http://psyberia.ru/zip/lingvonlp.zip (дата обращения: 10.05.2020).
3. Boonjing V., Pimchangthong D. Data mining for customers' positive reaction to advertising in social media // Proceedings of the Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems. 2017. V. 11. P. 945–948. DOI: 10.15439/2017F356
4. Brendan McMahan H., Holt G., Sculley D. Ad click prediction: A view from the trenches // KDD'13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, P. 1222–1230. DOI: 10.1145/2487575.2488200.
5. Griva A., Bardaki C., Pramatari K., Papakiriakopoulos D. Retail business analytics: Customer visit segmentation using market basket data // Expert Systems with Applications. 2018. V. 16. P. 10–16. DOI: 10.1016/j.eswa.2018.01.029.
6. Freund Y., Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of Computer and System Sciences. 1997. V. 55. P. 119–139. DOI: 10.1006/jcss.1997.1504.
7. Hu Y., Shin J., Tang Z. Incentive problems in performance-based online advertising pricing: Cost per click vs cost per action // Management Science. 2016. V. 62. P. 2022–2038. DOI: 10.1287/mnsc.2015.2223.
8. Richardson M., Dominowska E., Ragno R. Predicting Clicks: Estimating the Click-Through Rate for New Ads // WWW '07: Proceedings of the 16th international conference on World Wide Web. 2007. P. 521–530.DOI:10.1145/1242572.1242643.
9. Steinhaus H. Sur la division des corps materiels en parties // Bulletin de l'Académie Polonaise des Sciences. 1956. V. 4. P. 801–804.
10. Xinran H., Junfeng P., Ou J., et al. Practical lessons from predicting clicks on ads at Facebook //ADKDD'14: Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. 2014. P. 1–9. DOI: 10.1145/2648584.2648589.
11. Yang H., Zhu Y., He J. Local algorithm for user action prediction towards display ads //KDD'17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017. P. 2091–2099. DOI: 10.1145/3097983.3098089.
12. Импортданных: руководство для разработчиков. URL: https://developers.google.com/analytics/devguides/config/mgmt/v3/mgmtDataImport (дата обращения: 10.05.2020).
13. Начало работы с Google Аналитикой. URL: https://support.google.com/analytics/answer/1008015?hl=ru (дата обращения: 10.05.2020).
14. Google Analytics Spreadsheet Add-on. URL: https://developers.google.com/analytics/solutions/ google-analytics-spreadsheet-add-on (дата обращения: 10.05.2020).
15. Sales Conversion Optimization. URL: https://kaggle.com/loveall/clicks-conversion-tracking (дата обращения: 15.04.2020).
16. Wu M.-L., Chang C.-H. Aggregate two-way co-clustering of ads and user data for online advertisements // Journal of Information Science and Engineering. 2012. V. 28. P. 83–97.

УДК 004.89:368

ДВУХЭТАПНАЯ МОДЕЛЬ МАШИННОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ ДЕРЕВЬЕВ LIGHTGBM ДЛЯ ПРОГНОЗИРОВАНИЯ СТРАХОВЫХ РЕЗЕРВОВ

Соловьев Владимир Игоревич (VSoloviev@fa.ru)
Феклин Вадим Геннадьевич
Жукова Анастасия Сергеевна
Финансовый университет при Правительстве Российской Федерации

Рассматривается проблема прогнозирования страховых резервов на микроуровне без агрегирования данных для анализа с использованием деревьев решений с бустингом. Предлагается использование деревьев LightGBM в двухэтапной модели. На первом этапе определяется, есть ли претензии по контракту, которые возникли, но не были представлены (IBNR), или нет, а на втором этапе прогнозируется страховой резерв для случаев IBNR. Показано, что предложенная методика более эффективна, чем традиционные методы.

Ключевые слова: страхование, расчет страховых резервов, оценка размера отдельных страховых выплат, LightGBM.

Введение

Прогнозирование страховых резервов является крайне важной актуарной задачей, поскольку важно быть уверенными в достаточности резервов для покрытия убытков. При этом наиболее сложной проблемой является оценка претензий, которые возникли, но не были представлены (*Incurred But Not Reported, IBNR*) [1].

Традиционно для прогнозирования резервов используются метод цепной лестницы, метод Борнхьюттера – Фергюсона,

различные регрессионные модели. В последние годы популярными стали модели копул. Однако методы машинного обучения, в частности, деревья с градиентным бустингом, для оценки резервов не используются вовсе.

