Деревья решений для классификации демографических последовательностей[1]

Муратова Анна Александровна, Национальный исследовательский университет Высшая школа экономики, Москва amuratova@hse.ru

Игнатов Дмитрий Игоревич, Национальный исследовательский университет Высшая школа экономики, Москва dignatov@hse.ru

Аннотация

Анализ демографических последовательностей становится все более востребованным в последние годы. В ходе работы был осуществлен анализ демографических последовательностей с помощью деревьев решений. Для демографических и социоэкономических событий были определены первые и последние значимые события в жизни человека с учетом всех признаков и уже произошедших событий. Также определена зависимость событий для признаков пол и поколение. Полученные результаты помогут экспертам полнее оценить социально-демографическую ситуацию в нашей стране. Полученные результаты могут быть интересны для демографических исследований.

Проблема

В настоящее время исследователям разных стран доступно большое количество демографических данных о наступлении важных демографических событий и их последовательностей. Демографы исследуют взаимосвязь между такими событиями и выявляют часто встречающиеся последовательности событий на жизненном пути (life trajectory) людей [1,2,4]. Это не только помогает понять, как менялось демографическое поведение разных поколений в разных странах, но и позволяет отследить изменения в том, как люди расставляют приоритеты между семьей и работой, сопоставить этапы взросления мужчин и женщин.

Анализ демографических последовательностей становится все более востребованным в последние годы. В России исследований в этой сфере намного меньше, чем в зарубежных странах, поэтому данная работа является весьма актуальной. Полученные результаты помогут экспертам полнее оценить социально-демографическую ситуацию в нашей стране.

Целью данной работы был анализ демографических последовательностей с использованием различных методов машинного обучения. В частности, одной из задач являлось определение первого и последнего события в жизни человека с учетом всех признаков и уже произошедших событий. Также необходимо было определить зависимость событий для признака пол.

[1] 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 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

Анализ демографических и социоэкономических последовательностей, непосредственное предшествование

Одним из направлений, которым занимаются демографы, является изучение первых значимых событий в жизни человека, которые означают переход во взрослую жизнь. Институтом демографии НИУ ВШЭ была предоставлена база данных, содержащая результаты опроса 4943 человек (1582 мужчин и 3361 женщин). Для каждого человека указаны:

- 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)
- Дата рождения

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

Целью работы являлся анализ демографических и социоэкономических событий в отдельности. Для каждых из этих событий необходимо было на основе данных определить пол и поколение. Также нужно было определить первое и последнее событие на каждом из типов событий.

Данная работа представляет интерес и значимость как для демографов и социологов, так и для людей, занимающихся анализом данных.

Демографическими событиями являются партнерство, брак, рождение ребенка, а социоэкономическими – отделение от родителей, работа, окончание образования.

Для начала был произведен сравнительный анализ таких методов классификации, как: деревья решений, метод ближайшего соседа и машины опорных векторов. Так как деревья решений дали наиболее точный результат среди этих методов, было решено использовать их.

Для наиболее точного представления последовательностей событий было произведено 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 второе событие наступило до первого
- = – 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 события
    - < – если первое событие наступило до второго
    - > – если второе событие наступило до первого
    - = – если события произошли в одном месяце
    - n – если ни одно из событий еще не наступило

Результаты

1. Предсказание пола и поколения

1.1. Демографические события

Таблица 1. Предсказание пола

	CA	Sensitivity	F1	Precision
Бинарная	0.6808	0.9566	0.8029	0.6918
Временная	0.6662	0.9176	0.7889	0.6919
Пары событий	0.6808	0.9566	0.8029	0.6918
Непосредств. предш.	0.6800	0.9554	0.8023	0.6916
Признаки+Бинарная	0.6785	0.9176	0.7952	0.7015
Признаки+Временная	0.6996	0.8605	0.7957	0.7400
Признаки+Пары событий	0.6684	0.8872	0.7844	0.7030
Признаки+Непоср. предш.	0.6783	0.8813	0.7884	0.7132

Тут по двум признакам лучшими являются бинарная таблица, таблица пар событий и временная таблица, учитывающая признаки. Посмотрим на все три дерева решений.

