

Отчёт о разработке предиктивной модели на базе логистической регрессии

«**@str!model\_name@**»

Москва, **@str!year@**

**ОГЛАВЛЕНИЕ**

[1. ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 3](#_Toc146804804)

[2. ОБЩИЕ ПОЛОЖЕНИЯ 4](#_Toc146804805)

[2.1. Цель документа 4](#_Toc146804806)

[2.2. Структура документа 4](#_Toc146804807)

[3. ДАННЫЕ ДЛЯ МОДЕЛИРОВАНИЯ 5](#_Toc146804808)

[3.1. Источники данных для моделирования 5](#_Toc146804809)

[3.2. Определение Целевого события 5](#_Toc146804810)

[3.3. Подготовка данных для моделирования 5](#_Toc146804811)

[3.4. Формирование выборки для разработки модели 5](#_Toc146804812)

[3.5. Длинный список факторов 5](#_Toc146804813)

[3.6. Анализ качества исходных данных 5](#_Toc146804814)

[3.7. Разделение выборки 5](#_Toc146804815)

[4. РАЗРАБОТКА МОДЕЛИ 5](#_Toc146804816)

[4.1. Описание архитектуры модели 5](#_Toc146804817)

[4.2. Подход к однофакторному анализу 5](#_Toc146804818)

[4.3. Результаты однофакторного анализа 7](#_Toc146804819)

[4.4. Подход к многофакторному анализу 7](#_Toc146804820)

[4.5. Результаты многофакторного анализа 8](#_Toc146804821)

[4.6. Экспертная корректировка списка отобранных факторов 8](#_Toc146804822)

[5. КАЛИБРОВКА МОДЕЛИ 8](#_Toc146804823)

[5.1. Общие положения 9](#_Toc146804824)

[6. ИТОГОВАЯ МОДЕЛЬ 10](#_Toc146804825)

[ПРИЛОЖЕНИЕ 11](#_Toc146804826)

[Приложение 1: Алгоритм формирования Витрины для разработки 11](#_Toc146804827)

[Приложение 2: Анализ качества исходных данных 11](#_Toc146804828)

[Приложение 3: Длинный список факторов модели 11](#_Toc146804829)

[Приложение 4.1: Результаты однофакторного анализа 11](#_Toc146804830)

[Приложение 5: Результаты многофакторного анализа 11](#_Toc146804831)

[Приложение 6: Анализ PSI Итоговой модели 11](#_Toc146804832)

1. ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

**Банк** – Банк ВТБ (ПАО).

**Отчет** – настоящий отчет о разработке

**Выборка** – подмножество данных из общего набора данных, которое используется для обучения, валидации или тестирования модели. Выборка представляет собой конкретные примеры или наблюдения на которых модель будет обучаться.

**Целевое событие (ЦС)** – событие, которое модель пытается предсказать на основе входных параметров (признаков) (напр. дефолт)

**Нецелевое событие (НЦС)** – событие, которое не является предметом предсказания (напр. не-дефолт)

**Дискретные факторы** – факторы с ограниченным количеством вариантов значений

**Непрерывные факторы** – факторы с неограниченным количеством вариантов значений

**Преобразование факторов** – замена значений факторов на расчётные величины (номера групп, скоры, WOE), связанные с оценкой вероятности дефолта, относящейся к значению фактора

**Однофакторный анализ** – анализ данных по выборке сделок с целью определения предсказательной силы отдельных факторов

**Многофакторный анализ** – анализ комбинаций факторов с целью определения модели, наиболее точно дифференцирующей “положительные” и “отрицательные” классы

**Предиктивная модель** – математическая модель которая используется для оценки риска или вероятности наступления определенных событий на основе набора данных и характеристик.

**Балл модели** – рейтинг, присвоенный предиктивной моделью

1. ОБЩИЕ ПОЛОЖЕНИЯ
   1. Цель документа

Целью Отчета является описание процесса разработки предиктивной модели на базе логистической регрессии.

Отчет не является практическим руководством, не содержит полномерную теоретическую основу по разработке моделей количественной оценки риска, а описывает основные шаги и предпринятые действия в процессе разработки.

