Модель кредитного скоринга на основе поведенческих и макроэкономических факторов

А. Р. Абубакиров, Н. О. Никитин, А. В. Калюжная
Университет ИТМО, Санкт-Петербург
mr.azat.abubakirov@gmail.com, nicl.nno@gmail.com, kalyuzhnaya.ann@gmail.com

***Аннотация*. Данная статья посвящена вопросам построения модели предсказания дефолтов по кредитам физических лиц с учетом поведенческих факторов и введения дополнительных макроэкономических факторов. Где под поведенческими факторами понимаются факторы, характеризующие финансовое поведение заемщика и выделенные на основе анализа его транзакционной истории. В качестве метода, лежащего в основе скоринговой модели используется логистическая регрессия. В рамках экспериментальных исследований оценивается влияние макроэкономических факторов на качество предсказания дефолта по кредиту. Также, приводятся результаты исследования зависимости качества и устойчивости результата предсказания при выборе значимого набора переменных и варьирования долей обучающей и тестовой выборок.**

Ключевые слова: поведенческий скоринг; макроэкономические факторы; логистическая регрессия; кредитный дефолт

I. Введение

Классический кредитный анкетный скоринг основывается на данных, которые описывают непосредственно клиента банка и тот финансовый продукт, которым он заинтересовался. Недостатком моделей, которые используют для обучения такие данные, является то, что они статичны: они не учитывают изменение во времена финансового состояния клиента и, соответственно, кредитоспособности. К динамическим характеристикам клиента можно, например, отнести информацию, которая описывает оборот финансовых средств на счету клиента за последний год или количестве покупок, оплаченных кредитной картой. Описанная проблема решается в рамках задачи поведенческого (транзакционного) скоринга, который включает в себя использование агрегированных данных о транзакциях клиентов.

Используя упомянутые модели, финансовые институты предполагают, что внешняя экономическая ситуация не оказывает влияние на платежеспособность клиентов. Целью текущего исследования является демонстрация улучшения качества обучения моделей путем добавления макроэкономических показателей, то есть доказательство влияния макроэкономики на кредитоспособность заемщиков.

II. Обзор литературы

Использование макроэкономических факторов может быть реализовано разными способами. Например, они могут быть добавлены в модель на основе логистической регрессии [1]. Другим из решений является модель, основанная на анализе выживаемости [2][3]. Эта модель, которая изначально использовалась в медицине для оценивания эффективности лечения, естественным образом описывает процесс дефолта заемщика. Crook и Bellotti добавили в модель экономические факторы, зависящие от времени: индекс безработицы, процентная ставка, индекс покупательной способности и другие. Новые данные позволили незначительно улучшить точность прогнозов, а также показать зависимость дефолта заемщика от экономической ситуации.

Аналогично макроэкономическими фактором в модель выживаемости могут быть добавлены данные, основанные на транзакциях клиента [4]. Внедрение таких данных позволило улучшить предсказательную способность модели: наилучший результат был получен при использовании поведенческих и макроэкономических факторов за последние 12 и 3 месяца соответственно.

Классическая модель выживаемости предполагает [5], что интересующее событие в конечном итоге наступит. Однако область кредитного скоринга большое количество цензурированных данных, к тому же лишь малая часть заемщиков переходит в категорию должников. Модели, носящие название *Mixture cure models*, с макроэкономическими параметрами [6] позволяют сначала определить на основе статических данных к какой категории относится клиента, а после внедрения динамических данных вычислить вероятность выживаемости. Так как банки заинтересованы не только в том, чтобы клиент исполнил все свои обязательства, но и сделал это в определенное время, то модель может быть расширена [8] для прогнозирования досрочного погашения кредита.

Наличие в модели внешнеэкономических факторов позволяет не только повысить точность прогнозирования, но и моделировать риски в различных экономических условиях [4][7].

