**Регрессия, как инструмент контролируемого искусственного интеллекта**

В одной из статей *Forbes* рассказывается о том, что компания *Target* создала модель искусственного интеллекта (ИИ). Модель способна определить, является ли покупательница беременной, чтобы затем использовать эту информацию для целевого маркетинга товаров и предложений, связанных с ожиданием и рождением ребенка.[[1]](#footnote-1) ИИ способен найти применение практически в любой современной отрасли. Компания, выпустившая вашу кредитку, использует его, чтобы определить несвойственные вам операции. Враг в вашей любимой «стрелялке» на Xbox работает на ИИ. Фильтрация спама, обнаружение налогового мошенничества, автокоррекция орфографии и рекомендации друзей в социальных сетях — все это тоже ИИ.

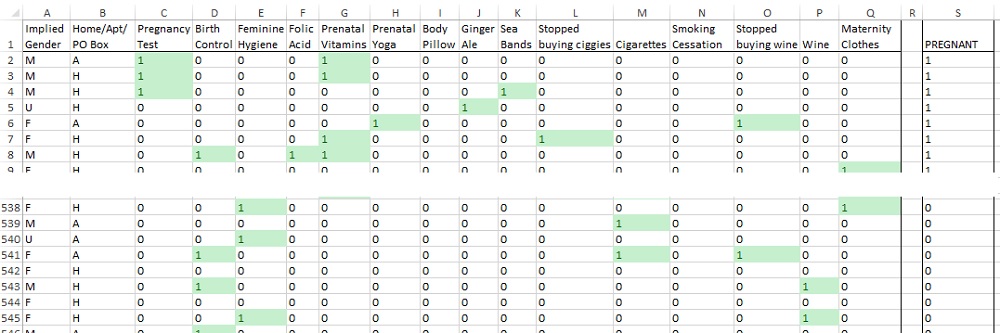


Рис. 1. Исходные данные для обучения

Но все же модель ИИ — не волшебница. Она просто превращает данные о произошедших событиях в формулу или набор правил, которые затем использует для предсказания событий в будущем. Как мы видели на примере [наивного Байеса](http://baguzin.ru/wp/?p=13844), эффективной модель делает ее способность вспоминать эти данные и ассоциированные с ними правила принятия решения, а также вероятности и коэффициенты. В основе ИИ лежат регрессионные модели.

### Определение беременных покупателей РитейлМарта с помощью линейной регрессии

Представьте, что вы — менеджер по рекламе. Из данных фокус-групп вы знаете, что у супругов, ожидающих прибавления, и родителей новорожденных покупка детских товаров легко входит в привычку. Вы хотите, чтобы РитейлМарт стал для этих людей магазином детских товаров №1. Но для этого необходимо, чтобы родители увидели вашу рекламу до того, как купят свою первую упаковку подгузников где-то еще. Вам нужна прогностическая модель, способная определить потенциально беременных покупательниц для дальнейшего целевого маркетинга.

У вас есть данные о покупках, и вам нужны маркеры, определяющие семьи, в которых ждут пополнения. Например, тест на беременность. Если покупательница заказывает тест, то ее вероятность оказаться беременной выше, чем у среднестатистического клиента онлайн-магазина. Такие признаки часто называются *features* — отличительные черты, или независимые переменные, ведь, то, что мы пытаемся определить — «Беременна (да/нет)?» является зависимой переменной, в том смысле, что ее значение будет зависеть от данных, содержащихся в независимых переменных (всех покупках), которые мы и загружаем в модель. Таких маркеров может быть несколько десятков. Все они не идеальны, но их совокупность вполне может сработать.

**Сборка обучающих данных.** По данным проведенных компанией опросов, 6% семей покупателей РитейлМарта ждут прибавления. Вам нужно выбрать несколько образцов из этой группы в базе данных РитейлМарта и собрать из их историй заказов признаки для модели, причем еще до того, как родятся дети. Таким же образом вы должны собрать эти признаки для примеров покупателей, не ожидающих прибавления. Набрав с помощью этих отличий по пачке семей ожидающих и не ожидающих детей, вы можете их использовать как примеры для обучения модели ИИ. Таким образом, просматривая историю заказов покупателей до первой покупки подгузников, можно выбрать отличительные признаки «беременных» семей. Представьте, что вы выбрали 500 таких семей и извлекаете информацию об их отличиях из базы данных РитейлМарта. Вы также выбрали 500 «не-беременных» семей (рис. 1).

Теперь вы знаете, что беременными в любой момент являются всего 6%, хотя в обучающей последовательности, это соотношение – 50/50. Это называется выборкой с запасом. Главное – найти точку пересечения «очков» по беременности, выданных моделью, между положительными и ложноположительными результатами. В первых двух столбцах обучающих данных стоят категорийные данные пола и типа адреса. Остальные признаки бинарны. Так, например, взглянув на первую строку таблицы, вы увидите, что этот покупатель считается беременным (столбец S). Это как раз тот столбец, значения в котором и должна научиться предсказывать ваша модель. Заглянув в историю заказов этого покупателя, вы найдете в ней тест на беременность и пару банок витаминов для беременных. Также эти семьи не заказывали в последнее время вино и сигареты.

Пролистав данные, вы увидите все типы покупателей: у кого-то множество индикаторов, а у кого-то всего пара. Как и ожидалось, семьи в ожидании детей все же покупают время от времени вино и сигареты, а бездетные — заказывают товары, ассоциирующиеся с беременностью.

**Создание фиктивных переменных.** Для начала вам нужно превратить категорийные данные в числовые. Обычно используется техника под названием *фиктивная переменная*: берется один столбец с категорийными данными (Implied Gender — пол) и разбивается на несколько бинарных столбцов. Вместо одного столбца появляется три: один для мужчин, другой для женщин и третий — для не указавших свой пол. Если в исходном столбце значение ячейки было «М», то теперь вместо этого у нас есть 1 в столбце Male, 0 в столбце Female и 0 в столбце Unknown. На самом деле третий столбец избыточен, потому что если и в столбце Male, и в столбце Female у вас 0, то уже подразумевается, что пол не указан. При использовании фиктивной переменной для кодирования категорийных данных вам всегда нужно на один столбец меньше, чем количество имеющихся у вас категорий (рис. 2).

