ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 1(25)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.8 DOI: 10.54708/SIIT-2026-no1-p93 EDN: BDSIHW

Прогнозирование страховых случаев на основе стекинг-ансамблирования

М. И. Морозов

Югорский государственный университет

Выявление рисков является неотъемлемой задачей любой финансовой организации, в том числе и автомобильного страхования. В настоящее время в бизнесе используются множество методов выявления рисков, в число которых входит построение моделей машинного обучения и гибких систем на основе деревьев решений. Системы принятия решений показывают высокую эффективность в выявлении рисков, однако такие системы имеют ряд проблем, которые вносят неопределенность в прогнозирование убыточности страховой компании: чувствительность к изменению данных, наличие линейных границ принятия решений, чувствительность к пропущенным данным. В решении ряда сложных бизнес-задач высокую эффективность показывают не только нейросети и методы машинного обучения, но и методы, позволяющие объединять прогнозы отдельных моделей – стекинг-ансамблирование. В данной работе мы ставим перед собой задачу преодолеть ограничения деревьев решений на основе ансамбля моделей машинного обучения в задаче предсказания вероятности наступления страхового случая. В качестве исходного набора данных используется информация о страховании автомобилей в России за 2023–2025 г. Проведено сравнение эффективности индивидуальных моделей машинного обучения и методов стекинг-ансамблирования. Полученные результаты работы позволяют судить об эффективности применённых методов в предсказании вероятности страхового случая

Деревья решений; нейронные сети; машинное обучение; автомобильное страхование; убыточность; стекинг-ансамблирование.

Введение

Множество финансовых организаций традиционно полагаются на премии и сборы как на основной источник дохода. Для поддержания своей финансовой стабильности и прибыльности страховые компании должны гарантировать, что застрахованные лица своевременно оплачивают свои страховые взносы. С другой стороны, страховая организация обязана делать прогнозирование расходов на страховые выплаты. Цель обоих ситуаций – минимизировать риск невыполнения обязательств как со стороны клиента, так и со стороны компании [Ахв23, Rit23]. Одним из методов решения данной проблемы является предсказание вероятности наступления страхового случая, что позволяет не только снизить финансовые риски, но и предложить клиентам более справедливые условия страхования.

Страховые компании используют различные методы выявления риска, включая актуарные расчеты и анализ данных о прошлом поведении клиентов. Анализ исторических данных позволяет оценивать вероятность страховых случаев с использованием статистических методов, машинного обучения и искусственного интеллекта, что позволяет страховой компании оставаться конкурентоспособной на страховом рынке [Бро23, Кир24]. Стоит упомянуть также активное развитие современных технологий предиктивной аналитики как на основе машинного обучения [Sob24], так и на основе трансформерных нейросетей [Sin23, Мор25].

Рекомендовано к публикации программным комитетом XI Международной научной конференции ITIDS’2025 «Информационные технологии интеллектуальной поддержки принятия решений», Уфа, 13–15 ноября 2025 г.

Морозов М. И. Прогнозирование страховых случаев на основе стекинг-ансамблирования // СИИТ. 2026. Т. 8, № 1(25). С. 93-100. DOI: 10.54708/SIIT-2026-no1-p93. EDN: BDSIHW .

Morozov M. I. “Insurance claims prediction based on stacking ensembles” // SIIT. 2026. Vol. 8, no. 1(25), pp. 93-100. DOI: 10.54708/SIIT-2026-no1-p93. EDN: BDSIHW (In Russian).

В настоящее время системы принятия решений с древовидной структурой являются распространенным методом выявление рисков, однако они имеют недостатки классических деревьев решений, среди которых можно выделить:

- Чувствительность к шуму и выбросам: небольшие изменения в исходных данных могут привести к значительным изменениям в структуре дерева.
- Склонность к переобучению: в процессе построения дерева решений может возникать сложная конструкция, которая недостаточно точно представляет данные.
- Ограниченная способность к обобщению: деревья решений могут плохо справляться с задачами, требующими сложных нелинейных зависимостей.

Исследование стремится определить наиболее эффективную методику для точной оценки страхового риска на основе несбалансированных данных, включающих детальную информацию о страховании автомобилей в России за 2023–2025 гг. В качестве способа преодоления ограничений классических деревьев решений мы обращаемся к методам прогнозирования вероятности наступления страхового случая на основе машинного обучения, нейронных сетей, а также методам объединения прогнозов моделей на основе стекинг-ансамблирования.

