**ОАО «СБЕР БАНК»**

**ОТЧЕТ**

**о построении интегральной модели для розничных клиентов с учетом скор балла нейронной сети и скор балла бустинговой модели на среднесрочном таргете**

Минск

2024

**Оглавление**

[1. Общие положения 3](#_Toc181114642)

[1.1. Цель документа 3](#_Toc181114643)

[1.2. Структура документа 3](#_Toc181114644)

[2. Структура модели 4](#_Toc181114645)

[2.1. Описание структуры модели 4](#_Toc181114646)

[3. Сбор данных 5](#_Toc181114647)

[3.1 Общая схема процесса сбора данных и источники данных 5](#_Toc181114648)

[3.2 Процесс сборки данных из нескольких файлов. Недостатки предоставленных данных. 5](#_Toc181114649)

[4 Проведение однофакторного анализа 7](#_Toc181114650)

[4.1 Оценка дискриминационной силы факторов 7](#_Toc181114651)

[4.2 Анализ корреляций 7](#_Toc181114652)

[5 Построение модели 7](#_Toc181114653)

[5.1 Описание многофакторного анализа 7](#_Toc181114654)

[5.2 Проведение многофакторного анализа 8](#_Toc181114655)

[6 Калибровка модели 8](#_Toc181114656)

[7 Тестирование модели 9](#_Toc181114657)

[7.1 Определение выборки для тестирования 9](#_Toc181114658)

[7.2 Результаты тестирования 9](#_Toc181114659)

[Приложение 10](#_Toc181114660)

1. Общие положения
   1. **Цель документа**
      1. Данный отчет создан с целью обеспечения непрерывности бизнес-процессов, сохранения и распространения внутренних экспертных знаний внутри Группы, а также для подтверждения обоснованности и корректности разработанной модели.
      2. Для достижения этих целей Отчёт подробно описывает основные методологические подходы к моделированию, а также предпосылки и решения, которые принимались на каждом из этапов разработки модели.
      3. Данный отчет написан в соответствии с «Регламентом управления моделями количественной оценки рисков».
      4. Цель модели – расчет интегрального скор-балла для принятия решения по заявке на потребительский кредит, кредитную и овердрафтные карты, а также оценка вероятности выхода клиента на просрочку 90 дней в первые 12 месяцев с даты выдачи продукта.
      5. По требованию бизнеса требуется разработать интегральную модель, учитывающую скор-балл по модели БКИ (нейронная сеть) и скор-балл бустинговой модели.

## Структура документа

* + 1. Данный отчет имеет следующую структуру:
* Раздел «Структура модели» описывает общую логику получения и применения итогового скорингового балла для сделки
* Раздел «Сбор данных» содержит описание процедуры формирования выборки для построения модели, а также описание подходов к фильтрации данных и обработке проблемных данных
* В разделе «Построение модели» приводится построение интегральных моделей, их калибровка, тестирование калибровок.
  + 1. Основные идентификационные данные модели собраны в Таблице 1.

Таблица 1 – Идентификаторы модели

|  |  |
| --- | --- |
| Название модели / ID |  |
| Заказчик модели | *ОАО «Сбер Банк»* |
| Название участника Группы и структурного подразделения верхнего уровня | *ОАО «Сбер Банк»* |
| Разработчик модели (в случае привлечения внешнего разработчика, необходимо указать название компании) | *ОАО «Сбер Банк»:*  *Разработка:*  *Иванович А. И.* |

* + 1. Сегментная спецификация модели представлена в Таблице 2.

Таблица 2 - Описание сегментной специфики модели

|  |  |
| --- | --- |
| **Параметр** | **Значение** |
| Сегмент / Тип портфеля | Розничное кредитование |
| Потребительские кредиты |
| Тип продукта | Кредиты |
| Тип клиента | Физические лица |
| Тип модели | PD |

* + 1. Проведение однофакторного анализа в сводной форме иллюстрировано в Таблице 3.

Таблица 3 – Сводное описание однофакторного анализа для модели

|  |  |  |
| --- | --- | --- |
| ***Этап построения модели*** | ***Структура документа*** | ***Описание*** |
| *Однофакторный анализ* | *Оценка дискриминирующей силы факторов* | *Оценка прогнозной силы каждого фактора путём расчёта индекса Джини для каждого фактора.* |
| *Анализ корреляций* | *Анализ парных корреляций.* |

* + 1. Проведение многофакторного анализа в сводной форме иллюстрировано в Таблице 4.