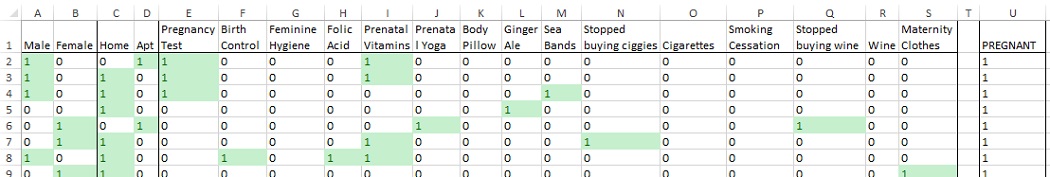


Рис. 2. Категорийные переменные в исходных данных заменены на фиктивные переменные

Если вы никогда не сталкивались с регрессией, рекомендую, прежде, чем двигаться дальше, изучить заметку [Простая линейная регрессия](http://baguzin.ru/wp/?p=6078).

Подготовим исходные данные для проведения линейной регрессии (рис. 3). На новый лист в Excel вставьте данные с листа «Рис. 2», начиная со столбца В и строки 8 — нужно оставить место для коэффициентов линейной модели и других оценочных данных, за которыми вы будете следить. В столбце U добавьте заголовок Intercept (точка пересечения с осью OY), потому что ваша линейная модель будет начинаться не в нуле. Более того, чтобы было проще вставить в модель свободный член, заполните столбец Intercept (U8:U1007) единицами. Это позволит вам обсчитать модель, применяя функцию СУММПРОИЗВ к строке коэффициентов и строке данных, что непременно потребует участия свободных членов. Все коэффициенты модели мы поместим в строку 2, так что озаглавьте ее Model Coefficients и запишите стартовые значения, равные единице, в каждую ее ячейку. Вы также можете воспользоваться условным форматированием строки коэффициентов, чтобы увидеть изменения, когда они появятся. После установки коэффициентов в строке 2 можно вычислить линейную комбинацию коэффициентов (формула СУММПРОИЗВ) со строкой данных покупателя и получить прогноз о беременности.

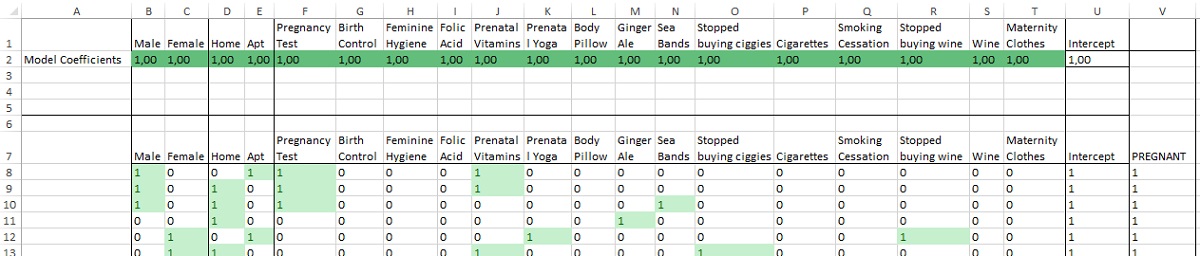


Рис. 3. Настройка таблицы для линейной модели

Получилось слишком много столбцов, чтобы рисовать график линейной регрессии, поэтому обойдемся без визуализации. Добавьте столбец W с прогнозом, и столбец Х с квадратом отклонения, который есть не что иное, как квадрат расстояния от значения Pregnant (беременный, столбец V) до прогнозируемого значения (столбец W). Возведение в квадрат позволяет каждому отклонению быть положительным, чтобы вы смогли затем сложить их все вместе и получить общее отклонение для модели (ячейка Х1).

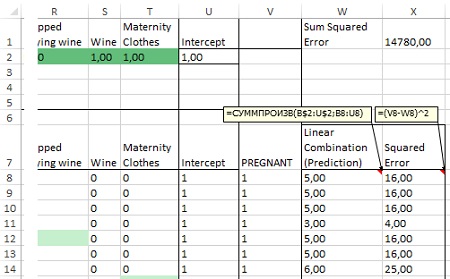


Рис. 4. Прогнозы и сумма квадратов отклонений

**Обучение *Поиском решения*.** Запустите *Поиск решения* и заставьте компьютер найти для вас лучшие коэффициенты. Вашей целевой функцией будет сумма квадратов отклонений в ячейке Х1, которую вы хотите минимизировать, изменяя значения переменных в ячейках В2:U2. Поскольку квадрат отклонения — это квадратичная функция, поэтому для решения вы не можете воспользоваться линейным симплекс-методом. Симплекс-метод работает очень быстро и гарантированно находит наилучшее решение, но модель понимает только линейные комбинации решений. Вам потребуется эволюционный алгоритм. Чтобы эффективно использовать эволюционный алгоритм, нужно установить нижнюю и верхнюю границы. Чем ближе они друг к другу (но не слишком близко!), тем лучше работает алгоритм. Я установил их на «–1» и «1» (рис. 5). Кликните на параметры, и установите *Максимальное время (в секундах)* = 600.

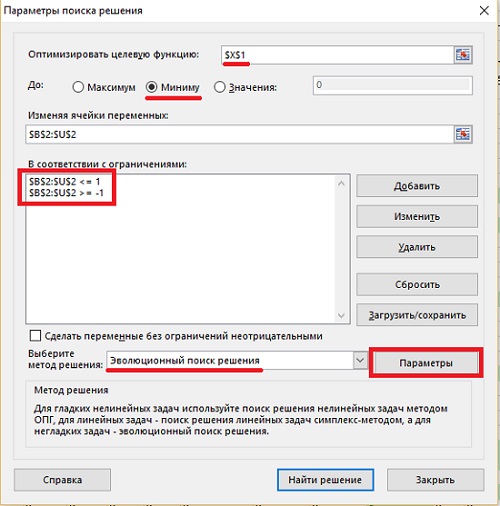


Рис. 5. *Поиск решения*, настроенный для регрессионной модели

Нажмите *Найти решение*. Пока *Поиск решения* пробует различные коэффициенты для модели, вы увидите, как меняется значение целевой функции (в самой нижней строке окна Excel). По окончании работы *Поиск решения* сообщит, что задача оптимизирована. Нажмите *ОК* и возвращайтесь к модели. В нашем примере *Поиск решения* закончил с результатом 139,2 (рис. 6). Имейте в виду: два разных запуска эволюционного алгоритма приводят к разным значениям.

