ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
2025. Т. 7, № 2 (21). С. 96–102 http://siit.ugatu.su
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.89 DOI 10.54708/2658-5014-SIIT-2025-no2-p96 EDN FEFPBE

Предсказание наступления страхового случая с помощью трансформерных нейросетей

М. И. Морозов

Аннотация. Одной из актуальных проблем рынка страхования является проблема выявления рисков при заключении договора со страхователем. Страховые организации используют множество методов выявления рисков страхования, такие как: скоринг, обогащение данных из внешних ресурсов, построение гибких систем принятия решений на основе деревьев решений. Системы принятия решений с течением времени показали свою высокую эффективность в процессе сокращения рисков при заключении договоров страхования, однако такие системы имеют ряд проблем, которые вносят неопределенность в прогнозирование убыточности страховой компании. Из основных ключевых проблем деревьев решений можно выделить: чувствительность к изменению данных, наличие линейных границ принятия решений, чувствительность к пропущенным данным. В последнее время нейросетевые технологии показали свою высокую эффективность в решении ряда сложных бизнес-задач. В данной работе мы ставим перед собой задачу преодолеть ограничения деревьев решений с помощью нейронной сети трансформера – TabTransformer в задаче предсказания вероятности наступления страхового случая. В качестве исходного набора данных была взята информация о страховании транспортных средств Эфиопии за 2011–2018 годы. Проведено сравнение эффективности трансформерных нейросетей с методами построения деревьев решений CART и Random Forest. Полученные результаты работы могут быть использованы в дальнейшем для реализации рекомендательной андеррайтинговой системы, целью которой будет помощь в предсказании возможных убытков по страховому полису.

Ключевые слова: деревья решений; трансформерные нейронные сети; машинное обучение; автомобильное страхование; страховой случай.

Введение

Страхование является одной из ключевых индустрий, которая играет решающую роль в обеспечении финансовой стабильности и безопасности общества. В процессе работы страховой сектор сталкивается с рядом сложных проблем, одна из которых – проблема выявления рисков [1]. Одним из способов преодоления этой проблемы является прогнозирование вероятности наступления страхового случая с помощью нейронных сетей, что может помочь в регулировании показателя убыточности страховой компании [2].

В настоящее время страховые компании активно используют различные методы выявления рисков при заключении договоров страхования. Кредитный и социальный скоринг позволяет оценить надежность клиента, обогащение данных из внешних ресурсов позволяет получить дополнительную информацию [3] и гибкие системы принятия решений, имеющие древовидную структуру, которые позволяют учитывать сложные зависимости в данных страхового полиса для решения бизнес-задач [4], что позволяет страховым компаниям минимизировать риски в процессе принятия решений.

В данной статье мы обратим свое внимание к системам проверок на основе деревьев решений, которые имеют ряд ключевых недостатков в процессе страхования, что вносит неопределенность в процесс принятия решений и прогнозирование убыточности страховой компании.

Рекомендовано к публикации программным комитетом X Международной научной конференции ITIDS’2024 «Информационные технологии интеллектуальной поддержки принятия решений», Уфа, 12–14 ноября 2024 г.

Постановка задачи и описание проблемы

В последние годы в финансовой сфере все больший интерес представляют нейросетевые технологии, которые уже показали свою высокую эффективность в решении сложных задач [5, 6]. В частности, в данной работе мы рассмотрим применение технологий машинного обучения – RandomForest [7], CART (Classification and Regression Trees) [8] и нейронной сети трансформера – TabularTransformer (далее TabFormer) [9, 10] для предсказания вероятности убытка по страховому случаю.

Системы принятия решений с древовидной структурой также имеют недостатки классических деревьев решений [11], 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 [12].