III. Используемые данные

A. *Описание данных*

Для проверки идеи о влияния макроэкономики на кредитоспособность клиентов были использованы транзакционные данные клиентов банка Санкт-Петербург, которые брали потребительский кредит за 2014–2016 года. Используемая выборка состояла из 31 тысяч записей. Данные были предоставлены в агрегированном виде и содержали в себе следующие показатели: дата заключения кредитного договора, количество транзакций, средняя сумма, которая приходилась на одну транзакцию, средняя сумма, которая обналичивалась, доли транзакций, которые относятся к здравоохранению, покупке продуктов питания, покупке предметов гардероба и финансовой сфере, и многие другие. В качестве целевой переменной был взят индикатор, показывающий допустил ли заемщик просрочку хотя бы одного платежа более чем на 30 дней.

В качестве макроэкономических параметров были взяты следующие показатели:

- курс доллара по отношения к рублю;
- цена баррели нефти Brent;
- средняя сумма заявки на потребительский кредит;
- количество заявок на потребительский кредит за месяц;
- средний размер нового депозита;
- количество новых депозитов за месяц.

Значения последних четырех параметров были получены из информационного ресурса «Открытые данные Сбербанк» [9]. Динамика изменений рассматриваемых показателей представлены на рис. 1–3.

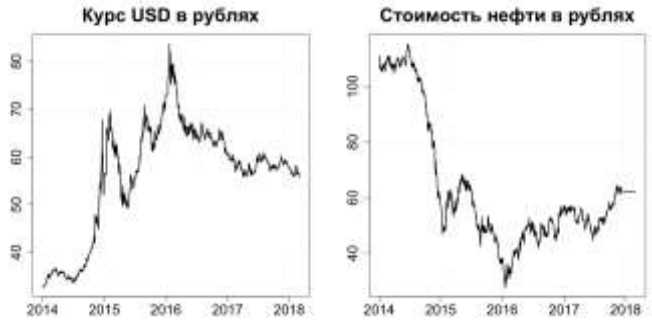

Рис. 1. Динамика курса доллара и стоимости нефти

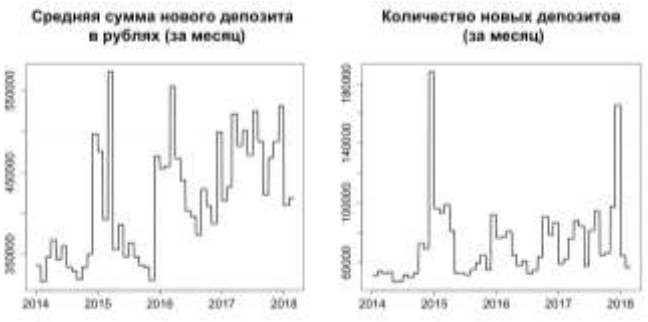

Рис. 2. Динамика открытие новых депозитов

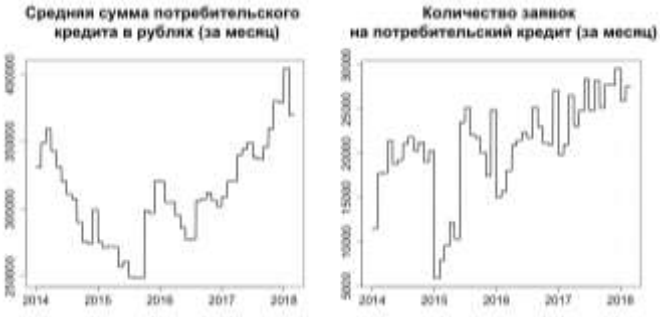

Рис. 3. Динамика заявок на потребительский кредит

Недобросовестных заемщиков существенно меньше, чем тех, возвращает кредит просрочек. В рассматриваемой выборке был около 600 записей о «плохих» клиентах, то есть около 2 %.

B. *Предобработка данных*

Во временных рядах, описывающих такие макроэкономические параметры, как курс доллара и стоимость нефти, были пропущенные значения. Например, в выходные и праздничные дни курс доллара не обновляется, поэтому в используемой выборке соответствующие значения были пропущены. Такие пропуски были заполнены последними непустыми значениями.