В данной работе предлагается подход к прогнозированию резервов, основанный на использовании деревьев решений с бустингом.

Дальше статья структурирована следующим образом. В следующем разделе приводится обзор литературы по прогнозированию страховых резервов. Затем описываются исходные данные и методология исследования, представляющая собой использование двухэтапной модели машинного обучения, которая на первом шаге с помощью модели классификации, основанной на алгоритме *LightGBM* [2], определяет, есть ли по данному договору *IBNR*-претензия, а на втором шаге с помощью модели регрессии, использующей *LightGBM*, прогнозирует величину резерва для случаев *IBNR*. При обсуждении итогов работы сравниваются результаты прогнозирования резервов с помощью предложенной техники, а также с помощью деревьев *AdaBoost* и гребневой регрессии.

Обзор литературы

Актуарная практика резервирования при страховании не жизни традиционно основана на совокупных данных о претензиях, структурированных в виде треугольников.

На данный момент существует несколько статистических методов, позволяющих оценить объем резервов. Одними из самых популярных методов, хорошо зарекомендовавших себя на практике, являются метод цепной лестницы [3, 4, 5, 6, 7] и метод Борнхьюттера – Фергюсона [7, 8]. Однако эти подходы эффективны лишь в том случае, когда анализируемые претензии имеют высокую вероятность и низкий уровень воздействия на размер резерва. Эти подходы использовались актуариями при оценке размера резервов в условиях ограниченной входной информации. Однако в настоящее время ограниченность информации больше не является основным препятствием, поэтому все больше и больше исследователей склоняются к тому, чтобы осуществлять резервирование на микроуровне на основе информации об отдельных выплатах [9, 10, 11].

В последние годы начали появляться работы, связанные с применением новых методов для прогнозирования страховых выплат и резервов, в том числе и методов машинного обучения.

Одним из подходов к оценке страховых выплат и резервов в медицинском страховании является подход, основанный на прогнозировании затрат граждан на здравоохранение. В работе [12] для прогнозирования затрат граждан Японии на медицинское обслуживание используются вариации лассо-регрессии. В работе [13] для оценки расходов на здравоохранение применяется двухкомпонентная модель: одна часть модели оценивает частоту определенных событий (посещение поликлиник, количество больниц и т. п.), а другая позволяет дать прогноз для расходов, связанных с каждым из этих событий.

Для непосредственной оценки страховых выплат применяются различные методы. Так, в работе [14] используется комбинация моделей понесенных и оплаченных убытков, позволившая существенно снизить ошибку прогноза по сравнению с традиционными подходами. Авторы работы [15] обнаружили существенную зависимость между количеством требований и их размером. Для оценки частоты подаваемых требований и их размера они применяли байесовский подход, а параметры оценивали методом Монте-Карло на основе цепей Маркова.

В работе [16] исследовано распределение годовых сумм требований с учетом нулевых требований. Для этого применялись связки дискретных и непрерывных копул, что позволило хорошо аппроксимировать непрерывные копулы, описывающие годовые суммы требований. В работе [17] построена двумерная модель копулы Клейтона для расчета требований *IBNR* и исследования связи между величиной требования и временем с момента, когда это требование возникло, до момента, когда был осуществлен платеж.

Для прогнозирования резервов по требованиям *IBNR* в работе [18] предложена полупараметрическая модель агрегированных требований с оценками по методу максимального правдоподобия.

Для решения многих задач успешно применяются обобщенные линейные регрессионные модели.

Так, в работе [19] такие модели используются для оценки предельного распределения требований, а в работе [20] с использованием обобщенных линейных регрессионных модели ослаблено традиционное в страховании условие независимости количества требований и их размера путем включения в модель рейтинговых факторов.

В работе [21] также применяется регрессионный подход к оценкам величины требований и их количества при условии, что требования являются зависимыми. Для учета этой зависимости используются двумерные копулы. Авторы показали, что явное включение данной зависимости в модель оказывает глубокое влияние на оценки и предложили алгоритм нахождения оптимального семейства копул.

В работе [22] предложено два подхода для описания зависимости количества требований и их размера. Первый подход основан на декомпозиции условной вероятности и рассматривает число требований как ковариату в регрессионной модели для среднего размера требований, второй подход использует модель копул для описания совместного распределения количества и размера требований. Для сравнения этих подходов был проведен имитационный эксперимент, который показал преимущество второго подхода. В частности, индекс Джини для модели копул оказался равен 42,14 против 38,64 для модели, основанной на первом подходе и 38,23 для традиционной модели Твиди.

