Review of Accounting Studies (2020) 25:1098–1134
https://doi.org/10.1007/s11142-020-09546-9

Machine learning improves accounting estimates: evidence from insurance payments

Kexing Ding[1,2] · Baruch Lev[3] · Xuan Peng[1,2] · Ting Sun[4] · Miklos A. Vasarhelyi[2]

Published online: 7 July 2020

Abstract

Managerial estimates are ubiquitous in accounting: most balance sheet and income statement items are based on estimates; some, such as the pension and employee stock options expenses, derive from multiple estimates. These estimates are affected by objective estimation errors as well as by managerial manipulation, thereby harming the reliability and relevance of financial reports. We show that machine learning can substantially improve managerial estimates. Specifically, using insurance companies' data on loss reserves (future customer claims) estimates and realizations, we document that the loss estimates generated by machine learning were superior to actual managerial estimates reported in financial statements in four out of five insurance lines examined. Our evidence suggests that machine learning techniques can be highly useful to managers and auditors in improving accounting estimates, thereby enhancing the usefulness of financial information to investors.

Keywords Machine learning · Accounting estimates

✉ Miklos A. Vasarhelyi
miklosv@business.rutgers.edu

Kexing Ding
dingkx@swufe.edu.cn

Baruch Lev
blev@stern.nyu.edu

Xuan Peng
pengxuan@swufe.edu.cn

Ting Sun
sunt@tcnj.edu

[1] Southwestern University of Finance and Economics, Newark, NJ, USA

[2] Rutgers the State University of New Jersey, New Brunswick, NJ, USA

[3] Stern School of Business, New York University, New York, NY, USA

[4] The College of New Jersey, Ewing Township, NJ, USA

1 Introduction

The PCAOB recently introduced a new standard for auditing accounting estimates, stating:

> Accounting estimates are an essential part of financial statements. Most companies' financial statements reflect accounts or amounts in disclosures that require estimation. Accounting estimates are pervasive in financial statements, often substantially affecting a company's financial position and results of operations… The evolution of financial reporting frameworks toward greater use of estimates includes expanded use of fair value measurements that need to be estimated (PCAOB 2018, p. 3).

Indeed, most financial statement items are based on subjective managerial estimates: fixed assets are presented net of depreciation—an estimate—and accounts receivables, net of estimated bad debts. Liabilities, like pensions and post-retirement benefits are estimates, and revenues from long-term projects or from contracts with future deliverables include estimates. Many expenses, such as the stock options or warranty expenses, also require estimates. Some items, like the pension expense, are based on multiple estimates, some of which, such as the expected gain on pension assets, are essentially guesses. Generally effective audit procedures, such as obtaining third-party confirmations of assets and liabilities, are inapplicable to estimates, which are opinions rather than facts. By and large, accounting estimates are very difficult to audit. There is therefore an urgent need to provide both managers and auditors an alternative or complementary generator of estimates.

Machine learning, quickly spreading into diverse areas of managerial practice, has the potential to provide such an independent estimates generator. When used as a predictive tool, machine learning techniques have applications in many domains. Researchers and practitioners have exploited the ability of machine learning to learn data patterns and have applied it to different contexts. As an antecedent to our study, there is a growing literature in accounting applying machine learning tools to predict the quality of accounting numbers. The earlier studies by Perols (2011) and Perols et al. (2017) are among the first in accounting to predict accounting fraud. Two recent studies, by Bao et al. (2020) and Bertomeu et al. (2020), used various accounting variables to improve the detection of ongoing irregularities. There is another strand of research that investigates the prediction of corporate bankruptcies or defaults using machine learning techniques. For example, Barboza et al. (2017) compared several machine learning models with traditional models and found that boosting, bagging, and random forest algorithms provide better prediction performance. The promising findings in this area encourage the development of new methods to enhance the performance of machine learning tools.

Overall, these studies largely complement ours: while they capture excessive managerial discretion or credit risk anomalies through analysis of financial statement numbers, our approach shows how machine learning can directly improve the estimate of an account balance, thus revealing the mechanisms through which machine learning may alleviate both intentional and unintentional errors. In particular, we demonstrate, using insurance companies' data on estimates and realizations of loss reserves (estimates of future claims related to current policies), that loss estimates derived from machine learning are, with a few exceptions, superior to the actual managerial loss estimates underlying financial reports. We thus establish, for the first time, the potential of machine learning to independently assess the reliability of estimates underlying financial reports, thereby improving the quality and usefulness of financial information. Furthermore, machine learning has the potential to substantially improve auditors' ability to evaluate accounting estimates, thereby enhancing the usefulness of financial information to investors.

The paper's order of discussion is as follows. Section 2 provides a background overview of the insurance claims loss estimation process. Section 3 and Section 4 discuss the machine learning algorithms used in this study and the application of machine learning to generate insurance companies' loss estimates, respectively. Section 5 presents our sample, while Section 6 provides the empirical results concerning the machine learning estimates, compared with managers' estimates. Section 7 presents additional analyses on estimation errors. Section 8 concludes.

2 Insurance claims loss estimation

Insurance companies provide protection to policyholders from certain risks that occur within a predefined period. While insurers receive the policy premium payments before or early during the period of coverage, the full costs of their activities—the total losses or claims by policyholders—usually remain unknown for several years after the coverage period ends. Insurance regulations require insurers to provide "management's best estimate" for these future claims in financial reports and to disclose the gradual settlement of loss claims in the following years. In other words, insurers match the payoffs directly to the year in which the initial estimate was made and the related insurance premium revenue recognized.

The unpaid component of the estimated future losses (claims) is the insurance loss reserve. The loss reserve is often the most significant component in property and casualty insurance firms' liabilities: on average, loss and loss adjustment expense reserves make up approximately two-thirds of an insurer's liabilities (Zhang and Browne 2013). Managers' loss estimation process is obviously subjective and requires considerable judgment, because not all claims for accidents that occur during a year are filed and settled by year-end. A substantial amount of losses incurred may be "Incurred But Not Reported Claims," in which case the insurance policyholders do not report the losses to insurance firms by the end of the current year but file the claims in later years. In addition, after the claims are filed, the final cash settlement may take years to complete. For example, injuries in a car accident may lead to several years of treatment and result in extended payments. Thus insurance firms must estimate the costs of claims filed during the year as well as claims that relate to the current year but will be filed in subsequent years.

Given the material impact of loss estimation on insurance firms' financial results and condition, auditors, investors, and regulators are naturally concerned with the quality of the estimates reported by managers. Studies have already established that managers may manipulate loss reserves to achieve various goals (e.g., Grace 1990; Petroni 1992; Weiss 1985; Beaver and McNichols 1998; Gaver and Paterson 2004; Browne et al.

2009).[1] Anderson (1971) analyzed the insurance industry from 1955 through 1964 and documented that insurers over-reserved losses heavily in early times but gradually reduced the degree of over-reserving to slightly under-reserving. However, Bierens and Bradford (2005) found that insurance firms from 1983 to 1993 tended to over-reserve. Grace and Leverty (2012) used a more recent sample (1989 to 2002) and found that firms generally overestimated losses but that there was considerable variation in insurers' practices.

3 Machine learning techniques

3.1 Machine learning algorithms

We compared four popular machine learning algorithms to predict insurance losses for five business lines and selected the algorithm with the best accuracy among those examined. The model-generated predictions were then compared to managers' estimates in financial reports. The four algorithms we used are linear regression, random forest, gradient boosting machine, and artificial neural network. We briefly discuss each machine learning model used.

3.1.1 Linear regression

Within the language of machine learning, linear regression is a supervised learning method that makes predictions based on the linear relationship between the numeric output and numeric input attributes (Friedman 2001; Bishop 2006). The learning process estimates the coefficients of the input attributes and aims to produce a prediction model that minimizes the mean squared error between the prediction and the true value. To select data attributes for linear regression, we used the M5 method: in each attempt, the attribute with the smallest standardized coefficient is selected and removed, and another regression estimation is performed (Witten et al. 2011). If there is an improvement in the model predictive accuracy in terms of the Akaike information criterion, the attribute is eliminated. This process is repeated until no improvement is observed.

Machine learning algorithms learn the hidden patterns of data in a way governed by a specific combination of hyper-parameters.[2] The determination of the optimal combination of hyper-parameters that produces a model with the most accurate prediction relies on trial and error. We used the Cartesian grid search to configure the optimal hyper-parameters for the model development in linear regression as well as the other three algorithms in this research. Cartesian grid search methodically builds and

[1] The literature has identified various managerial incentives may impact insurers' reserve manipulation, including tax deferral (Grace 1990; Petroni 1992; Nelson 2000), income smoothing (Anderson 1973; Smith 1980; Weiss 1985; Beaver, McNichols, and Nelson 2003), solvency and regulatory concerns (Forbes 1970; Petroni 1992; Nelson 2000; Gaver and Paterson 2004; Hoyt and McCullough 2010), and executive compensation incentives (Browne et al. 2009; Eckles and Halek 2010;). We provide more detailed discussions on managerial incentives in Section 7.

