



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Деревья решений для классификации демографических последовательностей

Муратова Анна Александровна
Игнатов Дмитрий Игоревич

Статья подготовлена в результате проведения исследования № 16-05-0011 «Разработка и апробация методик анализа демографических последовательностей» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016-2017 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

Постановка задачи

- В настоящее время доступно большое количество демографических данных о наступлении важных демографических событий и последовательностей
- Современные методы майнинга данных позволяют найти неявные взаимосвязи между событиями в последовательностях
- Полученные результаты помогут полнее проанализировать социально-демографическое поведение людей в стране



Постановка задачи

- Анализ демографических данных методами и программными средствами машинного обучения и майнинга данных
- Задачи:
 - Предсказание пола по демографическим и социоэкономическим событиям и признакам
 - Предсказание первого и последнего события для демографических и социоэкономических событий
- Используемые инструменты:
 - Orange
 - Программы собственной разработки на Python



Данные

- Институтом демографии НИУ ВШЭ была предоставлена база данных, содержащая 4943 человек (1582 мужчин и 3361 женщин)

Признаки:

- Пол (мужской, женский);
- Поколение (1 – 1930-1939, 2 – 1940-1949, 3 – 1950-1959, 4 – 1960-1969, 5 – 1970-1979, 6 – 1980–1986);
- Уровень образования (общее, профессиональное, высшее);
- Тип населенного пункта (город, поселок городского типа, село);
- Религиозность (да, нет);
- Частота посещения религиозных служб (несколько раз в неделю, раз в неделю, минимум раз в месяц, несколько раз в год и реже);

Также указаны даты рождения и даты наступления демографических событий (брак, партнер, ребенок) и социоэкономических событий (образование, отделение от родителей, работа) с точностью до месяца



Кодирование данных

1. Временные таблицы (время наступления событий в месяцах)
2. Бинарные таблицы
 - 1 – событие наступило
 - 0 – событие еще не произошло
3. Таблицы пар событий
 - < – если первое событие наступило до второго
 - > – если первое событие наступило после второго
 - = – если события произошли в одном месяце
 - n – если ни одно из событий еще не наступило
4. Таблицы пар событий с непосредственным предшествованием
 - l – первое событие произошло непосредственно до второго
 - r – первое событие произошло непосредственно после второго
 - < – если первое событие наступило до второго
 - > – если первое событие наступило после второго
 - = – если события произошли в одном месяце
 - n – если ни одно из событий еще не наступило



Демографические события

Предсказание признака «пол»

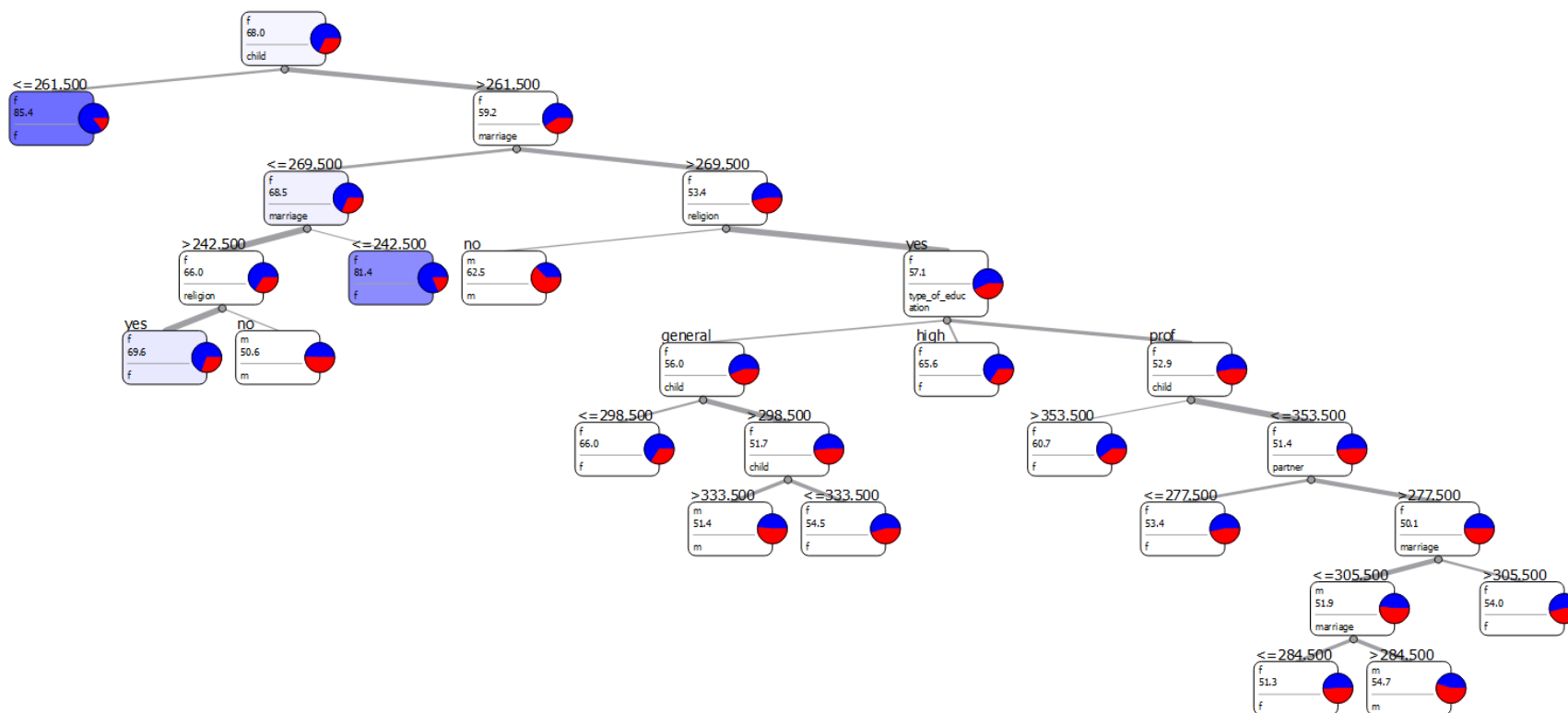
Кодирование (шкалирование)	СА (точность)
Бинарное	0.6808
Временное	0.6662
Пары событий	0.6808
Непосредственное предшествование	0.6800
Признаки + бинарное	0.6785
Признаки + временное	0.6996
Признаки + пары событий	0.6684
Признаки + непосредственное предшествование	0.6783



