Методы, технологии и приложения искусственного интеллекта

УДК 004.85:004.8:368.2
DOI:10.25729/ESI.2026.43.3.002

Нейросетевой стекинг для прогнозирования убытков в страховании

Морозов Михаил Ильич

Югорский Государственный Университет,
Россия, Ханты-Мансийск, *mikhailmorozov99@bk.ru*

Аннотация. В работе раскрываются основные ограничений древовидных систем принятия решений в страховании – чувствительность к шуму и выбросам, склонность к переобучению, ограниченная способность к обобщению. Предлагаются способы их преодоления за счет методов машинного обучения, нейронных сетей и методов стекинг-ансамблирования. Ставится гипотеза о улучшении качества прогнозирования убытков, где прогнозы от нескольких базовых моделей (XGBoost, CatBoost, LogisticRegression, RandomForest, TabTransformer), в совокупности с исходными признаками набора данных, объединяются в более сильный ансамбль моделей, где строится метамодель на основе полносвязной нейронной сети с дополнительным входом исходным признаков. На данных страхования автомобилей в России за 2023-2025 годы, проводится исследование, где описывается порядок эксперимента с обучением разнородных моделей машинного обучения и нейронных сетей. Описывается алгоритм предобработки набора данных с акцентом на специфичное кодирование категорий (европейская классификация автомобилей по сегменту люксовости), делается акцент на несбалансированности целевого класса – менее 3% убытков во всем объеме данных. На основе обученных моделей строятся модели ансамблей, используя пять методов стекинг-ансамблирования: Simple Average, LogisticRegression-Constraint, Adaptive Regression by Mixing, ARM-Tweedie, Multilayer Perceptron (MLP). Проводится качественное сравнение базовых моделей с ансамблями моделей на основе классификационных (ROC-AUC, PR-AUC) и вероятностных метрик (Logloss, BrierScore). По результатам исследования лучшей базовой моделью является TabTransformer (BrierScore = 0.037, LogLoss = 0.185, ROC-AUC = 0.635, PR-AUC = 0.054). Ансамбль на основе полносвязной нейронной сети (MLP) превзошел ее по всем ключевым метрикам (BrierScore = 0.026, LogLoss = 0.131, ROC-AUC = 0.649, PR-AUC = 0.055). Полученный результат исследования позволяют подтвердить поставленную гипотезу: ансамбль на основе полносвязной нейронной сети позволил получить прирост по основным метрикам оценки. Полученный результат доказывает методологическую ценность подхода, даже в условиях ограниченного сигнала данных нейросетевой ансамбль в совокупности с исходными признаками обеспечивает более эффективное объединение прогнозов по сравнению с методами предиктивной аналитики и методами стекинг-ансамблирования.

Ключевые слова: деревья решений, нейронные сети, машинное обучение, автомобильное страхование, убыточность, стекинг ансамблирование

Цитирование: Морозов М.И. Нейросетевой стекинг для прогнозирования убытков в страховании / М.И. Морозов // Информационные и математические технологии в науке и управлении, 2026. – № 3(43). – С. 22-30. – DOI:10.25729/ESI.2026.43.3.002.

Введение. В страховании одной из важных задач является задача предсказания вероятности наступления страхового случая для оценки риска и формирования страховой премии [1, 2]. Делается это с целью обеспечить экономическую стабильность страховой организации и формирования бюджета на выплату страховых случаев [3,4].

В оценке рисков в страховании используют как актуарные оценки данных клиентов, так и системы принятия решений на основе древовидной структуры. Подобные системы чаще всего используют вербальные правила с условими «если…то…», что позволяет настраивать необходимую сегментацию по портфелю страхования. Ключевой особенностью использования деревьев решений в страховании является интерпретируемость и прозрачность, страховая организация может объяснить, почему было принято то или иное решение. Однако такие системы строятся и поддерживаются пользователями вручную, что неизбежно приводит к многим проблемам деревьев решений: чувствительность к шуму и выбросам, склонность к переобучению, ограниченная способность к обобщению.