Бинарная:

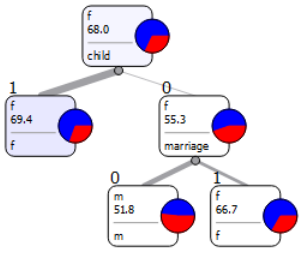

Рис. 1 Дерево решений, бинарная кодировка

Из дерева видим, что для женщин характерны событие ребенок (вероятность 69.4%) и событие брак без ребенка (вероятность 66.7%).

Для мужчин более вероятно отсутствие событий брак и ребенок (вероятность 51.8%).

Пары событий:

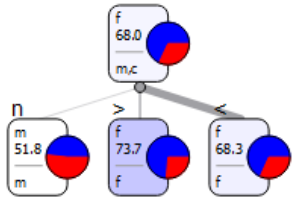

Рис. 2 Дерево решений, кодировка пары событий

У женщин с вероятностью 73.7% ребенок был до брака, а с вероятностью 68.3% брак до ребенка.

Как и в предыдущем дереве для мужчин более вероятно отсутствие событий брак и ребенок (вероятность 51.8%).

Признаки и временная кодировка:

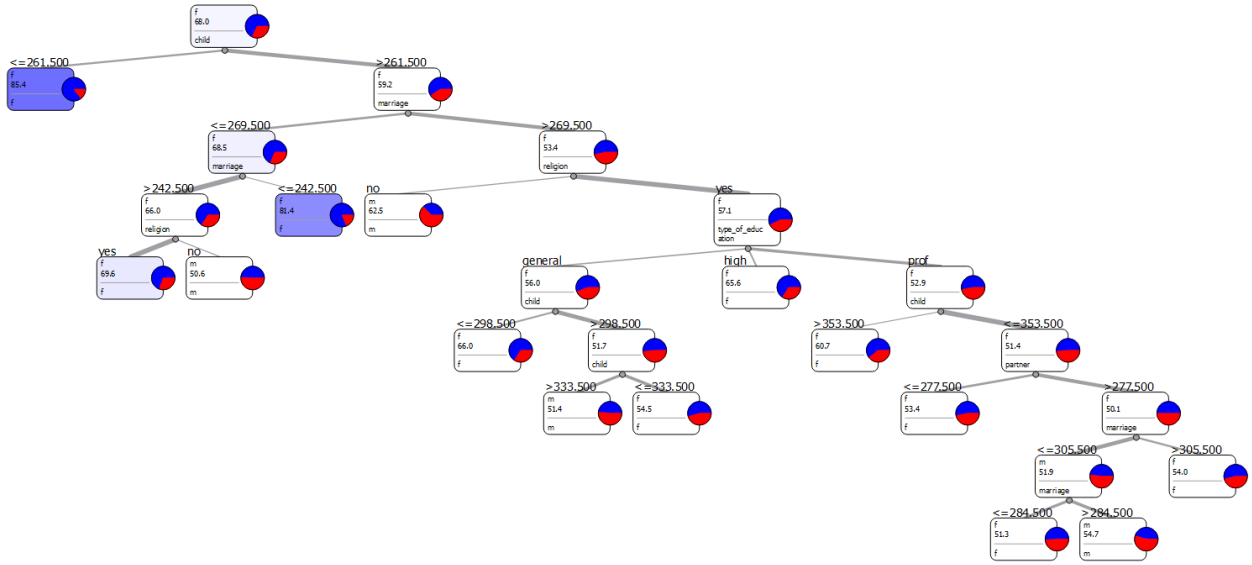

Рис. 3 Дерево решений, временная кодировка

Женщины:

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

Мужчины:

- Ребенок после 22, брак после 23, неверующий (вероятность 62.5%)
- Ребенок от 22 до 30 лет, партнер после 23, брак от 23 до 26, верующий, тип образования профессиональный (вероятность 54.7%)
- Ребенок после 22, брак после 23, ребенок после 27, верующий, тип образования общий (вероятность 51.4%)

- Брак от 20 до 23, ребенок после 22, неверующий (вероятность 50.6%)