[2] Hyper-parameter is a parameter whose value is set before the learning process begins. Setting up the value of a hyper-parameter controls the process of defining the model.

evaluates a model through each possible combination of a specified subset of hyper-parameters. For linear regression, the hyper-parameters are the learning rate and the number of iterations. Specifically, the learning is an iterative process of continuously updating the values of model weights (parameters): the entire training data needs to be passed through and learned by the algorithm multiple times. Each time the data is passed through the algorithm, the weights will be updated. This is called one training epoch. In other words, an epoch is the complete cycle of an entire training data learned by a model. Because we cannot always pass through the entire data into the algorithm at once, the data is divided into batches. The number of iterations is the number of batches needed to complete one epoch. The learning rate is the extent to which the parameters are adjusted during the learning process. Lower learning rates require more training epochs, given the smaller adjustment made to the model weights each update, whereas higher learning rates result in rapid changes and require fewer training epochs. The grids of hyper-parameter values we have tried are listed in Table 1.

3.1.2 Random forest

The random forest algorithm is derived from decision trees, which is a machine learning technique that extracts information from data and displays it in a tree-like structure. A decision tree consists of three components: a node, a branch, and a leaf. Each root node of the tree denotes an input attribute; the tree splits into branches based on the input attributes, with each branch representing a decision. The end of the branch is called a leaf, and each leaf node leads to a prediction of the target value. Decision trees can be applied to either regression or classification problems. We employed regression trees because the target variable (actual losses) has continuous values. A single decision tree could have limited capabilities to learn the data, whereas random forest improves the accuracy of the decision trees with the ensemble technique (Breiman 2001, 2002). Specifically, the random forest algorithm first generates a "forest" of decision trees; each tree uses a subset of randomly selected attributes. It then aggregates over the trees to yield the most efficient predictor. For a regression problem, the output is the mean prediction of individual trees. In general, the higher the number of individual trees, the better the predictive performance of the random forest.[3] However, as adding too many trees can considerably slow the training, after a certain point, the benefit in prediction performance from using more trees will be lower than the cost of computation time for these additional trees (Ramon 2013).

The hyper-parameters for grid search are the number of trees, the maximum depth of the tree, and the minimum leaf size. The maximum value of tree depth represents the depth of each tree in the random forest; a deeper tree will have more branches from the node to the root node and capture more information about the data. However, as the tree depth increases, the model may suffer from the overfitting problem because it captures too many details (Brownlee 2016). The third hyper-parameter refers to the minimum number of observations for a leaf. Although a smaller leaf makes the model more prone to capturing noises in training data, a too-small leaf size may result in overfitting. On

[3] It is generally believed that adding more trees in random forest will not introduce overfitting because each individual tree has limited depth (Breiman 2001). The performance of the random forest tends to stay stable at a certain value after a certain number of trees.

Table 1 Tuning details for machine learning algorithms

Linear regression	
Learning rate	0.1, 0.01, 0.001, 0.0001
Number of iterations	10, 50, 100, 500, 1000
Random forest	
Number of trees	50, 100
Maximum depth of the tree	20, 30, 50
Minimum leaf size	1, 5, 10, 50
Gradient boosting machine	
Number of trees	50, 100
Maximum depth of the tree	5, 10, 20, 30, 50
Minimum leaf size	1, 5, 10, 50
Artificial neural networks	
Activation function	Rectifier, Tanh, Max out, Rectifier with Dropout, Tanh with Dropout, Max out with Dropout
Number of hidden layers	2,3, 4
Number of nodes	The first layer had a number of nodes equal to the number of independent variables, in each additional layer the number of nodes decreased by approximately 50%
Number of epochs	10, 50, 100
Learning rate	Examined at 100 possible learning rates, scaled from 0.0001 to 0.000001.

the other hand, if we choose a leaf size that is too large, the tree will stop growing after a few splits, resulting in poor predictive performance. Table 1 provides the grids of values that have been tried in the random forest.

3.1.3 Gradient boosting machine

Gradient boosting machine uses an ensemble technique termed boosting to train new prediction models with respect to the errors of the previous models, and convert weak prediction models to stronger ones (Schapire 1990). The objective of boosting is to minimize model errors by adding weak learners (i.e., regression trees). After adding new trees, the learning procedure subsequently corrects for errors made by previous trees and improves predictions to reduce the residuals in a gradient descent manner (Friedman 2001; Mason et al. 2000). After the number of trees reaches a limit point, adding more trees will no longer improve the prediction performance (Brownlee 2016). The grid search approach for gradient boosting machine was the same as that for the random forest, except that we started with five as the maximum tree depth to ensure that the learners were weak but can still be constructed in the gradient boosting machine. The grids of values tried are reported in Table 1.

3.1.4 Artificial neural networks

An artificial neural network consists of multiple layers of interconnected nodes between the input and output data. Each layer transforms its input data into a more abstract representation, which is then used as the input data by the next layer to produce representation. An artificial neural network has three types of layers: input, hidden, and output layers. The input layer receives the raw data of explanatory variables, and the number of nodes in the input layer equals the number of the explanatory variables. Next, the input layer is connected to hidden layers, which apply complex transformations to the incoming data and transmit the output to the next hidden layers. The output will be transmitted only if it exceeds a certain threshold determined by an activation function.[4] The data processing is performed through a system of weighted connections: the values entering a hidden node are multiplied by certain predetermined weights, and the weighted inputs are then added to produce a single output. There may be one or more hidden layers. A neural network is called deep when more than two hidden layers exist. The output of the final layer (called the output layer) represents the extracted high-level information from the raw data (Sun and Vasarhelyi 2017). We normalized all input variables to improve the model performance and used several different values of hyper-parameters for the grid search. The hyper-parameters include activation function, number of hidden layers, number of nodes, number of epochs, and learning rate. The details are provided in Table 1.

3.2 Data splitting and performance validation

We employed a training, validation, and testing approach in this study, with the last year in the sample as the holdout set. For each algorithm, we developed machine learning models using the fivefold cross-validation method and used the holdout set to evaluate the practical usefulness of the models. The fivefold cross-validation method is a widely used resampling procedure in machine learning to estimate a model's performance on a limited data sample. Specifically, the observations in the cross-validation sample are shuffled and randomly split into five equal groups. Each group is used, in turn, as a validation set, and the remaining four groups combined as a training set. A model is developed based on the training set and evaluated on the validation set using various evaluation metrics. The results are averaged to produce a unique evaluation of the model performance. Thus all observations are used for both training and validation, with each observation used only once for validation. We then applied the models developed from the cross-validation process to produce loss estimates for the holdout period that follows the training and validation period. Figure 1 illustrates the splits. Observations in the holdout set have not been used in the model development process. Therefore the holdout set was used only for evaluation rather than model selection purposes.

This design has two advantages in our setting. First, cross-validation generally results in a less biased and more robust estimate of the model accuracy than a simple training and testing split method (Brownlee 2018). Second, the training, validation, and testing approach alleviates the potential problems introduced by the inherent time-series order of the data.

[4] Activation functions are mathematical equations that determine the output of a neural network. The function is attached to each node in the neural network to determine whether the node should be activated. A node will be activated when the output exceeds a certain threshold based on the activation function. When the node is activated, the output in the node will be transmitted to the next layer in the neural network.

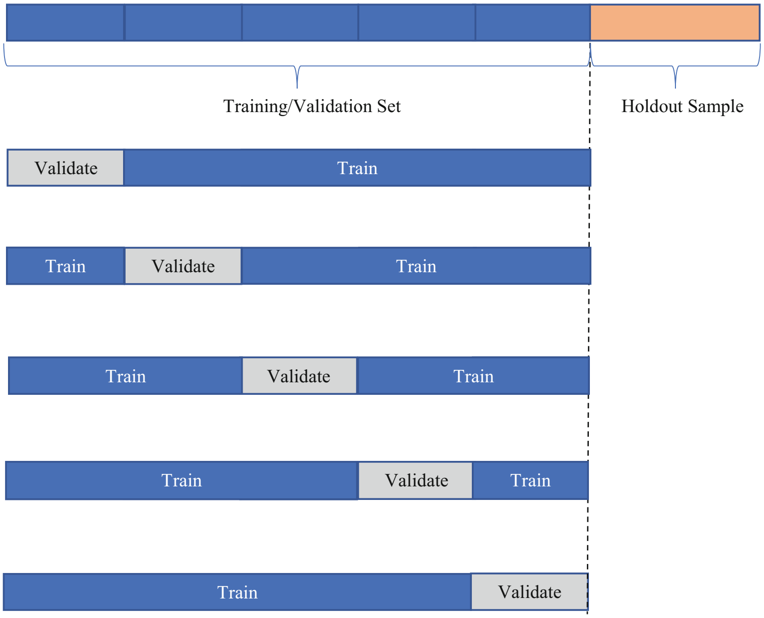

Fig. 1 Illustration of the training/validation/testing approach

3.3 Application to insurance loss estimation

We now explore the use of machine learning techniques in generating loss estimates for insurance companies. We conducted two sets of tests for each line of insurance (private passenger auto liability, commercial auto liability, etc.). The first set of tests did not include managers' loss estimates as an input attribute, keeping only variables that are not directly affected by managerial judgments. In other words, the variables used in these tests are based on verifiable facts, such as the number of outstanding claims, the number of loss claims closed with and without payment, etc. In the second set of tests, we added managers' initial estimates to the machine learning models. This design enables us to evaluate the performance of machine learning techniques on a standalone basis as well as when managers' inputs are incorporated into the algorithms. Moreover, because firms may experience different loss claim and payment patterns during the financial crisis period, we constructed three test periods: models developed from training samples of 1996–2005, 1996–2006, and 1996–2007 were applied to the holdout sets in 2006, 2007, and 2008, respectively.[5]

[5] For example, we first used data from 1996 to 2005 to develop machine learning models and applied the best-performing model to generate predictions for the year 2006. We have also tried dividing the sample in other ways, such as using the years 1996–2007 as the training sample and the final year 2008 as the testing set or using the entire data as the sample for cross-validation without holdout. The results were mostly consistent with our reported findings. We report the design of three holdout periods because it provides a more robust evaluation of model performance, given the size of our sample, while accounting for the inherent time-series of the data.

