Структурирование многофакторных статистических моделей на основе полиномов Жегалкина с приложением в медицине

Назиров Айдар Зуфарович, гр. 21.Б04-мм

Научный руководитель: кандидат физико-математических наук, доцент Алексеева Нина Петровна

Рецензент: специалист ФГБУ «НМИЦ ПН им. В. М. Бехтерева» Скурат Евгения Петровна

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Санкт-Петербург, 2025

Введение

Мотивация: При нынешних объемах медицинских данных потребность в разработке эффективных методов для их анализа и прогнозирования не теряет своей актуальности.

Цель работы: решение многофакторной задачи классификации — построение моделей для прогнозирования исхода по предоставленным данным.

Задачи:

- Исследование, реализация и оценка методов решения многофакторной задачи,
- Анализ полученных результатов и сравнение методов.

Предоставленные данные

Для работы были предоставлены медицинские данные о недоношенных младенцах, которые включают в себя:

- 84 наблюдения (индивида),
- 92 признака (60 категориальных и 32 количественных),
- целевую переменную — "исход" (выжил/умер).

№пп	Gestation_Early_VeryEarly	Одноплм ногопл	пол	АД_масса_тела_грам	...	Рб_ТС_ИВЛ_дни	Рб_Возраст_ден	Рб_АД_ИСХОД_80_yes_no
17	2	1	0	860	...	75	152	0
18	2	NA	0	790	...	7	73	0
83	1	1	0	NA	...	3	97	0
84	2	2	0	650	...	33	NA	NA
...	...	...	...	...	...	...	...	...
76	2	2	0	780	...	7	103	0

Figure: Фрагмент исходной таблицы

Метод I: теория

Определение (Информационное разнообразие)

Пусть $X = (x_1, \dots x_n)$ — выборка наблюдений дискретной величины с m градациями, каждая из которых встречается a_i раз, $a_1 + \dots + a_m = n$. Тогда информационное разнообразие этой выборки измеряется формулой:

$$I(X) = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

Определение (Информационный выигрыш от объединения)

Пусть $X = (x_1, \dots x_n), Y = (y_1, \dots y_m)$ — две выборки наблюдений. Тогда информационный выигрыш от объединения измеряется формулой:

$$\Delta(X, Y) = I(X, Y) - I(X) - I(Y).$$

Метод I: деревья классификации

Алгоритм построения деревьев:

1. Упорядочиваются значения признака, рассматриваются середины интервалов между ними, по каждой середине наблюдения разделяются на 2 группы,
2. Вычисляется информационный выигрыш от объединения полученных групп, и выбирается значение признака с наибольшим информационным выигрышем, тем самым получается ветвь дерева разбивающая наблюдения на 2 группы,
3. Алгоритм продолжается до тех пор, пока не будет достигнута требуемая точность (90%).