Не менее популярным методом оценки рисков в страховых организациях являются построение моделей машинного [5] и нейросетевого обучения [6,7], что позволяет выявлять нелинейные зависимости в данных и чаще всего именно комбинация деревьев решений и моделей машинного обучения дает более точную оценку страхового риска по портфелю страхования. В рамках данной работы автор формулирует гипотезу, что комбинация нескольких индивидуальных моделей машинного и нейросетевого обучения позволит построить более сильный ансамбль на основе полносвязной нейронной сети для точной оценки страхового риска в задаче предсказания наступления страховых случаев. Для подтверждения гипотезы сравниваются несколько методов классического стекинга с простой полносвязной нейронной сетью, целью которой является объединение прогнозов индивидуальных моделей с дополнительным входом исходных данных для более точного прогнозирования вероятности наступления страхового случая.

Данное исследование является расширением предыдущей работы [8]. Дополнительно добавлен новый метод трансформерной нейросети TabTransformer в ансамбль, а также изменена предобработка данных для более качественного обучения моделей и сравнения полученных результатов

2. Модели и методы исследования. В рамках текущей работы обучаются пять моделей, которые лягут в основу ансамбля: XGBoost [9], CatBoost [10], LogisticRegression, RandomForest [11], TabTransformer [12]. В качестве методов ансамблирования используются: Simple Average (SA), LogisticRegression-Constraint (LR-C), Adaptive Regression by Mixing (ARM), Adaptive Regression by Mixing Tweedie [13], Multilayer Perceptron (MLP).

2.1. Описание набора данных. Набор данных содержит информацию о страховании автомобилей в РФ на период с 2023 по 2025 год. Он содержит данные полисов о страховании клиентов. Набор данных является несбалансированным, менее 3% убытков из всего датасета, содержит более 50 уникальных признаков. Набор данных основан на промышленных логах высоконагруженной системы, предоставленных партнерской организацией. В соответствии с соглашением о конфиденциальности, полный набор данных не подлежит публичному раскрытию.

2.2. Предварительная обработка данных. На основе данных проводится извлечение исходных признаков, где на основе гипотез строятся новые признаки набора данных. С помощью корреляции проверяются новые построенные признаки на основе вклада в целевую переменную. Специфические категориальные признаки кодируются особым способом для сохранения связи между признаками датасета:

1. Особым случаем кодирования является признак марки и модели транспортного средства, т.к. размер категорий марки и моделей может насчитывать более нескольких тысяч, что приводит к проблеме кодирования категорий. В зависимости от стоимости автомобиля и его комплектации может меняться и стоимость страхования, и, чтобы сохранить такую взаимосвязь, используется европейская классификация автомобилей, где на основе стоимости, марки и модели автомобиля определяется сегмент люксовости автомобиля [14].
2. На основе информации о водителях строятся как индивидуальные, так и географические признаки, позволяющие учесть различные взаимосвязи в данных.

При извлечении исходных признаков строится матрица корреляции, где рассматриваются взаимосвязи между целевым и исходными признаками. Удаляются признаки с высоким уровнем корреляции, выше значения 0.9, так как такие признаки не привносят в данные новый сигнал для прогнозирования целевой переменной.

Далее происходит кодирование набора данных. Даты кодируются на основе синусного и косинусного расстояния по месяцам и дням, для учета цикличности календарного года. После

проведения основной предобработки данных происходит разделение данные на 3 целевых выборки:

- Train – данные между 2024 и 2025 годом
- Valid – данные за последующие 3 месяца 2025 г.
- Test – данные старше 2025 г.

Последующим этапом заполняются пропуски, для бинарных признаков происходит заполнение нулевыми значениями, для числовых – медианное заполнение и масштабирование на основе MinMax.

В категориальных данных на уровне предобработки заполняются пропуски значениями ‘Unknown’ и кодируются они в последнюю очередь, т.к. различные методы построения моделей обрабатывают категориальные признаки по-разному.

Для всех методов, представленных в разделе 2.3, кроме метода TabTransformer, используется кодирование категорий с помощью метода OneHotEncoder. Для метода TabTransformer категории подаются в сыром виде, где сам метод строит эмбеддинги для кодирования категорий

На предварительном этапе обработки вычисляются отдельные веса для целевого признака на основе частоты встречаемого класса во всей выборке данных, что позволяет учесть редкость предсказываемого класса за счет использования полученных весов в функции потерь. Веса для классов в функции потерь позволяют вводить дополнительный штраф при обучении моделей, что позволяет обратить внимание обучаемой модели на редкий класс. Полный алгоритм предобработки данных можно видеть на рис. 1.

