

ИСПОЛЬЗОВАНИЕ МАШИННОГО ОБУЧЕНИЯ И КЛАССИЧЕСКИХ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ ПОСТРОЕНИЯ СКОРИНГОВОЙ МОДЕЛИ В АВТОСТРАХОВАНИИ

Ю. Н. Миронкина, Д. И. Тимофеев

*Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия*

E-mail: YMironkina@hse.ru, ditimofeev@edu.hse.ru

В автостраховании актуальна проблема финансовых убытков, вызванных неверной классификацией клиентов с точки зрения возможной убыточности договора страхования с ними. Работа направлена на разработку скоринговой системы на основе методов машинного обучения и статистики, использующей данные портфеля крупной страховой компании (более 30000 наблюдений). Цель – выявление переменных, предсказывающих надёжность клиента, и сравнение различных методов: градиентного бустинга, случайного леса, логистической регрессии, нейросетей. Применялись методы предобработки (заполнение пропусков, кодирование категориальных признаков, нормализация). Для оценки моделей использовались точность, F1-мера, ROC-кривые, матрицы ошибок. Рассмотрены интерпретация моделей с помощью векторов Шепли и влияние новых методов кодирования признаков на качество и интерпретируемость. Анализ показал, что машинное обучение значительно повышает эффективность андеррайтинга, снижает риски и способствует лучшей селекции рисков и снижению стоимости полисов. Работа вносит вклад в развитие скоринговых систем, объединяющих современные подходы анализа данных, и предлагает практические инструменты для снижения убыточности, повышения финансовой устойчивости страховых компаний и роста доверия клиентов.

APPLICATION OF MACHINE LEARNING AND CLASSICAL STATISTICAL METHODS FOR BUILDING A SCORING MODEL IN AUTO INSURANCE

Y. N. Mironkina, D. I. Timofeev

In the field of auto insurance, one of the most pressing challenges is the problem of financial losses arising from the misclassification of clients in terms of the potential unprofitability of their insurance contracts. This study is devoted to the development of a scoring system based on machine learning and classical statistical methods, using portfolio data from a major insurance company comprising more than 30,000 observations. The primary objective is to identify variables that reliably predict client risk and to evaluate and compare a range of methods, including gradient boosting, random forest, logistic regression, and neural networks. Data preprocessing procedures included missing value imputation, categorical feature encoding, and normalization. Model performance was assessed using accuracy, F1-score, ROC curves, and confusion matrices. Furthermore, the study addresses model interpretability through Shapley values and examines the impact of advanced encoding techniques on both predictive power and transparency. The results demonstrate that machine learning significantly enhances underwriting efficiency, improves risk segmentation, and contributes to reducing policy costs. Overall, the study advances the development of effective scoring systems that integrate modern analytical approaches and provides practical tools for minimizing losses, strengthening the financial resilience of insurance companies, and fostering greater customer trust.

Введение. Одной из ключевых задач страховых компаний для предотвращения наступления больших убытков является оценка уровня надёжности клиента при заключении договора автострахования [7]. Один из наиболее популярных и широко используемых методов – модели автоматического скоринга, которые на основе определённых параметров могут определить «хороший» клиент или «плохой» с точки зрения принятия его риска на страхование.

Актуальность работы заключается в том, что в последнее время у страховщиков достаточно остро стоит вопрос о том, стоит ли принимать риск потенциального клиента или нет, так как неверная классификация клиентов и формирование вероятно высокорискового портфеля могут привести к серьёзным финансовым убыткам.

Для минимизации финансовых рисков страховщики используют различные методы, в том числе – пришедшие из банковской практики скоринговые модели, основанные на внутренних правилах и данных клиентов компании. Однако не все страховщики используют сильные и качественные методы, поэтому проблема финансовых потерь вследствие недостаточно тщательного андеррайтинга до сих пор остаётся актуальна. В последнее время растёт популярность методов машинного обучения таких как: случайный лес, градиентный бустинг, нейронные сети, так как они позволяют увидеть более сложные нелинейные зависимости в данных. Такие модели позволяют аппроксимировать зависимость между целевой переменной и такими переменными, как пол, возраст страхователя, водительский стаж, марка автомобиля, возраст автомобиля и другими факторами, характеризующими клиента и его автомобиль.