Метод I: результаты

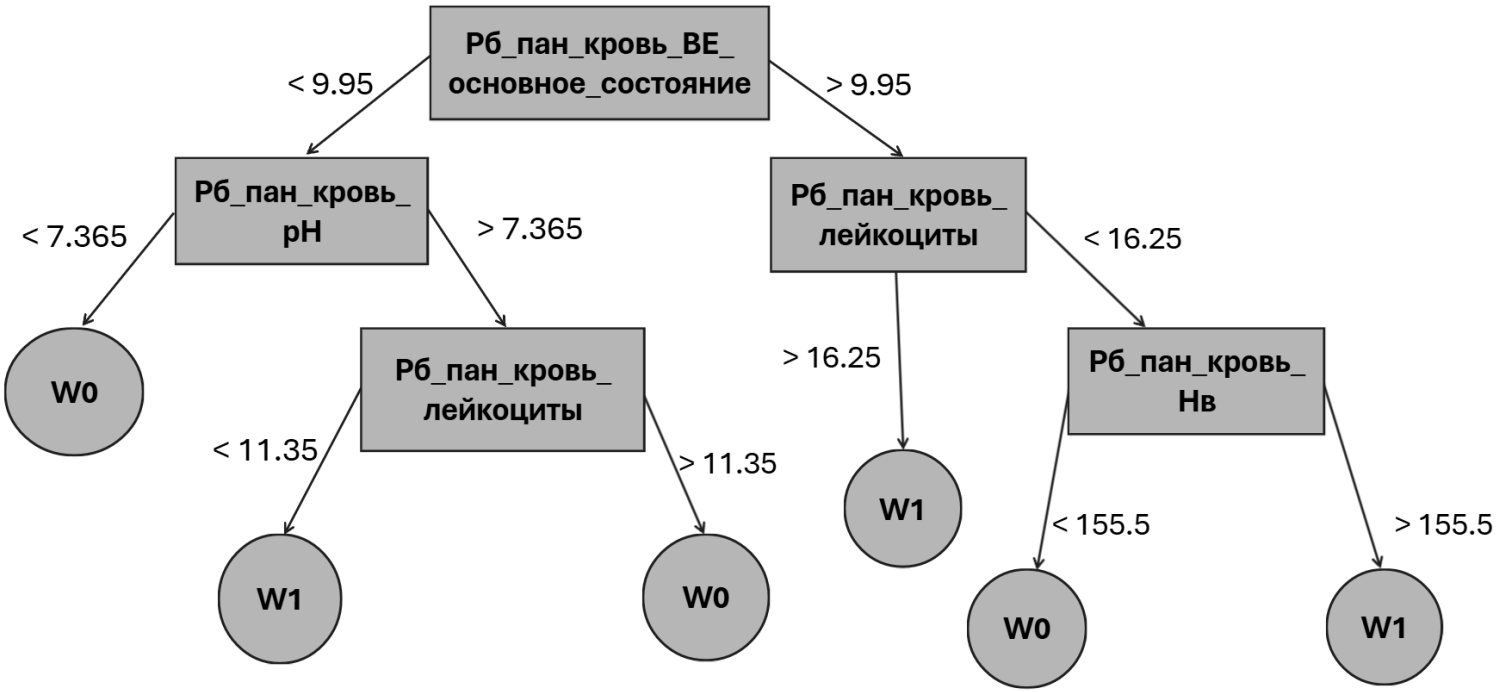


Figure: Дерево 1.1

Данное дерево имеет точность классификации равную 97,1%.

Интерпретация: явно представлены признаки и их значения влияющие на прогноз.

Метод I: результаты

Моделью классификации является система деревьев. Прогнозируемое значение целевой переменной — это среднее значение исхода для индивида, полученное по каждому дереву.

Table: Система классифицирующих деревьев

Дерево	Признаки	Наблюдения	Точность, %
1	4	69	97,1
2	3	62	93,5
3	4	57	98,2
4	3	68	91,1
5	6	67	95,5
6	6	83	92,7
7	7	74	91,8

Средняя точность классификации по дереву = 94,2%.