Модели и методы исследования

В данной работе для предсказания вероятности наступления страхового случая были использованы четыре метода предиктивной аналитики: Градиентный бустинг (XGBoost) [Che16], Логистическая регрессия (LogisticRegression), Случайный лес (RandomForest) [Bre01], CatBoost [Pro19]. Для объединения прогнозов и создания метамодели мы используем пять методов объединения прогнозов: простое среднее (Simple Average), адаптивная регрессия (ARM), адаптивная регрессия распределения Tweedie (ARM-Tweedie), Логистическая регрессия (Constraint Logistic Regression) [Che22] и нейронная сеть MLP. В данном разделе представлен обзор методов обучения индивидуальных моделей и метамоделей стекинга.

Описание набора данных

В качестве набора данных мы используем набор данных, содержащий анонимизированную информацию о страховании автомобилей в России за 2023–2025 гг. Набор данных основан на промышленных логах высоконагруженной системы, предоставленных партнерской организацией. В соответствии с соглашением о конфиденциальности, полный набор данных не подлежит публичному раскрытию. Набор данных содержит более 50 ключевых признаков и целевые метки о наличии или отсутствии убытка. Главной особенностью набора данных является несбалансированность целевого класса убытка.

Предварительная обработка данных

В качестве предварительной обработки данных проведены основные этапы, такие как проверка выбросов, пропусков, несоответствие типам данных. Подробный обзор методологии обработки страхового набора данных можно увидеть на примере предыдущего исследования на основе Эфиопского набора данных о страховании клиентов [Mop24].

Ключевыми отличиями в предварительной обработке данных является кодирование категориальных переменных с помощью метода Target Encoding [Mic01], который заменяет каждую категорию на агрегированное значение целевой переменной. Текущий метод был выбран с целью ускорить обучение, так как кодирование методами LabelEncoding и OneHotEncoding имеет недостатки в виде наличия линейных зависимостей и формирования слишком большого пространства признаков. Метод Target Encoding также имеет ключевую проблему в виде переобучения. К примеру, если вычислять среднее значение по всему набору данных, модель может узнать о будущих образцах данных. В текущей работе этот недостаток нивелируется с помощью кросс-валидации (Out-of-Fold) и разделением исходных данных на три выборки: обучающая (train), валидационная (valid) и тестовая (test).

В исходном наборе данных наблюдается дисбаланс целевого класса и является несбалансированным по ключевому признаку наличия убытка «is_claim». С целью преодолеть дисбаланс классов и обратить внимание моделей на класс с наличием убытка обоим классам были назначены веса в зависимости от частоты встречаемых классов. Полученные веса в дальнейшем используются в функциях потерь при обучении, чтобы обратить внимание моделей на более редкую категорию.

Методы предиктивной аналитики

Градиентный бустинг (XGBoost) является мощным методом, способным эффективно корректировать ошибки, и имеет возможность комбинировать преимущества предыдущих моделей, улучшая их общую производительность, позволяя учитывать вес каждого входного прогноза. В бизнесе данный метод служит хорошей основой для проверки гипотез, так в работе [Bar22] автор использует данный подход как один из методов для решения задачи прогнозирования убытков в страховании.

Логистическая регрессия (LR) используется в страховании для предсказания вероятности страховых случаев, в которой линейная комбинация входных переменных преобразуется логистической функцией, чтобы предсказать вероятность принадлежности к одному из двух классов.

Random Forest (RF) является ансамблевым методом, генерирующим множество деревьев решений для улучшения стабильности и точности прогнозов. Так, в работе [Bar22] метод Random Forest применятся для прогнозирования страховых случаев при несбалансированных наборах данных. Автор работы применяет метод балансировки SMOTE и проводит настройку гиперпараметров, что позволяет получить хорошие результаты, достигая значительного улучшения показателей AUC и F1-меры по сравнению с другими методами обучения в исследовании.

CatBoost является методом для работы с категориальными признаками и необработанными данными, что делает его особенно популярным в бизнес-приложениях, включая страхование, финансы и маркетинг.