Таблица 2. Предсказание поколения

	CA	Sensitivity	F1	Precision
Бинарная	0.2729	0.0667	0.1070	0.2707
Временная	0.2513	0.0680	0.1022	0.2058
Пары событий	0.2715	0.0667	0.1070	0.2707
Непоср. предш.	0.2764	0.0871	0.1335	0.2857
Признаки+Бинарная	0.2816	0.3497	0.3293	0.3111
Признаки+Временная	0.2602	0.1020	0.1461	0.2568
Признаки+Пары событий	0.2982	0.3619	0.3428	0.3256
Признаки+Непоср. предш.	0.3008	0.3469	0.3286	0.3121

По трем мерам точности лучший результат получается при рассмотрении всех признаков и событий с кодировкой «пары событий».

Таблица 3. Определение поколений, Orange

	1g	2g	3g	4g	5g	6g	
1g	266	54	253	99	47	16	**735**
2g	178	44	307	107	65	20	**721**
3g	161	52	571	252	147	39	**1222**
4g	118	34	449	264	198	46	**1109**
5g	69	23	275	193	216	57	**833**
6g	25	10	78	40	57	113	**323**
	817	**217**	**1933**	**955**	**730**	**291**	**4943**

Из таблицы выше можно заметить, что второе поколение плохо предсказывается, часто определяется в качестве третьего. Четвертое и пятое поколения часто принимается в качестве третьего. Лучше всего предсказываются первое и шестое поколения.

Посмотрим на дерево решений.

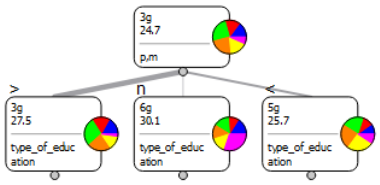

Рис.4 Дерево решений, кодировка пары событий

Посмотрим подробнее на каждую из ветвей дерева: партнер после брака, ни одно из этих событий не произошло и партнер до брака.

