(19) Евразийское патентное ведомство

(11) 038056 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2021.06.29

(21) Номер заявки
201700609

(22) Дата подачи заявки
2018.04.05

(51) Int. Cl. ***G06Q 40/02*** (2012.01)
G06N 5/00 (2006.01)

(54) КОМПЬЮТЕРИЗИРОВАННЫЙ СПОСОБ РАЗРАБОТКИ И УПРАВЛЕНИЯ МОДЕЛЯМИ СКОРИНГА

(31) **2017146235**
(32) **2018.04.04**
(33) **RU**
(43) **2019.10.31**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Травкин Олег Игоревич, Берестнев Дмитрий Алексеевич, Юдочев Дмитрий Владимирович, Жуковская Екатерина Сергеевна (RU)

(74) Представитель:
Астафьева С.А., Герасин Б.В. (RU)

(56) US-A1-20150019405
US-A1-20060212386
US-B1-8407139

(57) Данное изобретение в общем относится к области вычислительной техники, а в частности к способам автоматической разработки моделей кредитного скоринга и их автоматической имплементации в кредитный процесс. Компьютеризированный способ разработки и управления моделями скоринга, в котором получают данные за заданный период времени, содержащие факторы, влияющие на модель скоринга; после чего осуществляют разбиение полученных данных на выборки для разработки, валидации и тестирования модели скоринга; затем осуществляют трансформацию факторов посредством установления соотношений между группами значений преобразованного фактора и уровнями дефолтов; далее исключают из выборок по меньшей мере один преобразованный фактор, коррелирующий по меньшей мере с одним другим фактором; после чего формируют модель кредитного скоринга посредством обучения бинарной множественной логистической регрессии; и в итоге подбирают автоматически зоны отсечения по меньшей мере для одной модели скоринга для ее установки в кредитную процедуру. Технический результат - повышение качества создаваемых моделей кредитного скоринга.

038056 B1

038056 B1

Область техники

Данное изобретение в общем относится к области вычислительной техники, а в частности к способам автоматической разработки моделей кредитного скоринга и их автоматической имплементации в кредитный процесс.

Уровень техники

В настоящее время финансовые учреждения применяют стандартные статистические подходы к анализу исторических данных для описания возможных клиентов с точки зрения риска. Это позволяет классифицировать заемщиков на "хороших" и "плохих" и таким образом принимать окончательное решение о кредитовании. В большинстве кредитных учреждений созданы подразделения, разрабатывающие модели кредитного скоринга на основании собственной статистики с учетом специфики клиентского профиля. Однако данные кредитные учреждения часто обращаются в бюро кредитных историй, из-за чего процесс оценки кредитоспособности заемщика сильно затягивается и становится неточным, так как зависит от использованных алгоритмов бюро кредитных историй.

Сущность изобретения

Данное изобретение направлено на устранение недостатков, присущих существующим решениям, известным из уровня техники.

Технической проблемой (или технической задачей) в данном изобретении является осуществление автоматической разработки моделей кредитного скоринга с их последующей имплементацией в систему принятия решения и мониторингом.

Техническим результатом, проявляющимся при решении вышеуказанной задачи, является повышение качества создаваемых моделей кредитного скоринга. Дополнительным техническим результатом, проявляющимся при решении технической задачи, является увеличение скорости разработки моделей кредитного скоринга. Также снижается потребность в количестве ресурсов, необходимых для разработки и поддержки моделей, увеличение скорости и простоты внедрения моделей в промышленный контур, а также обеспечение мониторинга работы моделей и оперативной реакции на изменения.

Указанный технический результат достигается благодаря осуществлению способа разработки и управления моделями скоринга, в котором

получают данные за заданный период времени, содержащие факторы, влияющие на модель скоринга;

после чего осуществляют разбиение полученных данных на выборки для разработки, валидации и тестирования модели скоринга;

затем осуществляют трансформацию факторов посредством установления соотношений между группами значений преобразованного фактора и уровнями дефолтов;

далее исключают из выборок по меньшей мере один преобразованный фактор, коррелирующий по меньшей мере с одним другим фактором;

формируют модель кредитного скоринга посредством обучения бинарной множественной логистической регрессии;

подбирают автоматически зоны отсечения для по меньшей мере одной модели скоринга для ее установки в кредитную процедуру.

В некоторых вариантах осуществления получают данные за заданный период времени с мобильного устройства связи пользователя.

В некоторых вариантах осуществления при осуществлении разбиения полученных данных на выборки получают непересекающиеся во времени части исходной совокупности или случайные подвыборки.

В некоторых вариантах осуществления факторами, влияющими на модель скоринга, являются годовой доход, и/или размер непогашенного долга, и/или владение недвижимостью, и/или владение автомобилем, и/или стаж работы на последнем месте, и/или возраст.

В некоторых вариантах осуществления факторы, влияющие на модель скоринга, являются дискретными или непрерывными.

В некоторых вариантах осуществления при осуществлении трансформации факторов определяют степень отклонения уровня дефолтов по группе данных от среднего уровня дефолтов по всей выборке.

В некоторых вариантах осуществления при осуществлении трансформации факторов по факторам, попавшим в список исключенных, запускают алгоритм разбиения значений факторов с новым набором настроек.

В некоторых вариантах осуществления при исключении из выборок преобразованных факторов формируют таблицу со значениями коэффициентов парных корреляций преобразованных факторов.

В некоторых вариантах осуществления при исключении из выборок преобразованных факторов в цикле отбирают фактор, который имеет наибольшее количество коррелированных с ним факторов.

В некоторых вариантах осуществления при формировании модели кредитного скоринга строится логистическая модель с использованием пошаговой регрессии для отбора итогового набора факторов.

Краткое описание чертежей

Признаки и преимущества настоящего изобретения станут очевидными из приводимого ниже под-

робного описания изобретения и прилагаемых чертежей.

На фиг. 1 показан пример осуществления способа разработки и управления моделями скоринга в виде блок-схемы.

На фиг. 2 показана верхнеуровневая примерная схема осуществления способа разработки и управления моделями скоринга.

Основное ядро составляют два блока - это переобучение и подбор/корректировка зон отсечения, причем без адаптации зон отсечения невозможно организовать автоматическое внедрение модели в систему принятия решения. Результаты двух этих блоков интегрируются в промышленную среду (в данном варианте осуществления в SAS RTDM). Кроме того, каждый из этих двух блоков подвергается регламентным проверкам в виде ежедневного мониторинга целевого показателя, зависящего от зон отсечения (уровень одобрения) и ежемесячной валидации моделей.