Объединение прогнозов моделей

В качестве ансамблевых методов объединения прогнозов моделей – стекинга – в данной работе используются ряд методов, позволяющих наиболее подробно описать способы объединения прогнозов моделей выделив из них ряд лучших в задачах страхования.

Простое среднее (SA) – прогноз от каждой модели суммируется и делится на количество моделей, что позволяет агрегировать оценку от предикторов. В отличии от классического бинарного предсказания в текущей работе суммируются вероятности принадлежности к каждому классу, что позволяет получит среднюю оценку моделей по предсказанным вероятностям. В качестве преимуществ метода можно выделить его простоту, интерпретируемость и устойчивость к переобучению. В качестве недостатков, что слабые или переобученные модели влияют на итоговый прогноз так же сильно, как и сильные, что может снижать общую производительность ансамбля.

LinearRegression-Constraint (LR-C) использует схожую логику с методом простого среднего, с единственным отличием, что каждой индивидуальной модели назначается вес, в зависимости от степени доверия к каждой из моделей.

ARM (Adaptive Regression Mixtures) – усложнённая версия метода LinearRegression-Constraint с ключевым отличием в наличии динамической адаптации весов в зависимости от характеристик входных данных. В отличие от простого назначения весов, ARM подбирает оптимальные коэффициенты для каждой модели в зависимости от конкретного экземпляра данных ARM-Tweedie – продолжение метода ARM с ключевым отличием, он учитывает специфику распределения целевой переменной, где данные имеют несбалансированную структуру в виде множества нулевых, но довольно значимых значений. Это хорошо ложится

на страховые случаи, которые значимы и происходят редко. В отличие от ARM, который адаптирует веса моделей, опираясь только на их точность на конкретных примерах, ARM-Tweedie дополнительно оценивает, насколько корректно предсказывается не только значение, но и форма распределения (к примеру: частоту нулевых значений, разброс больших значений, асимметрию).

В качестве нейросетевого метода объединения прогнозов мы используем простую нейронную сеть MLP [Hor89] с тремя скрытыми слоями (128, 64, 32). MLP позволяет извлекать нелинейные зависимости и взаимодействия между прогнозами базовых моделей благодаря своей универсальной способности к аппроксимации.

Методы оценки модели

Особое внимание в исследовании уделено метрикам оценки моделей. Для решения задачи точности предсказания вероятности страхового случая необходимо, чтобы вероятностная оценка наиболее точно соответствовала реальности, так как точная оценка риска важна при формировании цены на услуги страхования. В данной работе используются три метода оценки моделей, что позволяет отобразить не только умение моделей различать классы, но и показать способность к точной оценке вероятности страхового случая.

В качестве методов оценки вероятности используются Log-loss и BraerScore. Log-loss позволяет узнать, насколько предсказанные моделью вероятности близки к реальности за счет штрафной функции, наказывающей модель за неверные прогнозы. Braer Score описывает среднеквадратическую ошибку вероятности и показывает, насколько сильно модель ошиблась от истинного значения. В отличии от Log-loss, данная функция потерь не штрафует модель за промахи, что позволяет делать более уверенные, хоть и менее точные прогнозы.

Для оценки способности модели различать классы используется метрика ROC-AUC, которую можно интерпретировать как способность модели сортировать объекты от более рискованных к менее рискованным, что позволяет отобразить, насколько хорошо модель умеет различать целевые классы.

Экспериментальная установка

В качестве инструмента учета истории эксперимента использовался инструмент Jupyterlab, позволяющий быстро настроить и поставить эксперимент. В качестве основных библиотек создания моделей машинного обучения использовались «sklearn»[1], «XGBoost»[2], «torch».

Основной Pipeline эксперимента делится на несколько ключевых шагов:

1. Производится предобработка набора данных, удаление пропусков, сортировка по временным данным. Набор данных делится на три временных промежутка: данные старше 2024 г. используются в обучении индивидуальных моделей (train). Следующая выборка используется исключительно для подбора гиперпараметров моделей, и данная выборка находится в промежутке между 2024–2025 гг. Данные старше 2025 г. используются как тестовая выборка для окончательной проверки результатов обучения моделей.