Трансформерные архитектуры нейросетей, изначально разработанные для обработки последовательных данных, таких как текст, показали свою высокую эффективность и в табличных данных. Архитектура TabFormer, являющейся разновидностью моделей трансформеров, предлагает ряд преимуществ, по сравнению с традиционными методами построения деревьев решений. TabFormer менее подвержена изменениям в данных благодаря своей архитектуре трансформера, основанной на многослойных механизмах самовнимания. Трансформеры способны моделировать сложные нелинейные зависимости между признаками, что позволяет более точно прогнозировать наступление страховых случаев. Они могут более эффективно обрабатывать пропуски в данных, что улучшает качество модели. Использование TabTransformer для предсказания вероятности наступления страховых случаев на основе данных о страховании транспортных средств позволяет преодолеть указанные недостатки традиционных методов и может значительно улучшить точность и надежность прогнозов.

Таким образом, целью данной работы является преодоление ограничений деревьев решений с помощью трансформерной нейросети TabTransformer в задаче предсказания вероятности убытка по страховому полису.

Краткий обзор методов

В настоящей работе были использованы два метода построения моделей машинного обучения на основе деревьев решений: RandomForest и DecisionTree CART и один метод построения модели на основе трансформерной архитектуры нейронных сетей – TabFormer. Ниже представлен краткий обзор методов, используемых в данном исследовании.

RandomForest представляет собой ансамбль деревьев принятия решений, где каждое дерево обучается на случайном подмножестве исходных данных и признаков. Этот подход позволяет уменьшить переобучение и улучшить обобщающую способность модели. Среди преимуществ Random Forest можно выделить его устойчивость к шуму и способность работать с большими наборами данных. Из ограничений данного метода можно выделить сложность интерпретации результатов и необходимость оптимизации гиперпараметров под конкретные данные.

CART является методом построения деревьев принятия решений, который делит данные на подмножества, основываясь на значениях признаков, и строит дерево решений, минимизирующее ошибку классификации или регрессии. Преимущества CART включают простоту интерпретации и легкость визуализации. Тем не менее метод также страдает от проблем, связанных с переобучением и чувствительностью к изменению данных.

TablularTransformer представляет собой метод построения модели машинного обучения, основанный на трансформерной архитектуре, специально адаптированной для табличных данных. В отличие от традиционных методов, работающих с деревьями решений, TabFormer использует механизм самовнимания (self-attention) для захвата сложных зависимостей между признаками. Это позволяет модели эффективно обучаться на данных с высокой размерностью и учитывать взаимодействия между признаками, которые могут быть упущены методами, основанными на деревьях решений. Одним из ключевых преимуществ TabFormer является его способность обрабатывать как числовые, так и категориальные данные, интегрируя их в единое представление. Это достигается с помощью эмбеддингов категориальных признаков, что позволяет модели лучше понимать и использовать информацию из различных типов данных. В дополнение к этому TabFormer обладает высокой гибкостью и адаптивностью, что делает его пригодным для широкого спектра задач. Однако как и у любого метода, у TabFormer есть свои ограничения. Во-первых, его обучение требует значительных вычислительных ресурсов и времени, особенно на больших наборах данных. Во-вторых, интерпретация результатов модели может быть сложной из-за высокой степени абстракции, присущей трансформерным архитектурам.

В совокупности использование таких различных методов, как RandomForest, CART и TabFormer, позволяет получить более полное представление о данных и выбрать наиболее подходящий подход для решения задачи предсказания убытка по договору страхования.

Набор данных

В качестве исходного набора данных был взят открытый эфиопский датасет «Vehicle Insurance Data 2018» [13] с информацией о страховании автомобилей за 2011–2018 годы. Данные в необработанном виде содержат около 800 тыс. записей и 16 ключевых признаков, которые представлены в табл. 1.

