# Non-life Insurance Reserve Prediction Using LightGBM Classification and Regression Models Ensemble

**Vladimir Soloviev and Vadim Feklin**

**Abstract** To predict insurance reserves at the micro-level without data aggregation, a two-stage machine learning model based on enhanced LightGBM decision trees is proposed. The first stage is a classification task: whether there are claims that have arisen but have not been submitted under the contract (IBNR). In the second stage, the insurance reserve is predicted using a regression model in the case of IBNR, or it is assumed to be equal to the sum insured otherwise. It is shown that the proposed method is more effective than traditional methods of insurance reserves forecasting.

**Keywords** Insurance · Reserve prediction · Individual claims · LightGBM

## 1 Introduction

Forecasting insurance reserves is a vital actuarial task since it is important to be sure that reserves are sufficient to cover losses. The most challenging problem is assessing claims that are incurred but not reported (IBNR) [1].

Traditionally, for forecasting insurance reserves, the chain ladder method, the Bornhuetter-Ferguson method, and various regression models. In recent years copula models have become popular.

However, machine learning methods, in particular, boosted gradient trees, are not used at all to assess reserves.

This chapter proposes an approach to forecasting insurance reserves based on LightGBM boosted decision trees [2].

The remainder of the chapter is structured as follows.

Section 2 reviews related works on forecasting insurance reserves.

V. Soloviev (✉) · V. Feklin
Financial University Under the Government of the Russian Federation, 4 Fourth Veshnyakovsky, Moscow 109456, Russia
e-mail: VSoloviev@fa.ru

V. Feklin
e-mail: VFeklin@fa.ru

A. G. Kravets et al. (eds.), *Cyber-Physical Systems: Intelligent Models and Algorithms*, Studies in Systems, Decision and Control 417,
https://doi.org/10.1007/978-3-030-95116-0_15

Section 3 describes the initial data and the research methodology, which is the use of a two-step machine learning model, which in the first step, using the LightGBM classification model, determines is there an IBNR claim under this contract, and in the second step, using the LightGBM regression model, predicts the reserve value for IBNR cases.

Section 4 discusses the outcome of the work. Using our technique for reserves forecasting, results are compared with using AdaBoost trees and ridge regression.

## 2 Literature Review

The actuarial practice of insurance reserving is traditionally based on a body of claims data structured in triangles.

There are several statistical methods for estimating insurance reserves. Classical techniques of using the chain ladder method [3–7] and the Bornhutter–Ferguson method [7, 8] have proven themselves in practice.

However, these techniques are effective only when the analyzed claims have a high probability and a low impact on the reserve volume.

This approach was used in assessing the reserve volumes in conditions of limited incoming information.

Currently, limited information is not a limitation. Therefore, more and more researchers are inclined to conduct studies at the micro-level based on individual pay-offs [9–11].

Recently, works have begun to appear related to applying new methods for forecasting insurance payments and reserves, including machine learning methods.

One of the approaches to assessing insurance payments and reserves related to health insurance is an approach based on forecasting the costs of citizens for health care.

In [12], Takeshima, Keino, Aoki, Matsui, and Iwasaki used variations of lasso regression to predict the costs of Japanese citizens for medical care. In [13], Frees, Gao, and Rosenberg, a two-component model is used to estimate health care costs: one part of the model estimated the frequency of certain events (visits to clinics, the number of hospitals, etc.), and the other predicted the costs associated with a particular of these events.

Various methods are used to assess insurance benefits directly. So, Taylor et al. [14] used a combination of models (paid and incurred models), which made it possible to reduce the prediction error compared to traditional approaches significantly.

Gschlößl and Czado [15] found a significant relationship between the number of requirements and their size. To assess the frequency of requirements and their size, they used the Bayesian approach, and the parameters were estimated using the Monte-Carlo method based on Markov chains.

Erhardt and Czado [16] studied the distribution of annual amounts of claims with zero requirements. To do this, they used bundles of discrete and continuous copulas, which allowed them to approximate continuous copulas that describe the annual sums of requirements.