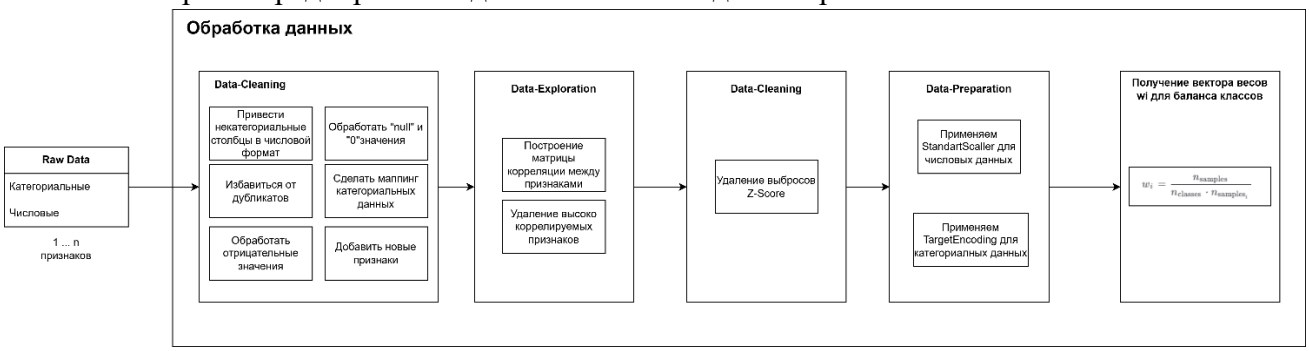

Рис. 1. Алгоритм предобработки данных

2.3. Методы предиктивной аналитики. В рамках текущей работы используются пять методов предиктивной аналитики, которые определяют основу ансамбля моделей. Выбор этих методов обусловливается их популярностью в задачах бинарной классификации и разнородностью самих методов в способах обучения моделей, а также популярностью использования этих методов в задачах предсказания убытков в страховой сфере [15,16]:

- **XGBoost** – устойчивый алгоритм градиентного бустинга, который позволяет эффективно обрабатывать пропуски и обеспечивать высокую точность прогнозирования в шумных данных, что является критичной и важной особенностью для страховых данных с естественной вариативностью.
- **CatBoost** – более продвинутая версия градиентного бустинга для более удобной работы с категориальными признаками. Метод использует способ кодирования категорий TargetEncoding и симметричные деревья, что снижает переобучение и повышает стабильность прогнозов на данных с большим количеством категорий.
- **LogisticRegression** – стандартная линейная модель, оценивающая вероятность наступления через логистическую функцию. С регуляризацией L1/L2 позволяет обеспечить устойчивость к переобучению, выделить ключевые факторы риска.

- **RandomForest** – ансамбль из несвязанных деревьев решений, основанный на методах бэггинга и случайного отбора признаков. Обладает устойчивостью к выбросам и шуму, позволяет обеспечить оценку важности признаков и служит базовым методом деревьев решений для сравнения с более сложными методами.
- **TabTransformer** – трансформерная архитектура обучения моделей, адаптированная для работы с табличными данными. Особенностью данного метода является кодирование категориальных признаки в виде эмбеддингов, наличие механизма самовнимания, что позволяет выявлять сложные зависимости между исходными признаками в данных. Указанные особенности позволяют улавливать более сложные взаимодействия в данных, что полезно при анализе редких событий.

2.4 Объединение прогнозов моделей. Для объединения прогнозов моделей, которые будут строится на основе методов предиктивной аналитики, применяется метод стекинга. Стекинг позволяет объединить прогнозы нескольких базовых моделей с помощью финальной метамодели, которая, в свою очередь, позволит усилить общую предсказательную способность ансамбля за счет компенсации индивидуальных особенностей каждой из моделей. Указанные ниже методы широко встречается в исследованиях [13] для построения метамоделей стекинга:

- **Simple Average (SA)** – метод простого среднего, полученные вероятности от базовых моделей складываются и делятся на количество моделей. Метод не учитывает особенности каждой из моделей, предполагается, что мы можем доверять каждой из моделей одинаково, когда на практике степень доверия может зависеть от различных факторов в данных.
- **LogisticRegression-Constraint (LR-C)** – метод позволяет строить взвешенную комбинацию моделей, что позволяет исключить доминирование отдельных моделей в ансамбле.
- **Adaptive Regression by Mixing (ARM)** – адаптивный смешиватель, где вес каждой базовой модели динамически корректируется в зависимости от ее точности.
- **Adaptive Regression by Mixing Tweedie** – модификация метода ARM с использованием распределения Tweedie в функции потерь, что позволяет учитывать специфику страховых данных с распределением убытков на хвосте, что позволяет повысить качество оценки как вероятности.
- **Multilayer Perceptron (MLP)** – простая полносвязная нейронная сеть, используемая как мета-классификатор. На вход принимает базовые прогнозы моделей и дополнительный поток исходных признаков, что, на основе гипотезы, должно позволять выявлять нелинейные зависимости между ошибками разных базовых моделей и исходными данными. В текущем исследования используется нейронная сеть, состоящая из 3-х слоев (128, 64, 32) с регуляризацией (dropout, L2) для предотвращения переобучения ансамбля.

2.5 Методы оценки моделей. В качестве методов оценки используется ряд ключевых метрик, которые позволяют оценить эффективность моделей как на основе калибровки вероятностей, так и способности выявлять редкие события с учетом несбалансированности данных:

- **BrierScore** – стандартная среднеквадратическая ошибка между прогнозируемой вероятностью и фактическим результатом в данных. Оценивает калибровку модели, где низкое значение говорит о соответствии прогнозируемых вероятностей реальным частотам наступления страховых случаев, что является критичным для расчета тарифа страхования на основе страховой премии. Теоретическим минимумом для данной задачи при 3% целевого класса является значение = 0.03.

- **LogLoss** – штрафует модель за неуверенные прогнозы, чем дальше прогноз от истинного класса, тем выше штраф. Метрика очень чувствительна к качеству оценки редких событий. Теоретическим минимумом для данной задачи при 3% целевого класса является значение, равное 0.13.
- **ROC-AUC** – оценивает способность модели ранжировать клиентов по риску независимо от порога бинаризации (threshold). Используется, как вспомогательная метрика для сравнения дискриминативной силы моделей, но является менее информативной при сильном дисбалансе классов.
- **PR-AUC** – предпочтительная метрика для задач предсказания случаев на несбалансированных данных, фокусируется на качестве прогнозов для редкого класса, отражает баланс между выявленными рисками и ложными срабатываниями. Высокое значение указывает на способность модели концертировать внимание на рискованных клиентах.

3. Экспериментальная установка. В качестве инструментов для проведения эксперимента использовалось программное обеспечение JupyterHub – для обучения моделей, MLFlow – для логирования результатов эксперимента. Порядок эксперимента состоит из нескольких ключевых шагов (Рис. 2).

После разделения данных на тренировочную, тестовую и валидационную выборки выполняется обучение моделей на полной тренировочной выборке. На основе методов предиктивной аналитики с помощью кросс валидации формируется датасет, который участвует в построении и обучении ансамблей. Датасет состоит из пяти колонок, где каждая колонка – это название метода предиктивной аналитики, каждая строка – предсказанная вероятность по экземпляру данных. На основе полученного датасета строится матрица корреляции обученных моделей, для выявления сильно коррелированных моделей выполняется построение ансамблей с помощью методов объединения прогнозов моделей.

В качестве финального этапа с помощью методов оценки моделей выполняется сравнение каждого из методов предиктивной аналитики с методами объединения прогнозов моделей. Для честной оценки и сравнения моделей обученных на основе методов предиктивной аналитики и полученных ансамблей финальная калибровка вероятностей на основе изотонической регрессии не проводится.