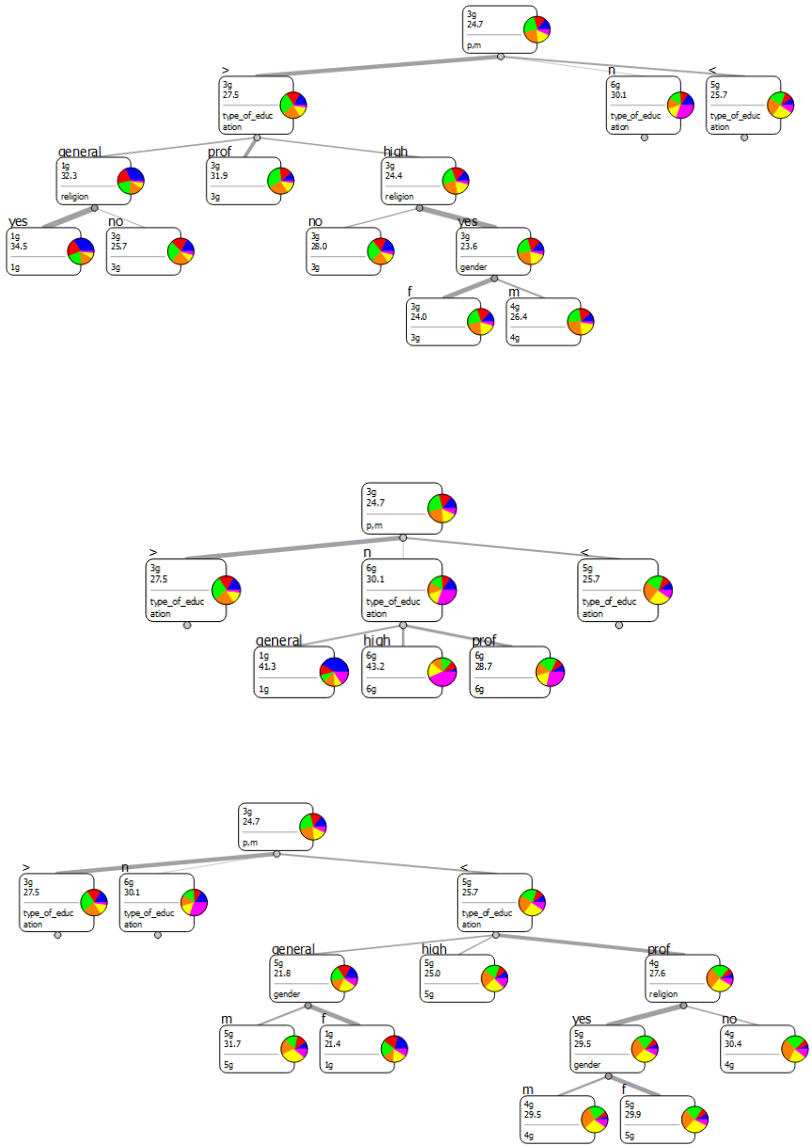

Рис. 5 Дерево решений, кодировка пары событий

Вероятности в дереве решений около 30 процентов, что мало, однако посмотрим на интересные закономерности в дереве.

1 поколение:

- Брак до партнера, тип образования общий, верующий (вероятность 34.5%)
- События брак и партнер еще не произошли, тип образования общий (вероятность 41.3%)
- Партнер до брака, тип образования общий, пол женский (вероятность 21.4%)

3 поколение:

- Брак до партнера, тип образования общий, неверующий (вероятность 25.7%)
- Брак до партнера, тип образования профессиональный (вероятность 31.9%)
- Брак до партнера, тип образования высший, неверующий (вероятность 28%)
- Брак до партнера, тип образования высший, верующий, пол женский (вероятность 24%)

4 поколение:

- Брак до партнера, тип образования высший, верующий, пол мужской (вероятность 26.4%)
- Партнер до брака, тип образования профессиональный, верующий, пол мужской (вероятность 29.5%)
- Партнер до брака, тип образования профессиональный, неверующий (вероятность 30.4%)

5 поколение:

- Партнер до брака, тип образования общий, пол мужской (вероятность 31.7%)
- Партнер до брака, тип образования высший (вероятность 25%)
- Партнер до брака, тип образования профессиональный, верующий, пол женский (вероятность 29.9%)

6 поколение:

- События брак и партнер еще не произошли, тип образования высший или профессиональный (вероятности 43.2% и 28.7% соответственно)

1.2. Социоэкономические события

Таблица 4. Предсказание пола

	CA	Sensitivity	F1	Precision
Бинарная	0.6800	1	0.8095	0.6800
Временная	0.6931	0.9280	0.8044	0.7098
Пары событий	0.6800	1	0.8095	0.6800
Непоср. предш.	0.6777	0.9961	0.8078	0.6794
Признаки+Бинарная	0.6800	1	0.8095	0.6800
Признаки+Временная	0.6925	0.9316	0.8047	0.7082
Признаки+Пары событий	0.6711	0.9548	0.7979	0.6852
Признаки+Непоср. предш.	0.6771	0.8625	0.7841	0.7188

Так как Признаки+Бинарная дает такие же точности, как и Бинарная, значит добавление признаков не влияет на точность. Рассмотрим дерево с временной кодировкой, так как оно дает более полный результат.

Временное:

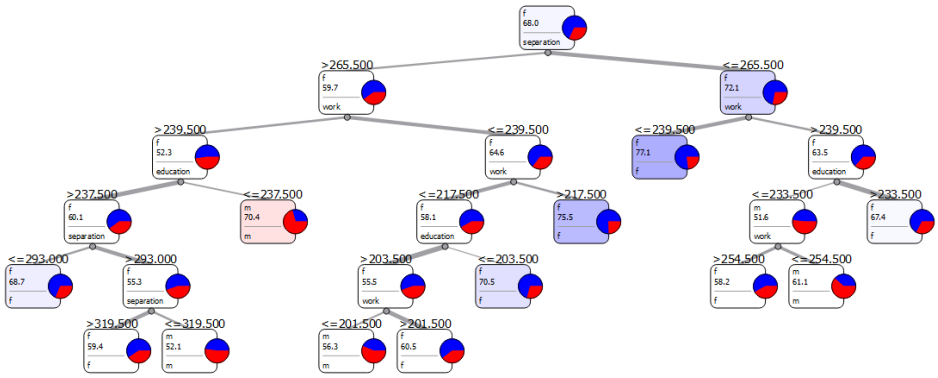

Рис. 6 Дерево решений, временная кодировка

Женщины:

- Работа до 20 лет, отделение до 22 (вероятность 77.1%)
- Работа от 18 до 19 лет, отделение после 22 (вероятность 75.5%)
- Образование до 17 лет, работа до 18, отделение после 22 (вероятность 70.5%)
- Работа после 19 лет, образование после 19, отделение от 22 до 25 (вероятность 68.7%)
- Образование после 19 лет, работа после 20, отделение до 22 (вероятность 67.4%)

Мужчины:

- Образование до 19 лет, работа после 19, отделение после 22 (вероятность 70.4%)
- Образование до 19 лет, работа после 20, отделение до 22, работа до 22 (вероятность 61.1%)
- Работа до 17 лет, образование после 17, отделение после 22 (вероятность 56.3%)

Таблица 5. Предсказание поколения

	CA	Sensitivity	F1	Precision
Бинарная	0.2626	0	-	-
Временная	0.2331	0.2898	0.2946	0.2996
Пары событий	0.2444	0	-	-
Непоср. предш.	0.2735	0.0463	0.0825	0.3820
Признаки+Бинарная	0.2840	0.3565	0.3548	0.3531
Признаки+Временная	0.2383	0.3347	0.3269	0.3195
Признаки+Пары событий	0.2650	0.3061	0.3210	0.3373
Признаки+Непоср. предш.	0.2798	0.3306	0.3425	0.3553

По трем мерам точности лучшей получилась таблица со всеми признаками с бинарным кодированием событий.

Таблица 6. Определений поколений, Orange

	1g	2g	3g	4g	5g	6g	
1g	261	34	279	130	31	0	**735**
2g	175	39	329	138	40	0	**721**
3g	147	40	647	297	84	7	**1222**
4g	99	39	550	314	89	18	**1109**
5g	61	17	380	239	115	21	**833**
6g	20	7	118	78	47	53	**323**
	763	**176**	**2303**	**1196**	**406**	**99**	**4943**

Первое поколение часто принимается в качестве третьего. Второе поколение очень плохо предсказывается, чаще всего принимается за третье или первое. Четвертое поколение часто принимается за третье. Пятое поколение чаще всего принимается за третье или четвертое. Шестое поколение чаще всего принимается за третье. То есть, получилось, что все поколения очень часто ошибочно принимаются за третье. Кроме того, по таблице третье поколение предсказывается лучше всего.

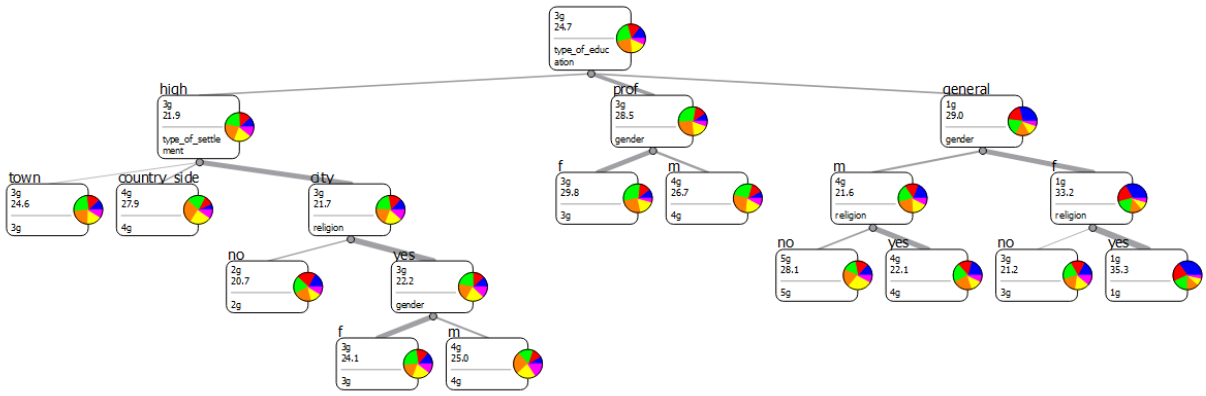

Рис. 7 Дерево решений, признаки и бинарная кодировка

1 поколение:

- Тип образования общий, пол женский, верующий (вероятность 35.3%)

2 поколение:

- Тип образования высший, место проживания город, неверующий (20.7%)

3 поколение:

- Тип образования профессиональный, пол женский (вероятность 29.8%)
- Тип образования высший, место проживания поселок городского типа (вероятность 24.6%)
- Тип образования высший, место проживания город, верующий, пол женский (вероятность 24.1%)
- Тип образования общий, пол женский, неверующий (вероятность 21.2%)

4 поколение:

- Тип образования профессиональный, пол мужской (вероятность 26.7%)
- Тип образования общий, пол мужской, верующий (вероятность 22.1%)
- Тип образования высший, место проживания деревня (вероятность 27.9%)
- Тип образования высший, место проживания город, верующий, пол мужской (вероятность 25%)

5 поколение:

- Тип образования общий, пол мужской, неверующий (вероятность 28.1%)

2. Предсказание первого события

2.1. Демографические события

Предскажем для каждой из таблиц первое событие на основе признаков.

Таблица 7. Меры точности, Orange

Method	CA	Sens	F1	Prec
Classification Tree	0.6565	0.0000	N/A	N/A

	ch_	mar_	mar_ch_	par_	par_ch_	
ch	0	340	0	37	0	**377**
mar	0	2905	0	133	0	**3038**
mar ch	0	10	0	0	0	**10**
par	0	1087	0	193	0	**1280**
par ch	0	6	0	8	0	**14**
	0	**4348**	**0**	**371**	**0**	**4719**

Первые события ребенок, брак и ребенок и партнер и ребенок предсказываются очень плохо. Однако таких первых событий очень мало (10 и 14 человек из 4719). Первое событие партнер часто принимается за брак. Событие брак предсказывается лучше всего.

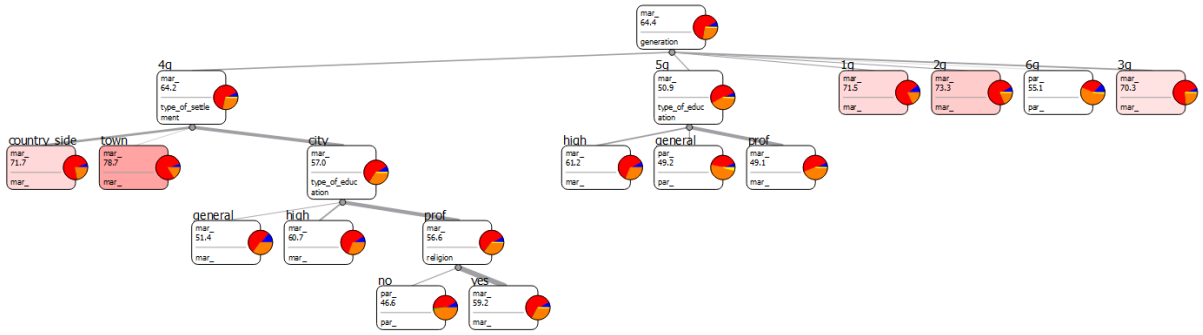

Рис. 8 Дерево решений, признаки

Первое событие больше всего зависит от поколения.

В 1, 2 и 3 поколении с вероятностями 71.5%, 73.3% и 70.3% первым событием является брак. В 6 поколении с вероятностью 55.1% первым событием является партнер.

В 4 поколении при проживании в деревне или поселке городского типа с вероятностями 71.7% и 78.7% соответственно первым событием будет брак, а при проживании в городе, если тип образования профессиональный и неверующий, то с вероятностью 46.6% первым событием будет партнер.

В 5 поколении при типе образования высшее и профессиональное первым событием с вероятностями 61.2% и 49.1% соответственно будет брак, а при типе образования общем, с вероятностью 49.2% первым событием будет партнер.