Таблица 1

Описание признаков для набора данных «Vehicle Insurance Data 2018»

№ п/п	Наименование	Тип данных	Описание
0	SEX	int64	0 – юридическое лицо, 1 – мужчина, 2 – женщина
1	INSR_BEGIN	object	Дата начала действия страховки
2	INSR_END	object	Дата окончания действия страховки
3	EFFECTIVE_YR	object	Год вступления страхового полиса в силу
4	INSR_TYPE	int64	Тип страхования: 1201 – частный, 1202 – коммерческий,
5	INSURED_VALUE	float64	1204 – автомобильный риск
6	PREMIUM	float64	Стоимость объекта страхования
7	OBJECT_ID	int64	Страховая премия
8	PROD_YEAR	float64	Идентификатор застрахованного объекта
9	SEATS_NUM	float64	Год производства автомобиля
10	CARRYING_CAPACITY	float64	Количество сидений в автомобиле
11	TYPE_VEHICLE	object	Грузоподъемность автомобиля
12	CCM_TON	float64	Тип автомобиля
13	MAKE	object	Объем двигателя для легковых автомобилей или вес в тоннах для грузовых автомобилей
14	USAGE	object	Марка автомобиля
15	CLAIM_PAID	float64	Цель использования автомобиля

Данные были предварительно обработаны, было произведено приведение данных к числовому формату, удаление дубликатов, обработка пропусков, удаление выбросов, кодирование категориальных переменных и масштабирование числовых признаков. В обработанном виде набор данных содержит 220 тыс. записей о страховании транспортных средств.

Для подачи данных в алгоритмы машинного обучения обработанный набор данных был разделен на обучающую и тестовую выборку в соотношении 80 % к 20 %, в которой был соблюден баланс целевого класса CLAIM_PAID. Финальный набор данных для обучения содержит 12 663 записи с отсутствием убытка CLAIM_PAID = 0 и 12 663 записи с наличием убытка CLAIM_PAID > 0. Для записей с наличием убытков был проведен биннинг переменной, путем разделения данных на 3 типа убытков: small, medium, big (маленький, средний, высокий) на основе признака частоты. В итоговом формате целевая переменная CLAIM_PAID была закодирована как целевой класс, содержащий 4 необходимых класса для предсказания убытков: no_claim, small, medium, big.

Экспериментальная установка

В качестве среды для проведения эксперимента был выбран облачный сервис Google Collab [14], в основе которого лежит язык программирования Python. В качестве исходной библиотеки для создания моделей машинного обучения RandomForest и CART была выбрана библиотека sklearn [15], которая содержит необходимые настройки классов для обучения моделей. Исходными библиотеками для создания и обучения модели трансформера TabFormer были взяты Keras [16] и Tensorflow [17].

Методы машинного обучения RandomForest и CART

Процесс обучения моделей машинного обучения RandomForest и CART проводился с помощью классов инициализаторов RandomForestClassifier и DecisionTreeClassifier путем вызова метода классов «fit(X_train, y_train)», где X_train – обучающая выборка, y_train – тестовая выборка набора данных.

В качестве гиперпараметров обучения моделей Random Forest и CART были взяты исходные настройки классов RandomForestClassifier и DecisionTreeClassifier библиотеки sklearn [15].

Оценка качества работы моделей производилась путем вызова метода «predict(X_test)», где X_test является тестовым набором данных. В качестве метрики оценки моделей был использован метод f1_score библиотеки sklearn.

Трансформерная нейросеть TabTransformer

Процесс построения и обучения модели на основе метода TabTransformer проводился на основе инструкций/рекомендаций Keras путем создания функций и процедур для инициализации итоговой модели.

В качестве гиперпараметров для обучения модели использовались следующие значения:

- LEARNING_RATE = 0.001 – отображает скорость обучения модели и то, насколько сильно изменяются веса модели на каждом шаге обучения.
- WEIGHT_DECAY = 0.0001 – коэффициент регуляризации L2, что позволяет добавлять штраф в модель для предотвращения переобучения.
- DROPOUT_RATE = 0.2 – отображает вероятность выключения нейронов во время обучения модели.
- BATCH_SIZE = 265 – количество примеров из набора данных, использующихся на каждом шаге обучения.
- NUM_EPOCHS = 15 – количество эпох обучения.
- NUM_TRANSFORMER_BLOCKS = 3 – количество блоков трансформера в модели, который состоит из механизма самовнимания и слоев линейного преобразования.
- NUM_HEADS = 4 – количество указателей в механизме самовнимания.