Демографические события

Предсказание признака «пол»

Дерево решений, признаки респондента и временная кодировка:



Демографические события

Предсказание признака «пол»

Мужчины:

Посылка	Вероятность
Ребенок после 22, брак после 23, неверующий (вероятность 62.5%)	62.5%
Ребенок от 22 до 30 лет, партнер после 23, брак от 23 до 26, верующий, тип образования профессиональный (вероятность 54.7%)	54.7%

Женщины:

Посылка	Вероятность
Ребенок до 22 лет	85.4%
Брак до 20 лет, ребенок после 22	81.4%
Брак от 20 до 23, ребенок после 22, верующий	69.6%
Ребенок от 22 до 25 лет, брак после 23, верующий, тип образования общий	66%
Ребенок после 22, брак после 23, верующий, тип образования высший	65.6%



Социальноэкономические события

Предсказание признака «пол»

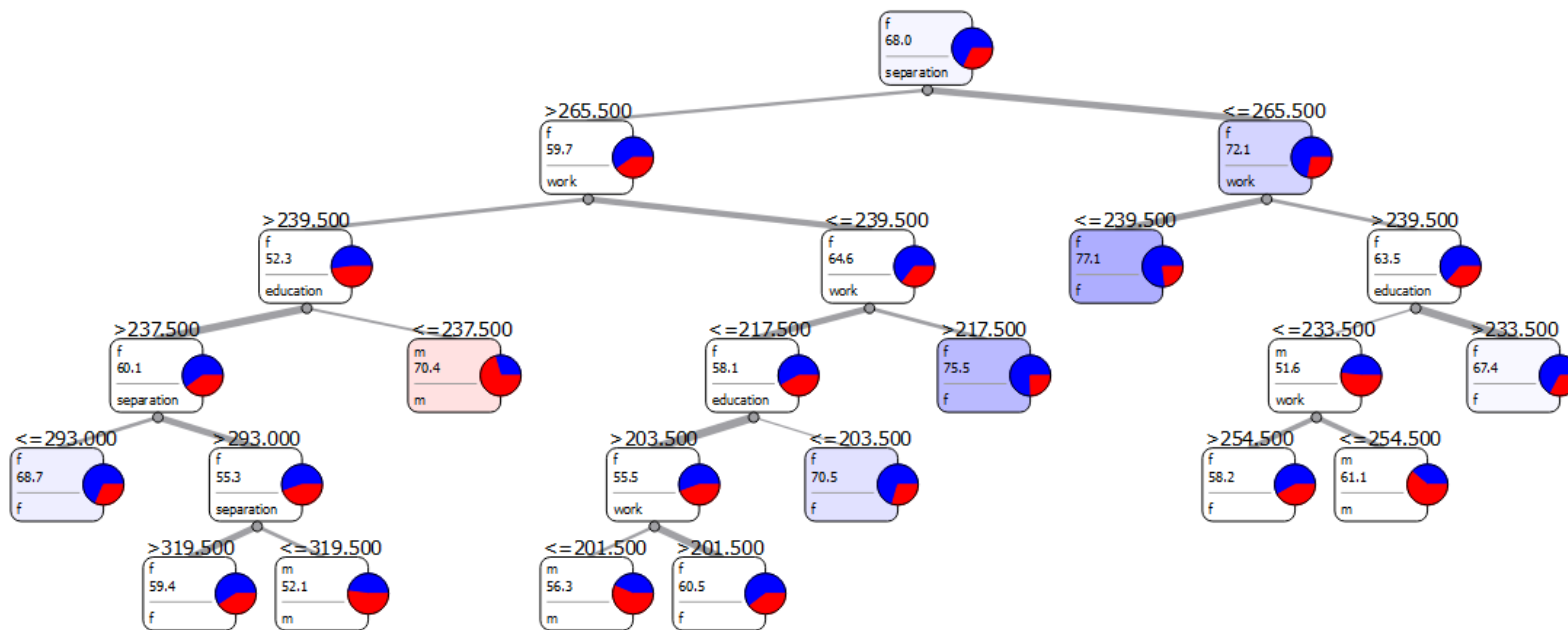
Кодирование (шкалирование)	СА (точность)
Бинарное	0.6800
Временное	0.6931
Пары событий	0.6800
Непосредственное предшествование	0.6777
Признаки + бинарное	0.6800
Признаки + временное	0.6925
Признаки + пары событий	0.6711
Признаки + непосредственное предшествование	0.6771