2.2. Социоэкономические события

Таблица 8. Меры точности, Orange

Method	CA	Sens	F1	Prec
Classification Tree	0.4283	0.3401	0.3802	0.4309

	edu_	sep_	sep_edu_	sep_wor_	sep_wor_edu_	wor_	wor_edu_	
edu	468	543	0	0	0	365	0	**1376**
sep	234	1079	0	0	0	341	0	**1654**
sep edu	11	24	0	0	0	12	0	**47**
sep wor	13	52	0	0	0	26	0	**91**
sep wor edu	4	8	0	0	0	2	0	**14**
wor	308	685	0	0	0	570	0	**1563**
wor edu	48	84	0	0	0	66	0	**198**
	1086	**2475**	**0**	**0**	**0**	**1382**	**0**	**4943**

Первые события отделение и образование, отделение и работа, отделение работа и образование, работа и образование в один и тот же месяц предсказываются очень плохо. События образование, отделение и работа предсказываются достаточно хорошо.

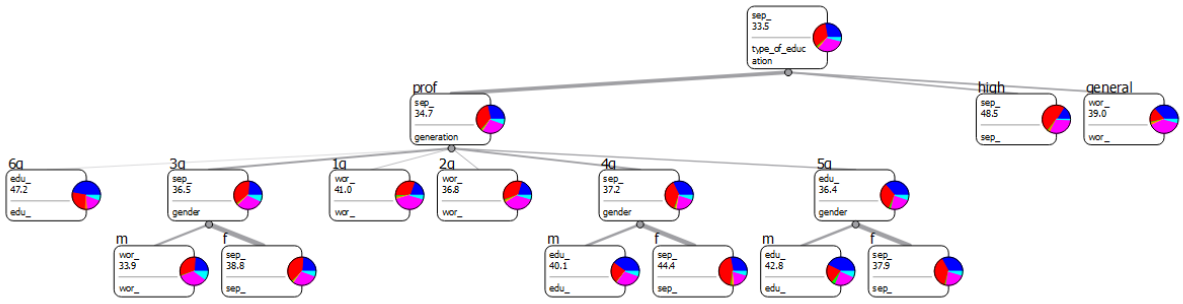

Рис. 9 Дерево решений, признаки

Первое событие больше всего зависит от типа образования.

При типе образования высшее, с вероятностью 48.5% первым событием будет отделение от родителей, а при типе образования общее, с вероятностью 39% первым событием будет работа.

При типе образования профессиональном, первое событие также зависит и от поколений. В 1, 2 и в 3 поколении у мужчин наиболее вероятным первым событием будет работа (с вероятностями 41%, 36.8% и 33.9% соответственно). В 4 поколении у мужчин, в 5 поколении у мужчин и в 6 поколении наиболее вероятным первым событием будет образование (с вероятностями 40.1%, 42.8% и 47.2% соответственно). В 3 поколении у женщин, в 4 поколении у женщин и в 5 поколении у женщин наиболее вероятным первым событием будет отделение от родителей (с вероятностями 38.8%, 44.4% и 37.9% соответственно).

3. Предсказание последнего события

3.1. Демографические события

Посмотрим, что дает большую точность: просто таблица с признаками или таблица с признаками и уже произошедшими событиями.

Таблица 9. Меры точности для признаков

	CA	Sensitivity	F1	Precision
Признаки	0.7542	1	0.8599	0.7542
Признаки+События	0.7830	1	0.8807	0.7869

Лучший результат дает таблица с признаками и уже произошедшими событиями, посмотрим на дерево.