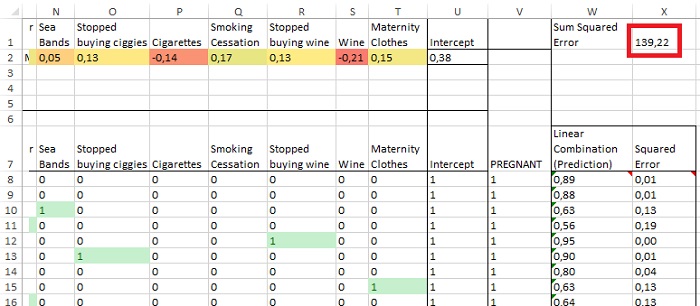


Рис. 6. Оптимизированная регрессионная модель

Пару замечаний. У Excel есть функция линейной регрессии – ЛИНЕЙН. Она вполне может сделать ту же работу, что и *Поиск решения*. Однако при 64 параметрах она вылетает, так что действительно большие регрессии вам все равно придется делать самим. В случае с линейной регрессией выбросы вашего обучающего набора данных довольно ощутимо «оттягивают» на себя модель в процессе подгонки тренда. Можно применить медианную регрессию, которая минимизирует сумму абсолютных значений отклонений, вместо суммы их квадратов. Просто замените квадрат отклонения в столбце Х абсолютным значением (функция ABS).

Если взглянуть на прогнозы в столбце W, то, на первый взгляд, они кажутся вполне похожими на правду. Например, беременная покупательница в строке 27, заказавшая тест на беременность, витамины для беременных и одежду для будущих мам, получила 1,06 балла, а покупатель из строки 996, который заказывал только вино, получил 0,18. Могут возникнуть вопросы:

* насколько регрессия в действительности подходит данным в количественном смысле, а не приблизительно?
* Все это — лишь случайное совпадение или статистически значимый результат?
* Какой вклад вносит каждый отличительный признак в конечный результат?

Чтобы ответить на эти вопросы о линейной регрессии, вам нужно рассчитать R-квадрат, критерий Фишера и критерий Стьюдента для каждого из коэффициентов модели.

**R-квадрат — оценка качества подгонки.** Если бы вы ничего не знали о покупателе из набора обучающих данных (не было бы столбцов от В до Т), но сделать прогноз о беременности все равно было бы необходимо, то вы бы получили следующие данные. При распределении 500/500 среднее (столбец W) получилось бы равным 0,5. А так как каждое отдельное значение — это либо 0, либо 1, то каждое отклонение равнялось бы 0,5, а его квадрат — 0,25 (столбец X). Для 1000 прогнозов эта стратегия прогнозирования среднего дает сумму квадратов, равную 250. Excel имеет функцию для расчета суммы квадратов отклонений КВАДРОТКЛ. Поместим следующую формулу в ячейку Х2: =КВАДРОТКЛ(V8:V1007).

Но если сложение значений по каждому прогнозу даст нам сумму квадратов отклонений, равную 250, то сумма квадратов отклонений, вычисленная с помощью линейной модели, гораздо меньше — всего 139,2. Т.е., 250 – 139,2 = 110,8 – есть сумму квадратов с выясненным происхождением (они обусловлены моделью), а происхождение отклонений в размере 139,2 остается невыясненным, даже после подгонки регрессии (в таком контексте сумма квадратов отклонений часто называется остаточной суммой квадратов).

Насколько хороша наша модель? Ответ на этот вопрос можно получить, взглянув на отношение «понятной» суммы квадратов к общей сумме. Это значение называется R-квадрат. Мы можем вычислить его в Х4: =Х3/Х2 = 0,44 (рис. 7)

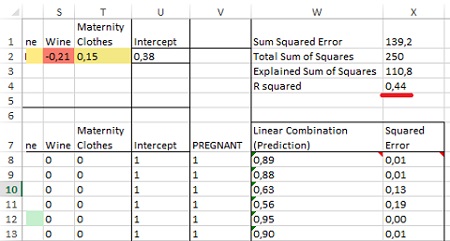


Рис. 7. R-квадрат линейной регрессии

Если модель подогнана идеально, то квадрат отклонения будет равен 0, «понятная» сумма квадратов — равняться общей, а R-квадрат будет в точности равен 1. Если модель совсем не подходит, R-квадрат будет ближе к 0. Не стоит забывать, что расчет R-квадрата работает только для нахождения линейных отношений между данными.

**Критерий Фишера: статистическая значимость подгонки.** R-квадрат показывает, насколько хорошо подогнана модель. Но он не дает никакой информации о статистической значимости подгонки. То, что ваша модель подходит — может быть простой случайностью. Такое предположение называется нулевой гипотезой. Обычно нулевая гипотеза отвергается, если есть вероятность подогнать модель хотя бы на 5%. Эта вероятность часто называется *величиной p*. Для ее вычисления мы проводим проверку на критерий Фишера. Мы берем три параметра модели и создаем с их помощью распределение вероятности (подробнее см. [Нормальное распределение. Построение графика в Excel. Концепция шести сигм](http://baguzin.ru/wp/?p=1170)). Вот эти три параметра:

* количество коэффициентов модели — в нашем случае 20 (19 отличительных черт плюс свободный член);
* степени свободы — это количество вариантов данных за вычетом количества коэффициентов модели;
* F-статистика — отношение «понятной» части суммы квадратов к «непонятной» (Х3/Х1 в таблице), умноженное на отношение количества степеней свободы к количеству зависимых переменных.

Чем больше F-статистика, тем ниже вероятность нулевой гипотезы. В ячейке Z1 (рис. 8) подсчитаем все значения из столбца V: =СЧЁТ(V8:V1007). В ячейке Z2 подсчитаем количество коэффициентов, суммируя строку 2: =СЧЁТ(B2:U2). В ячейке Z3 вычислим количество степеней свободы путем вычитания количества коэффициентов модели из количества записей: =Z1-Z2. В ячейке Z4 подсчитаем F-статистику, которая равна отношению «понятной» части суммы квадратов к «непонятной» (Х3/Х1), умноженному на отношение количества степеней свободы к количеству зависимых переменных (Z3/(Z2-1)): =(X3/X1)\*(Z3/(Z2-1)). В ячейке Z5 определим F-распределение вероятности: =F.РАСП(Z4;Z2-1;Z3;ЛОЖЬ). Вероятность получения такой подгонки, если принять нулевую гипотезу, равна нулю. Таким образом, нулевая гипотеза может быть отброшена, а произведенная подгонка — признана статистически значимой.