2. Для разных моделей машинного обучения используются несколько обработчиков, которые производят предобработку набора данных. Первый обработчик используется только для логистической регрессии. В качестве заполнения пропусков в данном обработчике используется метод SimpleImputer, числовые данные кодируются с помощью StandartScaler, а категориальные – с помощью TargetEncoder. Второй обработчик используется для всех остальных методов обучения индивидуальных моделей, и его единственным отличием является отсутствие метода масштабирования числовых данных, так как методы LR, CatBoost, XGBoost имеют свои обработчики числовых признаков.

---

[1] User Guide | scikit-learn. URL: https://scikit-learn/stable/user_guide.html. (дата обращения: 12.02.2026)
[2] XGBoost Documentation | xgboost 3.0.2 documentation. URL: https://xgboost.readthedocs.io/en/stable. (дата обращения: 12.02.2026)

3. Производится инициализация моделей машинного обучения и проводится обучение моделей на тренировочной выборке (train), используя кросс-валидацию с разделением данных на $K = 5$ выборок. Суть кросс-валидации состоит в обучении нескольких маленьких моделей на части train выборки и формировании итогового набора данных с вероятностными предсказаниями, который будут использоваться в обучении метамодели. К примеру, если мы имеем в общем случае 100 экземпляров данных, то мы обучаем модель на первых 80 экземплярах данных, а остальные 20 используются в качестве формирования прогнозов, и так пока не пройдем K раз по всем данным.

4. Проводится обучение индивидуальных моделей уже по целой выборке train, и формируется набор данных из вероятностных предсказаний, который будет использован как тестовая выборка в финальной оценке обученной метамодели.

5. На основе вероятностных предсказаний, полученных в п. 3, проводится обучение метамоделей, ее оценка на основе вероятностного набора данных, полученного в п. 4.

6. В качестве завершающего этапа проводится калибровки вероятностей с помощью изотонической регрессии (Isotonic Regression), где на выходе мы получаем точные и откалиброванные оценки вероятностей.

Интерпретация результатов

В данном разделе представлены результаты оценки обученных индивидуальных моделей на основе валидационной (табл. 1) и тестовой выборки (табл. 2), а также результаты объединения прогнозов моделей на валидационной (табл. 3) и тестовой выборке (табл. 4).

Оценка качества различных моделей показала различия как в способности различать классы, так и в калибровке вероятностных предсказаний.

Таблица 1

Результаты обучения индивидуальных моделей. Валидационная выборка

Модель / Метрика	XGBoost	LR	RF	CatBoost
BrierScore	0.2101	0.7895	0.0441	0.2115
LogLoss	0.6008	2.4757	0.1896	0.6059
ROC-AUC	0.6923	0.6787	0.6505	0.6932

Таблица 2

Результаты обучения индивидуальных моделей. Тестовая выборка

Модель/ Метрика	XGBoost	LR	RF	CatBoost
BrierScore	0.2115	0.8184	0.0130	0.2122
LogLoss	0.6027	2.5690	0.0805	0.6071
ROC-AUC	0.6895	0.6805	0.6415	0.6927

Среди индивидуальных моделей лучшую способность различать классы продемонстрировали модели XGBoost и CatBoost с оценкой ROC-AUC (0.69). При этом по метрикам оценки вероятности Brier Score и LogLoss они уступали методу случайного леса (RF).

Таблица 3

Результаты объединения прогнозов. Валидационная выборка

Модель/ Метрика	SA	LR-C	ARM	ARM-Tweedie	MLP
BrierScore	0.2153	0.1859	0.2216	0.2223	0.2027
LogLoss	0.6194	0.5509	0.6303	0.6313	0.5877
ROC-AUC	0.6938	0.6938	0.6914	0.6899	0.6937

Таблица 4

Результаты объединения прогнозов. Тестовая выборка

Модель/ Метрика	SA	LR-C	ARM	ARM-Tweedie	MLP
BrierScore	0.2153	0.1843	0.2252	0.2259	0.2026
LogLoss	0.6191	0.5471	0.6380	0.6388	0.5870
ROC-AUC	0.6916	0.6921	0.6892	0.6868	0.6915

Случайный лес показал рекордно низкие значения Brier Score (0.013) и LogLoss (0.08), что указывает на крайне аккуратные и хорошо калиброванные вероятностные предсказания, однако способность модели различать классы ROC-AUC (~0.64) оказалась хуже. Наименее успешной оказалась стандартная логистическая регрессия, где Brier Score и LogLoss были значительно выше (0.78–0.82 и 2.47–2.56), что отражает слабую калибровку модели.

