УДК 001.891.573

РАЗРАБОТКА СИСТЕМЫ РАННЕГО ОПОВЕЩЕНИЯ НА ОГРАНИЧЕННЫХ ИСТОРИЧЕСКИХ ДАННЫХ

V. Potseluev[1], I. Khemych [2]

[1] *Deloitte consultant*
[2] *Deloitte Senior consultant*

Аннотация

Логистическая регрессия в интеллектуальном моделировании известна как неотъемлемая часть большинства научных исследований, а также высокопрофессиональных предприятий. Из-за ограниченного количества исторически данных и пропущенных значений могут возникнуть трудности на этапе отбора и подготовки данных для моделей логистической регрессии. В данных материалах мы приводим некоторые рекомендации по улучшению качества подготовки модели и обработки данных с отсутствующими значениями и выбросами.

Введение

В последнее годы наблюдается значительный интерес и прорыв в применении логистических регрессий, как одного из типов прогнозирующего моделирования, которая позволяет предсказывать бинарное событие. К примеру, в модели для банковского сектора предсказывается событие введения временной администрирования, представляющую банк. Это двоичное событие, означающее, что банк может быть либо с временной администрацией, либо без неё. В отличие от классических банковских моделей, нет никаких просрочек по кредитной задолженности, чтобы решить, какой банк рассматривать как «хорошую» или «плохую» учетную запись. Поскольку основной целью такой модели является – как можно ранньше выявить повышенный кредитный риск, нецелесообразно устанавливать период исхода более 12 месяцев. Более того, чем больше продолжительность прогноза, тем выше его неопределенность.

Чтобы быть в соответствии с ранним оповещающим характером модели – мы изучили возможность использования в качестве независимых переменных, не сами значения, а их отношения как изменение дельт (например, месяц Т против месяца Т-1)[1]. Однако таким образом мы уменьшаем размер выборки. Объяснение довольно простое- этот подход требует представления двух отчетных дат. Естественно, чем больше дат, тем выше вероятность того, что некоторые будут отсутствовать. В нашем случае количество хороших записей сократилось на 20%. Именно поэтому подход был отброшен.

Так как «плохие» распределены в течение всего периода по одной точке каждый, в то время как финансовые отчеты о хороших доступны во многие моменты времени (в данном конкретном случае данные основаны на ежемесячной основе). Цель заключалась в том, чтобы выбрать даты для хороших счетов, чтобы окончательное распределение по датам максимально приближалось к наблюдаемому распределению плохих записей, обеспечивающих минимальное отклонение.

1. Размер выборки

Основываясь на эмпирических исследованиях, было достигнуто оптимальное значение, при котором достигается баланс между необходимыми данными и качеством модели. Такой набор соответствует приблизительно 16 -18 тыс. хороших записей и 4-5 тыс. плохих [5]. Ниже этих значений качество модели будет ухудшаться значительно быстрее, чем количество записей.

2. Балансировка

Бинарные модели рассчитываются лучше, когда в выборке для разработки отношение «хороший» «плохой» составляет примерно 1: 1. Когда какой-то класс преобладает, модель может быть менее качественной. Чтобы справиться с этим, общая стратегия заключается в применении различных алгоритмов формирования выборки – это оверсэмл (когда недопредставленный класс рассматривается, будто его было бы представлено больше раз); андерсэмл (когда чрезмерно представленный класс рассматривается будто он будет наблюдаться реже); дисбаланс выборки (когда целевая частота в выборке или плохая частота такая же, как и у населения).

Важно отметить, что различные статистические алгоритмы имеют разную чувствительность к применяемым размеру выборки и схеме балансировки. Логистическую регрессию относят к менее чувствительным моделям. Во всяком случае, всё же можно разработать хорошую модель, даже если соотношение «хороших» и «плохих» находится далеко от 1:1. Поскольку доля «плохих» составляет 41,6%, а статистический подход – логистическая регрессия, мы используем несбалансированную схему составления выборки [2].