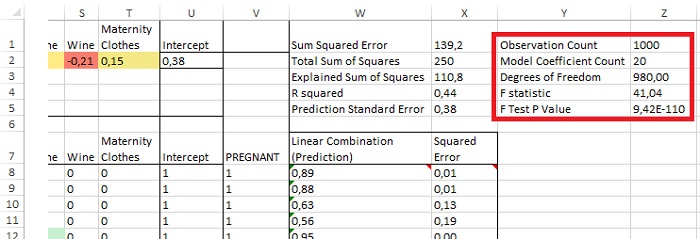


Рис. 8. Результат проверки критерия Фишера

**Проверка коэффициента Стьюдента (Т-тест) — какие переменные являются значимыми?** В Excel есть встроенные функции t-теста. Однако, чтобы понять, как работает эта статистика, полезно провести анализ «руками». Вы выполните [перемножение и преобразование матриц](https://ru.wikipedia.org/wiki/%D0%A3%D0%BC%D0%BD%D0%BE%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5_%D0%BC%D0%B0%D1%82%D1%80%D0%B8%D1%86). В Excel умножение матриц выполняется с помощью функции МУМНОЖ, а преобразование – с помощью МОБР (эти формулы являются формулами массива).

Если критерий Фишера проверял, насколько значима вся наша регрессия целиком, то Т-тест – есть способ проверить значимость отдельных переменных. Статистически незначимые переменные могут быть удалены. Проверка коэффициентов модели на значимость называется критерием Стьюдента, или t-тестом. Во время его выполнения, точно так же, как и во время проверки критерия Фишера, вы предполагаете, что данный коэффициент модели абсолютно бесполезен и должен быть равен 0. Приняв это предположение, t-тест рассчитывает вероятность получения коэффициента настолько далекого от 0, насколько вы видите в конкретной записи.

Для начала рассчитайте среднеквадратичное отклонение прогноза в ячейке Х5 (см. рис. 8), как квадратный корень из суммы квадрата отклонений (Х1), разделенный на количество степеней свободы (Z3): =КОРЕНЬ(X1/Z3) = 0,38. С помощью этой величины можно рассчитать стандартное отклонение каждого коэффициента. Сложным процесс вычисления стандартного отклонения делает необходимость понимать, как обучающие данные для коэффициентов варьируются и сами по себе, и взаимодействуя с другими переменными. Первый шаг в устранении этой неприятности — превращение обучающей последовательности в одну гигантскую матрицу (часто называемую матрицей плана) путем умножения ее на саму себя. Создайте новый лист в Excel (рис. 9). На нем показано произведение матрицы плана (диапазон B8:U1007 с рис. 8 или 12) самой на себя, что образует матрицу суммы квадратов и векторных произведений (СКВП). Во всем диапазоне В2:U21 используется одна и та же формула: {=МУМНОЖ(ТРАНСП('Рис. 12'!B8:U1007);'Рис. 12'!B8:U1007)}.

Обратите внимание на значения в матрице СКВП. По диагонали вы считаете совпадения каждой переменной с самой собой — то же, что и простое сложение единиц из каждого столбца матрицы плана. Например, в ячейке U21 свободный член равен 1000, потому что в исходных обучающих данных этот столбец состоит из 1000 ячеек.

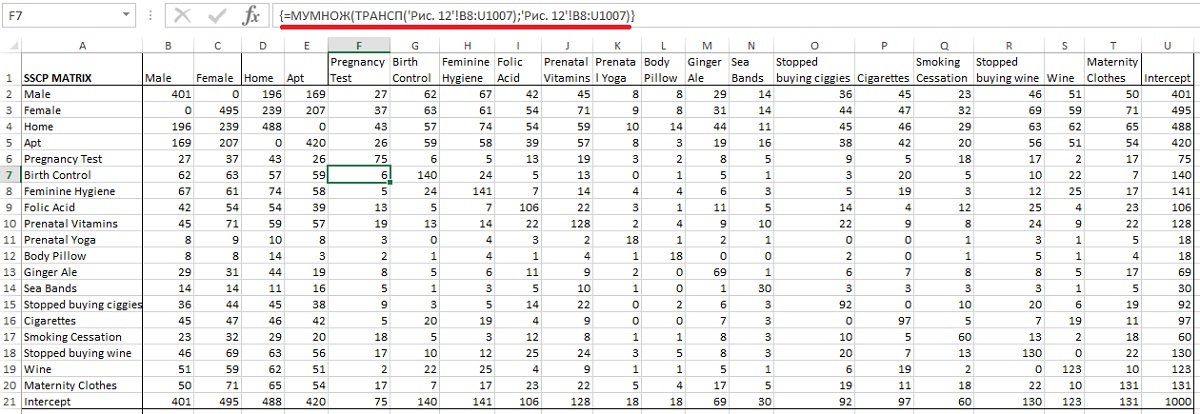


Рис. 9. Матрица суммы квадратов и векторных произведений (СКВП)

В ячейках, не входящих в диагональ, вас интересует число совпадений разных признаков. И если «мужчина» и «женщина» точно ни у кого не попадутся одновременно (В3=С2=0), то тест на беременность и противозачаточные появляются вместе в шести покупательских строках наших данных (F7). Матрица СКВП дает нам представление о величине каждой переменной и то, насколько они пересекаются и соотносятся между собой.

Вычисление стандартного отклонения коэффициента требует преобразования матрицы СКВП (рис. 10). Создайте таблицу ниже области, изображенной на рис. 9. Во всех ячейках диапазона В25:U44 используется формула массива: {=МОБР(B2:U21)}.

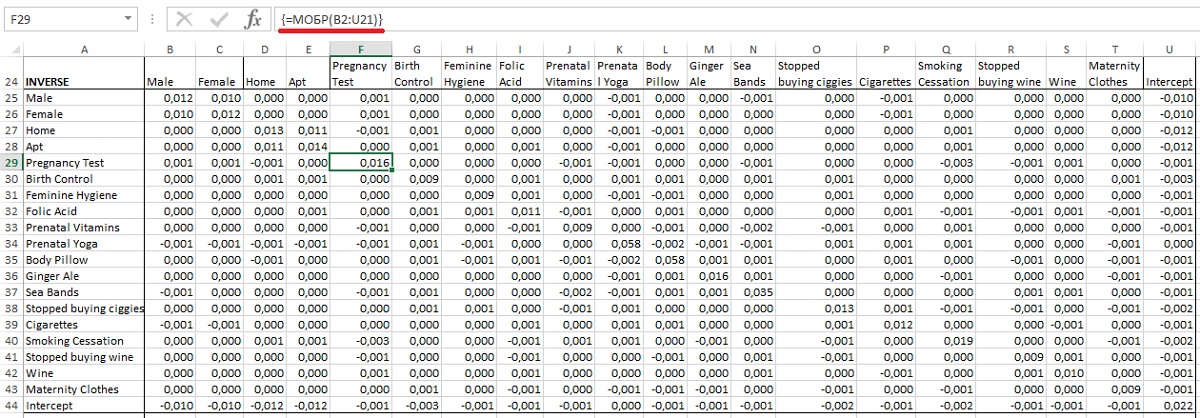


Рис. 10. Преобразование матрицы СКВП

Значения, требуемые для вычисления стандартного отклонения коэффициента, находятся в диагонали преобразованной матрицы СКВП. Например, стандартное отклонение коэффициента для параметра Male (мужчина) будет равным квадратному корню из значения совпадения его с самим собой в преобразованной матрице СКВП (корню из 0,0122, ячейка В25), умноженному на стандартное отклонение прогноза (рис. 11). Для извлечения этого значения можно использовать функцию ИНДЕКС. Например, в ячейке В47 формула: ='Рис. 12'!X5\*КОРЕНЬ(ИНДЕКС('Рис. 9-11'!B25:B44;'Рис. 9-11'!B46)) = 0,0417.