* 1. Структура документа
* Раздел «Данные для моделирования» содержит описание процедуры формирования выборки для построения модели, а также описание подходов к фильтрации данных и обработке проблемных данных;
* Раздел «Разработка модели» описывает процесс трансформации и отбора наиболее информативных факторов в модель;
* Раздел «Калибровка модели» посвящён описанию алгоритма калибровки модели;
* Раздел «Итоговая модель» содержит информацию об итоговой модели, а также анализ ее предиктивных качеств и стабильности

1. ДАННЫЕ ДЛЯ МОДЕЛИРОВАНИЯ
   1. Источники данных для моделирования
   2. Определение Целевого события
   3. Подготовка данных для моделирования
   4. Формирование выборки для разработки модели

**В Приложении 1 Отчета приведен алгоритм формирования витрины для разработки модели**

* 1. ****Длинный список факторов****

Длинный список факторов представляет собой набор показателей, на основе которого формируются будущие факторы предиктивной модели.

**В Приложении 2 Отчета приведен длинный перечень факторов**

* 1. Анализ качества исходных данных

**В Приложении 3 Отчета приложен файл с анализом качества исходных данных**

* 1. Разделение выборки

Для дальнейшего наализа выборка была разделена на сэмплы для обучения и тестирования. Статистика по сэмплам представлена в Таблице 1

Таблица 1. Статистика по сэмплам выборки для разработки

|  |
| --- |
| Характеристика |
| **@tbl!samples\_stat@** |

1. РАЗРАБОТКА МОДЕЛИ
   1. Описание архитектуры модели
   2. Подход к однофакторному анализу

Основной задачей однофакторного анализа является определение списка переменных, которые позволяют ранжировать заемщиков по доле ЦС. Однофакторный анализ включал в себя следующие этапы:

* трансформация факторов;
* оценка предиктивной способности факторов;
* оценка бизнес логики факторов;

Объектом исследования в процессе однофакторного анализ является длинный список переменных

* + 1. Алгоритм трансформации факторов

Трансформация всех факторов была произведена на базе *WOE* (*Weight of evidence*) преобразования. *WOE* характеризует степень отклонения уровня дефолтов по группе от среднего уровня дефолтов по всей выборке. Показатель *WOE* может принимать как положительные, так и отрицательные значения. Положительные значения *WOE* означают, что рассматриваемая группа имеет более низкую долю ЦС, чем выборка в целом, если значение *WOE* меньше нуля, то группа имеет более высокую долю ЦС, чем во всей выборке. Таким образом, значения фактора в каждой группе заменяются соответствующими значениями *WOE* по группам. Для группы с номером *i* величина вычисляется следующим образом:

,

где:

* - общее кол-во наблюдений,
* - общее кол-во наблюдений класса 1,
* - общее кол-во наблюдений класса 0,
* - общее кол-во наблюдений класса 1 в бине i ,
* - общее кол-во наблюдений класса 0 в бине i ,
* - коэффициент регуляризации,
* - корректировочная поправка

Добавление величины к показателям и производится для того, чтобы корректно произвести расчет в случае равенства их нулю. Для непрерывных факторов группировка осуществлялась таким образом, чтобы в каждый диапазон попадали наблюдения с сопоставимыми уровнями ЦС. Если доля ЦС является стабильной для некоторого интервала значений фактора, и существенно изменяется для отдельных значений, то все наблюдения из этого интервала могут быть объединены в одну группу. В результате процесса группировки непрерывный фактор делится на несколько групп, для каждого из которых возможно оценить уровень ЦС на базе наблюдений, попавших в этот диапазон. Группировка переменных с дискретным набором значений осуществляется аналогично группировке непрерывных факторов – на основании сопоставимых уровней ЦС в разных категориях переменной. В каждую группу может попадать одно или несколько значений фактора. Доля ЦС вычисляется по всем наблюдениям, входящим в группу.