Социальноэкономические события

Предсказание признака «пол»

Дерево решений, временная кодировка:



Социальноэкономические события

Предсказание признака «пол»

Мужчины:

Посылка	Вероятность
Образование до 19 лет, работа после 19, отделение после 22	70.4%
Образование до 19 лет, работа после 20, отделение до 22, работа до 22	61.1%
Работа до 17 лет, образование после 17, отделение после 22	56.3%

Женщины:

Посылка	Вероятность
Работа до 20 лет, отделение до 22	77.1%
Работа от 18 до 19 лет, отделение после 22	75.5%
Образование до 17 лет, работа до 18, отделение после 22	70.5%
Работа после 19 лет, образование после 19, отделение от 22 до 25	68.7%
Образование после 19 лет, работа после 20, отделение до 22	67.4%

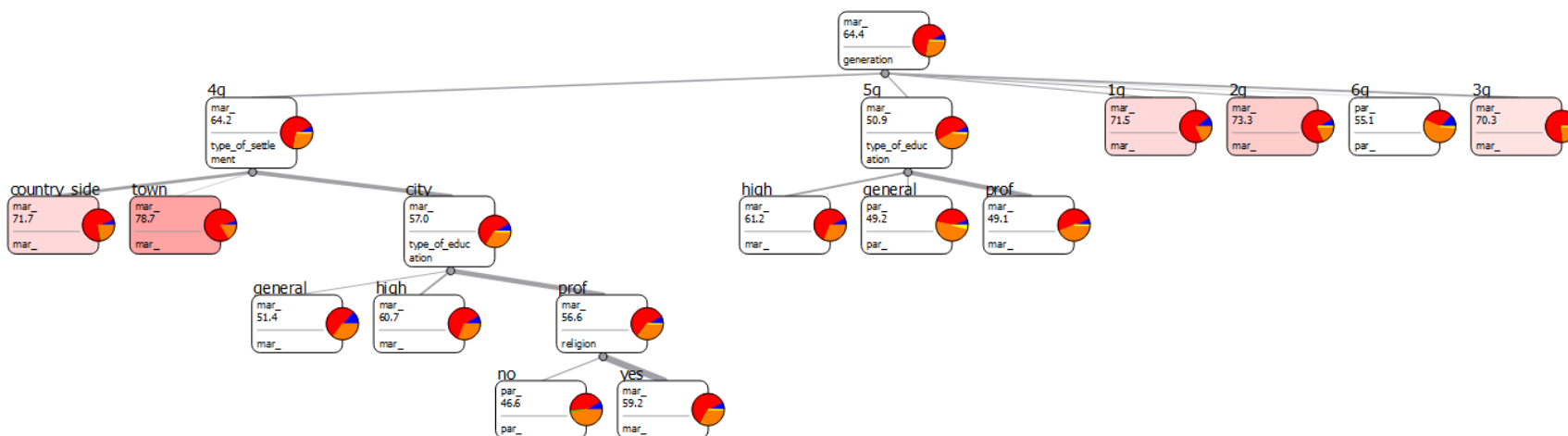


Демографические события

Предсказание первого события

Дерево решений, признаки

Точность дерева: 0.6565



Первое событие больше всего зависит от поколения.