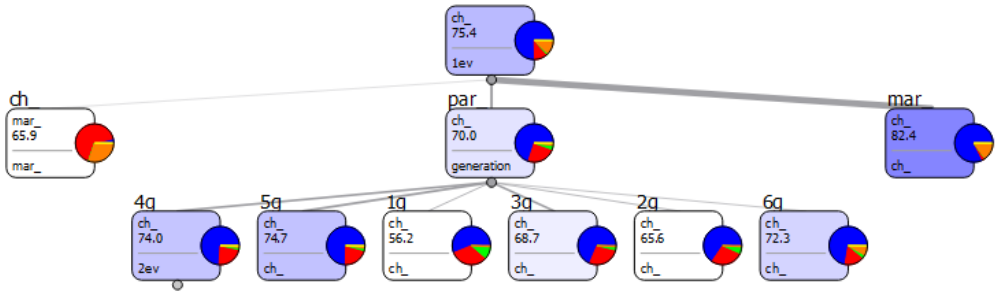

Рис. 10 Дерево решений, признаки и события

Если первым событием было ребенок, то с вероятностью 65.9% последним событием будет брак.

Если первым событием было брак, то с вероятностью 82.4% последним событием будет ребенок.

Если первым событием было партнер, то последнее событие зависит еще от поколения. В 1, 2, 3, 4, 5 и 6 поколении с вероятностями 56.2%, 65.6%, 68.7%, 74%, 74.7% и 72.3% соответственно последним событием будет ребенок.

3.2. Социоэкономические события

Посмотрим, что дает большую точность: просто таблица с признаками или таблица с признаками и уже произошедшими событиями.

Таблица 10. Меры точности для признаков

	CA	Sensitivity	F1	Precision
Признаки	0.4530	0.4499	0.4412	0.43429
Признаки+События	0.9421	0.9935	0.9469	0.9046

Лучший результат дает таблица с признаками и уже произошедшими событиями, посмотрим на дерево.

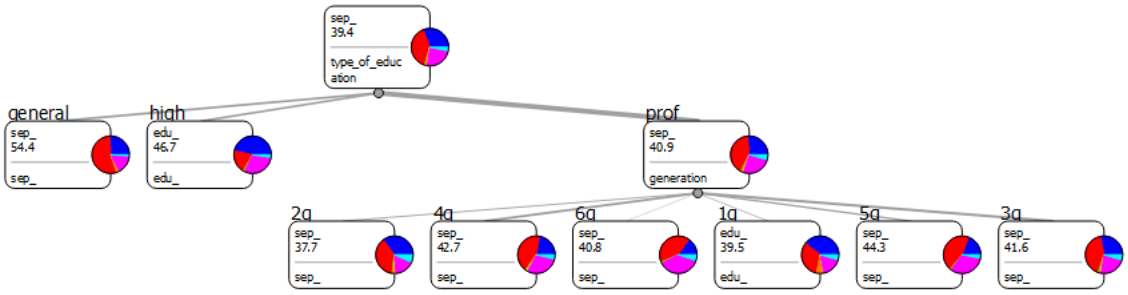

Рис. 11 Дерево решений, признаки и события

Если тип образования общий, то с вероятностью 54.4% последним событием будет отделение от родителей.

Если тип образования высший, то с вероятностью 46.7% последним событием будет образование.

Если тип образования профессиональный, то последнее событие также зависит от поколения. Для первого поколения с вероятностью 39.5% в этом случае последним событием будет образование, а

во 2, 3, 4, 5 и 6 с вероятностями 37.7%, 41.6%, 42.7%, 44.3% и 40.8% соответственно последним событием будет отделение от родителей.

Заключение

При предсказании пола и поколений для демографических и социоэкономических событий лучшие типы кодировок получились разными, то есть, нет одной кодировки, которая была бы лучшей для всех данных. На демографических данных для предсказания пола лучшей кодировкой получилась временная кодировка с рассмотрением всех признаков. Для предсказания поколения лучшей получилась кодировка пар событий с рассмотрением всех признаков. На социоэкономических данных для предсказания пола лучшей оказалась временная кодировка. Для предсказания поколений лучшей кодировкой оказалась бинарная с рассмотрением всех признаков. При предсказании последнего события как на демографических, так и на социоэкономических данных, лучший результат получился при рассмотрении всех признаков и уже произошедших событий.

Полученные результаты представляют интерес для демографов в качестве их дальнейшей интерпретации. Используемые методы для анализа данных могут быть использованы также для решения схожих задач в других областях.

Список литературы

1. 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.
2. 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.
3. 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.
4. 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.