Pettere and Kollo [17] used a two-dimensional Clayton copula model for calculating IBNR requirements and investigating the relationship between the requirement value and the time from when this requirement arose until the time the payment was made.

To forecast reserves for IBNR requirements, Zhao et al. [18] developed a semi-parametric model of aggregated requirements using the maximum likelihood method.

To solve many problems, generalized linear regression models are successfully used. So Frees and Wang [19] used them to estimate the limit of claims distribution, and Garrido et al. [20] weakened the traditional insurance condition for the independence of the number of claims and their size, including rating factors in the model, which are determined using generalized linear regression models.

Krämer et al. [21] also used a regression approach to estimating the magnitude of requirements and their quantity, provided that they are dependent. To account for this dependence, two-dimensional copulas are used. The authors showed that the explicit inclusion of this dependence in the model profoundly affects these estimates. They offered an algorithm for finding the optimal family of copulas.

Shi et al. [22] offered two approaches for describing the dependence of the number of requirements and their size. The first is based on the decomposition of conditional probability and considers the number of requirements as covariates in the regression model for the average size of requirements. The second uses the copula model to describe the joint distribution of the number and size of requirements. To compare these approaches, a simulation experiment was conducted that showed the advantage of the second approach. In particular, the Gini index for the copula model turned out to be 42.14 against 38.64 for the model based on the first approach and 38.23 for the traditional Tweedie compound Poisson model.

Hua [23] proposed the use of a mixed copula regression based on the GGS copula for modeling aggregate loss under conditions where there is a negative relationship between the loss frequency and its value. This allowed to significantly reduce the forecast error of the total losses. Thus, the total loss estimate based on the GGS copula for the panel data for 2010 turned out to be 4 362 626 with the actual value of 4 159 322, while the total loss estimate, which was based on the Tweedy independence model, turned out to be 6 147 354.

Lee and Shi [24] proposed an approach for periodic insurance claims modeling in a longitudinal installation using copulas to determine the dependence of the claims frequency on time, the dependence of the claims volume on time, and the relationship between the claims frequency and volume.

Heberle and Thomas [25] used fuzzy methods to estimate the requirements values. They built a fuzzy chain ladder (FCL) model, uses the TFN uncertainty to make predictions, and proposes a new approach to forecast prediction error.

de Andrés Sánchez [26] proposes a reserve estimation method based on a combination of Tanaki Ishibuchi and Nii fuzzy regression, with Sherman's reservation scheme.

Lally and Hartman [27] uses the Gaussian process regression with input warping and several covariance functions to estimate future claims to predict reserves. The authors showed that the Gaussian process regression models dominate the chain ladder and growth curve models in terms of accurate predictive accuracy as measured by RMSE. Several variants of Gaussian process models were also proposed, and it was shown that a model with a quadratic exponential covariance function works stably well for all data sets considered.

In contemporary literature, various variations of lasso and ridge regression, copula models, and fuzzy regression are mainly used to model individual reserves.

Machine learning tools that use all the available non-aggregated information can ultimately overcome the problems that arise with more traditional approaches.

But for various reasons, boosted decision tree models are not used to predict insurance reserves.

It seems that the use of such algorithms can significantly improve the quality of reserves estimation.

In this chapter, we consider the possibility of using LighntGBM boosted decision trees algorithms for insurance reserves forecasting.

# 3 Research Methods

## *3.1 Data*

The following features are used:

- AgreementNo—insurance contract number;
- AccidentDate—date of the insured event occurrence;
- ReportingDate—date of application of the client of the insurance company as a result of the insured event;
- PaymentDate—date of insurance compensation paid to the client by the insurance company;
- LoB—the line of business;
- Status—loss status (“Paid”–payout, “Reserve”–pending);
- SumInsured—the insurance amount equal to the maximum possible compensation amount specified in the insurance contract;
- ReserveAmount—the final amount of payment/reserve for an insured event revalued as a result of an internal investigation by the insurance company;
- ReportTime—time elapsed from the moment of the insured event occurrence until the moment the client contacted the insurance company, years;
- FinTime—time elapsed from the moment the client submitted an insured event statement to the insurance company until the final settlement by the insurance company, that is, the refusal of payment, years;