- EMBEDDING_DIMS = 16 – размерность векторов эмбеддингов для категориальных признаков. Эмбеддинги преобразуют категориальные признаки в плотные векторы фиксированной размерности, что позволяет модели лучше работать с категориальными данными.
- MLP_HIDDEN_UNITS_FACTORS = “[2, 1]” – количество нейронов в скрытых слоях MLP (Multi-layer Perceptron).
- NUM_MLP_BLOCKS = 2 – количество блоков MLP в базовой модели. Каждый блок включает в себя несколько слоев MLP, что позволяет модели захватывать нелинейные зависимости между признаками.

Оценка качества работы модели производилась путем вызова метода «evaluate». В качестве метрики оценки моделей был использован метод F1Score библиотеки Keras [16].

Интерпретация результатов

В качестве метрик оценки эффективности моделей использовались метрики полноты, точности и мера F1-score [18]. Итоги обучения моделей машинного обучения и трансформерной нейросети TabFormer представлены в табл. 2.

Таблица 2

Результаты обучения моделей

Модель / Метрика	RandomForest	CART	TabFormer
Accuracy	0.4454	0.4500	—
Precision	0.4264	0.4110	—
F1	0.4345	0.3916	0.4533

Основным преимуществом моделей машинного обучения является высокая скорость получения конечной модели (10 секунд) в сравнении с моделью трансформером TabFormer (8 минут). Однако для машинного обучения требуется предварительная подготовка исходных данных: кодирование категориальных переменных и масштабирование числовых признаков, что может вызывать сложности в работе с исходными данными. Для трансформерной нейросети TabFormer кодирование категориальных переменных осуществляется с помощью эмбеддингов.

Оценка моделей в табл. 2 по F1-мере не позволяет судить о преимуществе отдельных моделей. Низкая точность моделей объясняется недостаточным количеством признаков для предсказания события по убытку. Обусловливается это тем, что исходный датасет изначально был выбран некорректно для предсказания вероятности убытка, так как в автостраховании основными параметрами, влияющим на показатель убытков, являются телематические данные [19, 20], в том числе технические характеристики транспортных средств. Для подтверждения гипотезы о преимуществе трансформерных моделей в задаче предсказания вероятности убытка по договору страхования рекомендуется сменить набор данных на более релевантный.

Заключение

В процессе работы были достигнуты несколько ключевых результатов:

1) Основываясь на показателях F1-меры построенных моделей (табл. 2), проведенное исследование в данный момент не позволяет подтвердить гипотезу о преимуществе трансформерных нейросетей в задаче предсказания вероятности убытка по страховому полису.

2) Сравнение трех методов – CART, RandomForest и TabFormer – выявило, что традиционные методы имеют свои ограничения, такие как чувствительность к изменениям данных

и необходимость предварительной обработки. С другой стороны, TabFormer показал свои преимущества в работе с табличными данными, особенно в условиях наличия сложных зависимостей между признаками. Однако стоит отметить, что для достижения высоких результатов требуется значительное количество вычислительных ресурсов и времени на обучение модели.

3) Основной целью исследования было преодоление ограничений деревьев решений с помощью трансформерной нейросети TabTransformer. На данный момент цель пока нельзя считать достигнутой, так как необходимо окончательное подтверждение выводов и повышение точности моделей. Целесообразны дальнейшие исследования с использованием более релевантных для данной задачи наборов данных.

Благодарности и поддержка

Автор выражает благодарность научному руководителю д-ру техн. наук, профессору Андрею Витальевичу Мельникову за профессиональное руководство, помощь и активное участие в развитии научного исследования. Также автор считает целесообразным отметить работы по смежной тематике [21–25], полезные в контексте данного исследования.