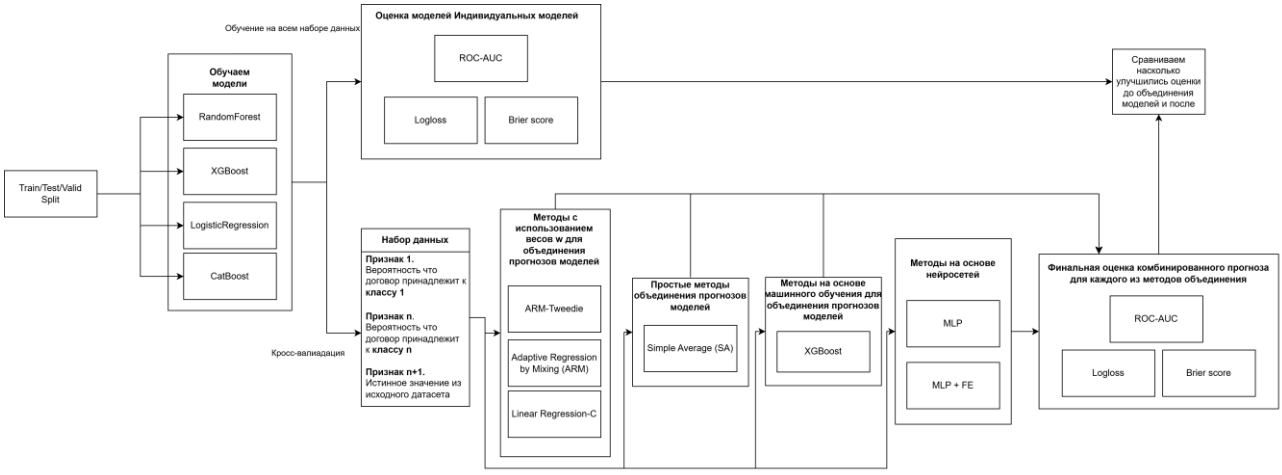

Рис. 2. Порядок проведения эксперимента

4. Интерпретация результатов. На основе предварительной обработки данных п. 2.2 и порядку эксперимента, описанному в п. 3, проводится обучение моделей. Ниже представлены результаты обучения на основе методов предиктивной аналитики (Таблица 1), матрица

корреляции обученных моделей (Рис. 3), результаты ансамблирования моделей (Таблица 2). Полученные метрики представлены на основе оценки моделей с помощью тестовой выборки.

Таблица 1. Результаты обучения моделей на основе методов предиктивной аналитики

Модель/Метрика	XGBoost	CatBoost	LR	RF	TabTransformer
BrierScore	0.170	0.134	0.252	0.215	**0.037**
LogLoss	0.521	0.435	0.701	0.624	**0.185**
ROC-AUC	0.617	0.621	0.643	**0.660**	0.635
PR-AUC	0.053	**0.055**	0.052	0.053	0.054

Таблица 2. Результаты обучения ансамблей

Модель/Метрика	SA	LR-C	ARM	ARM-T	MLP
BrierScore	0.114	0.208	0.232	0.232	**0.026**
LogLoss	0.401	0.608	0.658	0.657	**0.131**
ROC-AUC	0.648	0.644	0.647	**0.649**	**0.649**
PR-AUC	0.049	0.047	0.0506	0.051	**0.055**

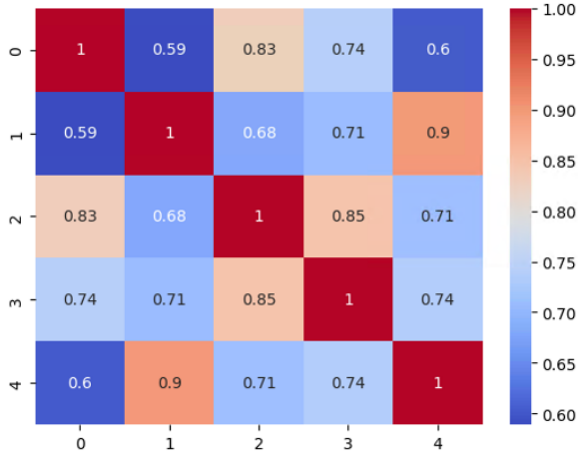

Рис. 3. Матрица корреляции моделей

Среди моделей (Табл. 1), обученных на основе методов предиктивной аналитики (п. 2.3), наилучшими показателями по метрикам BrierScore, LogLoss обладает метод TabTransformer, это означает, что данная модель точнее других предсказывает вероятности. По метрикам оценки способности модели различать классы, по метрике ROC-AUC лучшей моделью является RandomForest, по метрике PR-AUC – модель CatBoost. По комбинации вероятностных и дискриминационных метрик среди моделей предиктивной аналитики лидером является TabTransformer.

