

ОСОБЕННОСТИ И ПОДХОДЫ К ФОРМИРОВАНИЮ ВЫБОРКИ В КРЕДИТНОМ СКОРИНГЕ

О. С. Гичан

Белорусский государственный университет, г. Минск;

oleg.gichan@yandex.ru;

науч. рук. – Е. Г. Господарик, канд. экон. наук, доц.

Построение автоматизированных скоринговых систем позволяет банкам снизить кредитные риски, увеличить количество обрабатываемых заявок, принимать объективное решение о выдаче кредита. Этап определения цели и формирования соответствующей выборки для моделирования представляется одним из наиболее важных в разработке скоринговых систем. В работе систематизируется подход к разработке скоринговых систем, вариации к формированию выборки. Проводится сравнительный анализ современных подходов к формированию выборки.

Ключевые слова: кредитный скоринг; формирование выборки; дисбаланс классов; статус заемщиков; моделирование.

Важность достоверной и актуальной информации трудно переоценить. В сфере кредитования необходимо использовать все данные, доступные в соответствии с действующим законодательством – это залог успеха в конкурентной борьбе за лояльность клиентов.

В статье рассмотрим особенности формирования выборки данных (далее – выборки) для построения системы кредитного скоринга.

Стоит отметить, что при прочих равных разработка любой модели (оценка кредитоспособности клиентов банка, формирование релевантных рекомендаций клиентам и т.д.) ведется в трех пересекающихся плоскостях: определение цели и формирование соответствующей выборки; подготовка данных; построение модели и интерпретация полученных результатов.

Зачастую первый этап проходят слишком быстро, не уделяя должного внимания всем его особенностям, затем пытаются «выжать» все возможное из сформированной выборки. Но дело в том, что доступная для использования информация ограничена этой выборкой – это в конечном итоге влияет на эффективность работы организации. Пренебрежение любым аспектом при разработке «прод» модели может привести к потенциальным потерям или недополученной прибыли для организации. Давайте попробуем разобраться в тонкостях этого этапа.

Прогнозные модели основываются на предположении, что «будущее поведение можно предсказать исходя из прошлого опыта». Определение и выгрузка достаточного объема данных является залогом успеха/провала разработанной модели.

Для примера возьмем случай разработки скоринговой системы банка, в частности – для оценки кредитоспособности потенциального заемщика при подаче заявки на кредит (аппликационный скоринг).

Постановка цели и задач прямым образом влияет на валидацию разрабатываемых моделей (на каких данных будет протестирована, какой результат считать минимально удовлетворительным) и выбор наиболее подходящей. Цели могут быть выражены в общих чертах, например, сокращение потерь по новым кредитам, и в более конкретных, например, сокращение числа «плохих» кредитов по одобренным заявкам в течение девяти месяцев после принятия положительного решения. Далеко не редко встречается ситуация, когда преследуется не одна, а несколько целей (обновление действующих моделей и минимизация потенциальных рисков).

Далее будем рассматривать следующую ситуацию: в банке *N* необходимо обновить действующие скоринговые модели для оценки кредитоспособности потенциальных заемщиков; разработка проводится собственными силами банка. Как правило, в результате разработки ожидается скоринговая карта – набор характеристик, по которым рассчитывается оценка кредитоспособности клиента, и соответствующие им баллы [1, стр. 52-53].

Для того чтобы создать скоринговую карту требуется сформировать выборку, на которой модель будет обучаться и будет проводиться тестирование. Исходя из установленных ограничений (поставленная вначале цель, задачи, а также кредитная программа, целевая аудитория и другие) необходимо определить входные параметры выборки: кредитная программа и целевая аудитория; временной горизонт выборки; критерий определения статуса заемщика («платежеспособный», «неплатежеспособный», «не определено»); иные необходимые ограничения.

В нашем учебном примере (на реальных данных организации «Хоум Кредит Банк») будем считать, что разрабатываемая скоринговая модель будет использована на кредитной программе «кредитная карта». Скоринговые системы наиболее популярны в отношении физических лиц, поэтому установим данную аудиторию в качестве целевой.