Подробное описание изобретения

Данное изобретение может быть реализовано на компьютере, в виде автоматизированной системы (АС) или машиночитаемого носителя, содержащего инструкции для выполнения вышеупомянутого способа.

Изобретение может быть реализовано в виде распределенной компьютерной системы.

В данном решении под системой подразумевается компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность вычислительных операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных. В роли устройства хранения данных могут выступать, но, не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

Ниже будут описаны термины и понятия, необходимые для осуществления изобретения.

Кредитный скорииг - это метод моделирования кредитного риска заемщика, основанный на численных статистических методах. Назначение кредитного скоринга - принятие решений по выдаче кредитов физическим или юридическим лицам.

P-value - величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы (ошибки первого рода).

Репрезентативность - соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом.

Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определенной выборки на всю генеральную совокупность.

DR - уровень дефолтов. Рассчитывается как число дефолтных наблюдений в группе, деленное на число всех наблюдений в группе.

Бутстреп - практический компьютерный метод исследования распределения статистик вероятностных распределений, основанный на многократной генерации выборок на базе имеющейся выборки.

Вероятность дефолта - вероятность наступления дефолта по сделке в течение одного года с даты присвоения/корректировки рейтинга.

Выборка - набор сделок и их параметров, отвечающих заданным характеристикам и представляющим из себя часть анализируемой генеральной совокупности.

Выборка для обучения - набор сделок и их параметров, использующихся для оценки модели.

Выборка для оценки стабильности - набор сделок и их параметров, использующихся для оценки стабильности ранжирующей способности факторов и их разбиений.

Выборка для тестирования - данные по всем имеющимся договорам за все доступные отчетные даты. Определяется применительно к сегменту, на котором разрабатывается модель.

Генеральная совокупность - совокупность пар "сделка-дата", относящихся к выделенному сегменту.

Дискретные факторы - факторы с ограниченным количеством вариантов значений.

Непрерывные факторы - факторы с неограниченным количеством возможных вариантов значений.

Обучающая выборка - набор сделок и их параметров, использующихся для разработки модели.

Преобразование факторов - замена значений факторов на расчетные величины (скоры, WOE), связанные с оценкой вероятности дефолта, относящейся к значению фактора.

Скоринговый балл - значение показателя качества сделок с точки зрения вероятности их дефолта.

Тестовая выборка - выборка, используемая для проверки эффективности полученной модели (не участвует в разработке)

Трансформация факторов - то же, что и преобразование факторов.

PD - величина вероятности дефолта.

WOE (англ. weight of evidence) - величина, которая характеризует степень отклонения уровня дефолтов по группе от среднего уровня дефолтов по всей выборке.

Компьютеризированный способ разработки и управления моделями скоринга, схематично показанный на фиг. 1, включает следующие шаги.

Шаг 101: получают данные за заданный период времени, содержащие факторы, влияющие на модель скоринга.

Данные пользователя могут включать текущее состояние счетов (включая закрытые) - даты открытия, текущие остатки, срок, валюта, тип и название продукта, количество пролонгаций, текущий статус и так далее, не ограничиваясь.

Также полученные данные могут включать ежемесячные балансы (на конец каждого месяца) по каждому счету за последний промежуток времени (например, за последние полгода), все операции за тот же период с суммой, типом и подтипом, с признаком "дебет/кредит".

Вышеуказанные данные, которые представляют собой выборку, могут получать с мобильного устройства связи пользователя, например, такого как планшет, мобильный телефон, смартфон, или из автоматизированной системы финансово-кредитной организации, в которой хранятся данные.

На основе полученных данных о пользователях автоматически определяют кредитный скоринг, т.е. прогнозируют невозврат выданного кредита пользователем. Для этого используют обучающую выборку: набор объектов (пользователей), каждый из которых характеризуется набором признаков (таких как возраст, зарплата, тип кредита, состояние счетов, ежемесячные балансы, невозвраты в прошлом и т.д.), а также целевым признаком. Целевым признаком может быть, например, просрочка кредита. Если этот целевой признак - просто факт невозврата кредита (принимает значение 1 или 0, т.е. финансово-кредитная организация знает о своих клиентах, кто вернул кредит, а кто - нет), то это задача (бинарной) классификации. Если известно, насколько по времени клиент затянул с возвратом кредита, и хочется то же самое прогнозировать для новых клиентов, то это будет задачей регрессии.

Для каждой группы счетов (депозиты и прочие счета) могут учитываться следующие данные или факторы:

количество счетов;

количество счетов со статусом "Действующий";

количество счетов со статусом "Закрыт";

количество счетов со статусом "Счет арестован";

"Худший" статус по всем счетам клиента;

количество счетов в иностранной валюте;

количество счетов в драгоценных металлах;

минимальный срок по счетам;

средний срок по счетам;

максимальный срок по счетам;

минимальный срок по действующим счетам;

средний срок по действующим счетам;

максимальный срок по действующим счетам;

средневзвешенный по текущему остатку в рублях срок договора;

общая сумма текущих остатков;

максимальная сумма остатка по всем счетам;

средневзвешенный по текущему остатку доля валютных счетов;

средневзвешенный по текущему остатку доля счетов в драгоценных металлах;

время в днях, прошедшее с даты открытия самого раннего счета.

Специалисту в данном уровне техники очевидно, что представленный выше набор данных является примерным и в некоторых вариантах осуществления может отличаться от приведенного выше.

Далее осуществляют формирование по меньшей мере одной выборки для разработки модели скоринга. Для этого используются наиболее актуальные, выданные за один календарный год кредиты, находящиеся в портфеле не менее 12 месяцев. Поскольку модели скоринга разрабатываются для прогнозирования поведения всех заемщиков, ее разработка исключительно на выданных заявках может привести к неточным результатам. В таком случае модель будет обучена на смещенной выборке, поэтому осуществляют анализ заявок, по которым получены отказы предыдущей модели скоринга. В целях учета этих отказов к выборке для разработки модели скоринга добавляется некоторый процент худших заявок, по которым получен отказ предшествующей модели. Все такие заявки считаются по умолчанию дефолтными.

Шаг 102: осуществляют разбиение полученных данных на выборки для обучения, валидации и тестирования модели скоринга.