Среди моделей (Табл. 2), обученных на основе методов стекинг-ансамблирования (п. 2.4), по всем ключевым метрикам наблюдается превосходство полносвязной нейронной сети MLP, что позволяет подтвердить гипотезу о комбинации нескольких индивидуальных моделей как на основе машинного обучения, так и нейросетевого обучения с целью построить более сильный ансамбль для точной оценки страхового риска в задаче предсказания наступления страховых случаев. Однако низкий показатель PR-AUC во всех базовых моделях и ансамблях говорит об одном: в данных недостаточно сигнала для предсказания целевого события. Так как исходный набор данных содержит только информацию о договоре страхования, без обогащения данных из внешних источников, то такой информации недостаточно для предсказания целевого события на качественно значимом уровне, что и подтверждает метрика PR-AUC.

Полученные результаты не претендуют на практическую применимость в страховании без обогащения данных, однако доказывают методологическую ценность подхода: даже в

условиях ограниченного сигнала данных нейросетевой ансамбль в совокупность с исходными признаками данных обеспечивает более эффективное объединение прогнозов по сравнению с методами предиктивной аналитики и методами стекинг-ансамблирования.

Заключение. На основе полученных результатов исследования можно сделать следующие выводы:

- По результатам сравнения методов предиктивной аналитики лучшими являются: TabTransformer по метрикам BrierScore=0.037 и LogLoss=0.185, RandomForest по метрике ROC-AUC=0.66, CatBoost по метрике PR-AUC=0.055. По совокупности всех метрик TabTransformer лидирует, благодаря вероятностным метрикам и показателям PR-AUC, ROC-AUC, близким к моделям RandomForest и CatBoost.
- Среди методов стекинг-ансамблирования, лучшим является Multilayer Perceptron (MLP) с вероятностными метриками BrierScore=0.026, LogLoss=0.131, которые близки к теоретическому минимуму, дискриминационными метриками ROC-AUC=0.649, PR-AUC=0.055. Этот метод превосходит другие методы предиктивной аналитики и ансамблирования по общей совокупности и комбинации всех метрик.
- Ансамбль на основе полносвязной нейронной сети работает, поставленная гипотеза подтверждается. Исходные признаки датасета в совокупности с предсказаниями моделей предиктивной аналитики позволяют обучить более сильный ансамбль и добиться лучших результатов по метрикам.

Дальнейшим этапом исследования будет являться калибровка полученных моделей и ансамблей с последующей оценкой эффекта моделей по бизнес-метрикам на различных сегментах данных для проверки устойчивости поставленной гипотезы. Дополнительно планируется проверить гипотезу на данных с более сильным целевым сигналом, для осуществления практической применимости обученного ансамбля.

Благодарности. Автор выражает благодарность научному руководителю, заслуженному деятелю науки РФ, д-ру техн. наук, проф. Мельникову Андрею Витальевичу за профессиональное руководство, помощь и активное участие в развитии научного исследования.

Список источников

1. Ахвледиани Ю.Т. Актуальные направления развития страхового рынка в современных условиях / Ю.Т. Ахвледиани. – М.: МГУ, 2023. – С. 26–30.
2. 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, 2023, vol. 7, no. 4, pp. 22–41.
3. Бронская Т.А. Искусственный интеллект в страховой сфере как инструмент повышения конкурентоспособности / Т.А. Бронская. – Минск: БГУ, 2023. – С. 389–391.
4. Кириченко А.О. Прогнозирование страховых рисков с использованием искусственного интеллекта / А.О. Кириченко, А.Л. Золкин, И.А. Поскряков // Прикладные экономические исследования, 2024. – № 3. – С. 196–203.
5. So B. Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM. arXiv:2307.07771, 2024, DOI: 10.48550/arXiv.2307.07771.
6. Singh U. et al. Comparative analysis of transformers for modeling tabular data: a casestudy using industry scale dataset. arXiv:2311.14335, 2023, DOI: 10.48550/arXiv.2311.14335.
7. Морозов М.И. Предсказание наступления страхового случая с помощью трансформерных нейросетей / М.И. Морозов // Системная инженерия и информационные технологии, 2025. – Т. 7. – № 2(21). – С. 96–102.
8. Морозов М.И. Прогнозирование страховых случаев на основе стекинг-ансамблирования / М.И. Морозов // Системная инженерия и информационные технологии, 2026. – Т. 8. – № 1(25). – С. 93–100.
9. Chen T., Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.