Метод I: результаты

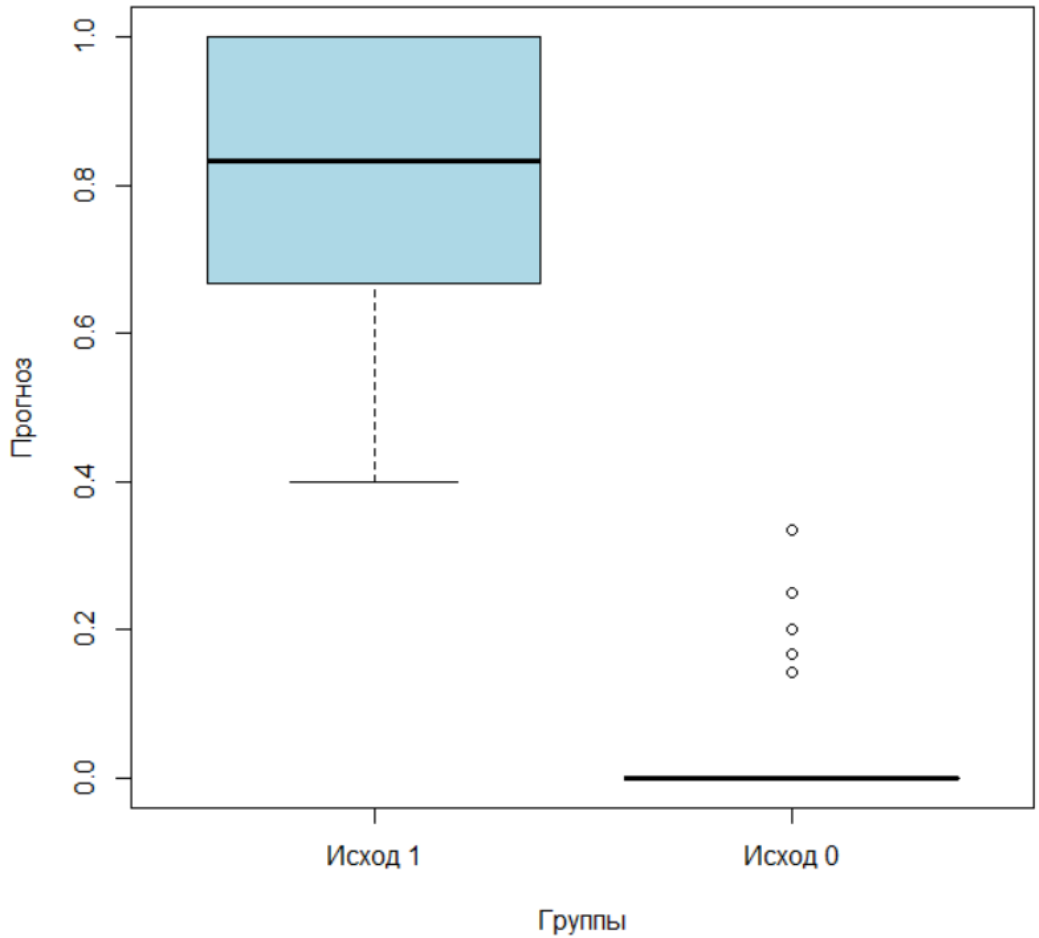


Figure: Прогноз по всей выборке

Метод II: теория

Определение (Полином Жегалкина)

Полином Жегалкина — полином с коэффициентами 0 и 1, где в качестве произведения стоит конъюнкция, а в качестве сложения — исключающее ИЛИ. Он имеет следующий вид:
$P = a_1 \oplus a_2 x_1 \oplus \ldots \oplus a_{n+1} x_n \oplus a_{n+2} x_1 x_2 \oplus \ldots \oplus a_m x_1 x_2 \ldots x_n$

Порядок параметризации симптомов полиномами определяет количество признаков, включаемых в один моном. В работе использованы полиномы степени не выше 3–й, и обусловлено это:

- предоставлением возможности моделирования нелинейных взаимосвязей между факторами,
- экспоненциальным ростом числа полиномов относительно используемых признаков.

Метод II: деревья по симптомам

Алгоритм построения деревьев аналогичен первому методу, однако, вместо простых признаков используются их логические комбинации парметризуемые полиномами (симптомами).

Процесс построения базы симптомов:

1. Строим всевозможные полиномы по всем комбинациям 3–х переменных, в качестве которых выступают исходные признаки,
2. Определяем наилучший прогнозирующий полином по информационному выигрышу от объединения для каждой комбинации.