Среди методов объединения прогнозов моделей особенно выделяется LogitsticRegression-Constraint (LR-C), который показал наилучшее сочетание низких значений Brier Score (0.184–0.186) и LogLoss (0.547–0.551) при сохранении ROC-AUC на уровне бустингов (~0.693). Это делает метод сбалансированным и надёжным выбором в задаче прогнозирования. Нейросетевой подход (MLP) также продемонстрировал стабильные результаты: умеренно низкие значения Brier и LogLoss при высоком ROC-AUC (~0.693). Напротив, MLP вместе с Feature-Enhance Stacking (MLP_FE) не дал ожидаемого прироста по основным метрикам, его ROC-AUC снизился (0.66–0.67), а качество по калибровке ухудшилось. Методы на основе ARM и Tweedie показали худшие результаты среди альтернативных моделей, уступив как по калибровке, так и по способности различать классы.

Эксперимент подтверждает, что выбор модели зависит от приоритетов задачи. Если важна корректная оценка вероятностей, наилучшие результаты дают RandomForest как индивидуальная модель и LogitsticRegression-Constraint как метамодель. Если же ключевой критерий – способность различать классы, то XGBoost, CatBoost демонстрируют наивысшие показатели. В то же время простая логистическая регрессия без модификаций показала наихудшие результаты, а ансамблирование как с помощью статистических методов, так и с помощью нейросети MLP в текущей реализации не дало преимуществ. Стоит также отметить скорости обучения моделей на основе машинного обучения и нейросетей, где методы на основе индивидуальных моделей машинного обучения обучились в течение ~30 мин, с другой стороны – нейросетевой метод MLP по скорости обучения занял ~ 6 ч без значительного прироста к качеству модели.

Заключение

В процессе работы были достигнуты несколько ключевых результатов:

- Основываясь на методах стекинг-ансамблирования, лучшим методом объединения прогнозов для данного набора данных является метод LogitsticRegression-Constraint (LR-C) с показателями Brier Score (0.184–0.186), LogLoss (0.547–0.551), ROC-AUC (~0.693).

- Метод стекинг-ансамблирования на основе нейронной сети MLP не показал существенного преимущества в объединении прогнозов моделей по метрикам Brier Score, LogLoss, ROC-AUC, даже в учет продолжительного времени обучения модели в 6 ч.
- Метод стекинг-ансамблирования на основе нейронной сети MLP с добавлением исходных данных (Feauture Enhance Stacking) в обучение метамодели не дал значительного преимущества, а только ухудшил оценки моделей по ROC-AUC (0.66–0.67), Brier Score (0.22), LogLoss (0.627).
- Среди индивидуальных моделей в качестве лучших можно считать CatBoost – по степени разделения целевых классов ROC-AUC (0.693) и RandomForest – по степени калибровки вероятностей Brier Score (0.013), LogLoss (0.08).
- Сравнивая нейронные сети, методы машинного обучения и методы объединения прогнозов моделей на основе стекинг-ансамблирования с методом на основе деревьев решений RandomForest, можно видеть прирост по метрике ROC-AUC, что позволяет судить о преодолении ограничений классических деревьев решений.

Благодарности и поддержка

Автор выражает благодарность научному руководителю засл. деят. науки РФ, д-ру техн. наук, проф. Андрею Витальевичу Мельникову за профессиональное руководство, помощь и активное участие в развитии научного исследования. Также автор считает необходимым отметить работы [Sob24, Che22, Bar22, Мак24, Кот23, Куз23], оказавшие влияние на данное исследование.

Список литературы | References

[Bar22] Baran S., Rola P. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. Apr. 2022. DOI: 10.48550/arXiv.2204.06109.

[Bre01] Breiman L. Random Forests // Machine Learning. Oct. 2001. Vol. 45, No. 1. Pp. 5–32. DOI: 10.1023/A:1010933404324. EDN: ARROTH.

[Che16] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. Aug. 2016. Pp. 785–794. DOI: 10.1145/2939672.2939785.

[Che22] Chenglong Ye, Lin Zhang, et al. Combining predictions of auto insurance claims // Econometrics. Apr. 2022. Vol. 10, No. 2. Pp. 19–19, DOI: 10.3390/econometrics10020019. EDN: JIGXLT.