На данном этапе исходная совокупность данных разбивается на обучающую, валидационную и тестовую выборку в заданном соотношении. В дальнейшем обучающая выборка используется на всех этапах процесса, валидационная применяется для отбора наиболее стабильных факторов и итоговой проверки качества модели скоринга, а тестовая - для комплексного независимого тестирования. Выборки в не-

которых вариантах осуществления могут формироваться как последовательные, непересекающиеся во времени части исходной совокупности или как случайные подвыборки.

Шаг 103: осуществляют трансформацию факторов посредством установления соотношений между группами значений преобразованного фактора и уровнями дефолтов.

В качестве факторов, используемых в качестве входных параметров для моделей скоринга и потенциально связанных с кредитоспособностью пользователя, могут быть, не ограничиваясь, такие как годовой доход, размер непогашенного долга, владение недвижимостью или автомобилем, стаж работы на последнем месте, возраст и т.п.

Среди факторов, описывающих данные кредитной заявки, большую часть обычно составляют дискретные (образование, пол, семейное положение, цель кредита, вид собственности на жилье, род деятельности и т.п.). При этом, если некоторые факторы поддаются некоторому упорядочению (например, образование - можно считать, что чем выше уровень, тем больше значение переменной), то для других не существует никакого осмысленного линейного порядка (например, семейное положение или цель кредита). Следовательно, такие переменные нельзя даже приблизительно считать непрерывными, поскольку их значения суть номера ответов на соответствующие вопросы, которые могут располагаться в произвольном порядке. Если используемая модель скоринга требует использования непрерывных переменных, то можно обойти дискретность переменных, заменив их на большее количество переменных, принимающих значения от 0 до 1.

Трансформация каждого рассматриваемого фактора заключается в замене его значений расчетной величиной - WOE.

WOE - weight of evidence, характеризует степень отклонения уровня дефолтов по группе данных от среднего уровня дефолтов по всей выборке. Таким образом, каждый фактор заменяется соответствующим ему WOE-фактором следующим образом:

$$\text{WOE-фактор}_i = \text{WOE}_i(f),$$

где f - рассматриваемый фактор, i - номер группы значений фактора f, $\text{WOE}_i(f)$ - значение WOE, соответствующее группе значений i. В некоторых вариантах осуществления показатель WOE может принимать любые значения. Положительные значения WOE говорят о том, что рассматриваемый сегмент имеет более низкое значение уровня дефолтов, чем выборка в целом (чем больше WOE, тем ниже уровень дефолтов). Значение WOE меньше нуля говорит о том, что рассматриваемый сегмент имеет более высокое значение уровня дефолтов, чем выборка в целом. Значения WOE по группе i может определяться следующим образом:

$$\text{WOE}_i = \ln\left(\frac{N_G(i)/N_G}{N_B(i)/N_B}\right),$$

где $N_G(i)$ и N_G - количество недефолтных наблюдений в группе i и по всей выборке соответственно, $N_B(i)$ и N_B - количество дефолтных наблюдений в группе i и по всей выборке соответственно.

Если $N_G(i)=0$ или $N_B(i)=0$, то значение WOE для группы определяется по формуле:

$$\text{WOE}_i = \ln\left(\frac{(N_G(i)+0.5)/N_G}{(N_B(i)+0.5)/N_B}\right).$$

Для непрерывных факторов группировка осуществляется таким образом, чтобы в каждый диапазон попадали наблюдения с сопоставимым уровнем дефолтов (DR). В результате процесса группировки непрерывный фактор делится на несколько групп, для каждого из которых возможно оценить уровень дефолтов на базе наблюдений, попавших в этот диапазон.

Группировка переменных с дискретным набором значений осуществляется аналогично группировке непрерывных факторов - на основании сопоставимого уровня дефолтов (DR). В каждую группу может попадать одно или несколько значений фактора. Уровень дефолтов вычисляется по всем наблюдениям, входящим в группу.

Использование WOE-факторов имеет следующие преимущества.

Линеаризация факторов в соответствии с предпосылками логистической регрессии.

Автоматическая обработка пропущенных значений: они либо объединяются с наиболее похожей по уровню дефолтов группой, либо выступают в качестве отдельной группы. В случае когда пропущенное значение не интерпретируемо или отсутствует в выборке, то оно относится в худшую по уровню риска группу.

Автоматическая обработка аномальных значений, так как они не способны негативно повлиять на модель и их фактическое значение не используется в модели. Они войдут в модель как элемент одной из крайних групп, характеризующейся своим WOE-значением, основанным только на соотношении дефолтных и недефолтных наблюдений в группе.

Возможность оценить и контролировать логичность направления связи значений фактора и уровня дефолтов (бизнес-логику), что позволяет гарантировать, что итоговые скоринговые баллы будут иметь смысл (например, люди старшего возраста обычно набирают больше баллов, чем молодые). Логичные связи подтверждают бизнес-опыт, поэтому позволяют получить более стабильную модель.

Позволяет снизить риск переобучения. В модель не включается каждое случайное изменение данных, что имело бы место в случае не сгруппированных атрибутов. Такая модель обладает большей гибкостью и способна выдержать некоторые изменения в популяции, что обеспечивает стабильность в течение более долгого периода времени.

Первоначальная группировка значений факторов может происходить с помощью однофакторных деревьев решений. Это позволяет увеличить дискриминирующую способность полученных факторов по сравнению с ручными группировками, так как полученные группы будут максимально однородны внутри и различны между собой на основании используемого статистического критерия.

Под дискриминирующей силой фактора понимают его способность дифференцировать дефолтные и недефолтные наблюдения. Для оценки дискриминирующей способности переменной может использоваться индекс Джини.

На основании практики, имеющейся в уровне техники, по интерпретируемости используемых в скоринге факторов необходимо обращать внимание не только на ранжирующую способность WoE-трансформированных факторов, но и на их бизнес-логику. По этой причине на данном этапе происходит не только автоматическое разбиение значений факторов и расчет для них WoE, но и проверка получившихся разбиений на бизнес-логику. Если полученное разбиение не проходит данную проверку, то алгоритм пытается получить новое разбиение, используя альтернативные настройки. Способ получения итоговых WoE-факторов включает шаги, приведенные ниже.

Сначала запускают разбиения значений факторов с указанным набором настроек.

Затем осуществляют слияние полученных групп по близости значений WoE в случае, если расстояние по WoE между группами не превосходит заданный порог. Для интервальных факторов также учитывается порядок следования групп, упорядоченных по значениям фактора. Факторы, у которых осталась всего одна группа после объединений, переходят в список исключенных.