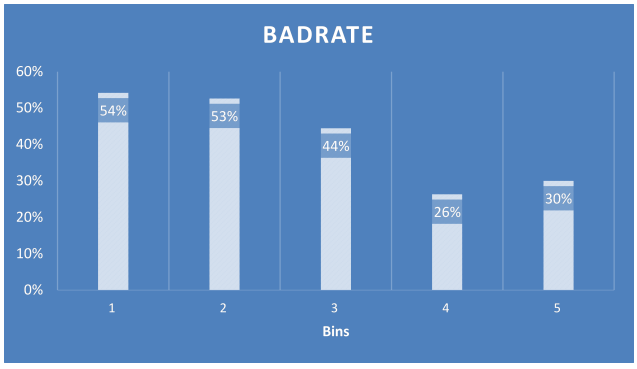
Обработка отсутствующих значений

Если, как в нашем случае, основная проверка данных не выявила какой-либо существенной проблемы или характера. При максимальном проценте отсутствующих значений для каждой переменной (вертикальный вид), равной 3%, это не существенно.

Аналогично, в самих данных нет ненормальных трендов (например, нечетные значения, различные типы значений в пределах одной переменной и т. д.). Часто аналитики предпочитают рассматрив ать отсутствующие значения просто путем замены его на среднюю / медианную для числовой переменной или по моде для наиболее частого категориального.

Тем не менее, может потребоваться исключить переменные с долей недостающих случаев выше 20-25%, если не будет получено надлежащего объяснения (так чтобы не было необходимости их кодировать). Причина в том, что во многих случаях такие отсутствующие значения появляются не случайно. Кроме того, когда есть причина, вызывающая это, и ее невозможно определить и объяснить должным образом, качество, а также стабильность модели могут демонстрировать плохие результаты при дальнейшем применении.

Другой популярный выбор – применить алгоритм кластеризации к набору данных. Этот подход был выбран на примере модели банков. Более точно, KNN-алгоритм (k-Nearest Neighbors Classification) применяется к выборке для разработки модели, содержащую только независимые переменные (пропущенные значения, закодированные как NA). Для каждого случая с любым значением NA (отсутствует) алгоритм будет искать его k наиболее похожих случаев и использовать значения этих случаев для заполнения неизвестных [4]. Иллюстративный пример алгоритма KNN показан на диаграмме:



Наконец, мы получили выборку без пропущенных значений. Все они были вменены KNN. Для будущих периодов, если отсутствующие значения снова появляются в конечных предикторах, их следует заменить средним или средним значением, оцениваемым по выборке для разработки после применения KNN:

Определение взаимосвязей

Выполняем одномерный анализ направленный на лучшее понимание связей, существующих между каждым предиктором и зависимой переменной.

Вот пример из переменной «ED2» (ROE) из примера разработки:

Bins	Cutpoint	CntRec	CntGood	CntBad	CntCumRec	CntCumGood	CntCumBad	PctRec	BadRate	GoodRate	Odds	LnOdds	WoE	IV
1	<= 0.0001	24	13	11	24	13	11	24%	54%	46%	1.1818	0.1671	0.4898	0.0587
2	<= 0.0005	19	10	9	43	23	20	19%	53%	47%	1.1111	0.1054	0.4281	0.0355
3	<= 0.0018	18	8	10	61	31	30	18%	44%	56%	0.8	-0.2231	0.0996	0.0018
4	<= 0.0085	19	5	14	80	36	44	19%	26%	74%	0.3571	-1.0296	-0.707	0.0865
5	> 0.0085	20	6	14	100	42	58	20%	30%	70%	0.4286	-0.8473	-0.525	0.0517
6	Missing	0	0	0	100	42	58	0%	NA	NA	NA	NA	NA	NA
	Total	100	42	58	NA	NA	NA	100%		42%	58%	72%	-0.3228	0.234

Рис. 1

Критерии для начального расщепления являются квантилями (с шагом 0,2). Фактически, общее количество корзин может быть и выше. Практически, рассматривают от 10 до 20, чтобы получить менее концентрированное распространение.

Тем не менее, количество корзин также зависит от размера выборки. Чем меньше шансов получить ложные отношения, которые на самом деле статистически незначимы. Поэтому рассмотрено 5 корзин одинакового размера (по 20% каждый).