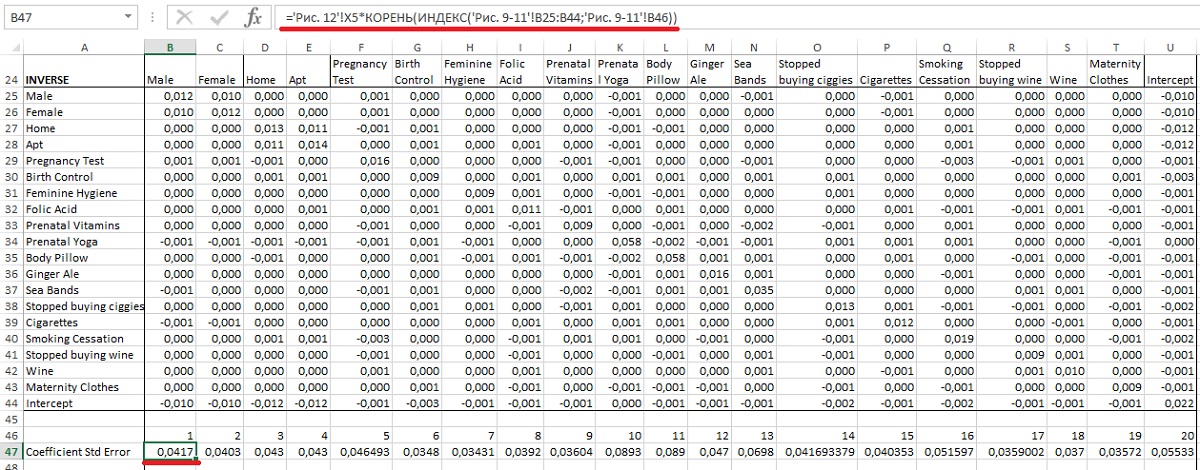


Рис. 11. Стандартное отклонение для каждого коэффициента модели

Теперь у вас есть все, что нужно для подсчета статистики каждого коэффициента (такого же, как и расчет F-статистики в предыдущем разделе). Мы будем производить проверку по двустороннему критерию, то есть рассчитывать вероятности получения коэффициента, по меньшей мере, такой же величины, положительной или отрицательной, если на самом деле между отличием и зависимой переменной нет никакой связи. Перенесите строку Coefficient Standard Error с листа «Рис. 11» в строку 3 на лист «Рис. 12» (рис. 12). T-статистика для этой проверки может быть рассчитана в строке 4 как абсолютное значение коэффициента, нормализированного по его стандартному отклонению. Для признака Male это будет выглядеть так: =ABS(B2/B3).

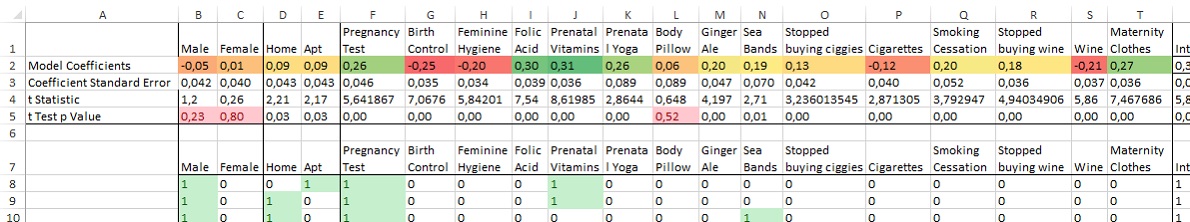


Рис. 12. Столбцы Male, Female, Body Pillow статистически незначимы по условиям теста

Проверку критерия Стьюдента теперь можно дополнить оценкой распределения Стьюдента (еще одного статистического распределения, вроде нормального распределения) относительно значения t-статистики (статистики Стьюдента) для ваших значений степеней свободы (строка 5). Например, в ячейке В5 используется формула =СТЬЮДРАСП(B4;$Z3;2) для расчета вероятности того, что коэффициент будет по меньшей мере таким, как при нулевой гипотезе. Двойка в этой формуле обозначает двусторонний критерий. Примените условное форматирование для ячеек, содержащих значения не меньше 0,05 (5%-ная вероятность), вы увидите, какие отличительные черты статистически не значимы. Ваши результаты могут отличаться из-за разницы в подгонке модели, а в таблице, показанной на рис. 12, столбцы Male, Female, Body Pillow представлены как статистически не значимые. А столбцы Home и Apt находятся на грани значимости.

**Делаем прогнозы на основании новых данных и измеряем результат.** Чтобы понять, как модель будет себя вести в реальном мире, нужно пропустить через нее несколько покупателей, которые не участвовали в предыдущем ее обучении. Этот отдельный набор примеров, используемый для тестирования модели, часто называется контрольной, или тестовой, выборкой, либо тестовой последовательностью.

Как отмечалось ранее, 6% покупателей РитейлМарта беременны, поэтому если вы выбрали 1000 случайных покупателей из базы, то примерно 60 из них окажутся беременными. Если для обучения модели вы делали выборку с запасом, то для тестирования стоит оставить отношение семей, ждущих ребенка, к остальным на уровне тех же 6%, чтобы измерения точности модели соответствовали тому, как модель будет себя вести в реальных условиях. Первые 60 покупателей беременны, а 940 остальных — нет (рис. 13).



Рис. 13. Тестовый набор данных

Точно так же, как и на листе «Рис. 12», пропустите этот набор через модель, подставляя новые данные в линейную комбинацию данных покупателей и коэффициентов, и затем прибавляя свободный член. Помещая этот прогноз в столбец V (рис. 14), в строке 56 вы получаете следующую формулу: =СУММПРОИЗВ('Рис. 12'!B$2:T$2;'Рис. 14'!A56:S56)+'Рис. 12'!U$2.