На следующем шаге осуществляют слияние групп маленького размера в соответствии с заданным пороговым значением с ближайшей по WoE группой. Для интервальных факторов также учитывается порядок следования групп, упорядоченных по значениям фактора. Факторы, у которых осталась всего одна группа после объединений, переходят в список исключенных. После каждого слияния необходимо вернуться ко второму пункту.

Также важно проводить слияние полученных групп по близости значений WoE в случае, если сформировано больше групп, чем изначально заданное максимальное количество для данного предиктора. Для интервальных факторов также учитывается порядок следования групп, упорядоченных по значениям фактора. Факторы, у которых осталась всего одна группа после объединений, переходят в список исключенных. После каждого слияния необходимо вернуться ко второму пункту.

В некоторых вариантах осуществления проверяют монотонность, условия немонотонности и направления риска для интервальных переменных в соответствии со справочником. Факторы, которые не соответствуют условиям из справочника, переходят в список исключенных.

В некоторых вариантах осуществления проверяют минимально допустимое количество групп. Если по переменной доступно меньше групп, чем изначально заданное минимально допустимое число, то она переходит в список исключенных.

В некоторых вариантах осуществления проверяют условия соотношения риска в различных группах для категориальных и бинарных переменных в соответствии со справочником (проверка бизнес-логики). Условия задаются с помощью специального языка, который позволяет описывать паттерны соотношения риска в группах любой сложности. Факторы, которые не соответствуют условиям из справочника, переходят в список исключенных.

В некоторых вариантах осуществления проверяют падение коэффициента Джини. Если данный коэффициент по предиктору на валидационной выборке меньше изначально заданного порогового значения либо падает по сравнению с коэффициентом Джини на обучающей выборке более чем на заданное число процентов, то такой фактор переходит в список исключенных.

В некоторых вариантах осуществления проверяют стабильность порядка следования групп, упорядоченных по WoE. Происходит сравнение обучающей выборки и 20 выборок, случайным образом отобранных из объединения обучающей и валидационной. Факторы, у которых выявлена нестабильность в порядке следования групп, упорядоченных по WoE, переходят в список исключенных.

По факторам, попавшим в список исключенных, необходимо запустить алгоритм разбиения значений факторов с новым набором настроек. Если доступных настроек нет или все они уже проверены, то формирование разбиений считается законченным. Количество настроек определяется возможностями используемого статистического пакета, например на основании SAS Enterprise Miner. Таким образом, по результатам применения алгоритма формируются WoE-факторы. Исходные факторы, которые не прошли проверку ни при одном наборе настроек разбиения, исключаются из процесса.

Шаг 104: исключают из выборок по меньшей мере один преобразованный фактор, коррелирующий по меньшей мере с одним другим фактором.

Анализ парных корреляций используется для выявления коллинеарных зависимостей между переменными. Наличие корреляций между факторами повышает стандартные отклонения коэффициентов регрессии, что снижает их устойчивость и надежность в многофакторном анализе. Для корреляционного анализа рассчитывается матрица корреляций - таблица со значениями коэффициентов парных корреляций преобразованных WOE-факторов. Анализ данной таблицы позволяет определить переменные, имеющие высокие линейные связи с другими факторами. Значение, начиная с которого коэффициенты корреляции признаются высоким, устанавливается в справочнике. Рекомендуемое значение, начиная с которого коэффициенты корреляции признаются высоким, находится в диапазоне от 0,5 до 1 по модулю. Из каждой пары коррелирующих факторов следует оставить только один на основании либо более высокой индивидуальной предиктивной способности, либо большей важности фактора с точки зрения бизнес-логики. В системе используется следующий алгоритм: в цикле отбирается фактор, который имеет наибольшее количество коррелированных с ним факторов (значение корреляции выше выбранного порога). Если таких несколько, то из них выбирается фактор с наименьшим значением индекса Джини. Такой фактор исключается из рассмотрения. После этого отбирается следующий фактор с наибольшим количеством коррелированных с ним оставшихся факторов и наименьшим значением индекса Джини. Таким образом, на выходе из цикла остаются факторы без корреляций выше выбранного порога. Данный подход обеспечивает наибольшее число некоррелированных факторов в итоговом списке факторов для моделирования.

Шаг 105: формируют модель кредитного скоринга посредством обучения бинарной множественной логистической регрессии, имеющий следующий вид:

$$PD(Y = 1|X_1, X_2, \dots, X_n) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n))},$$

где Y - зависимая переменная (признак дефолта), Y=1 - событие дефолта, $X_1,X_2,...,X_n$ - набор независимых, объясняющих WOE-факторов, $\beta_0,\beta_1,\beta_2,...,\beta_n$ - коэффициенты логистической регрессии, PD - вероятность дефолта.

Значения вероятности дефолта (PD - Probability of Default) располагаются в интервале [0, 1]. Она показывает вероятность дефолта для каждого рассчитанного рейтинга. В некоторых вариантах осуществления значения вероятности дефолта могут располагаться в интервале от 0 до 100 в процентном или численном эквиваленте. Несмотря на отсутствие коррелирующих пар, исключенных на предыдущем шаге, между факторами модели скоринга может возникать мультиколлинеарность, поэтому на этапе построения модели скоринга необходимо проверять ее отсутствие. Кроме того, т.к. модель скоринга разрабатывается на основе WoE-факторов, а чем больше WoE, тем меньше риск, необходимо проверять корректность знака коэффициента в модели скоринга (все коэффициенты регрессии должны быть отрицательными). Помимо этого, требуется обеспечить высокую стабильность модели, поэтому значимость каждого из входящих в нее факторов проверяется с помощью процедуры статистического бутстрэпа: каждый из факторов должен быть значим исходя из статистики Вальда минимум в 85% случаев. Способ формирования итоговой модели скоринга выглядит следующим образом.

На основе всех факторов, дошедших до данного этапа, строится логистическая модель с использованием пошаговой регрессии (stepwise) для отбора итогового набора факторов.

Для таких факторов происходит расчет фактора инфляции дисперсии (Variance Inflation Factor, VIF), Для определения VIF необходимо оценить линейную регрессионную модель, где в качестве зависимой переменной будет рассматриваемый фактор, а в качестве независимых переменных будут выступать оставшиеся факторы, включенные в модель. Итоговое значение VIF для фактора может быть найдено по формуле:

$$VIF = \frac{1}{1 - R^2},$$

где R^2 - коэффициент детерминации описанной выше модели. Переменная, значение VIF которой больше заданного значения и величина коэффициента Джини минимальна, исключается. Первый и второй шаги повторяются до тех пор, пока все включенные в модель факторы не будут иметь значение VIF ниже заданного.