Пример дерева классификации II метода

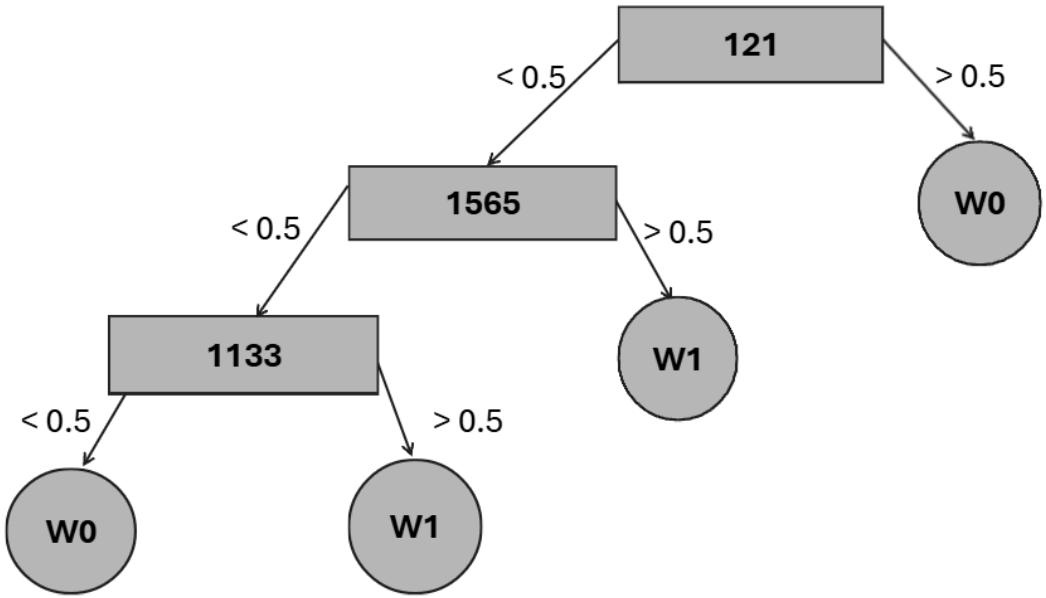


Figure: Дерево 2.1

Данное дерево имеет точность классификации равную 97,1%.

Интерпретация: $P_{121} = z \oplus yz \oplus xyz$ — симптом "угрозы" при значении признака $z = 0$ или значениях признаков x и $z = 1$.

Метод II: результаты

Table: Характеристики деревьев

Дерево	Признаки	Наблюдения	Точность, %
1	3	76	98,6
2	4	64	98,4
3	5	59	100,00
4	4	64	98,4
5	4	63	100,00
6	4	58	100,00
7	5	72	100,00

Средняя точность классификации по деревьям получилась равной 99,3%, что оказывается на $\approx 5\%$ выше, чем у деревьев первого метода.