В работе [23] предложено использовать регрессию на смешанных GGS-копулах для моделирования совокупных убытковв условиях, когда существует отрицательная зависимость между размером и частотой убытков. Это позволило существенно снизить ошибку прогноза совокупных потерь. Так, оценка совокупных потерь, полученная на основе GGS-копул, для рассматриваемых панельных данных за 2010 г. оказалась равной 4 362 626 при фактическом значении 4 159 322, в то время как оценка совокупных потерь, полученная с помощью модели независимости Твиди оказалась равной 6 147 354.

Авторы работы [24] предложили подход к моделированию периодических страховых требований в продольной установке с использованием копул для определения зависимости частоты и размера требований от времени, а также зависимости частоты и размера требований между собой.

В работе [25] для оценки величины требований применялись нечеткие модели. Построена модель нечеткой цепной лестницы (*FCL*) с использованием неопределенности *TFN* для построения прогноза. Также авторы предложили новый подход к оценке ошибки прогноза.

В работе [26] предложен метод оценки резервов, основанный на сочетании классического метода нечеткой регрессии Танаки Ишибучи и Нии со схемой резервирования заявок Шермана.

Применение для прогнозирования резервов регрессии на гауссовских процессах (*GP*) с несколькими ковариационными функциями для оценки будущих требований в работе [27] показало, что регрессионные *GP*-модели доминируют над моделями цепной лестницы и кривой роста с точки зрения точности прогноза, оцениваемой с помощью *RMSE*. Было предложено несколько вариантов *GP*-моделей и показано, что модель с квадратичной экспоненциальной ковариационной функцией стабильно хорошо работает по всем трем рассмотренным наборам данных.

В современной литературе для моделирования индивидуальных резервов в основном используются различные вариации лассо и гребневой регрессии, модели копул и нечеткой регрессии.

В то же время инструменты, которые используют всю доступную неагрегированную информацию, могут в итоге преодолеть проблемы, которые возникают при применении более традиционных подходов.

Но по разным причинам модели деревьев решений с бустингом не используются для прогнозирования страховых резервов. Представляется, что использование таких алгоритмов может существенно повысить качество оценивания резервов.

Данные

Для оценивания размера страховых выплат мы используем данные, которые включают период с 2004 по 2019 г., предоставленные авторам крупной страховой компанией из Центральной Европы.

Используются следующие признаки:

- *AgreementNo* – номер страхового договора клиента страховой компании;
- *AccidentDate* – дата наступления страхового случая;
- *ReportingDate* – дата заявления клиента страховой компании в результате наступления страхового случая;
- *PaymentDate* – дата выплаты страхового возмещения клиенту страховой компании;
- *LoB* – линия бизнеса;
- *Status* – статус убытка (*Paid*– выплата, *Reserve* – в рассмотрении);
- *SumInsured* – страховая сумма, равная максимально возможной сумме возмещения, указанной в договоре страхования;
- *ReserveAmount* – окончательная сумма выплаты/резерва по страховому случаю, переоцененной в результате внутреннего расследования компании.
- *ReportTime* – время, прошедшее с момента наступления страхового случая до момента обращения клиента в страховую компанию, выраженное в годах;
- *FinTime* – время, прошедшее с момента заявления клиента в страховую компанию о наступлении страхового случая до момента его окончательного урегулирования страховой компанией, то есть отказа или выплаты, выраженное в годах;
- *Label* – переменная, которая равна единице, если максимально возможная сумма страхового возмещения не равна окончательно рассчитанной сумме выплаты по страховому случаю, и нулю в противном случае;

- *Target* – разность между максимально возможной суммы страхового возмещения и окончательно рассчитанной суммы выплаты по страховому случаю.

Алгоритм

Для реализации решения поставленной задачи прогнозирования резервов были сформулированы две подзадачи:

- Определение, имеется ли по договору ситуация *IBNR*;
- Прогнозирование размера резерва для случая, когда договор находится в ситуации *IBNR* либо определение резерва равного страховой сумме в случае, если договор не является договором *IBNR*.

Для этого создаются синтетические признаки:

- *Label* – равна ли максимально возможная страховая сумма фактически оцененной сумме выплаты;
- *Target* –разница между максимально возможной страховой суммой и фактически оцененной сумме выплаты.