[Hor89] Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators // Neural Networks. Jan. 1989. Vol. 2, No. 5. Pp. 359–366, DOI: 10.1016/0893-6080(89)90020-8.

[Mic01] Micci-Barreca D. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems // SIGKDD Explor. Newsl. Jul. 2001. Vol. 3, No. 1. Pp. 27–32. DOI: 10.1145/507533.507538.

[Pro19] Prokhorenkova L., Gusev G., et al. CatBoost: unbiased boosting with categorical features. 20 Jan 2019. DOI: 10.48550/arXiv.1706.09516.

[Rit23] Ritho B. M., Simiyu E., Omagwa J. The Impact of Loss Ratio on the Financial Stability of Insurance Firms in Kenya // Journal of Finance and Accounting. Jun. 2023. Vol. 7, No. 4. Art. No. 4. DOI: 10.53819/81018102t4161. EDN: ACMIPP.

[Sin23] Singh U., Arora P., et al. Comparative Analysis of Transformers for Modeling Tabular Data: A Casestudy using Industry Scale Dataset. 24 Nov. 2023, arXiv: arXiv:2311.14335. DOI: 10.48550/arXiv.2311.14335.

[Sob24] So B. Enhanced Gradient Boosting for Zero-Inflated Insurance Claims and Comparative Analysis of CatBoost, XGBoost, and LightGBM. 18 Jun 2024. DOI: 10.48550/arXiv.2307.07771.

[Ахв23] Ахведиани Ю. Т. Актуальные направления развития страхового рынка в современных условиях // Роль управления рисками и страхования в обеспечении устойчивости общества и экономики: Сб. тр. XXIV Междунар. науч.-практ. конф., Москва, 01 июня 2023. М.: МГУ, 2023. С. 26–30. EDN: NXNJEF. [[Akhvlediani Yu. T. Current directions of development of the insurance market in modern conditions. Moscow State University, 2023. Pp. 26–30 (In Russian).]]

[Бро23] Бронская Т. А. Искусственный интеллект в страховой сфере как инструмент повышения конкурентоспособности // Тенденции экономического развития в XXI веке: Мат-лы V Междунар. науч.-практ. конф. Минск, 01 марта 2023. Минск: БГУ, 2023. С. 389–391. EDN: ECHUVY. [[Bronskaya T. A. Artificial intelligence in the insurance sphere as a tool for improving Competitiveness. Belarusian State University, 2023. Pp. 389–391 (In Russian).]]

[Кир24] Кириченко А. О., Золкин А. Л. и др. Прогнозирование страховых рисков с использованием искусственного интеллекта // Прикладные экономические исследования. 2024. № 3. С. 196–203. EDN SLHSLS. [[Kirichenko A. O., Zolkin A. L., et al. Forecasting insurance risks using artificial intelligence // Applied Economic Research, 2024. No. 3. P. 196–203 (In Russian).]]

[Кот23] Котельников В. А. Поддержка принятия решений при управлении услугами системы моментальных платежей с использованием интеллектуальных технологий // СИИТ. 2023. Т. 5, № 4(13). С. 111–122. EDN: KEDROK. [[Kotelnikov V. A. Support for decision-making in managing services of the instant payment system using intelligent technologies // SIIT. 2023. Vol. 5, No. 4(13). P. 111-122. (In Russian).]]

[Куз23] Кузнецова В. Ю. Информационная технология принятия решений в микрофинансовой организации // СИИТ. 2023. Т. 5, № 3(12). С. 27–41. EDN: PDZIIA. [[Kuznetsova V. Yu. Information technology of decision-making in a microfinance organization // SIIT. 2023. Vol. 5, No. 3(12). P. 27-41. (In Russian).]]

[Мак24] Макарова Е. А., Габдуллина Э. Р. и др. Алгоритм интеллектуального анализа региональных данных об инвестиционном риске // СИИТ. 2024. Т. 6, № 1(16). С. 77–86. EDN: EBASQU. [[Makarova E. A., Gabdullina E. R. et al. Algorithm for intellectual analysis of regional data on investment risk // SIIT. 2024. Vol. 6, No. 1(16). P. 77-86. (In Russian).]]