Список литературы / References

[1] Бронская Т. А. Искусственный интеллект в страховой сфере как инструмент повышения конкурентоспособности. Минск: Белорусск. гос. ун-т, 2023. С. 389–391. EDN ECHUVY. [[Bronskaya T. A. Artificial Intelligence in the Insurance Sector as a Tool for Increasing Competitiveness. Belarusian State University, 2023, pp. 389–391. EDN ECHUVY. (In Ruussian.]]

[2] Ritho B. M., Simiyu E., Omagwa J. “The impact of loss ratio on the financial stability of insurance firms in Kenya 4” // Journal of Finance and Accounting. 2023. Vol. 7. No. 4. Pp. 22–41. DOI 10.53819/81018102t4161. EDN ACMIPP.

[3] Авдеев Д. А. «Социальный скоринг» как фактор нарушения права на неприкосновенность частной жизни // Международный научно-исследовательский журнал. 2023. № 6 (132). DOI 10.23670/IRJ.2023.132.126. EDN MZQKFG. [[Avdeev D. A. "Social scoring as a factor in violation of the right to privacy” // International Research journal. 2023. No. 6 (132). DOI 10.23670/IRJ.2023.132.126. EDN MZQKFG. (In Ruussian.]]

[4] Agarwal P. et al. “A process-aware decision support system for business processes” // Proc. 28th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining. New York, NY: USA Association for Computing Machinery, 2022. Pp. 2673–2681. DOI 10.1145/3534678.3539088.

[5] Eling M., Nuessle D., Staubli J. “The impact of artificial intelligence along the insurance value chain and on the insurability of risks” // Geneva Pap Risk Insur Issues Pract 47. 2022. Pp. 205–241. DOI 10.1057/s41288-020-00201-7. EDN QQCAPW.

[6] Mahohoho, Brighton & Chimedza, Charles & Matarise, Florance & Munyira, Sheunesu. Artificial Intelligence Based Automated Actuarial Loss Reserving Model for the General Insurance Sector. 2023. 10.21203/rs.3.rs-3124884/v1.

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

[8] Breiman L. “Classification and Regression Trees“ // Wadsworth International Group. 1984.

[9] Padhi I. et al. “Tabular Transformers for Modeling Multivariate Time Series“. arXiv:2011.01843. arXiv, 2021. DOI 10.1109/ICASSP39728.2021.9414142.

[10] Huang X. et al. “TabTransformer Tabular Data Modeling Using Contextual Embeddings“ arXiv:2012.06678. arXiv, 2020.

[11] Субботин С. А. Построение деревьев решений для случая малоинформативных признаков // Радіоелектроніка, інформатика, управління. 2019. №1 (48). С. 122–131. [[Subbotin S. A. “Construction of decision trees for the case of uninformative features” // Radioelectronics, Informatics, Management. 2019. No. 1 (48), pp. 122–131. (In Ruussian.]]

[12] Bengio Y., Delalleau O., Simard C. “Decision trees do not generalize to new variations” // Computational Intelligence. 2010. Vol. 26. Pp. 449–467. DOI 10.1111/j.1467-8640.2010.00366.x.

[13] Vehicle Insurance Data [Online]. Available https://www.kaggle.com/datasets/imtkaggleteam/vehicle-insurance-data (Accessed 22.06.2024).

[14] Google Colab [Online]. Available https://colab.research.google.com/ (Accessed 18.09.2024).

[15] User Guide [Online] // scikit-learn. Available https://scikit-learn/stable/user_guide.html (Accessed 18.09.2024).

[16] Team K. “Keras documentation Developer guides“ [Online]. Available https://keras.io/guides/ (Accessed 18.09.2024).

[17] All symbols in TensorFlow 2 | TensorFlow v2.16.1 [Online] // TensorFlow. Available URL https://www.tensorflow.org/api_docs/python/tf/all_symbols (Accessed 18.09.2024).