Затем проводится проверка на наличие факторов с положительным знаком коэффициента регрессии. В случае их обнаружения происходит исключение фактора с минимальным значением коэффициента Джини, после чего необходимо вернуться к первому шагу. Если таких факторов нет, то следует перейти к следующему пункту.

Далее осуществляется объединение обучающей и валидационной выборок. Из их объединения случайным образом отбирается несколько десятков выборок того же размера, что и обучающая. На каждой из полученных выборок происходит обучение модели скоринга с текущим набором факторов. Если есть факторы, которые значимы, по статистике Вальда, менее чем в 85% случаев, то исключается тот из них, величина коэффициента Джини которого является наименьшей. После исключения необходимо вернуть-

ся к первому шагу. Если таких факторов нет, то скоринговая модель считается успешно построенной.

Таким образом, алгоритм позволяет в автоматическом режиме разрабатывать скоринговые модели, отвечающие всем разумным требованиям качества. Помимо этого он гарантирует, что каждый фактор будет соответствовать бизнес-логике, описанной в специальном справочнике.

В некоторых вариантах осуществления проводят автоматическую валидацию модели в соответствии с любой методикой валидации статистических моделей, известной из уровня техники. На данном этапе рассчитываются количественные тесты для оценки качества модели. Процесс валидации использует тестовую выборку, сформированную на шаге 102, и генеральную совокупность данных. В случае прохождения валидации переходим к шагу 107, иначе пользователю системы направляется уведомление о том, что валидация не пройдена, а также подробный отчет о выявленных недостатках. Варьируя настройки алгоритма, пользователь может скорректировать подходы к моделированию и обеспечить успешность следующей валидации.

Выбор оптимального значения порога отсечения зависит от цены совершения ошибки первого и второго рода при классификации. Модель должна точнее классифицировать "плохих" заемщиков, т.к. в кредитном скоринге цена ошибки перового рода выше. При снижении порога отсечения в модели будет увеличиваться чувствительность, т.е. способность модели правильно выявлять тех заемщиков, у которых будет просрочка платежа. За оптимальный порог отсечения можно взять точку баланса между чувствительностью и специфичностью.

Шаг 106: подбирают автоматически зоны отсечения для по меньшей мере одной модели скоринга для ее установки в кредитную процедуру.

Далее осуществляют автоматический подбор зон отсечения для моделей скоринга по скоринговым баллам для их установки в кредитную процедуру. Алгоритм подбора зон отсечения состоит из двух частей: внешней и внутренней. Внешняя часть отвечает за итеративный перебор уровней отсечения, внутренняя - за расчет ожидаемого уровня одобрения заявки на выдачу кредита, соответствующего текущему набору отсечений. Стоит отметить, что в качестве критерия для внутренней части алгоритма может выступать не только уровень одобрения, а любой интересующий показатель, зависящий от уровней отсечения, например, уровень риска или NPV портфеля. Алгоритм работает на исторической выборке данных по заявкам на кредиты. Ввиду того, что уровень одобрения характеризуется сезонностью в рамках недели, в данном изобретении речь идет о целевом уровне одобрения только в рамках семи дней, т.к. иначе придется определять его отдельно для каждого дня недели. Исходя из этого число дней, за которые рассматривается история по заявкам, должно быть кратно семи. Предположим, что в процессе принятия кредитного решения используется комбинация из трех моделей:

1) качества кредитной истории или скоринга бюро кредитных историй (БКИ-скоринга);

2) анкетных данных (заявочного скоринга);

3) склонности к мошенничеству или FDC-скоринга (Fraud Detection Card Scoring).

Предположим, что мы имеем комбинацию баллов отсечения по моделям заявочного, FDC- и БКИ-скоринга. Пусть (t1, t2, t3) - значение корректировок для отсечений по соответствующим моделям, а (n1, n2, n3) - число последовательных повторений корректировки для каждой из соответствующих моделей. Тогда внешний алгоритм подбора баллов отсечения будет следующим. Последовательно для каждой из моделей скоринга необходимо осуществить следующие действия:

1) прибавить соответствующую t корректировку из (t1, t2, t3) к уровню отсечения по этой модели;

2) запустить внутреннюю часть алгоритма, описанную далее, для подсчета ожидаемого уровня одобрения;

3) если отклонение ожидаемого уровня одобрения изменило направление, то выбрать такую комбинацию уровней отсечения по моделям заявочного, FDC- и БКИ-скоринга, при которой отклонение ожидаемого уровня одобрения является наименьшим (фактически выбор осуществляется из последних двух проверяемых комбинаций);

4) если отклонение ожидаемого уровня одобрения от целевого не изменило направления и первый пункт повторился менее n из (n1, n2, n3) раз, то перейти к первому пункту, т.е. к корректировке следующей модели скоринга.

В некоторых вариантах осуществления вышеописанная процедура повторяется до тех пор, пока не будет получен целевой уровень одобрения или достигнута верхняя/нижняя граница баллов по каждой из моделей.

В рамках внутренней части алгоритма оценивается изменение уровня одобрения при изменении баллов отсечения по работающим скоринговым моделям. Как уже отмечалось ранее, эффект от изменения зон отсечения может оцениваться на различные показатели, будь то риск или доходность, но в любом случае необходимо оценить, кто будет одобрен в рамках новых зон отсечения, а кто отказан (или вероятность одного из этих событий). В связи с этим будет рассмотрен алгоритм оценки изменения уровня одобрения.

Как правило, система принятия решения (СПР) в финансово-кредитной организации представляет собой последовательность проверок и применения правил и может включать следующие этапы прохождения заявок:

1) отказ по минимальным требованиям, на основе данных системы Hunter, стоп-листа и др.;

2) использование заявочного, БКИ- и FDC-скоринга;

3) применение моделей благонадежности;

4) андеррайтинг;

5) отказы на последующих этапах.

По этой причине для оценки уровня одобрения в случае переопределения фактических отказов скоринга по заявкам необходимо знать решение по ним на каждом из этапов, следующих за вторым этапом (использования скоринга). Для любой заявки, одобренной по всем работающим моделям скоринга (заявочного, БКИ-, FDC-скоринга и др.), доступна необходимая информация о процессе ее прохождения через последующие этапы СПР. Для заявок, по которым получен отказ хотя бы от одной из моделей, возникает неопределенность в отношении последующих этапов, т.к. такие заявки до этих этапов не доходят. Для того чтобы исключить данную неопределенность, в рамках алгоритма производится моделирование отказов после этапа скоринга для заявок, по которым ранее был получен отказ. Алгоритм можно представить как последовательность следующих действий.