Демографические события

Предсказание первого события

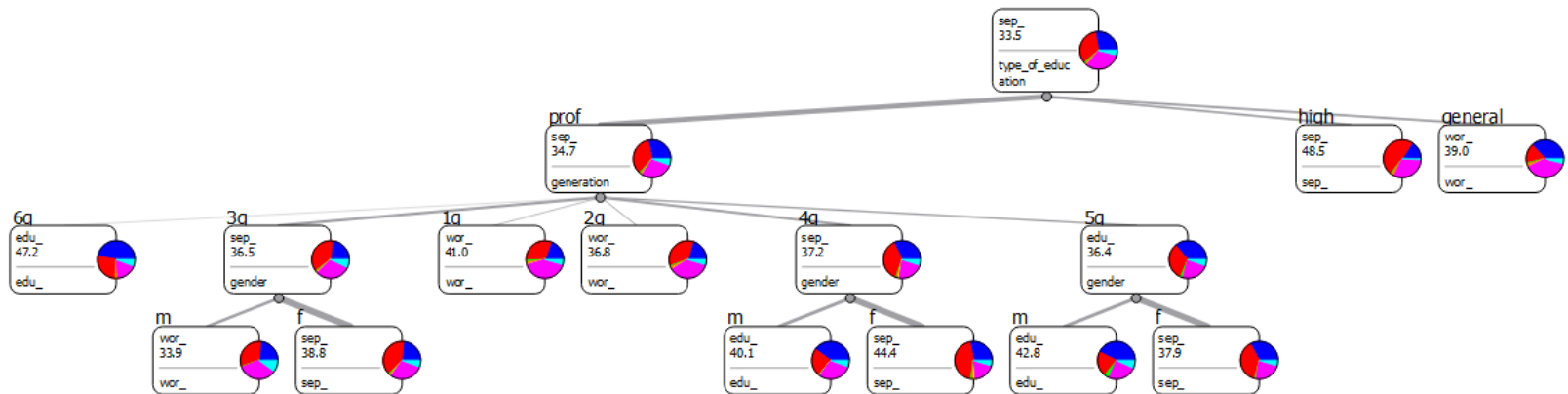
Поколения	Признаки	Вероятность	Первое событие
1 – 1930-1939 2 – 1940-1949 3 – 1950-1959		71.5% 73.3% 70.3%	брак
4 – 1960-1969	Место жительства деревня Место жительства поселок Место жительства город и тип образования высший	71.7% 78.7% 60.7%	брак
5 – 1970-1979	Тип образования высшее Тип образования профессиональное Тип образования общий	61.2% 49.1% 49.2%	брак брак партнер
6 – 1980–1986		55.1%	партнер

Социальноэкономические события

Предсказание первого события

Дерево решений, признаки

Точность дерева: 0.4283



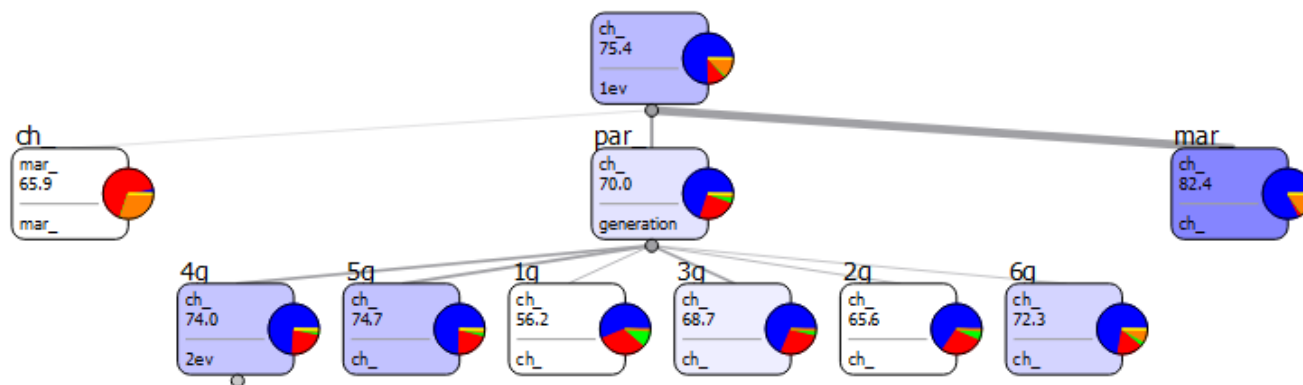
Первое событие больше всего зависит от типа образования.