Таблица 4 - Сводное описание многофакторного анализа для модели

|  |  |  |
| --- | --- | --- |
| ***Этап построения модели*** | ***Структура документа*** | ***Описание*** |
| *Многофакторный анализ и построение модели* | *Результаты многофакторного анализа* | *Вычисление весов факторов путём построения логистической регрессии на конечном наборе факторов и последующая нормировка баллов.* |

# Структура модели

## Описание структуры модели

* + 1. Данная модель представляет собой интегральную модель, учитывающую скор-балл по модели БКИ (НС) и скор-балл бустинговой модели.
    2. Перечень необходимых данных состоит из результатов применения моделей: логитов модели БКИ и скоров бустинговой модели.
    3. Собранная статистика по факторам была проанализирована для того, чтобы оценить дискриминирующий эффект каждого отдельно взятого фактора (однофакторный анализ). После этого были рассчитаны показатели корреляции между факторами. По результатам данного анализа были исключены факторы, которые имели низкую предсказательную силу и/или высокую корреляцию с другими факторами.
    4. Многофакторный анализ заключался в определении весов для каждого фактора. Процедура многофакторного анализа была направлена на то, чтобы из списка факторов включить в модель только такой набор, который с оптимальными весами даст наибольшую предсказательную силу модели.
    5. Построенная модель применена к данным: проанализированы дискриминационные свойства модели на тестовой выборке и принято решение о ее устойчивости и применимости модели к сделкам, составляющим текущий портфель.

# Сбор данных

## Общая схема процесса сбора данных и источники данных

* + 1. Запрос данных был разработан на основе опыта команды ОАО «Сбер Банк» с учетом наличия данных и структуры модели. Итоговый запрос был согласован с представителями ОАО «Сбер Банк» от бизнеса.
       1. Признак дефолта предоставлен ОАО «Сбер Банк» в привязке к номеру заявки.
    2. Логиты предсказаний модели БКИ переведены в неоткалиброванную вероятность дефолта по формуле ниже:

После преобразования выше требуется откалибровать вероятность по приведенном ниже выражению. Эта калибровка обеспечит, что откалиброванные значения вероятности будут оставаться в допустимых пределах и не приведут к отрицательным значениям скоров БКИ (отрицательный скор означает очень высокое значение PD).

Вероятности дефолта 0.8985 соответствует скор баллу равному 0.1. Значение 0.1 получена по выражению для преобразования PD в score ниже.

Для преобразования PD в Score воспользуемся следующей формулой (Z = logits):

* + 1. БКИ скор балл был отшкалирован, чтобы удовлетворять следующим условиям:

минимальный балл за каждый фактор – 0 баллов;

Скоринговый балл 444 соответствует шансу, равному 10;

Балл за удвоение шансов (Points to double odds) равен 69;

Решив систему уравнений, получим: A = 99, B = 216

* + 1. Скор балл БКИ обработан по следующей логике: если значение балла равно 0, для этой заявки скор балл БКИ не мог быть рассчитан в момент заведения заявки, нули были заменены на пропуски.
    2. Предсказания бустинговой модели также были преобразованы в скор по выражению выше.

## Процесс сборки данных из нескольких файлов. Недостатки предоставленных данных.

* + 1. Данные были получены из выборки для разработки модели БКИ, выборки для разработки бустинговой модели.
    2. Описание входных файлов.

Данные ОАО «Сбер Банк» и от разработки предыдущих моделей ОАО «Сбер Банк»:

1. «neuro\_bki\_predictions.csv» - данные со логитами предсказаний (логиты требуется перевести в вероятность, откалибровать и перевести в скор балл), номером заявки и таргетом.
2. «boosting\_predictions.csv» - данные со PD бустинговой модели (требуется перевести в SCORE балл), номерами заявок и таргетом.

Основным полем для склейки файла является ИД договора, типично имеющее следующий узнаваемый вид: «CRMPROD\_1-4HJ6MVQ\_00000000000000», обычно называемое «CRM\_ID», но может называться и по-другому: «Номер заявки», «APP\_ID», и так далее - названия разные в разных файлах.

* + 1. Полученная выборка была разбита на две подвыборки:

- обучающую (70% выборки),

- тестовую (30% выборки).