- Label—synthetic feature equal to 1 if the maximum possible compensation amount is not equal to the final calculated pay-off amount for an insured event, and zero otherwise;
- Target—the difference between the maximum possible compensation amount and the final calculated pay-off amount for an insured event. To estimate the insurance reserves, we use data that includes the period from 2004 to 2019 provided to the authors by a large insurance company from Central Europe.

## *3.2 Algorithm*

To implement the solution of the reserves forecasting problem, two subtasks were formulated:

- Determination of whether there is an IBNR situation under the contract;
- Forecasting the reserve volume for the case when the contract is in an IBNR situation or determining a reserve equal to the insurance amount if the contract is not an IBNR contract.

To do this, synthetic features are created:

- Label—is the maximum possible insurance amount equal to the actual estimated pay-off amount;
- Target—the difference between the maximum possible insurance amount and the actually estimated pay-off amount.

The model assumes the following assumptions:

1. Since the valuation date is the date of the report on the insured event, the model assumes that all information about the claim is known at that date, and therefore it can be used to predict future pay-offs;
2. The insurance company pays the policyholder the amount as soon as the final pay-off is established, thus generating a series of cash flows. The model takes into account one single aggregate payment for each claim payable at the closing date;
3. Settlement time is considered a discrete value expressed in years. Analysis of the available statistical data shows that the period for claim resolving does not exceed 6 years;
4. Forecasting is carried out for 2019. The model values obtained from the classification, indicating payment for already declared cases in the next 2020 and later, will relate to the reserve of declared but unresolved losses, not to the reserve of losses.

All amounts are calculated in Euros.

The training data set from 2004 to 2018 has 66,414 records, and the test dataset (2019) was 20 507 records.

## 4 Results and Discussion

As a result of the hyperparameters tuning for the LightGBM classification model, whether there is an IBNR situation under the contract, the following parameters were the best:

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

As a result of the hyperparameters tuning for the LightGBM regression model for reserve volume prediction in the case when the contract is in the IBNR situation, the following parameters were the best:

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

The final model produces, as the result, MAE = 30.37, MAPE = 0.16.

For comparison, the LightGBM model without preliminary IBNR/Non-IBMR classification shows MAE = 31.32, MAPE = 0.18, regression on AdaBoost trees–MAE = 39.59, MAPE = 0.26, ridge regression–MAE = 135.64, MAPE = 0.87.

We propose the use of LightGBM boosted decision trees in the algorithm, which first determines whether the contract has an IBNR situation, and then for the case when the contract is in the IBNR situation, another LightGBM regression model predicts the reserve volume. For this case, if the contract is not an IBNR contract, the reserve is determined to be equal to the insurance amount.

The technique proposed is more effective than traditional methods.

It also can be used as a method for assessing the size of individual insurance reserves.

## References

1. Jewell, W.: Predicting IBNYR events and delays: I Continuous time. ASTIN Bull. **19**, 25–56 (1989). 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. Adv. Neural Inf. Process. Syst. **30**, 1–9 (2017). https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree, Accessed 19 Apr 2020
3. Kaas, R., Goovaerts, M.J., Dhaene, J., Denuit, M.: Modern Actuarial Risk Theory Using R. Springer, New York (2008)
4. Wuthrich, M.V., Merz, M.: Stochastic Claims Reserving Methods in Insurance. Wiley, New York (2008)