10. Prokhorenkova L. et al. CatBoost: unbiased boosting with categorical features. arXiv:1706.09516, 2019, DOI: 10.48550/arXiv.1706.09516.
11. Breiman L. Random forests. Machine Learning, 2001, vol. 45, no. 1, pp. 5–32, DOI: 10.1023/A:1010933404324.
12. Huang X. et al. TabTransformer: tabular data modeling using contextual embeddings. arXiv:2012.06678, 2020, DOI: 10.48550/arXiv.2012.06678.
13. Chenglong Ye, Zhang L., Han M. et al. Combining predictions of auto insurance claims. Econometrics, 2022, vol. 10, iss. 2, 19, DOI: 10.3390/econometrics10020019.
14. Какие классы автомобилей существуют, и как их определить. – Major Auto. – URL: https://www.major-auto.ru/blog/132/ (дата обращения: 29.03.2026).
15. Baran S., Rola P. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. arXiv, 2022, DOI:10.48550/arXiv.2204.06109.
16. Mohamed Hanafy, Ruixing Ming. Machine learning approaches for auto insurance big data. Risks, 2021, vol. 9(2), no. 42, DOI: 10.3390/risks9020042.

***Морозов Михаил Ильич**. Аспирант спец. «Системный анализ, управление и обработка информации», Югорский Государственный Университет, SPIN: 7231-0420, ORCID: 0009-0004-1217-9439, mikhailmorozov99@bk.ru, Ханты-Мансийский автономный округ - Югра, г. Ханты-Мансийск, ул. Чехова, 1*

UDC 004.85:004.8:368.2
DOI:10.25729/ESI.2026.43.3.002

Neural network stacking for loss forecasting in Insurance

Mikhail I. Morozov

Yugra state university, Russia, Khanty-Mansiysk, *mikhailmorozov99@bk.ru*

Abstract. This study examines the primary limitations of tree-based decision models in the insurance sector, namely their sensitivity to noise and outliers, susceptibility to overfitting, and limited generalization capabilities. To overcome these challenges, we propose leveraging advanced machine learning techniques, neural networks, and stacking ensemble methods. We hypothesize that the quality of loss prediction can be slightly improved by combining the predictions of multiple base models (XGBoost, CatBoost, Logistic Regression, Random Forest, and TabTransformer), alongside the original dataset features, into a more robust ensemble. Specifically, the meta-model is constructed as a fully connected neural network with an additional input layer for the original features. The empirical study utilizes Russian auto insurance data from 2023 to 2025, detailing the experimental framework for training heterogeneous machine learning and neural network models. The data preprocessing pipeline is described, with a particular focus on a specific categorical encoding scheme based on the European car classification by luxury segment. Special attention is given to the severe imbalance of the target class, as losses account for less than 3% of the total dataset. Based on the trained base models, ensemble models are constructed using five stacking techniques: Simple Average, Constrained Logistic Regression, Adaptive Regression by Mixing (ARM), ARM-Tweedie, and Multilayer Perceptron (MLP). A comprehensive comparative analysis of the base models and the ensembles is conducted using both classification metrics (ROC-AUC, PR-AUC) and probabilistic metrics (Log Loss, Brier Score). The results indicate that the best-performing base model is TabTransformer (Brier Score = 0.037, Log Loss = 0.185, ROC-AUC = 0.635, PR-AUC = 0.054). The MLP-based ensemble outperformed it across all key metrics (Brier Score = 0.026, Log Loss = 0.131, ROC-AUC = 0.649, PR-AUC = 0.055). These findings confirm the initial hypothesis: the fully connected neural network ensemble yields a slightly improvement in the primary evaluation metrics. This demonstrates the methodological value of the proposed approach: even under conditions of a weak data signal, the neural network ensemble, augmented with original features, ensures a more effective aggregation of predictions compared to traditional predictive analytics and alternative stacking ensemble methods.