1. Для заявок, дошедших до этапа скоринга, производится симуляция отказов по трем видам моделей при новых баллах отсечения. Все заявки, по которым получены фактический отказ на этапе скоринга и одобрение по всем моделям во время симуляции, помечаются (для них необходимо отдельное моделирование вероятности отказа на последующих этапах СПР).

2. Осуществляется моделирование вероятности отказа на этапе применения модели благонадежности. Для построения модели используются заявки, которые успешно прошли процедуру скоринга до изменения баллов отсечения.

3. Производится моделирование вероятности отказа на этапе андеррайтинга. Для этого дополнительно из предыдущей выборки исключаются заявки, по которым получен отказ на этапе применения моделей благонадежности.

4. Осуществляется моделирование вероятности отказа на последующих этапах. Дополнительно исключаются заявки, по которым получен отказ на этапе андеррайтинга.

5. Рассчитывается вероятность отказа послепрохождения процедуры скоринга для помеченных заявок, требующих отдельного моделирования (п.1).

В целях определения вероятности отказа для заявок, по которым получен отказ на этапе скоринга, используется следующая формула:

$$\mathbf{P_{reject} = P_{blag} + \left(1 - P_{blag}\right) * P_{underr} + \left(1 - P_{blag}\right) * \left(1 - P_{underr}\right) * P_{next}},$$

где P_{blag} - вероятность отказа для заявки по модели благонадежности; P_{underr} - вероятность отказа для заявки на этапе андеррайтинга; P_{next} - вероятность отказа для заявки на последующих этапах.

Для определения вероятности одобрения по заявке вероятность отказа вычитается из единицы. После этого уровень одобрения рассчитывается как отношение количества одобренных заявок (суммы вероятностей одобрения) к числу всех заявок. При усреднении данного значения по рассматриваемому портфелю получают уровень одобрения при выбранных зонах отсечения.

Если выбрать за целевой показатель уровень риска, то полученное значение необходимо умножить на уровень риска, получаемый из модели PD. При усреднении данного произведения получают уровень риска в выданном при выбранных зонах отсечения портфеле.

После автоматического подбора зон отсечения происходит оптимизация этих зон по различным сегментам портфеля. Принцип работы алгоритма оптимизации построен на итеративном оптимизационном расчете оптимальных порогов принятия решения для отдельных сегментов клиентов с точки зрения соотношения "Уровень одобрения - уровень риска". Ниже приведены основные предпосылки, критичные для получаемых результатов работы алгоритма.

1. Уровень риска оценивается как средний уровень вероятности просрочки внутри каждого сегмента.

2. Прогноз вероятности просрочки делается на последних доступных данных с учетом сегментации.

Основная идея алгоритма расчета - итеративный сдвиг порога отсечения для отдельного клиентского сегмента, который в итоге приводит к повышению общего уровня одобрения при сохранении текущего уровня риска.

На каждой итерации алгоритма рассматривается оптимальный с точки возможного улучшения соотношения AR/DR клиентский сегмент, в рамках которого происходят операции "закрутка" - "раскрутка" в данной последовательности с предзаданным шагом в 15 баллов (данный шаг может наращиваться в соответствии с правилами формирования цикла, но не более чем до 60 баллов). Таким образом, ищется оптимальная окрестность базового порога отсечения, приводящая к улучшению общего соотношения AR/DR.

Далее происходит внедрение модели скоринга (или моделей) и зон отсечения в промышленную среду.

В результате автоматического подбора уровней отсечения целевой уровень одобрения может быть не достигнут с требуемой точностью. В результате необходимо адаптивно корректировать отсечения по скорингам для максимального приближения целевому AR. Для этого спустя 7 полных дней после по-

следнего изменения целевого уровня одобрения или сразу после корректировки баллов отсечения без изменения целевого уровня одобрения начинается адаптивная корректировка полученных баллов отсечения. Она продолжается до тех пор, пока фактический уровень одобрения не войдет в допустимые границы хотя бы раз. Корректировка производится по следующей схеме. Прибавляем ко всем зона отсечения следующую величину:

$$\min(20^i, \text{Корректировка} * \frac{\Delta}{high_{AR}{}^{i} - AR_{target}{}^{ii}}),$$

где Δ=AR - $high_{AR}$, если последний выход AR за установленные границы произошел в большую сторону;

$$-\min(20, \text{Корректировка} * \frac{\Delta}{AR_{target} - low_{AR}{}^{i}}),$$

где $\Delta=low_{AR}$ - AR, если в меньшую. Размер корректировки задается экспертно в справочнике эмпирическим путем.

Корректировки запускаются ежедневно до тех пор, пока уровень одобрения не вернется в допустимый интервал между значениями $high_{AR}$ и low_{AR}. В некоторых вариантах осуществления проводится ежемесячная автоматическая валидация модели скоринга в соответствии с принятой в финансово-кредитном учреждении методологией. Если модель не проходит валидацию, она направляется на переобучение.

В некоторых вариантах осуществления проводят ежедневный мониторинг уровня одобрения. Данная методология подходит для наблюдения не только за уровнем одобрения, но и другими показателями, например, таким как риск. В рамках мониторинга рассматривается средний скользящий уровень одобрения с окном в 7 дней как временной ряд, элементы которого моделируются с помощью независимых нормально распределенных случайных величин. Для того чтобы поддерживать уровень одобрения на каком-либо целевом уровне, прежде всего необходим критерий, с помощью которого можно понять, что изменение действительно произошло, так как данный показатель имеет естественные флуктуации. Исходя из этого, для выявления отклонений в целевом уровне одобрения может быть использован CUSUM-тест. Для этого определяют, что есть момент изменения уровня одобрения (разладки) - это момент, когда меняется закон распределения в потоке поступающих данных об уровне одобрения. В данном изобретении рассматривается изменение среднего значения. Пусть X_n, $n\geq1$ - последовательность наблюдений, которые моделируются с помощью независимых нормально распределенных случайных величин, $\theta \in [1,n]$ - неизвестный момент времени, в который меняется распределение наблюдений с $f_0 \sim N(\mu_0,\sigma^2)$ на $f_1 \sim N(\mu_1,\sigma^2)$, а n - текущий момент времени. Так как точный момент времени разладки неизвестен, то гипотеза H_0 - разладки на отрезке [1,n] нет, а H_1 - разладка произошла на отрезке [1,n]. Чтобы различить две этих гипотезы, необходимо определить обобщенный критерий отношения правдоподобия:

$$T_n = \max_{\theta \in [1,n]} \sum_{i=\theta}^{n} log \frac{f_1(x_i)}{f_0(x_i)} \geq C_0,$$