Данные, которые относятся банковским макроэкономических параметрам, описывают средние или суммарные значения за месяц: 15 числу каждого месяца соответствует одна запись в выборке. Эти значения были продублированы для всех остальных дней рассматриваемого месяца.

IV. Модель прогнозирования

Модель для прогнозирования построена на основе логистической регрессии, так как данный подход является классическим в задаче кредитного скоринга. В основе логистической регрессии лежит сигмоидная функция:

$$f(z) = \frac{1}{1+e^{-z}}, \text{ где}$$

$$z = a_0 + a_1 \times x_1 + ... + a_n \times x_n ,$$

$$a_i - \text{параметры регрессии}, i = \overline{1,n},$$

$$x_j - \text{значения независимых характеристик}, j = \overline{1,n}.$$

A. Отбор информативных признаков

Отбор наиболее информативных признаков был осуществлен с помощью метрики *Information Value*:

$$IV = \sum_{i=1}^{n} (DistrGood_i - DistrBad_i) \cdot \ln(\frac{DistrGood_i}{DistrBad_i}),$$

$$feature - \text{рассматриваемый признак},$$

$$value_i - \text{уникальные значения } feature, i = \overline{1,n},$$

$$target - \text{прогнозируемый признак},$$

$$DistrGood_i = P(feature = value_i \mid target = 1),$$

$$DistrBad_i = P(feature = value_i \mid target = 0).$$

Пороговые значения для классификации информативности признаков представлены в табл. 1.

ТАБЛИЦА I ПОРОГОВЫЕ ЗНАЧЕНИЯ ДЛЯ *INFORMATION VALUE*

Значение Information Value	***Информативность***
$0 \le$ ***IV*** < 0.02	Признак неинформативный
$0.02 \le$ ***IV*** < 0.1	Слабая
$0.1 \le$ ***IV*** < 0.3	Средняя
$0.3 \le$ ***IV*** $<+\infty$	Сильная

Для вычисления *Information Value* необходимо провести дискретизацию вещественных признаков, которые преобладают в рассматриваемой выборке. Разбиение на категории было реализовано в соответствии 0.25, 0.5 и 0.75 квантилями. Информативные признаки представлены в табл. 2.

ТАБЛИЦА II ИНФОРМАТИВНЫЕ ПРИЗНАКИ

Признак	***Значение Information Value***
Кредитная ставка по кредиту	0.268
Возраст	0.247
Mean_mon_pay	0.073
Mean_mon_trans	0.063
Mean_trans	0.049
Mean_day_trans	0.038
Mean_pay	0.036
Mean_day_pay	0.034
Сумма выданного кредита	0.033

B. Оценка качества модели

Ввиду несбалансированности классов в выборке для оценивания качества обучения модели была выбрана площадь под ROC-кривой.

V. РЕЗУЛЬТАТЫ

Данные были разбиты на обучающую и тестовые выборки. В ходе подбора оптимального соотношения размеров этих выборок были проведены эксперименты. Доля тестовой выборки изменялась от 0.05 до 0.95 с шагом 0.1. Для каждого из этих значений процесс обучения и прогнозирования на случайно выбранных подмножествах данных повторялся 100 раз. Усредненные результаты представлены на рис. 4.

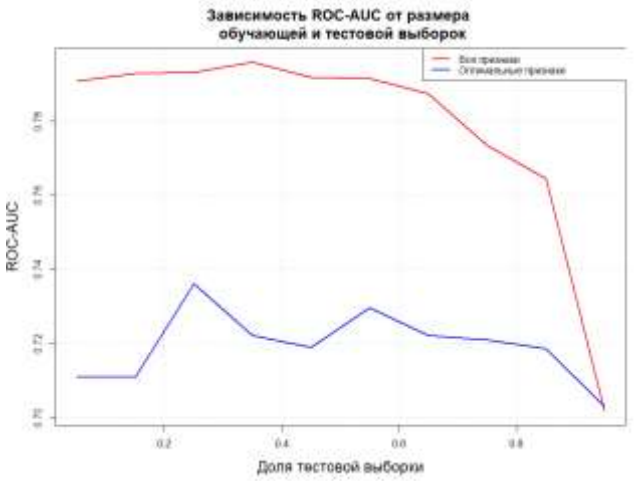

Рис. 4. Результаты подбора оптимального соотношения размеров тестовой и обучающей выборок. На основе полученных данных было принято решение выбрать для обучения 70% от всей выборки.