3.4 Insurance company data

U.S. insurance companies follow the statutory accounting principles (SAP)[6] to prepare statutory financial statements. Managers are required to provide the initial estimate for all losses (payment to insured) incurred in the current year (paid and expected to be paid in the future) as well as re-estimating the losses incurred in each of the previous nine years. In the statutory filings, insurance firms report the estimated total losses as "incurred losses" and the actual cumulative payments in each of the past 10 years (including the current year). The difference between the reported incurred losses and the cumulative paid losses for a given year is the reserve for future loss payment.

An advantage of the insurance industry data is that the extensive reporting requirements under SAP make it possible to match the actual payoffs to the insured directly to the year in which the initial estimate was made. Table 2 provides an example of this disclosure. Panel A shows the development of incurred losses for the National Lloyds Insurance Company. By the end of 2008, the estimates for the most recent 10 years (including the current year) are disclosed in Column 10. For example, in 2008, the current estimates for the losses incurred in the years 2007 and 2008 were \$24,334,000 and \$36,893,000, respectively. This is compared with the initial estimate for the losses incurred in 2007, which was \$24,226,000 (Column 9), meaning that Lloyds revised up the estimated losses incurred in 2007 by \$108,000 (\$24,334,000−\$24,226,000). Panel B reports the cumulative paid losses for each accident year, from 1999 to 2008, by the end of 2008. For example, by the end of 2008, the company has paid \$32,324,000 for the losses incurred during the year 2008, and \$23,585,000 for the losses incurred during 2007 (Column 10). In this example, the cumulative paid losses and the losses incurred for the year 1999 converge in 2006 and remain unchanged at \$4618,000 (Row 2), indicating that Lloyds likely paid off all the claims for accidents that occurred in 1999 by the end of 2006.

3.5 Dependent variable

Our dependent variable is the *ActualLosses*, which is the actual ultimate costs of insured events occurring in the year of coverage. It is the sum of losses already paid in the coverage year and the losses to be paid in the future. We chose total actual losses as our dependent variable because they can be directly compared to managers' "incurred losses" reported in annual reports. Studies that have examined insurance companies' loss reserve errors measured the "actual losses" as the cumulative losses paid during several subsequent years (Weiss 1985; Grace 1990; Petroni and Beasley 1996; Browne et al. 2009; Grace and Leverty 2012). We measured the *ActualLosses* for an accident year t as the 10-year cumulative payment of losses incurred in year t. This variable was extracted from the financial report in year $t + 9$, which is also the last time managers disclose the loss payment

[6] State laws and insurance regulations require that insurance companies operating in the United States and its territories prepare statutory financial statements in accordance with SAP. SAP is designed to assist state insurance departments in regulating insurance companies' solvency.

Table 2 Illustration of incurred losses and cumulative paid net losses reported in 2008 for National Lloyds Insurance Company

Years in which losses were incurred	Incurred net losses and defense and cost containment expenses reported at year end ($000 omitted)										Development	
	1	2	3	4	5	6	7	8	9	10	11	12
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	One year	Two year
1. Prior												
2. 1999	4583	4615	4614	4615	4615	4617	4617	4618	4618	4618		
3. 2000	XXX	4382	4450	4409	4407	4413	4411	4419	4422	4422		3
4. 2001	XXX	XXX	4845	4863	5012	5016	4909	4904	4905	4904	−1	
5. 2002	XXX	XXX	XXX	7463	7270	7064	7718	7169	7136	7147	11	−22
6. 2003	XXX	XXX	XXX	XXX	18,904	18,091	18,033	17,710	17,465	17,479	14	−231
7. 2004	XXX	XXX	XXX	XXX	XXX	18,201	15,408	15,301	14,754	14,727	−27	−574
8. 2005	XXX	XXX	XXX	XXX	XXX	XXX	24,097	20,611	23,627	24,554	927	3943
9. 2006	XXX	XXX	XXX	XXX	XXX	XXX	XXX	23,828	21,900	21,993	93	−1835
10. 2007	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	24,226	24,334	108	XXX
11. 2008	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	36,893	XXX	XXX
Years in which losses were incurred	Cumulative paid net losses and defense and cost containment expenses reported at year end ($000 omitted)										11 Number of claims closed with loss payment	12 Number of claims closed without loss payment
	1	2	3	4	5	6	7	8	9	10		
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008		
1. Prior	0											
2. 1999	3708	4548	4608	4614	4615	4617	4617	4618	4618	4618	1125	951
3. 2000	XXX	3486	4293	4393	4397	4403	4404	4404	4422	4422	1683	487
4. 2001	XXX	XXX	3736	4537	4716	4861	4903	4903	4904	4904	1154	777
5. 2002	XXX	XXX	XXX	5354	6884	6987	7045	7060	7111	7122	1753	721
6. 2003	XXX	XXX	XXX	XXX	15,926	17,281	17,141	17,326	17,424	17,438	5985	1432

Table 2 (continued)

Years in which losses were incurred	Incurred net losses and defense and cost containment expenses reported at year end ($000 omitted)										Development	
	1	2	3	4	5	6	7	8	9	10	11	12
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	One year	Two year
7. 2004	XXX	XXX	XXX	XXX	XXX	13,434	14,276	14,546	14,700	14,717	4828	1786
8. 2005	XXX	XXX	XXX	XXX	XXX	XXX	15,867	19,078	20,352	21,584	8211	1484
9. 2006	XXX	XXX	XXX	XXX	XXX	XXX	XXX	19,580	21,314	21,606	4930	1339
10. 2007	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	21,192	23,585	5217	2169
11. 2008	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	32,324	11,091	1724

for the initial accident year t.[7] For losses incurred in each business line during a given accident year, we generated only one estimate based on the information available at the end of this year, and the model estimates were compared to managers' initial predictions made in year t.

3.6 Independent variables

Our independent variables (predictors) consist of information already known at the time of estimation, that is, year t (no look-ahead bias). We included three sets of independent variables. First, operational variables (e.g., claims outstanding, premiums written, or premiums ceded to reinsurers) for each business line were obtained from Schedule P of the statutory filings. The second set of variables are company characteristics (e.g., total assets or state of operation) for the accident year. Finally, we added exogenous environmental variables (e.g., inflation or GDP growth) that reflect the macroeconomic factors that may influence the payment for the accident year. We use the lagged value of macroeconomic data because some information may be released after the insurers' financial statements are prepared. Definitions of all the independent variables are provided in Appendix Table 10.

3.7 Evaluation metrics

We used two metrics to compare estimates with actuals: the mean absolute error (MAE) and the root mean square error (RMSE). MAE is the average of the absolute differences between predictions and actual observations, and RMSE is the square root of the average of squared differences between predictions and actual observations. A smaller value of MAE or RMSE indicates higher prediction power. They are calculated as follows.

$$MAE = \frac{1}{n} \sum_{j=1}^{n} \left| TrueValue_j - ModelEstimate_j \right|. \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \left(TrueValue_j - ModelEstimate_j \right)^2} \tag{2}$$

We compared the machine learning loss predictions with managers' estimates to evaluate the machine's performance, using MAE or RMSE.

[7] SAP in general do not require firms to discount losses so that the estimates and actual payments are comparable. Some firms may choose to implicitly discount the future payments to reduce the reported reserve, especially for long-tail lines. However, as the inherent discount rate is not disclosed and discounting is not a standard procedure applied to all firms or all business lines, we calculated our dependent variable undiscounted. We also acknowledge the possibility that loss claims may take longer than 10 years to settle, in which case the cumulative payments in year$_{t+9}$ do not fully capture the actual losses. To alleviate the concern, we first discuss the payment patterns of each business line studied in the next section. Second, we used the manager estimates in year$_{t+9}$ as our proxy for actual losses instead, and the main inferences remained unchanged.

3.8 Our sample

We used the annual reports of US-based property and casualty insurance companies filed with the National Association of Insurance Commissioners (NAIC). The data were extracted from the SNL FIG website (S&P Global Market Intelligence platform), covering the period 1996 to 2017.

Property and casualty insurers offer a wide variety of insurance products that cover many different business lines, with each line having unique operating characteristics. The following procedures were separately performed on each business line to obtain the test samples.

a) For each business line, we first identified the insurance companies that had conducted business in this line. To be included in the sample, the firm must have started this line's business before 2008 and remained active until 2017, so that we can extract the ultimate payment (over 10 years) for at least one accident year.
b) Firm-years with missing or zero values for all operational variables were deleted from the sample. If total assets, total liabilities, net premiums written, or the direct premiums written were zero or negative, the firm-year was also excluded from the sample.
c) For each business line, we only kept observations that had positive total premiums and cumulative paid losses.