Keywords: decision trees, neural networks, machine learning, vehicle insurance, loss ratio, stacking ensemble

Acknowledgements: The author expresses gratitude to the scientific supervisor, Honored Scientist of the Russian Federation, Doctor of Technical Sciences, Professor Andrey Vitalievich Melnikov for his professional guidance, assistance and active participation in the development of scientific research.

References

1. Akhvlediani Yu.T. Aktual'nyye napravleniya razvitiya strakhovogo rynka v sovremennykh usloviyakh [Current directions of insurance market development in modern conditions]. Moscow, Moscow State University Publ., 2023, pp. 26–30.
2. 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, 2023, vol. 7, no. 4, pp. 22–41.
3. Bronskaya T.A. Iskusstvennyy intellekt v strakhovoy sfere kak instrument povysheniya konkurentosposobnosti [Artificial intelligence in the insurance sector as a tool for increasing competitiveness]. Minsk, Belarusian State University Publ., 2023, pp. 389–391.
4. Kirichenko A.O., Zolkin A.L., Poskryakov I.A. Prognozirovaniye strakhovykh riskov s ispol'zovaniyem iskusstvennogo intellekta [Forecasting insurance risks using artificial intelligence]. Prikladnyye ekonomicheskiye issledovaniya [Applied Economic Research], 2024, no. 3, pp. 196–203.
5. So B. Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM. arXiv:2307.07771, 2024, DOI: 10.48550/arXiv.2307.07771.
6. Singh U. et al. Comparative analysis of transformers for modeling tabular data: a casestudy using industry scale dataset. arXiv:2311.14335, 2023, DOI: 10.48550/arXiv.2311.14335.
7. Morozov M.I. Predskazaniye nastupleniya strakhovogo sluchaya s pomoshch'yu transformernykh neyrosetey [Prediction of the occurrence of an insured event using transformer neural networks]. Sistemnaya inzheneriya i informatsionnyye tekhnologii [System Engineering and Information Technology], 2025, vol. 7, no. 2(21), pp. 96–102.
8. Morozov M.I. Prognozirovaniye strakhovykh sluchayev na osnove steking-ansamblirovaniya [Forecasting insurance claims based on stacking ensemble]. Sistemnaya inzheneriya i informatsionnyye tekhnologii [System Engineering and Information Technology], 2026, vol. 8, no. 1(25), pp. 93–100.
9. Chen T., Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794, DOI: 10.1145/2939672.2939785.
10. Prokhorenkova L. et al. CatBoost: unbiased boosting with categorical features. arXiv:1706.09516, 2019, DOI: 10.48550/arXiv.1706.09516.
11. Breiman L. Random forests. Machine Learning, 2001, vol. 45, no. 1, pp. 5–32, DOI: 10.1023/A:1010933404324.
12. Huang X. et al. TabTransformer: tabular data modeling using contextual embeddings. arXiv:2012.06678, 2020, DOI: 10.48550/arXiv.2012.06678.
13. Chenglong Ye, Zhang L., Han M. et al. Combining predictions of auto insurance claims. Econometrics, 2022, vol. 10, iss. 2, 19, DOI: 10.3390/econometrics10020019.
14. Kakiye klassy avtomobiley sushchestvuyut, i kak ikh opredelit' [What car classes exist and how to determine them]. Major Auto. Available at: https://www.major-auto.ru/blog/132/ (accessed: 03.29.2026).
15. Baran S., Rola P. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. arXiv, 2022, DOI:10.48550/arXiv.2204.06109.
16. Mohamed Hanafy, Ruixing Ming. Machine learning approaches for auto insurance big data. Risks, 2021, vol. 9(2), no. 42, DOI: 10.3390/risks9020042.

Morozov Mikhail Iliich. *Post-graduate student of the specialty "Systems analysis, management and information processing., Yugra State University, SPIN: 7231-0420, ORCID: 0009-0004-1217-9439, mikhailmorozov99@bk.ru, Khanty-Mansi Autonomous Okrug - Yugra, Khanty-Mansiysk, Chekhov Street, 16.*

Статья поступила в редакцию 06.04.2026; одобрена после рецензирования 20.06.2026; принята к публикации 26.07.2026.

The article was submitted 04/06/2026; approved after reviewing 06/20/2026; accepted for publication 07/26/2026.