Описание моделей досрочных погашений и кластеризации

Москва

2019 г.

Оглавление

1. Введение 3

2. Досрочные погашения и кластеризация 4

2.1. Модели досрочных погашений и дефолтов 4

2.2. Спецификации моделей 4

2.3. Кластеризация портфеля закладных 6

2.4. Реализация в скриптах 6

3. Приложения 7

3.1 Описание моделей миграции 7

3.2 Описание процесса бинаризации 8

1. Введение

Данный отчет включает в себя информацию о моделях досрочного погашения и механизме их применения, а также результаты бэктестирования. Также наряду с информацией о моделях досрочных погашений в отчет включены сведения о кластеризации портфеля закладных.

Модели полного досрочного погашения и дефолта прогнозируют вероятность перехода кредита в терминальное состояние, после которого кредит более не наблюдается. Модель частичного досрочного погашения моделирует вероятность наступления события частичного погашения кредита. Модель, прогнозирующая размер частичного погашения, моделирует размер частичного погашения при условии, что произошло частичное погашение кредита.

Моделирование эволюции ИЦБ требует портфеля ипотечных закладных актуального на отчетную дату. Портфель содержит информацию о различных динамических и статических параметрах кредитов. К динамическим можно отнести такие параметры, как, например, остаток основного долга, текущая просроченная задолженность, а к статическим - первоначальный размер кредита, кредит/залог при выдаче. На основании этих факторов, а также смоделированной макроэкономики, модели досрочных погашений присваивают каждому кредиту вероятности дефолта, полного и частичного досрочных погашений.

Большая часть документа посвящена процедурам, реализованным в скрипте *model /model_lld.py*.

2. Досрочные погашения и кластеризация

2.1. Модели досрочных погашений и дефолтов

Модели полного досрочного погашения, дефолта и частичного досрочного погашения являются моделями logit (полиномиальная логистическая регрессия), где целевая переменная принимает значение 1, если событие произошло, и 0 - в обратном случае. Математическое описание моделей может быть найдено в Описание моделей миграции. Здесь же отметим, что факторы риска можно разделить на три категории:

- Постоянные факторы, $x_n, n = 1 \dots N$
- Факторы, зависящие от времени, $y_k, k = 1 \dots K$
- Смешанные факторы, $z_m, m = 1 \dots M$

В общем случае, риск досрочного погашения может быть записан, как

$$\ln \lambda(t) = \ln \lambda_x + \ln \lambda_y(t) + \ln \lambda_z(t) = \sum_{n=1}^{N} \beta_n x_n + \sum_{k=1}^{K} \beta_k y_k(t) + \sum_{m=1}^{M} \beta_m z_m(t) \qquad (1)$$

Каждая модель досрочных погашений (дефолт, полное досрочное погашение или частичное досрочное погашение) зависит от своего набора факторов, но общая схема верна для всех моделей. Факторы, применяющиеся в моделях досрочных погашений, подробнее описаны в пункте 2.2.

В формуле выше функция $\lambda(t)$ зависит от времени, так как от времени зависят переменные $y_k(t)$ и $z_m(t)$. Однако, постоянные факторы x_n от времени не зависят, и их можно вычислить на основании первоначального LLD. Остальные поправки вычисляются непосредственно во время моделирования эволюции кредитов.

2.2. Спецификации моделей

Ниже представлена таблица, в которой наглядно представлено, от каких переменных зависит каждая модель досрочных погашений в отдельности. Зеленым цветом выделены переменные, которые используются в данной модели. Цвет переменных определяет группу риска (из раздела 2.1), к которой относится данный фактор. Белая заливка ячеек соответствует постоянным факторам, светло-серая – факторам, зависящим от времени, темно-серая – смешанным факторам.