Метод II: результаты

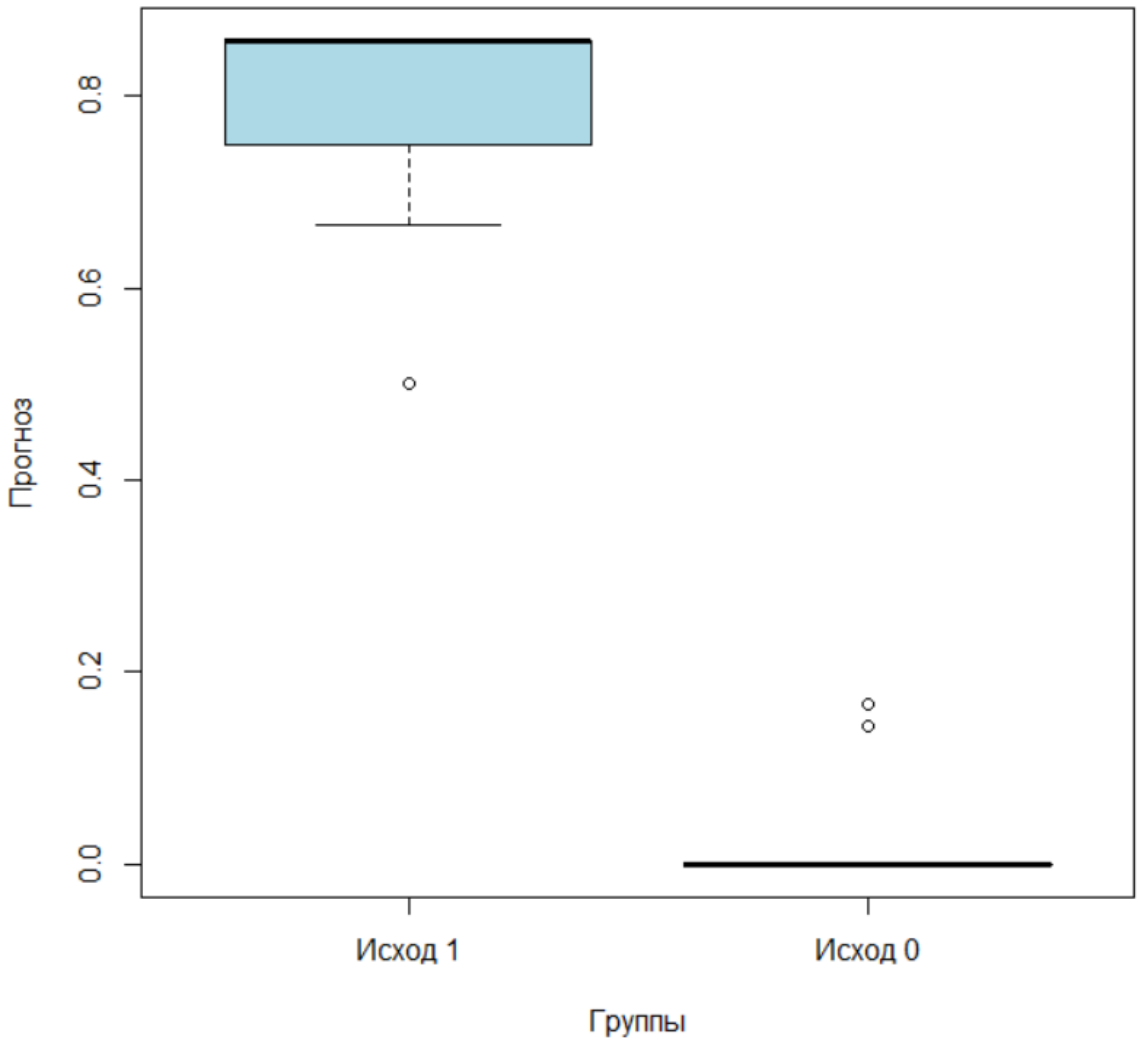


Figure: Прогноз по всей выборке

Кросс-валидация: определение

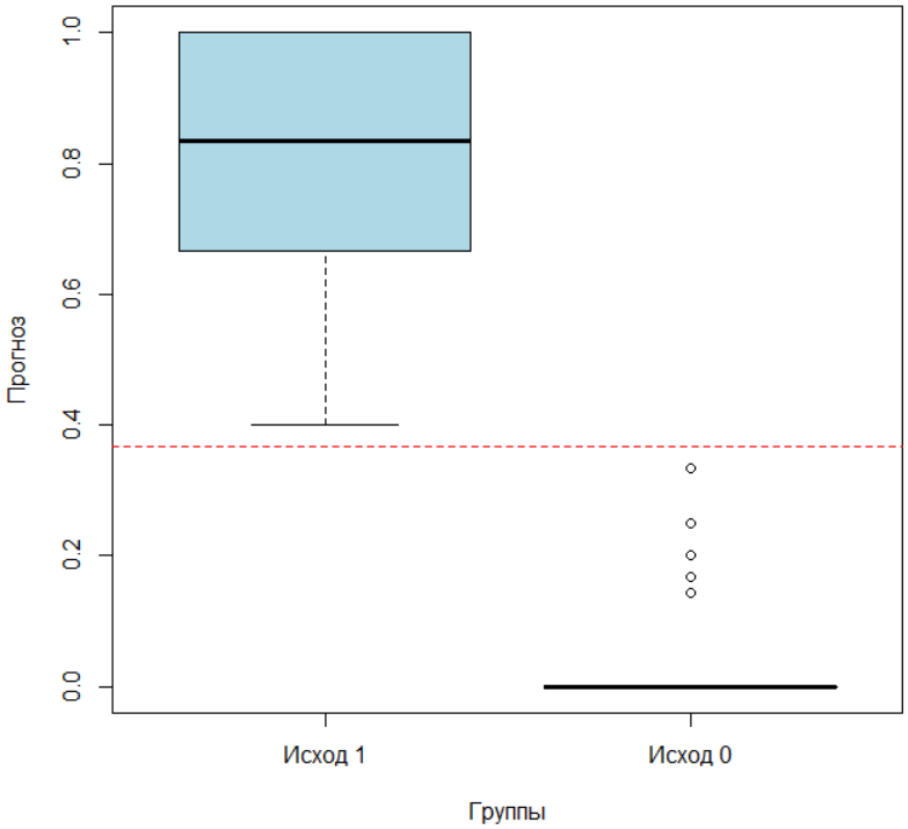
Определение (Стратифицированная кросс–валидация)

Стратифицированная кросс–валидация — метод оценки производительности модели, при котором в обучающей и тестовой выборках сохраняется одинаковое распределение исследуемой переменной.