* + 1. В таблице 5 приведено сводное описание выборок

Таблица 5 – сводное описание выборок

|  |  |  |
| --- | --- | --- |
|  | Выборка для обучения модели | Выборка для валидации (out-of-sample) |
| Временной период формирования выборки | 11.2022 – 08.2023 | 11.2022 – 08.2023 |
| Определение целевого признака | Выход в просрочку 90+ дней в течение 12 месяцев после выдачи кредита | |
| Количество записей в выборке | 26452 | 11337 |
| Количесвто записей с положительным целевым признаком | 512 | 220 |
| Уровень целевого признака | 1.93 % | 1.94 % |

* + 1. Данные выложены на внутренний сервер:

W:\WG-DataScience\BICC\Модели\Отчеты\Риски\Розничный риск\Интегральные модели\Интегральная модель НС БКИ + бустинг

* + 1. По данным обучающей выборки расчитаны средние значения для факторов score\_bki при отсутствии этих значений. Для этого было найдено среднее значение скор-балла по выборке с непустыми значениями для данного фактора.

|  |  |
| --- | --- |
| Параметр | Среднее значение |
| SCORE\_BKI | 295 |

# Проведение однофакторного анализа

## Оценка дискриминационной силы факторов

* + 1. В Таблице 6 содержится информация по дискриминационной силе факторов. Все факторы обладают высокой дискриминационной силой, выше порогового значения 5%, поэтому ни один фактор не был исключен.

Таблица 6 – Дискриминационная сила факторов

|  |  |
| --- | --- |
| Фактор | Gini на валидационной выборке |
| SCORE\_BKI | 70.88% |
| SCORE\_BOOSTING | 40.76% |

## Анализ корреляций

* + 1. Получившаяся матрица корреляций представлена в таблице 7:

Таблица 7 – Матрица корреляций между факторами

|  | **SCORE\_BKI** | **SCORE\_BOOSTING** |
| --- | --- | --- |
| **SCORE\_BKI** | 1.000000 | 0.29228 |
| **SCORE\_BOOSTING** | 0.29228 | 1.000000 |

* + 1. Пороговое значение корреляции Пирсона для включения в короткий список факторов обоих факторов – не более 0.7, поэтому ни один фактор не был исключен.

# Построение модели

## Описание многофакторного анализа

* + 1. Многофакторный анализ являлся следующим шагом после проведения однофакторного анализа. Объектом анализа являлись преобразованные факторы из короткого списка. Целью многофакторного анализа является выбор из короткого списка комбинации факторов и подбор соответствующих им весов, обеспечивающих наибольшую прогнозную силу и стабильность модели. Результатом анализа является линейная комбинация факторов с оптимальными весами.
    2. Многофакторный анализ реализовывался через применение многомерной логистической регрессии, которая позволила определить веса объясняющих факторов, обеспечивающую наибольшую предсказательную силу модели. Формула регрессии имеет вид:



β1, β 2, β 3, ..., β n – веса факторов,

X1, X2, X3,…,Xn – значения факторов,

Y – флаг дефолта.

* 1. **Проведение многофакторного анализа**
     1. Рассчитанные значения VIF представлены в Таблице 8. По результатам анализа мультиколлинеарности ни одного фактора исключено не было

Таблица 8 – Анализ мультиколлинеарности. VIF значения факторов

|  |  |
| --- | --- |
| Фактор | VIF, на обучающей выборке |
| SCORE\_BKI | 1.0863 |
| SCORE\_BOOSTING | 1.0863 |

* + 1. Построив логистическую регрессию, получаем ПД, его можно пересчитать в интегральный скор по формуле:
    2. Подставив выражение для ПД в виде логистической регрессии от скор-балла НС и скора бустинговой модели приходим к формуле для интегрированного скора.
    3. Значение индекса Джини полученной модели на всей обучающей выборке составило 72.09 Полученное значение индекса Джини свидетельствует о хорошей ранжирующей способности модели.

1. Калибровка модели
   1. Калибровка модели осуществляется с целью трансформации рассчитанных итоговых баллов в показатель вероятности дефолта. Трансформация балла в вероятность дефолта реализуется на базе логистической функции. Результатом процедуры калибровки на тренировочных данных является калибровочная функция
2. Тестирование модели
   1. **Определение выборки для тестирования**
      1. Тестовая выборка охватывает тот же период, что и обучающая выборка. Выборка представляет собой случайным образом отобранные 30% исходной выборки за данный период, не использованные при разработке модели. При этом, разбиение на тестовую и тренировочную выборки проходило с учетом пропорционального объема дефолтов. Итоговый объём данной тестовой выборки составил 11337 записи, из которых 220 являются дефолтными.
   2. **Результаты тестирования**
      1. Индекс Джини на тестовой выборке составил 72.09. Данное значение незначительно отличается от индекса Джини на выборке, использованной в разработке. Таким образом, можно сделать вывод об устойчивости качества полученного скорингового балла.
      2. Построение доверительных интервалов на тестовой выборке после проведения калибровки приведено на рисунке 1.

Рисунок 1 – доверительные интервалы по средним уровням дефолтов согласно кредитному рейтингу на тестовой выборке