где с помощью C_0 контролируется число ложных срабатываний. Данное выражение известно как CUSUM-тест. Полученная запись теста будет вычислительно неэффективна, но в случае независимых случайных величин статистика может быть представлена рекуррентным соотношением:

$$T_n = \max\{T_{n-1} + log \frac{f_1(x_n)}{f_0(x_n)}, 0\}$$

Так как мы предполагаем, что f_0 и f_1 распределены нормально:

$$T_n = \max\left\{T_{n-1} + \frac{\mu_1 - \mu_0}{\sigma^2}\left(x_n - \frac{\mu_1 + \mu_0}{2}\right), 0\right\}.$$

Пусть $\mu_1=\mu_0\pm\delta$, где δ - это допустимая погрешность, которая выбирается в зависимости от того, какое отклонение мы считаем приемлемым. Тогда выражение для вычисления CUSUM можно переписать в виде:

$$T_n^- = \max\left\{T_{n-1} - \frac{\delta}{\sigma^2}\left(x_n - \mu_0 + \frac{\delta}{2}\right), 0\right\}$$

для отклонений в сторону снижения и

$$T_n^+ = \max\left\{T_{n-1} + \frac{\delta}{\sigma^2}\left(x_n - \mu_0 - \frac{\delta}{2}\right), 0\right\}$$

для отклонений в сторону увеличения.

Итоговое решение находится из условия $\max(T_n^+, T_n^-) > h = e(C_0)$.

Описанный подход позволяет выявлять отклонения уровня одобрения от целевого уровня с минимальной задержкой и небольшим количеством ложных срабатываний.

Если смена целевого уровня одобрения произошла менее чем 7 дней назад, мы не можем проводить

CUSUM-тест, так как нет наблюдений скользящего среднего уровня одобрения за 7 дней, не включающих дни до корректировки. Кроме того, нужно застраховать себя от некорректной работы теста CUSUM. Для этого используется альтернативный более простой тест, основанный на установке границ допустимого диапазона для наблюдаемого показателя. Аспекты настоящего изобретения могут быть также реализованы с помощью устройства обработки данных, являющегося вычислительной машиной или системой (или таких средств как центральный/графический процессор или микропроцессор), которая считывает и исполняет программу, записанную на запоминающее устройство, чтобы выполнять функции вышеописанного варианта(ов) осуществления, и способа, показанного на фиг. 1, этапы которого выполняются вычислительной машиной или устройством путем, например, считывания и исполнения программы, записанной на запоминающем устройстве, чтобы исполнять функции вышеописанного варианта(ов) осуществления. С этой целью программа записывается на вычислительную машину, например, через сеть или со среды для записи различных типов, служащей в качестве запоминающего устройства (например, машиночитаемой среды). Устройство обработки данных может иметь дополнительные особенности или функциональные возможности. Например, устройство обработки данных может также включать в себя дополнительные устройства хранения данных (съемные и несъемные), такие как, например, магнитные диски, оптические диски или лента. Устройства хранения данных могут включать в себя энергозависимые и энергонезависимые, съемные и несъемные носители, реализованные любым способом или при помощи любой технологии для хранения информации, такой как машиночитаемые инструкции, структуры данных, программные модули или другие данные. Устройство хранения данных, съемное хранилище и несъемное хранилище являются примерами компьютерных носителей данных. Компьютерные носители данных включают в себя, но не в ограничительном смысле, оперативное запоминающее устройство (ОЗУ), постоянное запоминающее устройство (ПЗУ), электрически стираемое программируемое ПЗУ (EEPROM), флэш-память или память, выполненную по другой технологии, ПЗУ на компакт-диске (CD-ROM), универсальные цифровые диски (DVD) или другие оптические запоминающие устройства, магнитные кассеты, магнитные ленты, хранилища на магнитных дисках, или другие магнитные запоминающие устройства, или любую другую среду, которая может быть использована для хранения желаемой информации и к которой может получить доступ устройство обработки данных. Устройство обработки данных может также включать в себя устройство(а) ввода, такие как клавиатура, мышь, перо, устройство с речевым вводом, устройство сенсорного ввода, и так далее. Устройство(а) вывода, такие как дисплей, динамики, принтер и тому подобное, также могут быть включены в состав системы.

Устройство обработки данных содержит коммуникационные соединения, которые позволяют устройству связываться с другими вычислительными устройствами, например по сети. Сети включают в себя локальные сети и глобальные сети наряду с другими большими масштабируемыми сетями, включая, но не в ограничительном смысле, корпоративные сети и экстрасети. Коммуникационное соединение является примером коммуникационной среды. Как правило, коммуникационная среда может быть реализована при помощи машиночитаемых инструкций, структур данных, программных модулей или других данных в модулированном информационном сигнале, таком как несущая волна, или в другом транспортном механизме, и включает в себя любую среду доставки информации. Термин "модулированный информационный сигнал" означает сигнал, одна или более из его характеристик изменены или установлены таким образом, чтобы закодировать информацию в этом сигнале. Для примера, но без ограничения, коммуникационные среды включают в себя проводные среды, такие как проводная сеть или прямое проводное соединение, и беспроводные среды, такие как акустические, радиочастотные, инфракрасные и другие беспроводные среды. Термин "машиночитаемый носитель", как употребляется в этом документе, включает в себя как носители данных, так и коммуникационные среды. Последовательности процессов, описанных в этом документе, могут выполняться с использованием аппаратных средств, программных средств или их комбинации. Когда процессы выполняются с помощью программных средств, программа, в которой записана последовательность процессов, может быть установлена и может выполняться в памяти компьютера, встроенного в специализированное аппаратное средство, или программа может быть установлена и может выполняться на компьютер общего назначения, который может выполнять различные процессы.

Например, программа может быть заранее записана на носитель записи, такой как жесткий диск, или ПЗУ (постоянное запоминающее устройство). В качестве альтернативы программа может быть временно или постоянно сохранена (записана) на съемном носителе записи, таком как гибкий диск, CD-ROM (компакт-диск, предназначенный только для воспроизведения), MO (магнитооптический) диск, DVD (цифровой универсальный диск), магнитный диск или полупроводниковая память. Съемный носитель записи может распространяться в виде так называемого продаваемого через розничную сеть программного средства.