Демографические события

Предсказание последнего события

Кодирование (шкалирование)	СА (точность)
Признаки	0.7542
Признаки + события	0.7869

Дерево решений, признаки + события:



Последнее событие больше всего зависит от первого события.

Демографические события

Предсказание последнего события

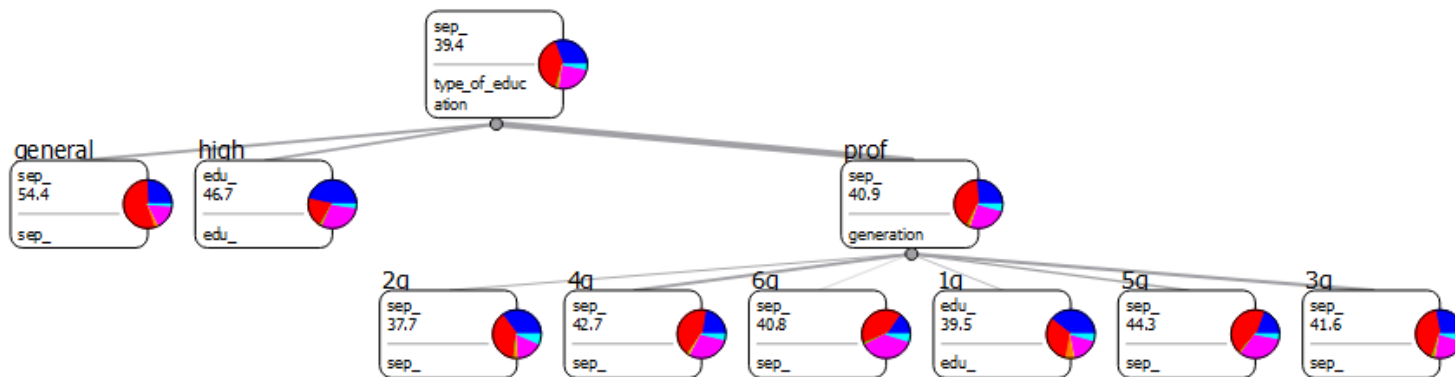
Первое событие	Признаки	Вероятность	Последнее событие
брак		82.4%	ребенок
ребенок		65.9%	брак
партнер	1 – 1930-1939	56.2%	ребенок
	2 – 1940-1949	65.6%	
	3 – 1950-1959	68.7%	
	4 – 1960-1969	74%	
	5 – 1970-1979	74.7%	
	6 – 1980–1986	72.3%	

Социальноэкономические события

Предсказание последнего события

Кодирование (шкалирование)	СА (точность)
Признаки	0.4530
Признаки + события	0.9421

Дерево решений, признаки + события:



Последнее событие больше всего зависит от типа образования.

Социальноэкономические события

Предсказание последнего события

Тип образования	Признаки	Вероятность	Последнее событие
общий		54.4%	отделение от родителей
высший		46.7%	образование
профессиональный	1 – 1930-1939	39.5%	образование
	2 – 1940-1949	37.7%	отделение от родителей
	3 – 1950-1959	41.6%	
	4 – 1960-1969	42.7%	
	5 – 1970-1979	44.3%	
	6 – 1980-1986	40.8%	



Выводы

- Методы классификации с помощью деревьев решений являются оптимальным средством, отвечающим текущим задачам демографов
- Полученные результаты представляют интерес для демографов в качестве их дальнейшей интерпретации
- Используемые методы анализа данных могут быть применены для решения схожих задач в других областях, в которых исследуются последовательности событий в поведении людей



Список литературы

- Billari F. C., Furnkranz J., and Prskawetz A. Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. *European Journal of Population*, 22:37–65. 2006.
- Ignatov D. I., Mitrofanova E. S., Muratova A. A., Gizdatullin D. [Pattern Mining and Machine Learning for Demographic Sequences](#). KESW 2015, Moscow, Russia. Springer International Publishing. 2015.
- Jensen, Anders Boeck et al. “Temporal Disease Trajectories Condensed from Population-Wide Registry Data Covering 6.2 Million Patients.” *Nature Communications* 5 (2014): 4022. PMC. Web. 13 Nov. 2016.
- Cees H. Elzinga, Aart C. Liefbroer. De-standardization of Family-Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis // *European Journal of Population*. 2007. Volume 23, Issue 3, P. 225-250.



Спасибо за внимание!