Для построения качественной скоринговой системы необходим большой набор данных с различными метриками клиентов. В качестве базы данных исследования используется страховой портфель из полисов нескольких крупных Российских страховых компаний из более чем 30000 наблюдений. Анализ набора данных с помощью методов машинного обучения и классических статистических моделей позволит подобрать наиболее значимые переменные, отражающие андеррайтинговый риск клиента и оценить их влияние.

В ходе исследования было рассмотрено несколько моделей машинного обучения и классических статистических методов. Более подробно остановимся на модели градиентного бустинга, которая показала наиболее выдающиеся результаты в решении поставленных задач. *Градиентный бустинг* — это самая мощная техника машинного обучения на табличных данных для решения задач классификации, регрессии и ранжирования, основанная на идее построения предиктивной модели в форме ансамбля слабых предсказательных моделей, обычно деревьев решений. Основная идея заключается в последовательном добавлении к ансамблю новых моделей, которые исправляют ошибки предыдущих моделей. Сам процесс итеративно повторяется до тех пор, пока не будет достигнута минимальная ошибка, то есть пока антиградиентная функция не спустится в минимум при заданных гиперпараметрах. Оптимизация проходит с помощью градиентного спуска.

Обучение моделей. В имеющемся признаковом пространстве необходимо

задать наиболее релевантную к оценке риска целевую переменную, которую решено было определить как относительный ущерб для каждого наблюдения, то есть отношение общего объема выплат к страховой сумме по договору, т.к. они весьма разнятся между договорами, и важно, именно какая часть из страховой суммы выплачена клиенту, ведь страховая сумма определяет и размер уплаченной им страховой премии. Таким образом, была построена бинарная целевая переменная как превышение относительного ущерба верхнего квартиля распределения относительных выплат по портфелю, т.е. следующим образом:

$$y_i = \begin{cases} 1 \text{ ("плохой")}, & \text{if } y_i > y_{3q}^{otn} \\ 0 \text{ ("хороший")}, & \text{if } y_i < y_{3q}^{otn} \end{cases}$$

Порогом разделения был выбран именно 3 квартиль после перебора всех квантилей алгоритмом с применением дисперсионного анализа, который показал наибольшее различие в дисперсиях между совокупностями, что говорит о том, что выбранная граница должна повысить потенциальную предсказательную способность моделей.

После проведения кластерного анализа (методом k-means) дисперсионным анализом можем оценить пространственное разделение признаков, важность переменных в разделении, а ранговая корреляция покажет согласие с квартильным разделением по относительной убыточности.

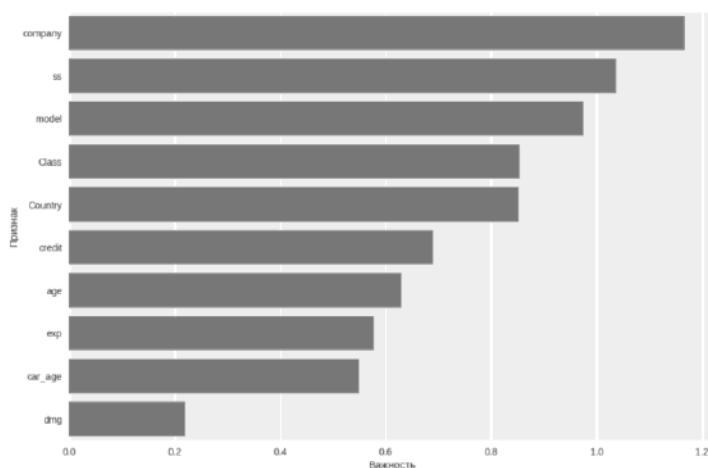


Рис. 1. Важность признаков при разделении на два кластера

В данном случае видно, что признаки страховая компания, страховая сумма и модель автомобиля вносят наибольший вклад, поскольку разница между их значениями в центроидах двух кластеров максимальна. Это указывает на то, что эти признаки являются основными дифференцирующими факторами между кластерами и впоследствии могут оказаться крайне важными при обучении моделей.