* + 1. Оценка качества трансформации и отбор факторов

Для оценки предиктивной способности факторов использовался индекс GINI. Пороговое значение индекса GINI для отбора переменных в короткий список было установлено на уровне 10%. Переменные с меньшей величиной ранжирующей способности исключались из дальнейшего анализа. Расчет значений показателя GINI производился по WOE-факторам. Так же в процессе однофакторного анализа для каждой переменной был проведено тестирование её бизнес логики – анализ характера изменения зависимости доли ЦС, при изменении значений переменной. В отношении каждого фактора было сделано предположение о характере такой зависимости – растущая, если доля ЦС растет при увеличении значения переменной или падающая, если доля ЦС уменьшается с ростом значения фактора, которое далее было сопоставлено с наблюдаемой зависимостью уровня риска от значения WOE-фактора. Если на практике вид зависимости совпадал с ожидаемым поведением уровня ЦС, то такие переменные отбирались в короткий список факторов для дальнейшего анализа, в ином случае переменные исключались из процесса разработки модели. Качество биннига фактора оценивается по результатам следующих проверок:

* на тренд WOE – он может быть либо монотонным, либо V (или /\)-образным
* стабильность тренда WOE – тренд должен сохраняться на всех тестовых сэмплах, и доверительные интервалы WOE в соседних бинах не должны пересекаться
* стабильность Gini – Gini на всех тестовых сэмплах не должно сильно отличаться относительно обучающей выборки и все значения должны попадать в доверительные интервалы, рассчитанные бутстрэпом.
  1. Результаты однофакторного анализа

Графические и аналитические результаты биннинга приведены в Приложении 4 к Отчету.

* 1. Подход к многофакторному анализу

Целью многофакторного анализа является отбор из короткого списка факторов совокупности переменных, которые статистически значимо связаны с целевой функцией, а также оценка весов коэффициентов отобранных факторов, для расчёта итогового скорингового балла модели. Объектом многофакторного анализа являются длинный список трансформированных переменных по каждому набору данных, который были получен на этапе однофакторного анализа. Наиболее распространенным статистическим алгоритмом, который устанавливает связь между бинарной зависимой переменной и набором независимых факторов, является множественная логистическая регрессия:

где – зависимая переменная, признак дефолта, – событие дефолта, – набор независимых, объясняющих факторов, – коэффициенты логистической регрессии, – вероятность ЦС. Данный алгоритм был использован в процессе многофакторного анализа для разработки скоринговых моделей по каждому набору переменных.

Многофакторный анализ включал в себя следующие этапы:

* Исключение факторов с низкой предиктивной способностью (gini < 15)
* Исключение факторов с нестабильным распределением во времени (максимальный индекс PSI > 0.25)
* анализ парных корреляций;
* пошаговыйотбор *(stepwise selection)* факторов в модель
* расчет эффективных весов переменных для итоговой модели.

Для оценки парных корреляций по каждому набору трансформированных *WOE*-переменных был проведен расчет матрицы корреляций. Анализ данной таблицы позволил определить переменные, имеющие высокие линейные связи с другими факторами. В качестве порогового значения, при котором переменные считались сильно коррелирующими друг с другом, был выбран уровень в 70% и более. При указанных значениях корреляций между двумя переменными для исключения выбирался тот фактор, который имел наименьшую ранжирующую способность.

Для отбора факторов в итоговые модели, в процедуре множественной логистической регрессии использовались пошаговыйотбор *(stepwise selection)* – метод, который заключается в последовательном включении факторов в модель. При этом, как только новый фактор входит в модель, производиться проверка на значимость всех переменных модели, и исключаются незначимые факторы, если такие имеются. Метод останавливается, когда ни один фактор не может быть включен или исключен из модели на основании выбранного порога значимости.

В каждой из описанных процедур отбора переменных использовался порог по уровню статистической значимости в 5%. Таким образом, факторы с уровнем значимости выше данного порога исключались из модели.

Также в процессе отбора переменных в итоговую модель проводился контроль знаков регрессионных коэффициентов. Все коэффициенты при *WOE*-факторах, должны принимать отрицательные значения. При смене знака коэффициента фактора с отрицательного на положительный, производилось исключение таких переменных.

На завершающем этапе многофакторного анализа для переменных, входящих в модель производился расчет значений их эффективных весов, согласно следующей формуле:

где – эффективный вес переменной модели с номером *i, –* вес переменной модели после многофакторного анализа.