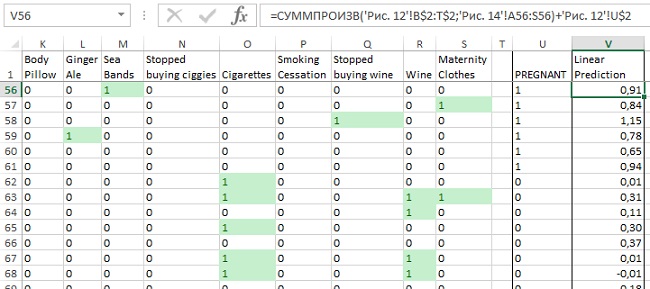


Рис. 14. Прогнозы тестового набора

Вы видите, что модель идентифицировала много семей, ожидающих прибавления, с прогнозами ближе к 1, чем к 0. Самые высокие значения прогноза получены для семей, которые покупали товары, четко относящиеся к беременности — фолиевую кислоту или витамины для беременных. С другой стороны, среди этих 60 «беременных» семей есть такие, которые не покупали ничего указывавшего на беременность. Конечно, они не покупали алкоголь или табак, но, как показывают их более низкие прогнозы, отсутствие покупки немного значит. И наоборот, посмотрев на прогнозы «не-беременных» семей, вы увидите несколько упущений. К примеру, в строке 154 «не-беременный» покупатель заказывает одежду для беременных и перестает заказывать сигареты, и модель дает ему прогноз в 0,77.

Таким образом, становится ясно, что если вы действительно собираетесь использовать эти прогнозы в собственном маркетинге, то вам придется установить некий порог для того, чтобы счесть покупателя беременным и начать отправлять ему соответствующие предложения. Возможно, нужно посылать рекламные материалы всем, у кого прогноз 0,8 и больше. Но, может быть, для вящей уверенности стоит поднять порог до 0,95. Перед тем как установить этот порог классификации, посмотрите на баланс плюсов и минусов в оценке качества работы модели. Измерения, больше всего влияющие на этот показатель, основаны на количествах и отношениях четырех величин, получаемых из нашей тестовой выборки:

* действительно положительные — отнесение беременного покупателя к беременным;
* действительно отрицательные — отнесение «не-беременного» покупателя к не беременным;
* ложноположительные (также называемые ошибками I рода) — отнесение «не-беременного» покупателя к беременным.
* ложноположительные (также называемые ошибками II рода) — неспособность определить беременного покупателя как такового.

*Установка граничных значений.* Создайте новый лист (рис. 15). Самое меньшее значение, которое можно использовать в качестве граничного между беременными и не беременными, — это значение наименьшего прогноза в тестовой выборке. Мы рассчитаем его в ячейке А2: =МИН('Рис. 14'!V2:V1001). Максимальной величиной может стать максимальное значение прогноза по тестовой выборке. Мы рассчитаем его в ячейке А5: =МАКС('Рис. 14'!V2:V1001). Ваша линейная регрессия способна делать предположения и ниже 0, и выше 1, потому что на самом деле это не вероятности попадания в определенный класс.

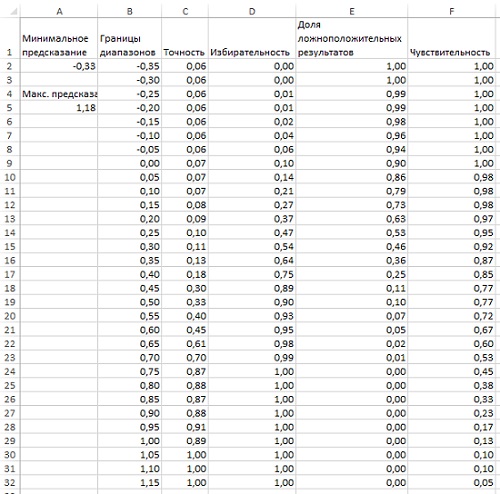


Рис. 15. Параметры качества модели

Добавьте в столбец В диапазоны от минимального значения до максимального с шагом 0,05. Рассчитаем следующие параметры качества модели для каждого из этих граничных значений: точность, избирательность, долю ложноположительных результатов, чувствительность.

*Точность* (столбец С) — это мера того, сколько семей в ожидании детей мы определили верно из всех семей, которые модель назвала «беременными». Допустим, граничное значение находится в В2 и равно -0,35. Какова точность нашей модели, если беременным мы считаем каждого с прогнозом -0,35 и выше? В ячейке С2 подсчитайте количество случаев, когда «беременная» семья получала прогноз больше или равный -0,35, а затем разделите это количество на общее число строк с прогнозом -0,35 и выше: =СЧЁТЕСЛИМН('Рис. 14'!$V$2:$V$1001;">=" & B2;'Рис. 14'!$U$2:$U$1001;"=1")/СЧЁТЕСЛИ('Рис. 14'!$V$2:$V$1001;">=" & B2).

Как видно на рис. 15, точность модели возрастает с увеличением граничного значения, а при границе, равной 1, модель становится идеально точной. Идеально точная модель считает беременными только беременных покупателей.

*Избирательность* (доля действительно отрицательных результатов, столбец D) выражается количеством покупателей, которые были верно отнесены к таковым (действительно отрицательный результат), разделенным на общее количество «не-беременных» семей. Например, в ячейке D2 используется формула: =СЧЁТЕСЛИМН('Рис. 14'!$V$2:$V$1001;"<" & B2;'Рис. 14'!$U$2:$U$1001;"=0")/СЧЁТЕСЛИ('Рис. 14'!$U$2:$U$1001;"=0"). Видно, что по достижении граничного значения 0,85 правильно определяется 100% «не-беременных» покупателей.

*Доля ложноположительных результатов* — это стандартный параметр, по которому определяется качество модели. А так как у вас уже есть доля действительно отрицательных результатов, то он может быть вычислен как разница между единицей и этой величиной (столбец Е). Видно, что по мере возрастания граничного значения, уменьшается количество ложноположительных результатов. Другими словами, вы совершаете меньше ошибок I рода (называя беременными покупателей, которые таковыми не являются).

*Доля действительно положительных значений или чувствительность* — это количество правильно определенных беременных женщин, разделенное на общее количестве таковых в тестовом наборе. Например, в ячейке F2 используется формула: =СЧЁТЕСЛИМН('Рис. 14'!$V$2:$V$1001;">=" & B2;'Рис. 14'!$U$2:$U$1001;"=1")/СЧЁТЕСЛИ('Рис. 14'!$U$2:$U$1001;"=1"). Вы можете убедиться, что с ростом граничного значения некоторые беременные женщины перестают определяться как таковые (ошибка II рода), а доля действительно положительных значений падает.