Таблица 1 Зависимости от переменных для различных моделей

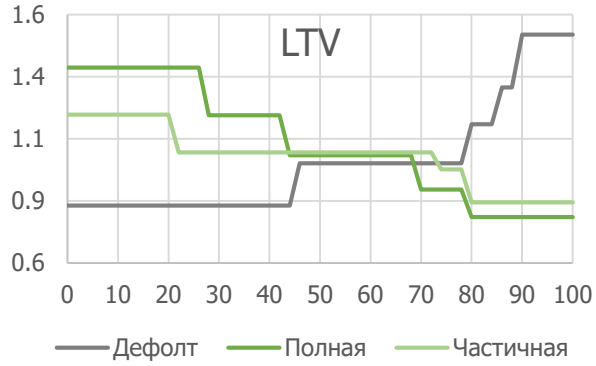
	LTV	Регион	Спред нач.	Доход/ Кредит	Образование	t	ОФЗ, лаг 6-мес.	ОФЗ, спот	Сезон	Спред тек.
Дефолт										
Полная										
Частичная										
Размер частичной										

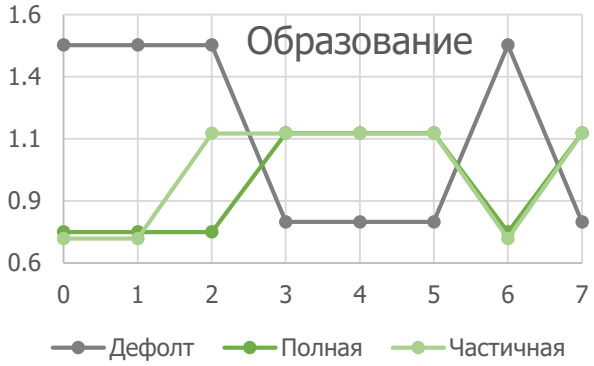
Ниже приведены полные названия переменных:

- t – время, прошедшее с момента выдачи кредита (в годах)
- LTV – отношение кредита к залогу
- Регион – код региона предмета ипотеки
- Спред нач. – разница между среднерыночной ставкой по ипотеке на момент выдачи кредита и ставкой по кредиту
- Доход/кредит – отношение дохода к сумме кредита
- Образование – код образования, где 0 – среднее, 1 – среднее специальное, 2 – неоконченное высшее, 3 – высшее, 4 – два и более высших, 5 – ученая степень, 6 – неполное среднее, 7 - MBA
- ОФЗ, спот – сценарная короткая ставка
- ОФЗ, лаг 6-мес. – шестимесячный лаг сценарной короткой ставки
- Сезон – месяц года (1 – январь, 2 – февраль, …, 12 – декабрь)
- Спред тек. – разница между среднерыночной ставкой по ипотеке на текущий момент и ставкой по кредиту

При моделировании эволюции кредита первоначальные значения переменных заменяются на значения, полученные из функций риска для каждой переменной. Эти функции получены в результате исследования зависимостей досрочных погашений на исторических данных с помощью применения WOE-преобразования. Подробнее процедура WOE преобразования описана в Описание процесса бинаризации.

В качестве примера, ниже приведены функции риска для LTV (численная переменная) и Образования (категориальная переменная) для высококлассной ипотеки.





Стоит отметить, что данные функции риска включают в себя не только значения WOE, но также коэффициент пропорциональности β_i, полученный при обучении модели на исторических данных.

2.3. Кластеризация портфеля закладных

Теоретически, моделирование возможно провести для каждого кредита, но на практике данная задача выполняется слишком долго. Поэтому проводится процедура кластеризации кредитов – объединение различных кредитов со сходными параметрами в группы, эволюция которых моделируется в дальнейшем.

Агрегированные параметры кластера вычисляются при помощи взвешивания на основной долг параметров, попавших в этот кластер кредитов. Такое вычисление производится для всех числовых переменных, за исключением тех, которые обозначают основной долг (первоначальный и текущий). В качестве первоначального и текущего долга используются средние значения первоначального и текущего долга кредитов, попавших в кластер.

Таблица 2 показывает пример того, как и какие параметры кредита получаются после кластеризации.

Таблица 2 Пример кластеризации и усреднения параметров

№ кредита	Остаток ООД	Срочность	Ставка, %	Поправка дефолта	Поправка полного досрочного погашения
1	100	10	9	1.2	0.8
2	150	20	11	1.1	1.1
3	250	50	13	0.8	1.6

Среднее — Взвешенное среднее

№ кластера	Остаток ООД	Срочность	Ставка, %	Поправка дефолта	Поправка полного досрочного погашения
1	167	33	11.6	0.97	1.29

2.4. Реализация в скриптах

Функции риска досрочных погашений могут быть найдены в папке *specs.* В публичной версии приведен пример такой спецификации. Конкретные значения функций приведены в качестве примера, и не отражают реальных спецификаций, используемых в ДОМ.РФ.