Для сравнения разделения на классы двумя методами была посчитана ранговая корреляция Спирмена – статистический метод, используемый для измерения силы и направления взаимосвязи между порядковыми переменными. Полученные результаты ($r \approx -0,2237$; $p - \text{value} \approx 1,2 \cdot 10^{-273}$) свидетельствуют о том, что между исходными метками классов и между присвоенными кластерами

присутствует статистически значимая средняя корреляция. Таким образом можно утверждать, что полученные кластеры интерпретируемы и отражают два класса с низким и высоким уровнем убыточности страхователей, а также сформировать гипотезы о том, что чем меньше страховая сумма транспортного средства, тем больше вероятность, что клиент окажется не убыточным для страховой компании; чем больше возраст клиента, тем ниже вероятность, что клиент окажется убыточным и т.п.

Теперь мы можем перейти к самому моделированию и обучению моделей. Как упоминалось выше, в работе сравниваются несколько типов моделей - классические статистические: логистическая регрессия и линейный дискриминантный анализ (LDA) и методы машинного обучения: гистограммный градиентный бустинг и случайный лес.

Самой главной метрикой в нашей работе является F1, она используется, когда полнота и точность имеют одинаково важное значение в работе. То есть F1 – это гармоническое среднее между *Precision* и *Recall*.

Таблица 1

Сравнение моделей по ключевым метрикам

Модель	Ассурасу	F1 (среднее)	F1 (ср. взв.)	Класс 0 (F1)	Класс 1 (F1)
Логистическая регрессия	0.8	0.68	0.78	0.88	0.48
Линейный дискриминантный анализ	0.8	0.67	0.77	0.88	0.46
Градиентный бустинг	0.96	0.95	0.96	0.98	0.93
Случайный лес	0.92	0.88	0.91	0.95	0.82

Сравнение показало существенные различия между классическими статистическими методами и современными алгоритмами машинного обучения. Логистическая регрессия и линейный дискриминантный анализ обеспечивают приемлемую точность (~0.80) и хорошо распознают «хороших» клиентов (F1=0.88), однако сильно теряют в качестве при классификации «плохих» клиентов (F1=0.46–0.48). Это делает их недостаточно надёжными для задач скоринга, где критично именно выявление рискованных клиентов. Случайный лес демонстрирует высокую общую точность (0.92) и хорошее качество по обоим классам (F1=0.95 для класса 0 и 0.82 для класса 1). Несмотря на снижение полноты по «плохим» клиентам, модель остаётся устойчивой и сбалансированной. Гистограммный градиентный бустинг показал наилучшие результаты (Ассурасу=0.96, F1=0.93–0.98, AUC=0.99), практически без потерь в идентификации как «хороших», так и «плохих» клиентов. Он является наиболее мощной моделью для построения скоринговой системы и рекомендован к использованию в качестве основной.

Перед тем как перейти к построению скоринговой карты, нам необходимо рассчитать WoE – веса влияния для переменных исследования [2]. *Weight of Evidence* – это понятие, широко используемое в статистическом анализе и моделировании рисков, особенно в банковской сфере для оценки кредитной способно-

сти. WoE измеряет силу предиктора в отношении целевой переменной. Оно рассчитывается для каждой категории независимой переменной и логарифмирует отношение доли хороших результатов к доле плохих.

$$WoE_j = \ln \left(\frac{\text{процент хороших}_j}{\text{процент плохих}_j} \right)$$

Information Value (информационный критерий) измеряет общую силу предиктора. Это агрегированная мера, которая суммирует взвешенные значения WoE по всем категориям, умноженные на разницу в процентах хороших и плохих.

$$IV_i = \sum (\text{Процент хороших}_i - \text{Процент плохих}_i) \cdot WoE_j$$

Эти метрики позволяют оценить вклад каждой категории в зависимую переменную. В качестве примера можно продемонстрировать значения для части признаков.

Таблица 2

WoE и IV значения для числовых признаков

	WoE	IV	«Плохой»	«Хороший»
Водительский стаж				
0-11	-1.7654	2.2531	16454	7871
12-23	-0.2499	0.0158	6568	17757
24-35	1.6854	0.2909	1153	23172
36-47	3.8755	0.8092	136	24189
48+	6.1603	1.3172	14	24311
Возраст				
18-30	-0.3503	0.0426	8926	15399
31-42	-1.0386	0.4442	11278	13047
43-55	0.4335	0.0314	3506	20819
56-67	2.3136	0.3981	555	23770
68+	5.4559	1.9919	60	24265

Из табл. WoE для числовых признаков (табл. 2) видно, что наибольший риск связан с водителями со стажем 0–11 лет (WoE = –1.7654), что подтверждает уязвимость начинающих и страховую практику. При стаже 24–35 лет риск снижается (WoE = 1.6854), а IV указывает на умеренное влияние этой группы. Наиболее высокие риски характерны для крайних возрастов: молодёжь 18–30 лет и особенно пожилые 68+ лет (WoE = 5.4559, IV = 1.9919).