We focused on five business lines: (1) private passenger auto liability, (2) commercial auto liability, (3) workers' compensation, (4) commercial multi-peril, and (5) homeowner/farmowner. The five lines were selected from the 20 business lines identified by NAIC primarily because these lines had a sufficiently large number of observations remaining after the sample selection process (more than 400 insurers), indicating significant operations in the business lines. We excluded minor lines, such as special liability, products liability, and international, together making up less than 5% of industry loss reserves (A. M. Best 1994). In the final sample, we have a total of 32,939 line-firm-year observations for all five business lines combined, with each line's number of observations provided in Table 4.

Table 3 reports the payment patterns for each business line. The "tail" is an insurance expression that describes the time between the occurrence of the insured event (accident) and the final settlement of the claim. A longer tail usually implies higher uncertainty regarding the estimation of ultimate losses. As indicated in Table 3, for the private passenger auto liability line, 40.64% of the ultimate losses are paid off during the initial accident year, and 99.82% of the total payments throughout the 10 years are made during the first five years. The homeowner/farmowner business line (bottom line) has a payment pattern that differs from the private passenger auto liability line (top line). For these policies, the insurers pay off 72.62% of the ultimate losses after the first year, and 93.50% of the loss payments are made by the end of the second year. This suggests that the homeowner/farmowner business line has a relatively short tail, compared to the other four lines, consistent with prior studies that investigated the tail characteristics of insurance business lines (Nelson 2000). Overall, for all five business lines, the majority of payments are made during the first five years after the initial accident year.

Table 3 Cumulative payment percentage in the first five years for each business line

Business Line	Year 0	Year 1	Year 2	Year 3	Year 4	Year 5
Private Passenger Auto Liability	40.64%	72.44%	86.76%	94.20%	97.61%	99.06%
Commercial Auto Liability	25.03%	50.74%	70.90%	85.57%	93.88%	97.70%
Workers' Compensation	24.99%	56.11%	72.90%	83.20%	89.09%	93.03%
Commercial Multi-Peril	44.52%	69.22%	80.03%	88.58%	93.85%	97.11%
Homeowner/Farmowner	72.62%	93.50%	96.83%	98.58%	99.48%	99.82%

Table 4 reports summary statistics for firms that operate in each business line. The private passenger auto liability insurance is the largest business line in dollar terms, with total premiums written of $142 million on average, followed by the homeowner/ farmowner line, and workers' compensation line. The total premiums written in the private passenger auto liability line during 2015 amount to $199.37 billion, making up 34% of all the property and casualty insurance business.[8] The commercial auto liability and workers' compensation are usually provided by larger insurance companies, with average assets of $1.467 billion and $1.437 billion, respectively. In general, managers overestimate the ultimate losses when they report the future loss projections for the first time, except in the commercial auto liability line.

3.9 Empirical results

In this section, we first report the fivefold cross-validation machine learning results for each business line and then present the holdout test results. For each machine learning algorithm, we developed models with and without managers' loss estimates as an input attribute and reported the results separately. Because our objective is to evaluate whether machine learning models outperform managers in terms of predictive accuracy, we compared the model-generated estimates to managers' predictions in financial reports. For example, the Sentry Select Insurance Company initially estimated the total losses incurred during 2006 in the private passenger auto liability line to be $49,127,000, while the actual losses were $41,787,000.[9] In the cross-validation process, the random forest algorithm without managers' loss estimates generated an estimate of $43,657,000, and the prediction was $41,990,000 with managers' estimates included in the model. Thus both machine models generated estimates that were substantially closer to the actual losses than managers' prediction. Moreover, in the holdout tests, where we used the model developed from 1996 to 2006 to predict the losses incurred in 2007, the random forest models with and without managers' estimates predicted $38,953,000 and $40,996,000, respectively. The managers' estimate was $43,650,000 for the same year, while the actual losses turned out to be $39,791,000, suggesting that the machine learning predictions were again more accurate than the estimate provided by the firm.

[8] Insurance Information Institute, *The Insurance Fact Book 2017.*

[9] The example was drawn from the period 1996–2007, with 1996–2006 as the cross-validation set and 2007 the holdout set.

Table 4 Summary Statistics for the five insurance business lines

	Private Passenger Auto Liability			Commercial Auto Liability			Workers' Compensation	
Variables	Mean	Std. Dev	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev
ActualLosses	90,246	562,776	7239	17,764	52,139	6549	38,836	119,348
ManagerEstimate	93,183	580,029	7239	16,864	45,651	6549	44,987	136,526
Assets	1,255,516	5,130,236	7239	1,466,545	5,597,631	6549	1,436,801	4,278,044
Liabilities	772,594	2,891,796	7239	903,094	3,178,499	6549	953,594	2,854,313
NPW	429,264	1,775,834	7239	462,630	1,864,671	6549	405,332	1,124,403
DPW	373,990	1,411,322	7239	404,979	1,480,464	6549	370,433	952,908
NPE	420,376	1,755,638	7239	451,707	1,840,323	6549	394,541	1,091,865
PaidClaim	14,601	86,669	7239	1472	4354	6549	3861	13,142
UnpaidClaim	7360	49,956	7239	773	2764	6549	1792	8858
OutstClaim	5977	32,178	7239	686	1897	6549	1976	5647
ReportedClaim	27,935	162,349	7239	2929	8508	6549	7542	25,867
LinePremiums	142,228	812,935	7239	31,912	87,021	6549	77,485	223,037
PremiumCeded	12,980	45,691	7239	6891	29,892	6549	16,297	64,708
LinePayment	39,751	245,636	7239	4547	12,484	6549	9889	29,312
PaymentCeded	3451	13,043	7239	851	4254	6549	1625	7218
LineDCC	619	2963	7239	203	889	6549	740	3350
DCC Ceded	89	580	7239	52	430	6549	116	961
SSR	581	4158	7239	81	2058	6549	42	1357
PaidLoss	37,182	245,663	7239	3964	10,980	6549	9220	28,873
GDP	427,042	373,103	7239	360,382	302,911	6549	358,224	312,374
Inflation	5.25	1.46	7239	5.23	1.43	6549	5.24	1.43
GdpChg	4.82	2.37	7239	4.73	2.33	6549	4.80	2.42

Table 4 (continued)

	Workers' Compensation	Commercial Multi-Peril			Homeowner/Farmowner		
Variables	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev	Obs
	Workers' Compensation	Commercial Multi-Peril			Homeowner/Farmowner		
Variables	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev	Obs
ActualLosses	5118	24,430	81,218	6409	41,256	256,340	7624
ManagerEstimate	5118	24,818	79,099	6409	41,432	257,541	7624
Assets	5118	1,275,870	4,302,558	6409	1,087,610	3,776,991	7624
Liabilities	5118	834,212	2,842,832	6409	695,520	1,289,327	7624
NPW	5118	393,667	1,414,870	6409	365,591	1,311,209	7624
DPW	5118	349,002	1,042,452	6409	328,399	977,791	7624
NPE	5118	383,353	1,387,281	6409	356,919	1,289,327	7624
PaidClaim	5118	1385	6097	6409	8823	55,271	7624
UnpaidClaim	5118	744	2378	6409	2831	18,026	7624
OutstClaim	5118	546	1547	6409	1103	5053	7624
ReportedClaim	5118	2670	9384	6409	12,668	76,276	7624
LinePremiums	5118	48,453	143,724	6409	73,833	397,783	7624
PremiumCeded	5118	10,757	41,518	6409	11,700	47,174	7624
LinePayment	5118	10,827	33,913	6409	32,754	191,933	7624
PaymentCeded	5118	2035	9569	6409	4652	34,432	7624
LineDCC	5118	269	885	6409	462	2356	7624
DCC Ceded	5118	56	427	6409	64	388	7624
SSR	5118	155	4829	6409	108	802	7624
PaidLoss	5118	9293	30,133	6409	28,671	182,795	7624
GDP	5118	392,843	334,997	6409	427,566	372,256	7625

Table 4 (continued)

	Workers' Compensation	Commercial Multi-Peril			Homeowner/Farmowner		
Variables	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev	Obs
Inflation	5118	5.26	1.41	6409	5.28	1.47	7624
GdpChg	5118	4.8	2.34	6409	4.80	2.40	7624

All variables are defined in Appendix Table 10, and all variables are in thousands of dollars, except the *GdpChg* and *Inflation*, which are in percentages