Зачастую заемщиков классифицируют на «хороших», или платежеспособных, «плохих», или неплатежеспособных, а также «не определен», или клиенты с недостаточной кредитной историей, имеющие допустимую просрочку платежа и др. [3, стр. 49]. Критерием к отнесению клиентов в определенную группу в основном выступает количество дней просроченной задолженности по кредитному договору (максимальное значение по основному долгу и процентам). В упрощенном виде, Наци-

ональный банк Республики Беларусь придерживается следующего разделение клиентов [3, стр. 49]:

- «плохими» признаются заемщики с максимальным количеством просроченной задолженности по кредиту превышающим 90 дней в течение 12 месяцев со дня подписания договора;
- «хорошими» признаются заемщики с максимальным количеством просроченной задолженности по кредиту не превышающим 30 дней в течение 12 месяцев со дня подписания договора.

Все остальные принимают статус «не определен» и в процессе обучения модели участия не принимают.

Базель определяет понятие «плохого» заемщика как существование просроченной задолженности на счету свыше 90 дней на протяжении 12 месяцев. Однако банк не ограничен в выборе критерия классификации заемщиков, и как правило, определяет оптимальные значения на основе анализа винтажей и ролл-рейтов [2, стр. 45].

Промежуток времени, в течении которого анализируется поведение заемщика, а именно выполняет ли он в полной мере взятые на себя кредитные обязательства (в вышеприведенном случае, 12 месяцев) представляет собой «окно мониторинга». Считается, что 12-13 месяцев – это достаточный срок для того, чтобы либо показать тенденцию к выполнению своих обязательств, либо к их нарушению. Окно выборки» подразумевает под собой некоторый промежуток времени, в течении которого оформлялись кредитные сделки.

Собранная информация (характеристики клиента как из внутренних источников, так и из внешних) вместе с установленным статусом («хороший», «плохой», «не определен») представляет собой выборку, на которой будет обучаться модель, тестировать свое качество и в некоторых случаях валидировать полученные входе обучения и теста результаты.

Важным этапом в формировании репрезентативной выборки является определение необходимого объема. Число кредитов с известным статусом должно быть достаточно большим для обеспечения необходимого объема выборки. Отдельно стоит отметить, что число «плохих» кредитов в большинстве кредитных портфелей в разы меньше кредитов, классифицированных как «хорошие». Вариации в формировании выборки происходят на основе следующих моментов: равномерность распределения зависимой переменной и ее доля в общей выборке (статус кредита), число независимых переменных в модели, максимально допустимая ошибка выборки, экспертные оценки [5, стр. 6].

В современных условиях цифровизации определение оптимального объема выборки становится менее значимой проблемой (т.к. базы данных о клиентах в большинстве белорусских банков исчисляются десят-

ками или сотнями тысяч кредитных договоров). Основная проблема заключается в явном дисбалансе классов.

Для решения этого вопроса Сорокин А. [5, стр. 7] рекомендует использовать следующую формулу:

$$n = \frac{z_y^2 w * (1-w)}{\Delta_w^2},$$

где n – минимальный объем выборки, z_y – значение стандартного нормального закона распределения, определяемое в зависимости от выбранного уровня надежности y , w – доля «плохих» клиентов по тестовой выборке, Δ_w – максимально допустимая предельная ошибка оценки доли «плохих» заемщиков.

Ковалев М. [2, стр. 45] и Raymond A. [7, стр. 350] предлагают использовать вариант, при котором соотношение «хороших» и «плохих» кредитов в выборке один к одному (грубо говоря, 1500 «плохих» и 1500 «хороших»). Siddiqi N. [8, стр. 63] рекомендует использовать все доступные и «очищенные» данные, оперируя тем фактом, что существующий баланс классов в выборке (предположим 4% в выборке принимают статус «плохие») уже отражает действующие вероятности отнесения кредита к тому либо иному классу.

Существуют действенные техники борьбы с дисбалансом классов в выборках, которые сводятся либо к искусственному занижению доли «хороших» кредитов, либо к искусственному увеличению доли «плохих» кредитов.

Далее на основе реальных данных проверим, какие техники показывают себя лучше остальных.

Размер выборки превышает 300 тыс. кредитов с известными статусами («хороший» либо «плохой»). Доля «плохих» во всей выборке равняется 8.07%. Для целей обучения модели выделено 70% выборки, для тестирования и выбора лучшей модели оставлено 30% выборки. Модель классификации кредитов построена на основе алгоритма дерева решений «Light GBM». Метрика качества модели – показатель AUC (от англ. «Area Under Curve») [5, стр. 20].

Результаты эксперимента представлены на рисунке. Наблюдается явное преимущество варианта, при котором выборка остается неизменной, баланс классов искусственно не корректируется. Действительно, при искусственном занижении или завышении пропорций того или иного класса выборки искажается реальный портрет клиента и модель не видит всех взаимосвязей, которые находятся в выборке.