* 1. Результаты многофакторного анализа

Список факторов, вошедших в модель по результатам многофакторного анализа приведен в таблице 2.

Подробные результаты многофакторного анализа приведены в приложении 5 к Отчету

* 1. Экспертная корректировка списка отобранных факторов

1. КАЛИБРОВКА МОДЕЛИ

**@case!start\_calibration@**

* 1. Общие положения

Калибровка предназначена для преобразования скорингового балла модели в вероятность наступления целевого события (PE). В процессе калибровки производиться поиск функции калибровки, которая имеет следующий вид:

где – коэффициент сдвига калибровочной кривой, – коэффициент наклона калибровочной кривой, – скоринговый балл базовой модели, – скоринговый балл базовой модели после процедуры калибровки. Далее, для каждого контрагента, вероятность целевого события вычисляется по формуле следующего вида:

Процедура калибровки заключается в определении таких значений калибровочных коэффициентов, при которых средний прогнозный уровень дефолта по портфелю контрагентов совпадет или очень близок к долгосрочному целевому уровню дефолтов, который задается значением центральной тенденции. При этом верхняя граница 99% доверительного интервала итоговой средней вероятности дефолта по всем клиентам в каждом грейде мастер-шкалы должна быть выше среднего наблюдаемого DR в этом же грейде, либо, если это условие невозможно удовлетворить во всех грейдах, то кол-во его нарушений должно быть минимальным. Далее нарушение этого условия будем называть «красное пробитие» и обозначать K в сигналах биноминального теста на калибровку. Так же возможны еще сигналы «желтое пробитие» (Ж) - выход наблюдаемого DR за верхнюю границу 95% доверительного интервала, «консервативное красное пробитие» (K\*) - выход наблюдаемого DR за нижнюю границу 99% доверительного интервала и «консервативное желтое пробитие» (Ж\*) - выход наблюдаемого DR за нижнюю границу 95% доверительного интервала. Последние три типа пробития не считались критически важными и соответственно не учитывались при расчете калибровочных коэффициентов.

Так как подбор функции калибровки проводиться на выборке данных, где соотношение дефолтных наблюдений и наблюдений не в дефолте не соответствует значению центральной тенденции, то для приведения этого соотношения к целевому значению вводятся поправочные веса наблюдений, значения которых может быть определено следующим образом:

где – вес для наблюдений в дефолте, – вес для наблюдений не в дефолте, , – количество наблюдений в дефолте и не в дефолте в выборке для калибровки соответственно, – целевое значение центральной тенденции. Для приведения уровня дефолтов в выборке для калибровки к значению центральной тенденции достаточно использовать вес либо для наблюдений в дефолте, либо вес для наблюдений не в дефолте. Для калибровки модели был использован вариант, при котором взвешиваются наблюдения в дефолте.

Результаты калибровки приведены в Приложении 7 к Отчету

**@case!end\_calibration@**

1. ИТОГОВАЯ МОДЕЛЬ

Итоговая модель приведена в Таблице 3 к Отчету

|  |
| --- |
| Фактор |
| **@tbl!report\_table@** |

Анализ стабильности популяции PSI факторов и итогового скора модели приведен в приложении 6.

ПРИЛОЖЕНИЕ

Приложение 1: Алгоритм формирования Витрины для разработки

Приложение 2: Анализ качества исходных данных

Анализ качества исходных данных приведен в файле *data\_check.xlsx*

Приложение 3: Длинный список факторов модели

|  |  |
| --- | --- |
| Фактор | Описание |
| **@tbl!factors\_table@** |  |

Приложение 4.1: Результаты однофакторного анализа

Графическое представление биннингов вошедших в модель переменных:

**@img!binnings@**

**Приложение 4.2.: Скорката**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| feature | group | values | woe | special\_bins |
| **@tbl!scorecard@** |  |  |  |  |

Приложение 5: Результаты многофакторного анализа

Результаты многофакторного анализа приведены в файле *mfa.xlsx*

Приложение 6: Анализ PSI Итоговой модели

Графическое представление анализа PSI итогового скора модели и всех факторов:

**@img!psi@**