*Оценка соотношения измерений и кривая ошибок.* При выборе порогового значения для бинарного классификатора важно выбрать наилучший баланс параметров качества его работы. Чем выше порог, тем выше точность модели, но ниже выборка. Одна из наиболее популярных визуализаций, которые используются для просмотра этой оценки, — это кривая ошибок (рис. 16) или график зависимости действительно положительных значений от ложноположительных (столбцы E и F на рис. 15). Можно увидеть, что модель способна идентифицировать 40% беременных покупателей без единого «ложного срабатывания». И эта точка соответствует пороговому значению 0,85 Неплохо! А если вас устраивает редкая отсылка семьям, не ожидающим детей, купонов, связанных с беременностью, модель может достичь доли действительно положительных значений в 75% при всего лишь 9% «ложных срабатываний».



Рис. 16. Кривая ошибок линейной регрессии

Кривая ошибок также может служить инструментом выбора между двумя моделями. В идеале кривая подскакивает до 1 по оси Y с максимально возможной скоростью и остается там до конца графика. Так что модель, график которой более похож на описанный выше (также стоит отметить, что она должна иметь наибольшую площадь под кривой), часто считается лучшей.

**Предсказание беременных покупателей РитейлМарта с помощью логистической регрессии.** Чтобы с чем-то сравнить качество нашей модели, рассмотрим еще одну.

Если взглянуть на предсказанные нашей линейной регрессией значения, становится ясно: модель полезна для классификации, но сами прогностические значения не являются вероятностями класса (поскольку значения выходят за диапазон от 0 до 1). Существует модель, чьи прогнозы, являются вероятностями класса – *логистическая регрессия*. Чтобы перейти к ней, нужно пропустить нашу модель через функцию связи – exp(x)/(1+exp(x)), где х — значение предсказания нашей модели. Видно, что при х –> –∞, функция связи стремится к 0, а при х –> +∞, функция связи обращается в 1 (рис. 17). На практике эта функция связи оказалась настолько полезной, что ей дали собственное название – «логистическая» функция.

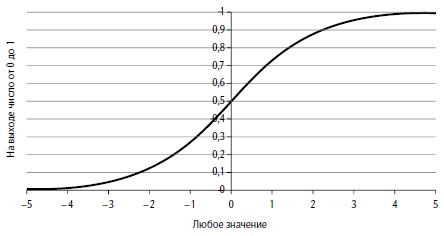


Рис. 17. Функция связи

Создайте новый лист на основе листа «Рис. 12». Удалите лишние данные, как показано на рис. 18. Добавьте столбец Х – Prediction (after Link Function) – Предсказание (после обработки функцией связи). Например, в D5 используется формула: =EXP(W5)/(1+EXP(W5)).

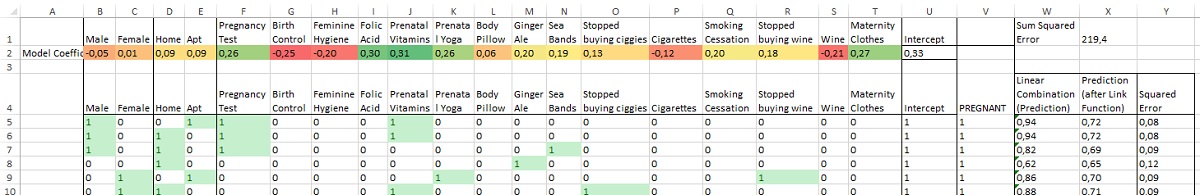


Рис. 18. Лист, готовый к построению логистической модели

Большая часть прогнозов оказывается близка к среднему, между 0,4 и 0,7. А все из-за того, что мы не оптимизировали коэффициенты (область B2:U2) для нашей новой модели. Оптимизацию придется повторить. Поэтому мы снова создаем столбец с квадратом отклонения (Y), но теперь в расчет отклонения включаем прогнозы линейной функции из столбца Х (а не W): =(V5-X5)^2. Ячейка Х1 по-прежнему является значением, которые мы хотим минимизировать: =СУММ(Y5:Y1004). Воспользуемся *Поиском решения* (рис. 19). По сравнению с предыдущей моделью мы расширили границы для коэффициентов.

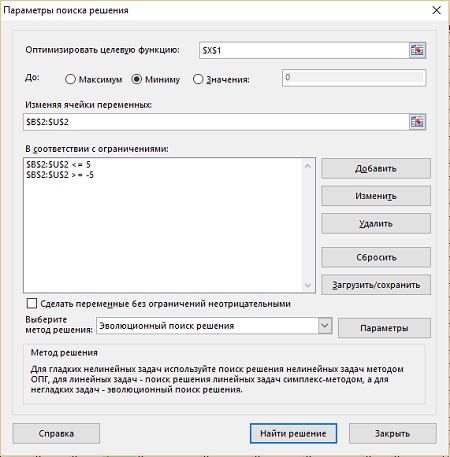


Рис. 19. Настройка Поиска решения для логистической модели

Закончив с реоптимизацией для новой функции связи, вы можете заметить, что теперь все ваши прогнозы по обучающим данным находятся между 0 и 1 (рис. 20, столбец Х), причем многие из них с уверенностью можно отнести к 1 или 0. Как вы видите на рис. 20, эстетически такие прогнозы выглядят лучше, чем результаты линейной регрессии. Кроме того, сумма квадратов отклонений (Х1) также уменьшилась, что говорит о более качественной модели.

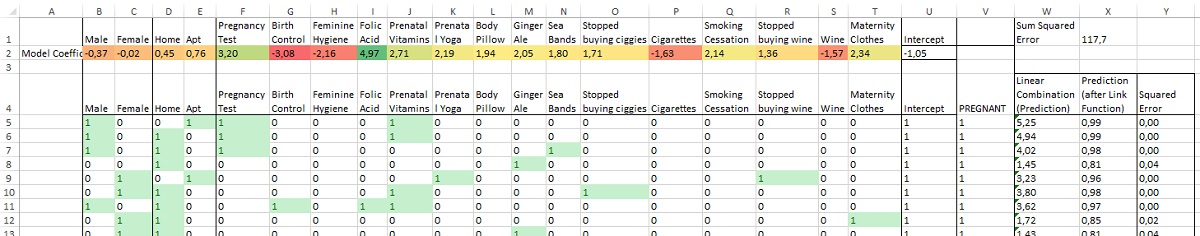


Рис. 20. Оптимизированная логистическая модель

На самом деле логистическая регрессии оптимизируется не минимизацией суммы квадратов, а путем нахождения коэффициентов, максимизирующих совместную вероятность извлечения этой обучающей выборки из базы данных РитейлМарта при том, что модель точно описывает реальность (подробнее см. [Наивный байесовский классификатор документов в Excel](http://baguzin.ru/wp/?p=13844)). Поэтому вместо квадрата отклонения в столбце Х (формула =(V5-X5)^2), следует использовать формулу вероятности:

*py(1-p)(1-y)*

где *р* выражает классовую вероятность, которую ваша логистическая модель выдает в столбце Х, а *у* – значение действительной беременности, располагающееся в столбце V. Для беременного покупателя (1 в столбце V) с прогнозом, равным 1 (1 в столбце Х), вычисление такой вероятности также дает 1. Но если прогноз у беременного покупателя был равен 0, то результатом этого вычисления будет уже 0. Таким образом, вероятность каждой строки максимальна, когда предположение и реальность сходятся.