[Мор24] Морозов М. И. Подготовка данных датасета "Vehicle Insurance Data 2018" для предсказания страховой премии // Информационные технологии и математическое моделирование (ИТММ-2024): Мат-лы XXIII Междунар. конф. им. А. Ф. Терпугова, Томск, 20–26 окт. 2024. Томск: ТГУ, 2024. С. 562–568. LQNWDB. [[Morozov M. I. Data Preparation of the Vehicle Insurance Dataset 2018 for Insurance Premium Prediction. 2024, pp. 562–568. (In Russian).]]

[Мор25] Морозов М. И. Предсказание наступления страхового случая с помощью трансформерных нейросетей // СИИТ. 2025. Т. 7, № 2(21). С. 96–102. EDN: FEFPBE. [[Morozov M. I. Prediction of the occurrence of an insured event using transformer neural networks: 2(21) // SIIT. 2025. Vol. 7, No. 2(21). P. 96–102. (In Russian).]]

ОБ АВТОРАХ | ABOUT THE AUTHORS

МОРОЗОВ Михаил Ильич
Югорский государственный университет, Россия.
mikhailmorozov99@bk.ru ORCID: 0009-0004-1217-9439.
Аспирант по спец. «Системный анализ, управление и обработка информации».

MOROZOV Mikhail Iliich
Yugra State University, Russia.
mikhailmorozov99@bk.ru ORCID: 0009-0004-1217-9439.
Post-graduate student of the specialty "Systems analysis, management and information processing".

МЕТАДАННЫЕ | METADATA

Заглавие: Прогнозирование страховых случаев на основе стекинг-ансамблирования.

Авторы: Морозов М. И.

Аннотация: Выявление рисков является неотъемлемой задачей любой финансовой организации, в том числе и автомобильного страхования. В настоящее время в бизнесе используются множество методов выявления рисков, в число которых входит построение моделей машинного обучения и гибких систем на основе деревьев решений. Системы принятия решений показывают высокую эффективность в выявлении рисков, однако такие системы имеют ряд проблем, которые вносят неопределенность в прогнозирование убыточности страховой компании: чувствительность к изменению данных, наличие линейных границ принятия решений, чувствительность к пропущенным данным. В решении ряда сложных бизнес-задач высокую эффективность показывают не только нейросети и методы машинного обучения, но и методы, позволяющие объединять прогнозы отдельных моделей – стекинг-ансамблирование. В данной работе мы ставим перед собой задачу преодолеть ограничения деревьев решений на основе ансамбля моделей машинного обучения в задаче предсказания вероятности наступления страхового случая. В качестве исходного набора данных используется информация о страховании автомобилей в России за 2023–2025 г. Проведено сравнение эффективности индивидуальных моделей машинного обучения и методов стекинг-ансамблирования. Полученные результаты работы позволяют судить об эффективности применённых методов в предсказании вероятности страхового случая.

Ключевые слова: Деревья решений; нейронные сети; машинное обучение; автомобильное страхование; убыточность; стекинг-ансамблирование.

Язык: Русский.

Статья поступила в редакцию 12 февраля 2026 г.

Title: Insurance claims prediction based on stacking ensembles.

Authors: Morozov M. I.

Abstract: Risk identification is an essential task for any financial organization, including vehicle insurance. Currently, businesses employ a variety of risk identification methods, among which are machine learning models and flexible systems based on decision trees. Decision-making systems demonstrate high effectiveness in risk detection, however, they suffer from several limitations that introduce uncertainty into the prediction of an insurer's loss ratio: sensitivity to data shifts, reliance on linear decision boundaries, and vulnerability to missing data. In addressing complex business problems, high performance is demonstrated not only by neural networks and machine learning methods but also by techniques that combine predictions from multiple individual models—specifically, stacking ensemble methods. In this study, we aim to overcome the limitations of decision trees by leveraging ensembles of machine learning models to predict the probability of an insurance claim occurrence. The dataset used comprises information on automobile insurance policies in Russia from 2023 to 2025. We compare the performance of individual machine learning models against stacking ensemble approaches. The results obtained demonstrate the effectiveness of the applied methods in predicting the likelihood of insurance claims.

Key words: Decision trees, neural networks, machine learning, vehicle insurance, loss ratio, stacking ensemble.

Language: Russian.

The article was received by the editors on 12 February 2026.