В модели заложены следующие предположения:

- поскольку дата оценки является датой заявления о страховом случае, в модели предполагается, что вся информация о претензии известна на эту дату, и поэтому она может использоваться для прогнозирования будущих платежей;
- страховая компания выплачивает страхователю сумму, как только окончательный размер выплаты установлен, таким образом генерируя серию денежных потоков. В модели учитывается один единый совокупный платеж по каждой претензии, подлежащей оплате на дату закрытия;
- время урегулирования считается дискретной величиной, выраженной в годах. Анализ доступной статистической информации показывает, что срок урегулирования претензии не превышает 6 лет;
- прогнозирование осуществляется для 2019 г. Модельные значения, полученные в результате классификации, указывающие на выплату по уже заявленным случаям в следующем 2020 г. и позже, будут относиться к резерву заявленных, но неурегулированных убытков, а не к резерву убытков.

Все суммы в расчетах представлены в Евро.

Объем обучающего набора данных с 2004 по 2018 г. составил 66 414 записей, объем тестового набора данных (2019 г.) составил 20 507 записей.

Результаты и обсуждение

В результате подбора гиперпараметров модели классификации для определения с помощью алгоритма *LightGBM*, имеется ли по договору ситуация *IBNR*, наилучшими оказались следующие:

- *Number of leaves:* 256;
- *Minimum leaf instances:* 50;
- *Learning rate:* 0.025;
- *Number of trees:* 500.

В результате подбора гиперпараметров модели прогнозирования на основе алгоритма *LightGBM* размера резерва для случая, когда договор находится в ситуации *IBNR*, наилучшими оказались следующие:

- *Number of leaves:* 256;
- *Minimum leaf instances:* 50;
- *Learning rate:* 0.025;
- *Number of trees:* 500.

*И*тоговая модель выдает ошибки на следующем уровне: *MAE* = 30.37, *MAPE* = 0.16.

Для сравнения, модель *LightGBM* без предварительной классификации на *IBNR* и не *IBNR*, показывает *MAE* = 31.32, *MAPE* = 0.18, регрессия на деревьях *AdaBoost* – *MAE* = 39.59, *MAPE* = 0.26, гребневая регрессия – MAE = 135.64, *MAPE* = 0.87.

На основе данного анализа можно сделать вывод, что использование деревьев решений с бустингом в алгоритме, в котором вначале определяется, имеется ли по договору ситуация *IBNR*, а затем для случая, когда договор находится в ситуации *IBNR*, прогнозируется размер резерва, а для случая, если договор не является договором *IBNR*, резерв определяется равным страховой сумме, является эффективнее традиционных способов.

Представленная в работе техника может быть использована в качестве метода оценки размера индивидуальных страховых резервов.

Литература

1. Jewell W.Predicting IBNYR events and delays. I. Continuous time// ASTIN Bulletin. 1989. V. 19. P. 25–56. DOI: 0.2143/AST.19.1.2014914.
2. Ke G., Meng Q., Finely T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y.LightGBM: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems.2017. V. 30. P. 1–9. URL:https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree (дата обращения: 26.05.2020).
3. Kaas R., Goovaerts M.J., Dhaene J., Denuit M.Modern Actuarial Risk Theory – Using R. NY: Springer, 2008. 382 p. URL:https://www.springer.com/gp/book/9783540709923
4. Wuthrich M.V., Merz M.Stochastic Claims Reserving Methods in Insurance. NY:Wiley, 2008. 424 p.