[18] Opitz J. “A closer look at classification evaluation metrics and a critical reflection of common evaluation practice” // Transactions of the Association for Computational Linguistics. 2024. Vol. 12. Pp. 820–836. DOI 10.1162/tacl_a_00675. EDN SQAYAZ.

[19] Peiris H. et al. “Integration of traditional and telematics data for efficient insurance claims prediction” // ASTIN Bulletin. The Journal of the IAA. 2024. Vol. 54. No. 2. Pp. 263–279. DOI 10.1017/asb.2024.6. EDN OSCGGA.

[20] Duval F., Boucher J.-P., Pigeon M. “How Much Telematics Information Do Insurers Need for Claim Classification“. arXiv:2105.14055. arXiv, 2021.

[21] Кучкарова Н. В. Оценка актуальных угроз и уязвимостей объектов критической информационной инфраструктуры с использованием технологий интеллектуального анализа текстов // СИИТ. 2024. Т. 6. № 2(17). С. 50–65. EDN NLDWBE. [[Kuchkarova N. V. “Assessment of current threats and vulnerabilities of critical information infrastructure objects using text mining technologies” // SIIT. 2024. Vol. 6, no. 2(17), pp. 50–65. EDN NLDWBE. (In Ruussian.]]

[22] Макарова Е. А., Габдуллина Э. Р., Юсупов М. М., Вагапова Г. Р. 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., Yusupov M. M., Vagapova G. R. “Algorithm for intellectual analysis of regional data on investment risk” // SIIT. 2024. Vol. 6, no. 1(16), pp. 77-86. EDN EBASQU. (In Ruussian.]]

[23] 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), pp. 111–122. EDN KEDROK. (In Ruussian.]]

[24] Шалфеева Е. А. Методология производства жизнеспособных систем доверительного искусственного интеллекта // СИИТ. 2023. Т. 5. № 4(13). С. 28–49. EDN CJTKQH. [[Shalfeeva E. A. “Methodology to produce viable systems of trustworthy artificial intelligence” // SIIT. 2023. Vol. 5, no. 4(13), pp. 28–49. EDN CJTKQH. (In Ruussian.]]

[25] 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), pp. 27–41. EDN PDZIIA. (In Ruussian.]]

Поступила в редакцию 22 января 2025 г.

МЕТАДАННЫЕ / METADATA

Title Prediction of an insurance claim using transformer neural networks.

Abstract One of the pressing issues in the insurance industry is the identification of risks when concluding contracts with the insured. Insurance companies use various methods to assess insurance risks, such as scoring, data enrichment from external sources, and the development of flexible decision-making systems based on decision trees. Over time, decision tree systems have proven to be highly effective in mitigating risks during the insurance contracting process. However, these systems have several limitations that introduce uncertainty in predicting an insurance company's loss ratio. The main issues with decision trees include sensitivity to data changes, the existence of linear decision boundaries, and susceptibility to missing data. Recently, neural network technologies have demonstrated their high efficiency in solving complex business problems. This paper aims to address the limitations of decision trees by using a transformer neural network—TabTransformer—in predicting the probability of an insurance claim. The dataset used consists of Ethiopian vehicle insurance information from 2011 to 2018. The performance of transformer neural networks is compared with the decision tree methods CART and Random Forest. The results obtained can be used in the future to implement a recommendation-based underwriting system, which will assist in predicting potential losses under an insurance policy.

Key words decision trees; transformer neural networks; machine learning; automobile insurance; insurance claim.

Язык статьи / Language Русский / Russian.

Об авторе / About the author

МОРОЗОВ Михаил Ильич
Югорский государственный университет, Россия.
Аспирант.
E-mail mikhailmorozov99@bk.ru
ORCID https://orcid.org/0009-0004-1217-9439

MOROZOV Mikhail Ilyich
Ugra State University, Russia.
Postgraduate student.
E-mail mikhailmorozov99@bk.ru
ORCID https://orcid.org/0009-0004-1217-9439