Table 5 Cross-validation results

Business line	Training/Validation Sample	Obs	Managers' estimates		Machine learning without manager estimates				Machine learning with manager estimates			
			MAE	RMSE	MAE	RMSE	Accuracy edge		MAE	RMSE	Accuracy edge	
							(MAE)	(RMSE)			(MAE)	(RMSE)
Private Passenger Auto Liability					**Random forest**				**Random forest**			
	1996–2005	5949	9461	37,494	8213	34,687	13%	7%	7758	36,071	18%	4%
	1996–2006	6298	9793	38,266	7848	34,547	20%	10%	7220	30,305	26%	21%
	1996–2007	6602	9575	37,940	7869	35,047	18%	8%	6902	30,220	28%	20%
Commercial Auto Liability					**Random forest**				**Random forest**			
	1996–2005	5383	4209	18,562	3565	14,051	15%	24%	3446	13,555	18%	27%
	1996–2006	5661	4155	18,375	3520	13,881	15%	24%	3266	13,583	21%	26%
	1996–2007	5957	4338	19,175	3575	13,671	18%	29%	3322	13,121	23%	32%
Workers' Compensation					**Random forest**				**Random forest**			
	1996–2005	4183	11,547	43,652	7518	29,418	35%	33%	7144	28,629	38%	34%
	1996–2006	4398	12,360	44,187	7434	29,387	40%	33%	6988	26,888	43%	39%
	1996–2007	4645	13,214	47,541	7298	29,468	45%	38%	6861	26,574	48%	44%
Commercial Multi-Peril					**Random forest**				**Random forest**			
	1996–2005	5235	5737	27,615	5103	22,060	11%	20%	4854	22,062	15%	20%
	1996–2006	5457	5871	27,931	5151	23,404	12%	16%	4968	22,308	15%	20%
	1996–2007	5846	6017	28,349	4963	22,556	18%	20%	4534	21,265	25%	25%
Homeowner/Farmowner					**Linear regression**				**Linear regression**			
	1996–2005	6121	3905	16,789	5674	22,069	−45%	−31%	4402	16,359	−13%	3%
	1996–2006	6544	3878	16,611	5687	21,070	−47%	−27%	4203	16,201	−8%	2%
	1996–2007	6946	3962	16,826	5548	21,269	−40%	−26%	4321	16,674	−9%	1%

We used the MAE and RMSE metrics to evaluate model performance.[10] In the cross-validation process, we found that the random forest algorithm produced good predictions for four out of five lines and the linear regression model performed well for the fifth—homeowner/farmowner line. Thus we report the linear regression prediction results for the homeowner/farmowner line and present the random forest prediction results for the other four lines.[11] We also report the model *accuracy edge*, relative to manager estimates. The accuracy edge of a machine learning model is computed as managers' estimation MAE (RMSE) minus model estimation MAE (RMSE), divided by managers' estimation MAE (RMSE).

3.10 Fivefold cross-validation

As illustrated in Section 4, we first used three samples to train and validate the machine learning models: 1996–2005, 1996–2006, and 1996–2007 samples. We selected the models based on the cross-validation results and report their performance in Table 5. For each of the five business lines examined, we report the MAE and RMSE of managers' and models' estimates as well as the corresponding accuracy edges. For example, managers' estimation MAE (RMSE) for the private passenger auto liability line (line No. 1, first row) during the period 1996–2005 was 9461 (37,494). The random forest algorithm without managers' estimates yielded an MAE (RMSE) of 8213 (34,687), having an accuracy edge of 13% (7%) over managers' estimates. In the samples of 1996–2006 and 1996–2007, the random forest had smaller MAE and RMSE than those of managers, exhibiting higher predictive accuracy.

The results of the commercial auto liability line (line No. 2), the workers' compensation line (line No. 3), and the commercial multi-peril line (line No. 4) also suggest that the random forest algorithm achieves an accuracy edge over managers' estimates. Specifically, in line No. 2, the random forest estimates were, on average, 16% (26%) more accurate than managers' predictions in terms of MAE (RMSE). When predicting losses for line No. 3, the random forest model exhibited a considerable accuracy edge: on average, its MAE (RMSE) was 40% (35%) lower than managers' estimation. Turning to line No. 4, the average accuracy edge of the random forest model measured by the MAE (RMSE) was 14% (19%), relative to managers. After including *ManagerEstimate* as an additional attribute, the performance of random forest algorithm was further improved. The results are reported in the right-hand side of Table 5. In line No. 1, the average accuracy edge of random forest increased to 24% (15%) in MAE (RMSE). The MAE (RMSE) comparisons in line No. 2–4 also indicate an enhancement of model accuracy after incorporating managers' estimates. The accuracy edge in No. 3 was the most significant— 43% (39%) based on MAE (RMSE), on average.

In the homeowner/farmowner line (No. 5), however, managers outperformed the machine learning models. Consulting industry experts about this exception, a possible explanation is that the homeowner/farmowner line contains unique types of losses, such as catastrophes, property damage, and bodily injury. These loss categories are not

[10] The data from S&P Global Market Intelligence platform is presented in thousands of US dollars. Therefore results are expressed in thousands.

[11] The cross-validation test results of other machine learning models are provided in Appendix Table 10.

Table 6 List of influential variables in machine learning algorithms

Private Passenger Auto Liability Line		Commercial Auto Liability		Workers' Compensation		Commercial Multi-Peril		Homeowner/Farmowner	
W/o ManagerEstimate	With ManagerEstimate	W/o ManagerEstimate	With ManagerEstimate	W/o ManagerEstimate	With ManagerEstimate	W/o ManagerEstimate	With ManagerEstimate	W/o ManagerEstimate	With ManagerEstimate
LinePremiums	LinePremiums	PaidLoss	ManagerEstimate	PaidLoss	PaidLoss	PaidLoss	ManagerEstimate	PaidLoss	ManagerEstimate
DPW	DPW	LinePremiums	PaidLoss	LinePayment	ManagerEstimate	LinePremiums	LinePremiums	LinePremiums	PaidLoss
LinePayment	LinePayment	LinePayment	LinePayment	LinePremiums	LinePayment	LinePayment	LinePayment	LinePayment	LinePremiums
PaidLoss	ManagerEstimate	ReportedClaim	LinePremiums	LineDCC	LinePremiums	SSR	SSR	UnpaidClaim	UnpaidClaim
NPE	PaidLoss	OutstClaim	ReportedClaim	PaidClaim	LineDCC	LineDCC	LineDCC	DPW	LinePayment
UnpaidClaim	UnpaidClaim	PaidClaim	OutstClaim	OutstClaim	PaidClaim	NPW	NPW	PaidClaim	DPW
ReportedClaim	ReportedClaim	SSR	PaidClaim	Liability	OutstClaim	NPE	NPE	OutstClaim	ReportedClaim
PaidClaim	PaidClaim	UnpaidClaim	State	NPW	State	State	Assets	SSR	PaidClaim
OutstClaim	OutstClaim	State	SSR	State	ReportedClaim	OutstClaim	State	ReportedClaim	OutstClaim
DCC Ceded	DCC Ceded	Liability	Liability	ReportedClaim	Liability	ReportedClaim	OutstClaim	NPE	NPE
Liability	Liability	Assets	Assets	Assets	NPW	Assets	ReportedClaim	LineDCC	SSR
LineDCC	LineDCC	NPE	UnpaidClaim	DPW	NPE	Liability	Liability	Assets	LineDCC
SSR	NPW	NPW	LineDCC	NPE	DPW	UnpaidClaim	UnpaidClaim	NPW	NPW
NPW	Assets	PremiumsCeded	NPW	GDP	Pricegrowth	DPW	DPW	Liability	Liability
Assets	NPE	PaymentCeded	PremiumsCeded	PaymentCeded	GDP	PaymentCeded	PaidClaim	PaymentCeded	Assets

This table provides the lists of the top 15 influential variables in predicting ultimate losses with random forecast algorithm for the 1996–2007 training/validation period differentiated in firms' financial reports. Mixing up different loss types is problematic because different loss categories have unique payment patterns, which are known to managers but not available in the development of the machine learning prediction models. In addition, the homeowner/farmowner line has a relatively short tail, implying that the majority of losses (claims) have been reported and paid off during the first year (See Table 3). In this case, the total losses for most homeowner/farmowner accidents are already known to managers by year-end, making it challenging for machine learning models to outperform managers.

Collectively, our results suggest that machine learning models generate more accurate loss predictions than managers in most circumstances. Furthermore, we found that, in general, models incorporating *ManagerEstimate* have higher predictive accuracy, compared to the models without it.

To provide insights into the machine learning procedure, we present in Table 6 the 15 most influential variables identified by the random forest algorithm for each business line.[12] For example, the premiums written in the accident year (*LinePremiums*) was the most powerful predictor for the random forest algorithm to estimate losses in line No. 1. Overall, we observed that several predictors, such as *LinePremiums*, *LinePayment*, and *ManagerEstimate* (when added to the model), consistently play an important role in generating model predictions.

3.11 Holdout tests

In this section, we apply the models developed from the cross-validation sample to predict losses for the holdout period. Specifically, the model established from the 1996–2005 sample was used to predict losses in 2006, and the model from the 1996–2006 (1996–2007) sample was employed to predict the losses in 2007 (2008).[13]

The holdout tests examined the predictive accuracy of machine learning models on holdout sets—a more demanding prediction test. Table 7 Panel A reports the holdout results. Overall, the findings are consistent with the cross-validation results, indicating that machine learning models have superior predictive accuracy. When *ManagerEstimate* was not included in the model, the random forest algorithm generated more accurate estimates than managers in most cases for lines No. 1–4.[14] After we added *ManagerEstimate* to the model, its performance further improved. The analyses of line No. 5 (homeowner/farmowner line) indicate that managers outperformed linear regression when *ManagerEstimate* was not included in the model. After we added *ManagerEstimate* as an independent variable, the model's performance was enhanced: its prediction error measured by RMSE was slightly smaller than managers' in all three holdout samples.