В рамках вычислительных экспериментов по прогнозированию дефолта макроэкономические показатели брались с различными отступами от даты заключения договора, так как динамика состояния экономики не отражаются на финансовом рынке мгновенно. Отступы принимали значения 3, 6 и 12 месяцев. Эксперименты проводились с учетом макроэкономических показателей (табл. 4) и без них (табл. 3), на полной выборке и на ее подмножестве, содержащей только наиболее информативные признаки.

ТАБЛИЦА III ЗНАЧЕНИЯ ROC-AUC, ПОЛУЧЕННЫЕ БЕЗ УЧЕТА МАКРОЭКОНОМИЧЕСКИХ ПАРАМЕТРОВ

Признаки	**Значение Information Value**
Все признаки	0.7953151
Информативные признаки	0.7342109

ТАБЛИЦА IV ЗНАЧЕНИЯ ROC-AUC, ПОЛУЧЕННЫЕ С УЧЕТОМ МАКРОЭКОНОМИЧЕСКИХ ПАРАМЕТРОВ

Признаки	**Отступы макроэкономических параметров[a], мес.**						***ROC-AUC***
	usd	***brent***	***avg crd***	***avg dps***	***cnt crd***	***cnt dps***	
Все	12	3	6	6	12	12	0.814
Информа-тивные	3	3	12	12	3	12	0.755

[a] В таблице отображены лучшие результаты, которые были получены в результате перебора отступов для макроэкономических параметров.

VI. Итоги

Добавление макроэкономических параметров в обучающую выборку положительно повлияло на качество обучения модели. На полной выборке метрика ROC–AUC возросла на 0.019, а на выборке, состоящей из информативных с точки зрения метрики *Information Value* признаков — на 0.021.

Отбор информативных признаков нельзя назвать успешным решением задачи уменьшения размерности. Большую роль в подсчете метрики *Information Value* для вещественных переменных играет выбранный способ дискретизации.

Список литературы

[1] Shen S.-W., Nguyen T.-D., & Ojiako U. (2013). Modelling the predictive performance of credit scoring. Acta Commercii, 13(1), 12–pages. https://doi.org/10.4102/ac.v13i1.189

[2] Bellotti T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. Journal of the Operational Research Society, 60(12), 1699–1707. https://doi.org/10.1057/jors.2008.130

[3] Im J.K., Apley D.W., Qi C., & Shan X. (2012). A time-dependent proportional hazards survival model for credit risk analysis. Journal of the Operational Research Society, 63(3), 306–321. https://doi.org/10.1057/jors.2011.34

[4] Bellotti T., & Crook J. (2013). Forecasting and stress testing credit card default using dynamic models. International Journal of Forecasting, 29(4), 563–574. https://doi.org/10.1016/j.ijforecast.2013.04.003

[5] Clark T.G., Bradburn M.J., Love S.B., & Altman, D.G. (2003). Survival Analysis Part I: Basic concepts and first analyses. British Journal of Cancer, 89(2), 232–238. https://doi.org/10.1038/sj.bjc.6601118

[6] Dirick L., Bellott, T., Claeskens G., & Baesens B. (2017). Macro-Economic Factors in Credit Risk Calculations: Including Time-Varying Covariates in Mixture Cure Models. Journal of Business and Economic Statistics, pp. 1–14. https://doi.org/10.1080/07350015.2016.1260471

[7] Simons D., & Rolwes F. (2008). Macroeconomic default modelling and stress testing, 1–31.

Dirick L., Claeskens G., & Baesens B. (2015). An Akaike information criterion for multiple event mixture cure models. European Journal of Operational Research, 241(2), 449–457. https://doi.org/10.1016/j.ejor.2014.08.