Основная идея при исследовании такой таблицы, как указано выше, заключается в рассмотрении следующих аспектов:

- Существует ли линейная зависимость между предиктором и зависимой переменной;
- Существует ли монотонная тенденция в отношении Weight of Evidence (WoE);
- Имеет ли наблюдаемая связь значимость с экономической точки зрения?;
- Насколько независимая переменная связана с зависимой (Information Value или IV)

Для данного интервала «WoE» рассчитывается как:

$$WoE = \left[\ln\left(\frac{Distr\,Goods}{Distr\,Bads}\right)\right] \cdot 100\% \qquad (1)$$

Это полезно, потому что это позволяет сравнивать интервалы с хорошими и плохими весами с средним значением выборки (среднее значение).

Положительные значения «WoE» указывают на более высокую плохую оценку по сравнению со средним, а отрицательное – противоположное.

Одним из ключевых свойств «WoE» является то, что он всегда отображает линейную связь с естественным логом хороших и плохих.

Один из самых популярных и широко используемых аналитической мерой ассоциации. «IV» выражается следующим образом:

$$IV = \sum_{i=1}^{N} [(Distr\,Goods_i - Distr\,Bads_i) \cdot \cdot \ln\left(\frac{Distr\,Goods}{Distr\,Bads}\right)]. \quad (2)$$

Проще говоря, «IV» показывает, насколько сильная ассоциация (отношение) предиктора с зависимой переменной. Он колеблется от 0 до бесконечности. Более высокое значение означает более сильные отношения. Существуют некоторые стандартные диапазоны классификации «IV». Ниже 0,05 указывает на небольшую связь. Значения от 0,05 до 0,25 указывают на умеренную связь. Значения между 0,25 и 0,5 показывают сильную связь. Значения выше указывают на очень сильные отношения. Обычно необходимо тщательно анализировать очень сильные отношения. Иногда слишком сильная переменная, включенная в модель, может привести к переопределению на этапе проверки и сделать модель со слишком небольшим числом предикторов (что не всегда целесообразно).

Стандартизация

Строго с точки зрения моделирования стандартизация не имеет большого значения. Однако, если, наконец, мы захотим узнать вклад каждой переменной, то все они должны быть стандартизированы до моделирования. Другими словами, если одна ковариация принимает значения других порядков, в сравнении с другими, она может иметь меньшую оценку параметра, имея аналогичный вес. Среди самых простых и популярных способов стандартизации переменной – вычесть среднее значение, а затем делить на его стандартное отклонение. Это позволило бы иметь все переменные в одном масштабе [3].

3. Выпадающие значения

Когда несколько наблюдений имеют значения, которые слишком далеки от остальных, это может привести к менее точной оценке параметров регрессии. Это привело бы к более высокому уровню ошибки. Существуют различные подходы к устранению выбросов. Рассмотрены, три наиболее распространенных – удалить эти наблюдения, заменить на +/- 3 Стандартные отклонения или заменить на 1.5 Inter Quantile. Числа 3 и 1,5, конечно, не являются определенными, и они могут варьироваться [7].

Очевидно, что первый вариант менее подходит при работе с ограниченным размером выборки. Это может привести к потере некоторой информации, представленной в виде данных.

Помимо всего прочего, следует применять и некоторые экспертные оценки. Не всегда выбросы являются чем-то ненормальным по своей природе. Поэтому, рассматривая выбросы, чистый статистический фон всегда должен быть взвешен против экономической интерпретации.

Перечень использованных источников

1. DChambers, J. M. and Hastie, T. J. Statistical Models in S, Wadsworth & Brooks/Cole. 120. — 1992. — p. 608.
2. David W. Hosmer Applied Logistic Regression. 34,. — Scripta Materialia (2001) — p. 128.
3. Joseph M. Hilbe Negative Binomial Regression. 47. — (2010) — p. 216.
4. Joseph M. Hilbe Logistic Regression Models. 129. — (2009) — p. 249.
5. Andrew P. Grieve and Shah-Jalal Sarker Simulation-based sample-sizing and power calculations in logistic regression with partial prior information. J. Ph. Statistics. 15. — (2016) — p. 507-516.
6. Wilson, P. W. Detecting outliers in deterministic nonparametric frontier models with multiple outputs, Journal of Business and Economic Statistics, 11, 319–323. — (2008) — p. 319–323.
7. Wilson, P. W. Detecting outliers in deterministic nonparametric frontier models with multiple outputs, Journal of Business and Economic Statistics, 11, 319–323. — (2008) — p. 319–323.