Заключительным этапом нашей работы является построение скоринговой модели через масштабирование. Классическим масштабом для коэффициентов является шкала от 0 до 1000, где 0 – самый потенциально ненадёжный клиент, а 1000 – это идеально надёжный клиент, однако из-за того, что могут получиться отрицательные баллы или же, категория настолько сильно повышает уровень доверия к клиенту, что баллы могут уходить за 1000, то справедливо будет использовать динамическую шкалу.

Использование кооперативной теории игр для интерпретации моде-

лей и построение скоринговой системы. В работе предпринята попытка решить наиболее критичную проблему при использовании методов машинного обучения, а именно отсутствие неких коэффициентов модели, численных результатов, и вследствие этого - интерпретации результатов по примеру классических статистических методов. Для этого использованы методы кооперативной теории игр, а именно значения Шепли.

Так как модель градиентного бустинга не поддается интерпретации в прямом её понимании, то мы будем использовать не классическую формулу с коэффициентами из логистической регрессии, как в [3], а значения вектора Шепли вместо них.

$$\text{балл} = - \left(WoE_j \cdot Shap_i + \frac{\left| \sum_{i=1}^n \frac{Shap_j}{n} \right|}{n} \right) \cdot R + \frac{A}{n}$$

где $R = \frac{D}{\ln(2)}$, D – количество баллов, удваивающих шансы, $A = B - R \cdot \ln(C)$, $Shap$ – значение вектора Шепли для конкретной переменной, n – количество независимых регрессоров и WoE – значения веса влияния категории переменной. Значения Шепли помогают ответить на вопрос «Какое влияние каждый признак вносит в предсказание модели?» Это делается путём оценки предсказательной ценности каждого признака путем рассмотрения всех возможных комбинаций признаков. Таким образом, можно понять, как изменение одного признака влияет на изменение предсказания модели.

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)),$$

где $\phi_i(v)$ – значение Шепли для i -ого признака, N – общее количество признаков, S – подмножество признаков без i -ого признака, $v(S)$ – предсказание модели с подмножеством признаков S , $v(S \cup \{i\})$ – предсказание модели с подмножеством признаков S , включая i -й признак, $|S|$ – количество признаков в подмножестве S , $|N|$ – общее количество признаков.

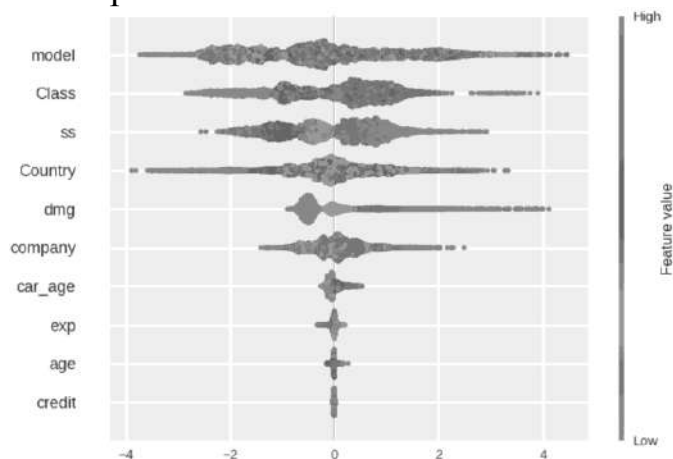


Рис. 2. Значения Шепли для Градиентного бустинга

На графике признаки упорядочены от самых важных сверху вниз, где размер и цвет точек показывают степень и направление воздействия признака на

результат. Модель и класс оказывают значительное воздействие на целевую переменную, показывая широкий разброс значений SHAP. Страховая сумма и страна также демонстрируют значимые вклады, но с меньшей вариативностью по сравнению с моделью и классом. Остальные переменные так же оказывают влияние, но менее значимое, так как обладают сложным распределением Шепли.

Для того, чтобы определить точные значения для каждого признака из всего пространства векторов Шепли, нужно посчитать средние абсолютные значения.