Принимая независимость каждой строки данных, можно вычислить логарифм совместной вероятности данных, логарифмируя каждую из этих вероятностей:

*y\*ln(p)+(1 – y)\*ln(1 – p)*

Вместо того, чтобы минимизировать сумму квадрата отклонения, вы можете вычислить значение логарифмической функции правдоподобия для каждого прогноза, и затем сложить их. Те коэффициенты модели, которые максимизируют совместную вероятность, будут наилучшими.

Поскольку, когда коэффициенты модели генерируют прогноз, близкий к действительному значению класса 0/1, вы можете получить деление на ноль, примените функцию ЕСЛИОШИБКА. Создайте новый лист на основе листа «Рис. 20», но в столбце Х используйте формулу: =ЕСЛИОШИБКА(V5\*LN(X5)+(1-V5)\*LN(1-X5);0). Оптимизируя, вы получаете набор коэффициентов, очень похожий на найденный с помощью суммы квадратов отклонений, с небольшими расхождениями тут и там (рис. 21). Не забудьте, что в этом случае целевую функцию следует максимизировать.

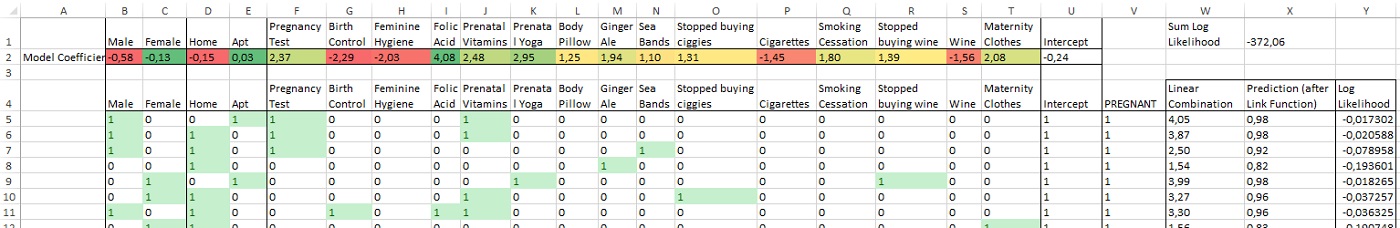


Рис. 21. «Настоящая» логистическая регрессия

**Сравнение работы линейной и логистической регрессий.** Теперь, когда у вас есть вторая модель, можно приступить к сравнению ее работы с работой линейной регрессии с помощью тестовых данных (рис. 22).



Рис. 22. Логистическая регрессия: прогнозы на основе тестовых данных

Выберите граничные значения от 0 до 1 с шагом в 0,05 (поменяйте 1 на 0,999, чтобы избавить формулу в столбце *Точность* от деления на 0). Рассчитайте параметры качества модели для логистической регрессии (рис. 23).

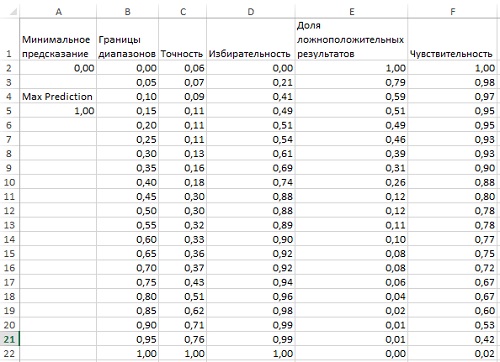


Рис. 23. Параметры качества модели логистической регрессии

Сравнивая таблицы 23 и 15 сложно сказать, какая модель лучше, поэтому визуализируем параметры с помощью кривой ошибок (рис. 24). Видно, что графики обеих моделей на границах диапазона сливаются, а в центральной части логистическая регрессия дает несколько лучшие результаты.

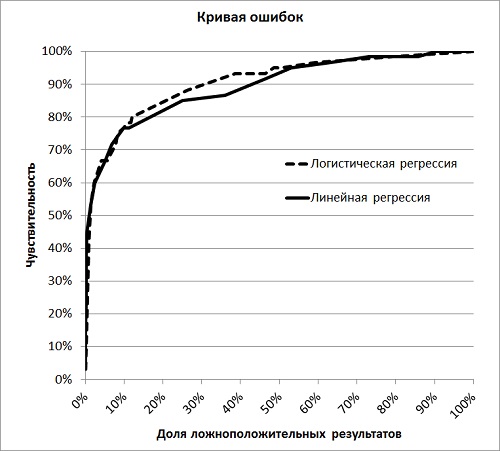


Рис. 24. Кривые ошибок линейной и логистической регрессий

Если вам понравился контролируемый искусственный интеллект, посмотрите:

* [Data Mining with R, Luis Torgo](http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/RDataMining/Data%20Mining%20with%20R-Kumar.pdf) прекрасно подходит для следующего шага. В этой книге дается описание машинного обучения на языке программирования R. R — любимый язык всех и каждого, кто работает со статистикой, и его несложно освоить в достаточном для моделирования ИИ объеме.
* [The Elements of Statistical Learning](https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf). Выражает академический взгляд на различные модели ИИ. Местами тяжеловатая для восприятия, эта книга может действительно поднять ваши интеллектуальные изыскания на новый уровень.
* Форум [CrossValidated](http://stats.stackexchange.com/) на StackExchange. Частенько оказывается, что интересующий вас вопрос уже давно кто-то задал.

1. Написано по материалам книги Джона Формана [Много цифр: Анализ больших данных при помощи Excel](http://baguzin.ru/wp/?p=13802). – М.: Альпина Паблишер, 2016. – С. 241–292 [↑](#footnote-ref-1)