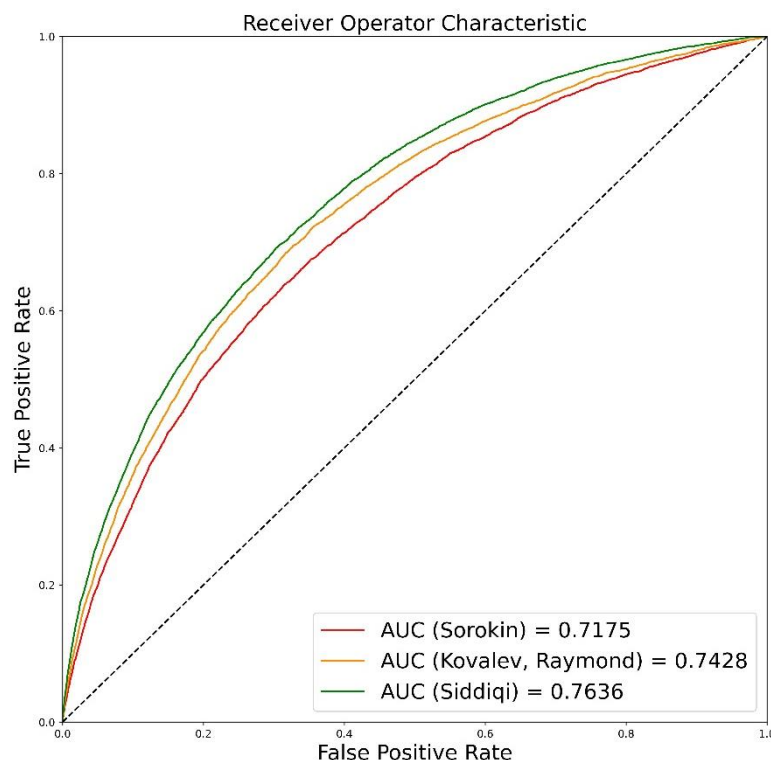


Рис. Сравнительный анализ моделей на основе показателей AUC.
Примечание: Источник – собственная разработка.

Проведенный эксперимент позволяет утверждать, что вмешательство в рассматриваемую выборку для разработки модели аппликационного скоринга не является желательным. Требуется принимать во внимание всех клиентов, доступных для моделирования. Таким образом будет построен реальный портрет клиента и будет достигнута максимальная эффективность модели при ранжировании заемщиков.

Библиографические ссылки

1. Гичан, О. С. Современная практика разработки скоринговых карт для розничных клиентов в белорусских банках / Гичан О.С., Господарик Е.Г. // Банковский вестник. – 2020. – №4/681. – С.49-59.
2. Ковалев, М. М. Методика построения банковской скоринговой модели для оценки кредитоспособности физических лиц // М.М. Ковалев, В. Корженевская // Банки Казахстана. – 2008. – № 1. – С. 43–48.
3. Мэйз Э. Руководство по кредитному скорингу / под ред. Элизабет Мэйз; пер. с англ. И.М.Тикота; науч. Ред. Д.И.Вороненко. – Минск: Гревцов Паблицер, 2008. – 464 с.
4. Пленкин, В. В. Скоринговая модель кредитных историй // Пленкин В.В., Тартенас М.В. // Банковский вестник. – 2015. – №6/623. – С.48-52.
5. Сорокин, А. С. Построение скоринговых карт с использованием модели логистической регрессии [Электронный ресурс] / Интернет-журнал «Науковедение» - Выпуск 2, март–апрель 2014. – Москва, 2014. – Режим доступа: <https://naukovedenie.ru/PDF/180EVN214.pdf>. Дата доступа: 20.01.2021.

6. *Finlay S.* Credit scoring, response modelling and insurance rating: practical guide to forecasting consumer behavior / Palgrave Macmillan, 2010.
7. *Raymond A.* The Credit Scoring Toolkit. Theory and Practice for Retail Credit Risk Management and Decision Automation // Oxford University Press Inc., New York, 2007.
8. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring // John Wiley & Sons: Hoboken, NJ, 2006.
9. *Sven F. Crone* Instance sampling in credit scoring: An empirical study of sample size and balancing / Sven F. Crone, Finlay S. // International Journal of Forecasting, 2012.