Она была выбрана для оценки методов, так как количество наблюдений в разных классах исхода значительно различается.

Для применения кросс-валидации исходная выборка была разделена на обучающую и тестовую в соотношении $\approx 4{:}1$.

Кросс-валидация: результаты



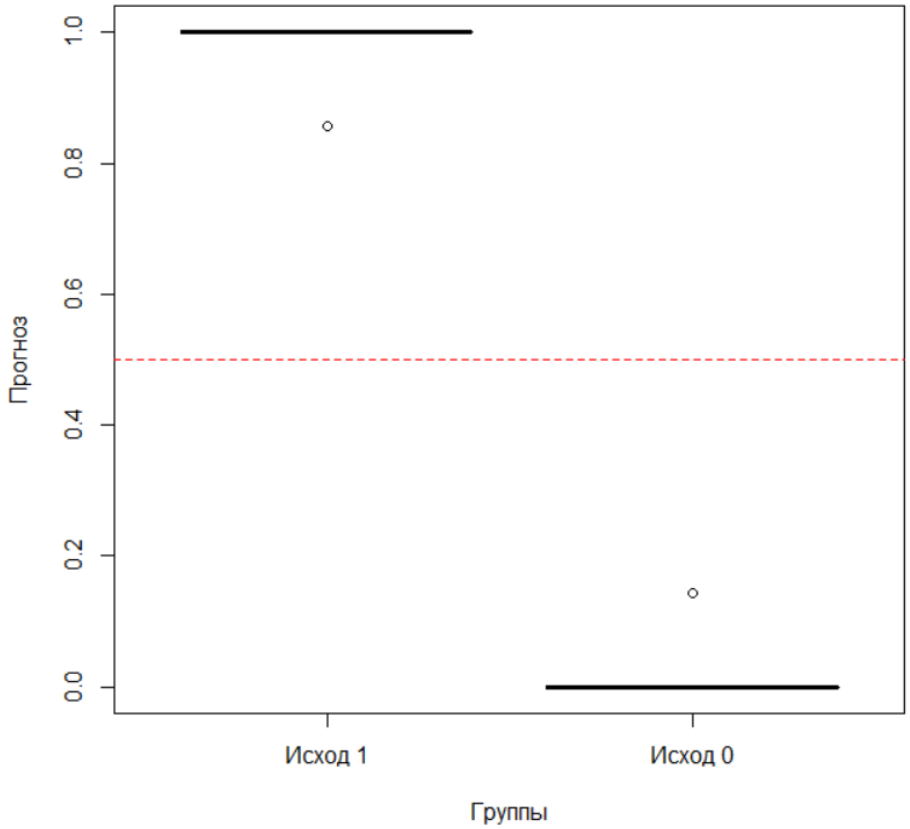


Figure: Полученные модели по обучающим выборкам. Слева представлена модель по первому методу, справа — по второму.

Сравнение методов между собой

- Первый метод, в отличие от второго работает как с категориальными, так и с количественными переменными,
- Несмотря на меньшую размерность данных, второй метод сохраняет точность.
- Структура деревьев модифицированного метода оказывается проще: среднее число симптомов в деревьях II метода 3,2 против 4,7 признаков в деревьях I метода, но сложнее в интерпритации по причине использования симптомов (комбинаций).

Сравнение с известными методами классификации

Table: Сравнение методов

Метрика	Метод I	Метод II	LR	RF
Accuracy,%	100	100	96,38	100
Sensitivity,%	100	100	88,88	100
Specificity,%	100	100	98,46	100
Интерпрети-руемость	Высокая	Средняя	Высокая	Средняя
Данные	Любые	Категори-альные*	Количест-венные*	Любые
Устойчивость мультикол-линеарности	Высокая	Высокая	Низкая*	Высокая

Заключение

Полученные результаты:

- Разработана программа для реализации предложенных методов классификации на языке R (код содержит порядка 1500 строк),
- Произведена оценка полученных моделей по обоим методам с помощью кросс-валидации,
- Осуществлено сравнение рассмотренных методов между собой и с уже известными способами решения задач классификации.