Таблица 3

Значения Шепли для Градиентного бустинга

Признак	Значение Шепли	exp(Shapley)
Модель	1.2144	3.3684
Класс	0.8863	2.4263
СС	0.7757	2.1721
Страна	0.6261	1.8704
Компания	0.3658	1.4417
Возраст авто	0.0969	1.1007
Опыт	0.0304	1.0309
Возраст	0.0166	1.0167
Кредит	0.0028	1.0028

Табл. 3 показывает значения Шепли для признаков модели градиентного бустинга. Наибольшее влияние оказывает **модель автомобиля** (exp=3.37), далее идут класс (2.43), страховая сумма (2.17) и страна (1.87). Наименьшее влияние имеют возраст, опыт и кредит. Экспоненцированные значения отражают, во сколько раз признак изменяет вероятность исхода относительно базового уровня.

Современные модели, включая градиентный бустинг, случайный лес и нейросети, чётко фиксируют различия между классами, что подчёркивает их способность выявлять сложные зависимости.

В качестве примера работы скоринговой системы можно взять двух случайных страхователей и посчитать для их договора скоринговые баллы.

Таблица 4

Демонстрация работы скоринговой системы для градиентного бустинга

Признак	Данные клиента 1	Данные клиента 2	Баллы 1	Баллы 2
Модель	Mitsubishi	Great Wall	-40,0294	-16,6025
Класс	SUV Medium	Large	268,4560	-73,9768
СС	3,4 млн	1,2 млн	147,2550	-226,9216
Страна	Япония	Китай	14,2947	22,0273
Компания	СК1	СК2	63,5615	36,6694
Возраст авто	4	2	-9,8249	197,913
Опыт	20	3	36,3779	-54,0418
Возраст	45	25	76,7956	76,7956
Кредит	Нет	Да	48,0319	53,8124
Сумма			852,5205	91,2957
Итог			«Хороший»	«Плохой»

Из табл. 4 видно, что скоринговая система оценивает страхователей по совокупности характеристик автомобиля и водителя. Автомобиль Mitsubishi SUV Medium стоимостью 3,4 млн из Японии получает более высокие баллы, чем Great Wall Large за 1,2 млн из Китая. Это отражает предпочтение системы к более дорогим и надёжным авто. Дополнительные баллы даёт больший возраст и опыт водителя Mitsubishi. Наличие кредита и выбор страховой компании также учитываются, но в меньшей степени. В итоге владелец Mitsubishi классифицируется как «Хороший» страхователь, а владелец Great Wall – как «Плохой». Порог перехода между классами для разных моделей находится в диапазоне 387–480 баллов, что подтверждает способность системы эффективно различать уровень риска.

Разработка скоринговых моделей на основе моделей продвинутого машинного обучения позволит страховым компаниям более точно и объективно оценивать страховые риски, связанные с каждым потенциальным клиентом и проводить более точную андеррайтинговую политику селекции рисков и назначения им адекватных страховых премий. Это, в свою очередь, будет способствовать минимизации финансовых убытков и повышению финансовой устойчивости и прибыли страховых компаний.

СПИСОК ЛИТЕРАТУРЫ

1. *Lars O. H., Petter E.* Explaining Deep Learning // Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data. *J. Risk Financial Manag.* 2023.
2. *Сорокин А. С.* К вопросу валидации модели логистической регрессии в кредитном скоринге // Вестник евразийской науки. 2014. № 2 (21).
3. *Сорокин А. С.* Построение скоринговых карт с использованием модели логистической регрессии // Вестник евразийской науки. 2014. № 2 (21).
4. *Swati T.* Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions // *American International Journal of Business Management (AIJBM)*. 2022.
5. *Zhirov V. K.* Neural network as a tool to solve credit scoring problems // *Journal of Physics: Conference Series*. 2021.
6. *Géron A.* Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow // O'Reilly Media. 2019. [Electronic resource]. URL: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/> (access date: 25.09.2025).
7. *Миронкина Ю. Н., Звездина Н. В., Скорик М. А., Иванова Л. В.* Актуарные расчеты: учебник и практикум для вузов / Москва: Изд-во Юрайт, 2024. 506 с.
8. *Wube H. D., Esubalew S. Z., Debelee T. G., Weldesellasie F. F.* Deep Learning and Machine Learning Techniques for Credit Scoring: A Review // *Communications in Computer and Information Science*. 2024. pp. 30-61.