[12] We thank an anonymous reviewer for suggesting this test. For brevity, we only report the important variables for models trained during the sample period 1996–2007 because the influential variables appear to be similar across different samples for each line.

[13] For completeness, we also show holdout prediction results for other models in Appendix Table 12. However, the results in appendix Table 12 are used for model selection.

[14] The model performance downturn in 2008 for line No. 4 might be caused by the economic turbulence during the financial crisis years, 2007–2009, which interrupted the patterns of insurance loss payments. The National Bureau of Economic Research (NBER) marks December 2007–June 2009 as a peak recession period.

Table 7 Holdout test results

Panel A Holdout test results

Business line	Holdout Sample	Obs	Managers' estimates		Machine learning without manager estimates				Machine learning with manager estimates			
			MAE	RMSE	MAE	RMSE	Accuracy edge		MAE	RMSE	Accuracy edge	
							(MAE)	(RMSE)			(MAE)	(RMSE)
Private Passenger Auto Liability					Random forest				Random forest			
	2006	670	9337	36,120	8434	35,271	10%	2%	8083	33,037	13%	9%
	2007	659	9435	36,221	8390	37,857	11%	−5%	7670	33,500	19%	8%
	2008	637	10,616	50,851	8507	41,440	20%	19%	8664	39,732	18%	22%
Commercial Auto Liability					Random forest				Random forest			
	2006	620	3852	17,287	3679	14,527	4%	16%	3475	14,468	10%	16%
	2007	609	3288	13,481	3056	10,413	7%	23%	2912	10,228	11%	24%
	2008	592	4219	19,361	3216	9638	24%	50%	3268	11,353	23%	41%
Workers' Compensation					Random forest				Random forest			
	2006	499	18,871	67,596	7675	28,922	59%	57%	6981	25,458	63%	62%
	2007	498	15,469	62,124	6477	24,704	58%	60%	6237	26,793	60%	57%
	2008	473	12,507	44,755	9081	36,131	27%	19%	8841	37,460	29%	16%
Commercial Multi-Peril					Random forest				Random forest			
	2006	582	6122	26,034	4543	14,943	26%	43%	3889	14,430	36%	45%
	2007	570	5725	27,361	4783	19,821	16%	28%	4567	19,633	20%	28%
	2008	563	6685	32,112	8400	39,315	−26%	−22%	7475	36,082	−12%	−12%
Homeowner/Farmowner					Linear regression				Linear regression			
	2006	697	2964	12,231	5219	14,457	−76%	−18%	3413	11,225	−15%	8%
	2007	692	3525	14,042	5202	15,297	−48%	−9%	4064	13,748	−15%	2%
	2008	678	5565	24,434	7968	23,580	−43%	3%	5628	20,881	−1%	15%

Panel B Bootstrap analyses: prediction error difference between machine learning models and managers

Table 7 (continued)

Business line	Holdout Sample	Difference in prediction error between managers and models					
		(Machine learning without manager estimates)			(Machine learning with manager estimates)		
		Mean	Standard error	Significance	Mean	Standard error	Significance
Private Passenger Auto Liability		**Random forest**			**Random forest**		
	2006	892	11.4	***	1241	12.0	***
	2007	1066	15.3	***	1757	12.3	***
	2008	2113	11.4	***	1947	8.9	***
Commercial Auto Liability		**Random forest**			**Random forest**		
	2006	173	5.3	***	381	3.7	***
	2007	228	3.3	***	381	3.3	***
	2008	998	5.7	***	952	5.0	***
Workers' Compensation		**Random forest**			**Random forest**		
	2006	11,208	23.7	***	11,905	23.9	***
	2007	9033	22.8	***	9214	22.1	***
	2008	3432	14.3	***	3665	14.1	***
Commercial Multi-Peril		**Random forest**			**Random forest**		
	2006	1574	7.1	***	2239	6.2	***
	2007	942	9.5	***	1166	9.4	***
	2008	−1715	10.6	***	−796	8.2	***
Homeowner/Farmowner		**Linear Regression**			**Linear Regression**		
	2006	−2257	3.5	***	−448	1.8	***
	2007	−1671	3.6	***	−534	2.1	***
	2008	−2406	4.6	***	−58	2.9	***

This panel reports the prediction error differences between managers and the machine learning model and standard errors in the bootstrap analysis. *** indicates the difference is significant at the 0.01 level

We used the bootstrap technique to examine the statistical significance of the difference in prediction performance between machine learning models and managers.[15] The bootstrap method uses an existing sample to create a large number of simulated samples that can be used to estimate the distribution of the performance difference. Specifically, we used the bootstrap to simulate 10,000 samples for each holdout sample and computed the differences between managers' and models' absolute prediction errors for each bootstrap sample. These differences varied across the simulated samples and formed a distribution. Table 7 Panel B reports the bootstrap mean of the prediction error differences and the standard errors. We test whether the differences are significant and report the significance levels. Overall, the bootstrap analyses provide supports to the conclusions drawn from the holdout tests.

Taken together, our results indicate the usefulness of machine learning models in estimating insurance losses, particularly for long-tail lines of insurance business. In addition, the random forest algorithm consistently shows superior predictive accuracy for long-tail business lines, and the linear regression performs better when the claims tail is short. Thus it is essential to understand the economics of a business line before applying a model to predict its losses. Furthermore, leveraging the information in managers' estimates enhances the prediction performance of machine learning models.

4 Additional analyses: Estimation errors

We now provide more detailed analyses of managers' and machine learning's estimation errors. Although machine learning process is usually presented as a black box and the models are challenging to interpret, in this section, we shed some light on the important question: what causes the advantage of machine learning models over managers?

Consistent with prior research, we defined managers' estimation error (*ManagerError*) as the reported loss estimate minus the actual loss, scaled by total assets. We focused on the signed estimation errors, instead of absolute errors, as in the previous section. Signed errors can give more insights into managers' reporting bias and are easier to interpret, whereas, in the previous section, our main objective is to evaluate the prediction accuracy. Similarly, the machine learning model estimation error (*ModelError*) is the model estimate minus the actual loss, scaled by total assets. Since the random forecast algorithm performed well in most business lines, we used the holdout prediction results generated by random forest models for the years 2006, 2007, and 2008 to calculate model estimation errors. Table 8 compares managers' estimates to models' estimates. On average, the model estimates were more accurate than manager estimates: the average absolute estimation error of machine learning models with (without) manager estimates as an input attribute was 0.0106 (0.0110), while the average manager estimation error was 0.0120. The difference is significant at 1% (5%) level. In addition, managers' signed estimation errors were larger than model errors on average, suggesting that managers tended to overstate insurance losses during our sample period. The results on managers' estimation errors are generally consistent with previous studies that investigated insurance loss estimation errors (e.g., Grace and

[15] We thank an anonymous reviewer for suggesting this test.

Table 8 Additional Analyses on Estimation Errors

	ManagerError		ModelError (Machine learning without manager estimates)					ModelError (Machine learning with manager estimates)					
	Mean	Median	Mean	Median	Mean Diff.	*p* value		Mean	Median	Mean Diff.	p value		Obs
Panel A: Comparing Manager Estimation Errors and Model Estimation Errors for Line Losses													
Unsigned Error	0.0120	0.0035	0.0110	0.0033	0.0010	0.02	**	0.0106	0.0032	0.0014	0.00	***	8247
Signed Error	0.0039	0.0015	0.0014	0.0006	0.0025	0.00	***	0.0015	0.0007	0.0010	0.00	***	8247
Panel B: Comparing Aggregated Manager Estimation Errors and Model Estimation Errors													
Unsigned Error	0.0286	0.0139	0.0257	0.0120	0.0029	0.01	***	0.0249	0.0118	0.0038	0.00	***	3126
Signed Error	0.0102	0.0088	0.0038	0.0031	0.0071	0.00	***	0.0040	0.0041	0.0063	0.00	***	3126

This panel presents differences in manager estimation errors and model estimation errors, both signed and unsigned, in the aggregate level. The model estimates were obtained from the hold-out sample predictions for the period 2006–2008. *ManagerError* is defined as the reported loss estimates minus true loss, divided by total assets. *ModelError* is the model estimate minus true loss, divided by total assets. We compared both signed estimation errors and unsigned estimation errors. The aggregated error is the sum of an insurer's estimation errors in all lines examined in the year. *P*-values reflect the differences in mean estimation errors. *, **, *** indicate significant difference in means at the 0.10, 0.05, and 0.01 levels, respectively, using a two-tailed t-test

Leverty 2012). To better understand the aggregate effect of estimation errors, we added the five lines' losses for the current accident year and compared the total to the corresponding true aggregate reserves.[16] The results in Panel B of Table 8 suggest that, at the aggregated level, managers' estimation errors were around 2.9% of the total assets on average, while machine learning models had an error percentage of 2.5%.