Программа может быть установлена со съемного носителя записи, описанного выше, на компьютер, или может быть передана по кабелю с сайта загрузки в компьютер, или может быть передана в компьютер по сетевым каналам передачи данных, таким как ЛВС (локальная вычислительная сеть) или Интернет.

Компьютер может принимать переданную таким образом программу и может устанавливать ее на носитель записи, такой как встроенный жесткий диск. Процессы, описанные в этом документе, могут

выполняться последовательно по времени, в соответствии с описанием, или могут выполняться параллельно или отдельно, в зависимости от характеристик обработки устройства, выполняющего процессы, или в соответствии с необходимостью. Система, описанная в этом документе, представляет собой логический набор множества устройств и не ограничивается структурой, в которой эти устройства установлены в одном корпусе.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Компьютеризированный способ разработки и управления моделями скоринга, включающий следующие шаги:

получают данные за заданный период времени, содержащие факторы, влияющие на модель скоринга;

осуществляют разбиение полученных данных на выборки для разработки, валидации и тестирования модели скоринга;

осуществляют трансформацию факторов посредством установления соотношений между группами значений преобразованного фактора и уровнями дефолтов;

исключают из выборок по меньшей мере один преобразованный фактор, коррелирующий по меньшей мере с одним другим фактором;

формируют модель кредитного скоринга посредством обучения бинарной множественной логистической регрессии;

подбирают автоматически зоны отсечения по меньшей мере для одной модели скоринга для ее установки в кредитную процедуру.

2. Способ по п.1, характеризующийся тем, что получают данные за заданный период времени с мобильного устройства связи пользователя.

3. Способ по п.1, характеризующийся тем, что при осуществлении разбиения полученных данных на выборки получают непересекающиеся во времени части исходной совокупности или случайные подвыборки.

4. Способ по п.1, характеризующийся тем, что факторами, влияющими на модель скоринга, являются годовой доход, и/или размер непогашенного долга, и/или владение недвижимостью, и/или владение автомобилем, и/или стаж работы на последнем месте, и/или возраст.

5. Способ по п.1, характеризующийся тем, что факторы, влияющие на модель скоринга, являются дискретными или непрерывными.

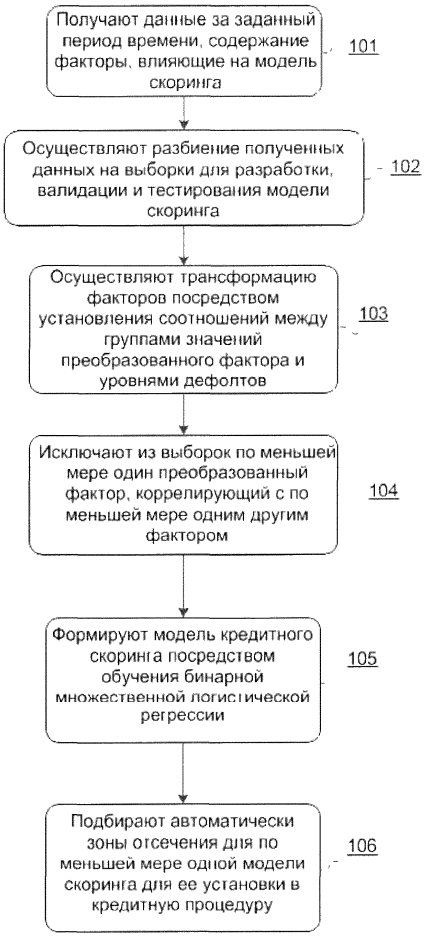
6. Способ по п.1, характеризующийся тем, что при осуществлении трансформации факторов определяют степень отклонения уровня дефолтов по группе данных от среднего уровня дефолтов по всей выборке.

7. Способ по п.1, характеризующийся тем, что при осуществлении трансформации факторов по факторам, попавшим в список исключенных, запускают алгоритм разбиения значений факторов с новым набором настроек.

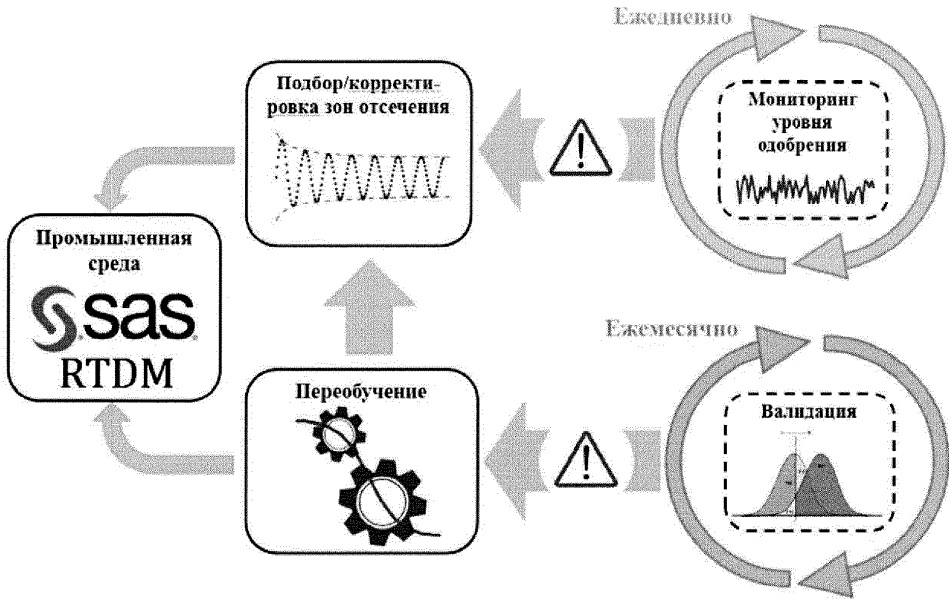
8. Способ по п.1, характеризующийся тем, что при исключении из выборок преобразованных факторов формируют таблицу со значениями коэффициентов парных корреляций преобразованных факторов.

9. Способ по п.1, характеризующийся тем, что при исключении из выборок преобразованных факторов в цикле отбирают фактор, который имеет наибольшее количество коррелированных с ним факторов.

10. Способ по п.1, характеризующийся тем, что при формировании модели кредитного скоринга строится логистическая модель с использованием пошаговой регрессии для отбора итогового набора факторов.



Фиг. 1



Фиг. 2

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2