5. Zhang, Y.: A general multivariate chain ladder model. Insur. Math. Econ. **46**(3), 588–599 (2010). https://doi.org/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. J. Comput. Appl. Math. **335**, 168–184 (2018). https://doi.org/10.1016/j.cam.2017.11.044
7. Martínez-Miranda, M.D., Nielsen, J.P., Verrall, R.: Double Chain Ladder and Bornhuetter-Ferguson. North Am. Actuar. J. **17**(2), 101–113 (2013). https://doi.org/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 Syst. Appl. **451**, 400–409 (2016). https://doi.org/10.1016/j.eswa.2015.09.021
9. Antonio, K., Denuit, M., Pigeon, M.: Individual loss reserving with the multivariate skew normal framework. ASTIN Bull. **43**(3), 398–428 (2013). https://doi.org/10.1017/asb.2013.20
10. Jessen, A.H., Samorodnitskiy, G., Mikosch, T.: Prediction of outstanding payments in a Poisson cluster model. Scand. Actuar. J. **2011**(3), 210–335 (2014). https://doi.org/10.1080/03461238.2010.481080
11. Plat, R., Antonio, K.: Micro-level stochastic loss reserving for general insurance. Scand. Actuar. J. **2014**(4), 648–670 (2014). https://doi.org/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 Health **21**(Supplement 2), s97 (2018). https://doi.org/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 Am. Actuar. J. **15**(3), 377–392 (2011). https://doi.org/10.1080/10920277.2011.10597626
14. Taylor, G., McGuire, G., Sullivan, J.: Individual claim loss reserving conditioned by case estimates. Ann. Actuar. Sci. **3**(1–2), 215–256 (2008). https://doi.org/10.1017/S1748499500000518
15. Gschlößl, S., Czado, C.: Spatial modelling of claim frequency and claim size in non-life insurance. Scand. Actuar. J. **2007**(3), 202–225 (2007). https://doi.org/10.1080/03461230701414764
16. Erhardt, V., Czado, C.: Modeling dependent yearly claim totals including zero claims in private health insurance. Scand. Actuar. J. **2012**(2), 106–129 (2012). https://doi.org/10.1080/03461238.2010.489762
17. Pettere, G., Kollo, T.: Modelling claim size in time via copulas. In: Transactions of 28th International Congress of Actuaries, Paris, France, 28 May–2 June 2006, 1–10 (2006). https://www.researchgate.net/publication/228883603_Modelling_claim_size_time_via_copulas. Accessed 19 Apr 2020
18. Zhao, X.B., Zhou, X., Wang, J.L.: Semiparametric model for prediction of individual claim loss reserving. Insur. Math. Econ. **45**(1), 1–8 (2009). https://doi.org/10.1016/j.insmatheco.2009.02.009
19. Frees, E.W., Wang, P.: Copula credibility for aggregate loss models. Insur. Math. Econ. **38**(2), 360–373 (2006). https://doi.org/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. Insur. Math. Econ. **70**, 205–215 (2016). https://doi.org/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. Insur. Math. Econ. **53**(3), 829–839 (2013). https://doi.org/10.1016/j.insmatheco.2013.09.003
22. Shi, P., Feng, X., Ivantsova, A.: Dependent frequency–severity modeling of insurance claims. Insur. Math. Econ. **64**, 417–428 (2015). https://doi.org/10.1016/j.insmatheco.2015.07.006
23. Hua, L.: Tail negative dependence and its applications for aggregate loss modeling. Insur. Math. Econ. **61**, 135–145 (2015). https://doi.org/10.1016/j.insmatheco.2015.01.001
24. Lee, G.Y., Shi, P.: A dependent frequency–severity approach to modeling longitudinal insurance claims. Insur. Math. Econ. **87**, 115–129 (2019). https://doi.org/10.1016/j.insmatheco.2019.04.004

25. Heberle, J., Thomas, A.: Combining chain-ladder claims reserving with fuzzy numbers. Insur. Math. Econ. **55**, 96–104 (2014). https://doi.org/10.1016/j.insmatheco.2014.01.002
26. de Andrés Sánchez, J.: Calculating insurance claim reserves with fuzzy regression. Fuzzy Sets Syst. **157**(23), 3091–3108 (2006). https://doi.org/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. Math. Econ. **82**, 124–140 (2018). https://doi.org/10.1016/j.insmatheco.2018.06.008