Overall, the results suggest that machine learning algorithms can provide more accurate insurance loss estimates than those reported by managers. Broadly speaking, three reasons may explain the model's edge. First, managers may be using low-quality information or fail to consider relevant information in their estimations. However, this is unlikely to be true, as all the input variables we included in the models were available before the initial accident year-end and the majority of variables were extracted from insurers' financial statements. The macroeconomic variables (e.g., *GDPLevel* and *Inflation*) from external sources had trivial influence in the model estimation process (see Table 6). Thus the larger errors in managers' estimates were not likely to be caused by inferior information quality. The second possible explanation is that insurance firms may apply erroneous estimation models. If so, we would expect managers' estimates to be of little value when incorporated in model estimation. However, we found that, in general, managers' estimates were among the top four most influential variables in predicting the ultimate losses (see Table 6), and incorporating managers' estimates into machine learning algorithms increased the predictive accuracy (see Table 5 and Table 7). This finding suggests that managers' procedure was overall effective in producing loss estimates. Third, various incentives may motivate managers to report biased estimates intentionally. It is well documented in the literature that various managerial incentives may lead to reporting bias. Reserving practice in the insurance industry provides significant flexibility and magnitude for managers to manipulate the numbers they report: management may increase or decrease the reported income by selecting a certain reserve level, which by nature is a subjective estimate of future cash payments. Moreover, the literature has found that auditors (Petroni and Beasley 1996) and regulators (Gaver and Paterson 2004) do not seem to detect insurers' earnings management effectively. Managers may actually adopt a conservative practice via over-reserving to maintain a positive reserve margin.[17] However, as long as other managerial incentives are also present in the reporting decision and information users cannot anticipate those biases perfectly, manipulation is likely to occur, reducing the value of financial reports (Fischer and Verrecchia 2000; Samuels et al. 2018). We started by examining the rationales identified by prior studies and then investigated whether the machine learning model estimation errors were affected by the incentives that cause managers' biases.

Because determining the taxable income involves loss estimates, over-reserving is more beneficial if more income is classified as a reserve. Thus insurers tend to overstate

[16] We thank an anonymous reviewer for suggesting the analysis. We acknowledge that this estimate is not equivalent to the loss reserves in insurers' financial statement, because the latter may include reserves for business lines other than the five examined as well as for prior years' losses. However, the other business lines are relatively minor, without sufficient data to support the model development, and estimating the loss reserves for prior years will impose much stricter requirements on our sample period, resulting in a sample size too small to conduct analyses. Therefore we added the five business lines to approximate the true reserve level.

[17] For more discussion on reporting conservatism in insurance firms, please see https://www.naic.org/sap_app_updates/documents/01-28.pdf and http://www.variancejournal.org/issues/01-01/120.pdf.

loss reserves to reduce current tax liabilities (Grace 1990). The decision variable here is the taxable income *before* reserves are determined, a higher value of which motivates managers to overstate reserves. Following this logic, research has captured insurers' tax incentive by adding the estimated reserves back to the reported level of taxable income to derive *TaxShield*, which takes a larger value if the insurer has a higher tax reduction incentive (e.g., Grace 1990). It is calculated as follows.[18]

$$TaxShield_t = \frac{Net\ Income_t + Loss\ Reserve_t}{Total\ Assets_t} \tag{3}$$

Second, we evaluated the impact of another well-recognized managerial incentive to distort the reported reserves: the income smoothing incentive (Weiss 1985; Grace 1990; Beaver et al. 2003). Firms, in general, are reluctant to report turbulent earnings that may indicate higher risks and discourage potential investors or bondholders from investing in the firm (e.g., Lambert 1984; Trueman and Titman 1988). Regulators are also concerned with a firm's income stability and include the change in surplus ratio in their solvency test (Grace 1990).[19] As mentioned above, the unique nature of insurance loss reserving practice provides managers great opportunities to smooth income. We followed prior research and used an insurer's average return on assets (ROA) during the past three years as our proxy for the income smoothing incentive (*Smooth*). The intuition is that following three previous good years, insurers tend to underestimate loss reserves in the current year to inflate earnings and continue the positive trend (Grace 1990). In addition to the *Smooth* variable, we used another indicator variable, *SmallProfit*. Beaver et al. (2003) have found that firms with small positive earnings are likely to have boosted reported income by understating loss reserves. Similar to Grace and Leverty (2012), we identified the firm-years in the bottom 5% of the positive earnings distribution. We expected these firms to have understated insurance loss reserves.

Financial distress also drives insurance firms to manage reserve estimates (Petroni 1992; Gaver and Paterson 2004). Regulators use the Insurance Regulatory Information System (IRIS) ratios to assess insurance firms' solvency. The NAIC provides a "usual range" for each ratio, and if the ratio falls out of the range, a ratio violation occurs. Regulatory intervention is involved when the number of ratio violations exceeds an acceptable threshold. Thus financially weak firms tend to under-reserve to appear adequate in capital and avoid regulatory scrutiny (Petroni 1992; Gaver and Paterson 2004). We set the variable *Violation* equal to 1 if the firm has at least one IRIS ratio violation and 0 otherwise. As managers may manipulate financial reports to avoid ratio violations (e.g., Petroni 1992; Gaver and Paterson 2004; Guttman and Marinovic 2018), we expected that firms with IRIS ratio violations would be more likely to understate losses because they are closer to triggering regulatory scrutiny than firms without violations. Also, insurers are required to maintain sufficient capital measured by the risk-based capital ratio. Regulators may take actions against the firm if the ratio

[18] We note that, while *Net Income* and *Loss Reserve* both involve managers' estimates in calculation, the error components cancel out when we added the two items together so that the total approximates the taxable income before reserves are determined.

[19] As net income is a primary component of surplus changes, regulators are implicitly encouraging smoother earnings (Grace 1990).

Table 9 The association between estimation errors and managerial incentives

		(1)			(2)					(3)				
	Pred. Sign	ManagerError			ModelError (Machine learning without manager estimates)					ModelError (Machine learning with manager estimates)				
Panel A: The association between estimation errors and managerial incentives														
		Coeff.	Std Err.		Coeff.	Std Err.		Diff.		Coeff.	Std Err.		Diff.	
TaxShield	+	0.055	0.0222	***	−0.048	0.0214		0.102	***	−0.022	0.0215		0.076	***
Smooth	–	0.013	0.0211		0.088	0.0509		−0.076		0.021	0.0208		−0.008	
SmallProfit	–	−0.005	0.0028	**	−0.002	0.0025		−0.002		−0.004	0.0028	*	0.000	
Insolvency	–	0.003	0.0031		0.002	0.0027		0.001		0.005	0.0037		−0.002	
Violation	–	−0.003	0.0011	***	0.001	0.0013		−0.003	***	−0.001	0.0010		−0.002	***
Liab	–	−0.001	0.0067		−0.005	0.0102		0.004		−0.003	0.0083		0.001	
Crisis	–	−0.002	0.0009	**	−0.001	0.0016		−0.001		−0.001	0.0010		−0.001	
Size	?	−0.001	0.0003	***	−0.002	0.0004	***	0.000		−0.001	0.0003	***	0.000	
SmallLoss	?	0.000	0.0046		−0.003	0.0047		0.002		0.000	0.0039		0.000	
Profit	+	−0.004	0.0018		0.001	0.0019		−0.004	**	−0.001	0.0016		−0.002	**
Loss	?	−0.002	0.0038		−0.004	0.0041		0.002		−0.004	0.0041		0.002	
LineSize	?	0.000	0.0001		0.000	0.0003		0.000		0.000	0.0002		0.000	
Reinsurance	?	0.003	0.0027		−0.003	0.0064		0.006		0.004	0.0032		−0.002	
Public	?	0.001	0.0015		0.003	0.0020		−0.002		0.001	0.0015		−0.001	
Mutual	?	0.002	0.0015		0.003	0.0014	**	−0.001		0.002	0.0014		0.000	
Group	?	−0.006	0.0108		0.003	0.0052		−0.009		−0.006	0.0028	**	0.000	
Line Fixed Effects		Yes			Yes					Yes				
Obs.		8247			8247					8247				
R-squared		2.02%			1.88%					0.89%				
Panel B: The association between aggregated estimation errors and managerial incentives														
		Coeff.	Std Err.		Coeff.	Std Err.		Diff.		Coeff.	Std Err.		Diff.	
TaxShield	+	0.095	0.0440	**	−0.092	0.0438		0.186	***	−0.044	0.0419		0.139	***
Smooth	–	0.018	0.0400		0.168	0.1010		−0.150		0.038	0.0388		−0.020	
SmallProfit	–	−0.012	0.0072	*	−0.005	0.0062		−0.006		−0.010	0.0071	*	−0.002	

Table 9 (continued)