В папке *model* находится скрипт *model_lld.*py, в котором производится:

1. Предобработка портфеля закладных
2. Вычисление постоянных и переменных поправок моделей досрочных погашений
3. Процедура кластеризации портфеля
4. Моделирование денежных потоков портфеля

3. Приложения

3.1 Описание моделей миграции

Математически модель логистической регрессии выражает зависимость логарифма шанса (логита) от линейной комбинации независимых переменных:

$$\ln \lambda = \beta X \tag{2}$$

где $\lambda = \frac{p}{1-p}$ – моделируемый риск

$\beta X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K$ – регрессионная модель для логарифма риска.

p - вероятность события миграции кредита в новое состояние в течении одного периода времени, $\lambda_0 = e^{\beta_0}$ — базовый уровень риска (base hazard), $\{x_i\}$ и $\{\beta_i\}$ — независимые переменные (ковариаты) и их вес в модели; величины $\lambda_i = e^{\beta_i x_i}$ также принято называть поправками к базовому риску от i-го признака

Уравнение (1) отражает линейную зависимость вероятности наступления события по кредиту в зависимости от значений независимых переменных. Константа в модели отражает базовый риск моделируемого события при равенстве всех независимых переменных нулю. Значения коэффициентов при независимых переменных отражают степень их влияния на риск события в логарифмической шкале. Для расчета вероятности события уравнение (1) преобразуется в:

$$p = \frac{1}{1 + e^{-\beta X}} = \frac{\lambda}{1 + \lambda} \tag{3}$$

В роли переменных модели выступают характеристики кредита и макроэкономические показатели.

Здесь надо объяснить, что модели выживаемости рассматриваются как кусочно-постоянные, т.е. в течении периода t риск считается постоянным и равным λ_t, однако от периода к периоду риск может меняться не только в силу изменения ковариат (time-varying), но и в силу изменения базового уровня риска с течением времени (time-dependent baseline hazard). Обычно, в таком случае рассматривается переменный уровень риска $\lambda_0(t) = e^{\beta_0(t)}$, что в кусочно-постоянной модели эквивалентно введению в качестве одной из ковариат $\{x_i\}$ качественного преобразования переменной времени t.

3.2 Описание процесса бинаризации

Процесс биннинга переменной – это процесс разбиения диапазона, которой может принимать переменная, на группы с присвоением им различных информационных весов. Веса различных групп передают влияние группы на целевую переменную. Таким образом зависимость между целевой и объясняющей переменных выражается в виде кусочно-постоянной функции. Количество таких групп может быть любым, но на практике их число обычно не превышает 20, при этом в каждой группе должно содержаться не менее 5% наблюдений. Отметим, что процесс биннинга для качественных и количественных переменных немного отличаются.

Для качественных переменных сначала рассчитывается значение WoE для каждой группы и доля в общем количестве наблюдений. Показатели WoE для каждой группы рассчитываются по формуле:

$$WoE = \ln\left(\frac{B_i}{G_i}\right) \tag{4}$$

где B_i, G_i – относительные частоты «событие произошло» и «событие не произошло» кредитов соответственно в i-ой группе переменной.

Группы, близкие по значениям WoE и с количеством наблюдений более 500 или с долей наблюдений более 5%, объединяются в общую категорию. Категории с количеством наблюдений менее 500 или с долей наблюдений менее 5% объединяются с категорией с наименьшим значением WoE, чтобы иметь минимальный вес в результате моделирования. Такой подход позволяет избежать присваивания больших весов категориям, которые статистически незначимы.

Количественные переменные преобразуется иначе. Сначала переменная делится на 10-20 групп примерно равных по количеству наблюдений по шкале анализируемой переменной. После этого рассчитывается значение WoE для каждой группы. Соседние группы, близкие по значениям WoE, объединяются в общую группу. При этом предполагается монотонность вероятности наступления события (значения WoE группы) с увеличением номера группы.