy 2: *For all remaining individuals, this model targets persons who are likely to get their flu shot, therefore herd immunity can be established by promoting flu vaccination in non-targets.*

The final classification model incorporates a decision node to segregate by age 60 and then classifies on an "expert" Random Forest trained over that age group. Our method is applicable to survey data in future years given its performance over the recently released 2014 dataset (released March 16th, 2016). Leveraging data from previous years (2009 - 2013), our classifier improved with incorporation of new data, indicating that our method is expected to improve with subsequent data.

V - Future Applications

Given the Canadian government's recent budget allocating $25 million dollars to supporting vaccination efforts, our model can be used to efficiently deploy these resources.

Acknowledgements & References

This research would like to acknowledge Olga Baysal and Boyan Bejanov for their feedback throughout DATA 5000.

[1] Government of Canada: Budget 2016 (2016). *Chapter 5 – An Inclusive and Fair Canada*. Retrieved from: http://www.budget.gc.ca/2016/docs/plan/ch5-en.html
[2] Statistics Canada. (2016). Canadian Community Health Survey, 2009-2014: Annual component [Data file and code book]. Retrieved from http://www.odesi.ca/
[3] Statistics Canada. Table 1, Immunization coverage by antigen for two-year-old children in Canada, 2013. CANSIM. Last updated July 21, 2015. http://www.statcan.gc.ca/daily-quotidien/150721/t001c-eng.htm (accessed March 23, 2016).
[4] Statistics Canada. Table 2, Age-standardized flu vaccination rates by one or more selected chronic conditions, population aged 12 to 64, Canada and regions, 2003 and 2013-2014. CANSIM. Last updated November 27, 2015. http://www.statcan.gc.ca/pub/82-624-x/2015001/article/14218-eng.htm (accessed March 23, 2016).

Carleton UNIVERSITY