5. Zhang Y.A general multivariate chain ladder model // Insurance: Mathematics and Economics. 2010.V. 463.P. 588–599. DOI: 10.1016/j.insmatheco.2010.03.002.
6. Denuit M., Trufin J.Collective loss reserving with two types of claims in motor third party liability insurance // Journal of Computational and Applied Mathematics. 2018. V. 335, P. 168–184. DOI: 10.1016/j.cam.2017.11.044.
7. Martínez-Miranda M.D., Nielsen J.P. Verrall R.Double chain ladder and Bornhuetter – Ferguson// North American Actuarial Journal. 2013. V. 172. P. 101–113. DOI: 10.1080/10920277.2013.793158.
8. Hiabu M., Margraf C., Martínez-Miranda M.D., Nielsen J.P.Cash flow generalisations of non-life insurance expert systems estimating outstanding liabilities//Expert Systems with Applications. 2016. V. 451. P. 400–409. DOI: 10.1016/j.eswa.2015.09.021.
9. Antonio K., Denuit M., Pigeon M. Individual loss reserving with the multivariate skew normal framework //ASTIN Bulletin. 2013. V. 433. P. 398–428. DOI: 10.1017/asb.2013.20.
10. Jessen A.H., Samorodnitskiy G., Mikosch T.Prediction of outstanding payments in a Poisson cluster model //Scandinavian Actuarial Journal. 2011. V. 2011. P. 214–237. DOI: 10.1080/03461238.2010.481080.
11. Plat R., Antonio K. Micro-level stochastic loss reserving for general insurance // Scandinavian Actuarial Journal. 2014, V. 2014. P. 649–669. DOI: 10.1080/03461238.2012.755938.
12. Takeshima T., Keino S., Aoki R., Matsui T., Iwasaki K.PRM23: Development of Medical Cost Prediction Model Based on Statistical Machine Learning Using Health Insurance Claims Data // Value in Health. 2018. V. 21. Supplement 2. P. S97. DOI: 10.1016/j.jval.2018.07.738.
13. Frees E.W., Gao J., Rosenberg M.A. Predicting the frequency and amount of health care expenditures // North American Actuarial Journal. 2011. V. 153. P. 377–392. DOI: 10.1080/10920277.2011.10597626.
14. Taylor G., McGuire G., Sullivan J.Individual claim loss reserving conditioned by case estimates // Annals of Actuarial Science. 2008. V. 3. P. 215–256. DOI: 10.1017/S1748499500000518.
15. Gschlößl S., Czado C.Spatial modelling of claim frequency and claim size in non-life insurance // Scandinavian Actuarial Journal. 2007. V. 2007. P. 202–225. DOI: 10.1080/03461230701414764.
16. Erhardt V., Czado C.Modeling dependent yearly claim totals including zero claims in private health insurance // Scandinavian Actuarial Journal. 20122. V. 2012. P. 106–129. DOI: 10.1080/03461238.2010.489762.
17. Pettere G., Kollo T.Modelling claim size in time via copulas // Transactions of the 28th International Congress of Actuaries, Paris, France, 28 May – 2 June 2006. P. 1–10. URL: https://researchgate.net/publication/228883603_Modelling_claim_size_time_via_copulas (дата обращения: 26.05.2020).
18. Zhao X.B., Zhou X., Wang J.L.Semiparametric model for prediction of individual claim loss reserving // Insurance: Mathematics and Economics. 2009. V. 45. P. 1–8. DOI: 10.1016/j.insmatheco.2009.02.009.
19. Frees E.W., Wang P. Copula credibility for aggregate loss models // Insurance: Mathematics and Economics. 2006.V. 38. P. 360–373. DOI: 10.1016/j.insmatheco.2005.10.004.
20. Garrido J., Genest C., Schulz J. Generalized linear models for dependent frequency and severity of insurance claims // Insurance: Mathematics and Economics. 2016. V. 70. P. 205–215. DOI: 10.1016/j.insmatheco.2016.06.006.
21. Krämer N., Brechmann E.C., Silvestrini D., Czado C.Total loss estimation using copula-based regression models // Insurance: Mathematics and Economics. 2013. V. 53. P. 829–839. DOI: 10.1016/j.insmatheco.2013.09.003.
22. Shi P., Feng X., Ivantsova A.Dependent frequency–severity modeling of insurance claims // Insurance: Mathematics and Economics. 2015. V. 64. P. 417–428. DOI: 10.1016/j.insmatheco.2015.07.006.
23. Hua L. Tail negative dependence and its applications for aggregate loss modeling // Insurance: Mathematics and Economics. 2015. V. 61. P. 135–145. DOI: 10.1016/j.insmatheco.2015.01.001.
24. Lee G.Y., Shi P.A dependent frequency–severity approach to modeling longitudinal insurance claims // Insurance: Mathematics and Economics. 2019. V. 87. P. 115–129. DOI: 10.1016/j.insmatheco.2019.04.004.
25. Heberle J. Thomas A.Combining chain-ladder claims reserving with fuzzy numbers// Insurance: Mathematics and Economics. 2014. V. 55. P. 96–104. DOI: 10.1016/j.insmatheco.2014.01.002.
26. de Andrés Sánchez J.Calculating insurance claim reserves with fuzzy regression// Fuzzy Sets and Systems 15723, 3091–3108 2006. DOI: 10.1016/j.fss.2006.07.003.
27. Lally, N., Hartman, B. Estimating loss reserves using hierarchical Bayesian Gaussian process regression with input warping // Insurance: Mathematics and Economics. 2018. V. 82. P. 124-140. DOI: 10.1016/j.insmatheco.2018.06.008