		(1)			(2)					(3)				
	Pred. Sign	ManagerError			ModelError (Machine learning without manager estimates)					ModelError (Machine learning with manager estimates)				
Insolvency	–	0.000	0.0064		0.002	0.0057		−0.002		0.007	0.0078		−0.006	*
Violation	–	−0.010	0.0027	***	0.000	0.0027		−0.010	***	−0.003	0.0025		−0.007	***
Liab	–	0.013	0.0146		−0.003	0.0204		0.016		0.001	0.0179		0.012	
Crisis	–	−0.005	0.0020	***	−0.003	0.0037		−0.002		−0.003	0.0024		−0.002	**
Size	?	−0.002	0.0007	**	−0.004	0.0009	***	0.002	**	−0.003	0.0007	***	0.001	*
SmallLoss	?	−0.004	0.0106		−0.008	0.0099		0.005		−0.003	0.0090		−0.001	
Profit	+	−0.008	0.0044		0.001	0.0045		−0.009	**	−0.003	0.0040		−0.005	*
Loss	?	−0.005	0.0088		−0.009	0.0096		0.004		−0.009	0.0094		0.004	
Reinsurance	?	−0.002	0.0029		−0.016	0.0109		0.014		−0.001	0.0034		−0.001	
Public	?	0.002	0.0040		0.008	0.0054		−0.007		0.003	0.0039		−0.002	
Mutual	?	0.006	0.0039		0.009	0.0035	**	−0.002		0.005	0.0036		0.001	
Group	?	−0.014	0.0225		0.003	0.0074		−0.017		−0.012	0.0072	*	−0.002	
Intercept	?	0.022	0.0098	**	0.053	0.0137	***	−0.031		0.042	0.0093	***	−0.019	
Obs		3126		3126						3126				
R-squared		2.18%		3.34%						1.21%				

This panel reports regression analysis results of the aggregate signed estimation errors by managers and machine learning models. The regressions are estimated based on 3126 firm-year observations. The model estimates are the hold-out sample prediction results based on random forests models for 2006–2008. P-values are based on one-sided tests for coefficients with predicted signs and two-sided otherwise. Firm-clustering adjusted standard errors are used to calculate p-values. Column (1) presents the regression results for the full sample of manager estimates during 2006–2008 that can be directly compared to model estimates. Columns (2) and (3) show the regression results when the dependent variable is *ModelError*, without and with managers' estimates as model input attribute. *Taxshield* is calculated as the sum of net income and total reserves, divided by total assets. *Smooth* is defined as the average return on assets over the previous three years. *SmallProfit* is an indicator variable which equals 1 if the earnings fall in the bottom 5% of the positive earnings distribution. Insolvency is set to 1 if the risk capital ratio is smaller than 2 and 0 otherwise. Violation is an indicator variable that is equal to 1 if the insurer has more than 2 IRIS ratio violations and 0 otherwise. Other variable definitions are provided in Appendix Table 10. The table also presents the coefficient differences between manager estimation errors and model estimation errors, using two-sided t tests. *, **, *** indicate statistical significance at the 0.10, 0.05, and 0.01 levels, respectively is less than two. We therefore included another indicator variable, *Insolvency*, which equals one if the ratio is smaller than two and 0 otherwise. The ratio is defined as the total adjusted capital divided by the authorized control-level risk-based capital, a hypothetical minimum capital level determined by the risk-based capital formula.

We also included a series of firm and business line characteristics as control variables. Detailed control variable definitions are provided in Appendix Table 10. The following regression model was used to examine the incentives related to manager estimation errors.

$$\begin{aligned} ManagerError = \alpha_0 &+ \alpha_1 TaxShield + \alpha_2 Smooth + \alpha_3 SmallProfit \\ &+ \alpha_4 Insolvency + \alpha_5 Violation + \alpha_6 Liab + \alpha_7 Crisis + \alpha_8 Size \\ &+ \alpha_9 SmallLoss + \alpha_{10} Profit + \alpha_{11} Loss + \alpha_{12} Linesize \\ &+ \alpha_{13} Reinsurance + \alpha_{14} Public + \alpha_{15} Mutual + \alpha_{16} Group \\ &+ LineFixedEffects + \epsilon. \end{aligned} \tag{4}$$

The regression results are reported in column (1) of Table 9. Consistent with prior research, we found that firms with a stronger tax reduction incentive were more likely to over-reserve, as indicated by the significant and positive coefficient of *TaxShield* (*Coeff.* = 0.055, $p < 0.01$). The significantly negative values for the coefficients of *Violation* (*Coeff.* = −0.003, $p < 0.01$) and *SmallProfit* (*Coeff.* = −0.005, $p < 0.05$) suggest that financially weak firms and firms with income smoothing incentives tended to report underestimated insurance losses. The findings support the argument that managerial incentives, including tax reduction, income smoothing, and financial strength concerns, affect insurance firms' reserve levels. We also found that the coefficient for *Crisis* is significantly negative, indicating that, during the financial crisis period, managers were more likely to underreport loss estimates, presumably as a response to the negative macroeconomic shock.

We next examine whether our machine model estimates are less affected by managers' incentives, as one would expect. Therefore we re-run eq. (3) but replaced the dependent variable with *ModelError*. If the model estimates were less affected by the incentive biases than manager estimates, we would expect the coefficients of the incentive variables in the model regressions to be insignificant. Column (2) in Table 9 Panel A reports the regression results of the models without managers' estimates as an input variable. The results indeed indicate that the incentives that drive managerial estimation biases did not affect the model estimates, as none of the incentive variables were statistically significant. However, when we used the models including manager estimates, the coefficient of *SmallProfit* became marginally significant at 10% level, suggesting that incorporating managers' estimates might also bring their biases into the models. The regression results for the aggregated estimation errors are reported in Panel B of Table 9. We found that the aggregated manager estimation errors were significantly influenced by various managerial incentives, which did not seem to impact the aggregated model estimates: the coefficients of the incentive variables are mostly insignificant, except for the *SmallProfit* when we incorporated managers' estimates into the models.

Overall, the results indicate that the influence of managerial incentives is hardly present in the model estimation, which explains, in part, the model's superior performance.[20] Because reserving practice provides managers ample discretion in manipulation with relatively low costs, information users will find it difficult to efficiently evaluate the relevance of the reports, due to the noise in reporting (Fischer and Verrecchia 2000). Machine learning techniques discussed in this study provides a potential solution to help improve the quality of financial statements by providing estimates that are less affected by managerial biases.

5 Conclusion

Managerial subjective estimates are endemic to financial information, and their frequency and impact are constantly increasing, mainly by the expansion of fair value accounting by standard setters. The adverse impact of managerial estimation errors, both intentional and unintentional, on the quality of financial information is largely unknown and unresearched, but it is likely high. Undoubtedly, improvement in the quality and reliability of accounting estimates will substantially enhance the relevance and usefulness of financial information.

Accounting estimates generated by machine learning are potentially superior to managerial estimates because they may use the archival (training) data more consistently and systematically than managers. On the other hand, managers may include in their estimates (forecasts) forward-looking information (e.g., on expected inflation or the state of the economy) that machines obviously ignore. Accordingly, we assess the superiority of machines over humans in generating accounting estimates.

Our results, based on a large set of insurance companies' loss (future claim payments) estimates, revisions, and realizations, indicate that, with one exception (homeowner/farmowner insurance), loss estimates generated by machine learning are more accurate than managers' actual estimates underlying financial reports. This is a surprising and very encouraging finding, given the urgency to improve accounting estimates. At this early stage of applying machine learning to estimating accounting numbers, we don't know how generalizable our findings are. More research is needed to establish and generalize the use of machine learning for other types of accounting estimates, such as bad debts and warranty reserves.

Accounting estimates generated by machine learning may have multiple uses in practice. They can be used by managers and auditors as benchmarks against which managers' estimates will be compared, with large deviations suggesting a reexamination of managers' estimates. Alternatively, machine learning could be used to generate managers' estimates in the first place, enhancing the reliability (no manipulation) and consistency of accounting estimates. In any case, the potential of machine learning, whose use is fast expanding in many other fields, to improve financial information should be further researched.

[20] We have re-estimated the regressions excluding the workers' compensation line and using linear regression algorithm predictions for the homeowner/farmowner line, and the empirical inferences remain unchanged.

Acknowledgements The authors would like to thank Richard Sloan (editor), Patricia Dechow, Stephen Ryan, Jeremy Bertomeu (our discussant at the 2019 RAST Annual Conference), Russell Lundholm, Richard Crowley, Ganapathi Narayanamoorthy, Alexander Kogan, Michael Alles, two anonymous referees, and conference participants at RAST 2019 Annual Conference for helpful comments. All authors contributed equally to this work. Kexing Ding and Xuan Peng are thankful to Rutgers Continuous Auditing Research Lab for financial support.

Appendix

Table 10 Variable definition

Insurance Claims Loss Prediction	
ActualLosses	The cumulative payment on all events for a given accident year cumulative in the next ten years (including the accident year)
ManagerEstimate	Managers' initial estimates of the losses incurred during the accident year
Business Line Operational Variables	
Outstclaim	Cumulative claims outstanding for the current accident year
Reportedclaim	Cumulative reported claims for the current accident year
PaidClaim	Number of loss claims closed with payment
UnpaidClaim	Number of loss claims closed without payment
LinePremiums	Premiums written in the current accident year on the business line
PremiumsCeded	Premiums ceded to reinsurers
LinePayment	Total payments for the current accident year
PaymentCeded	Payments ceded to reinsurers
LineDCC	Defense and cost containment payments direct and assumed
DCC Ceded	Defense and cost containment payments ceded
SSR	Salvage and subrogation received
PaidLoss	Total losses paid for the current accident year
Company Characteristics	
State	Categorical variable that represents the state to which the insurer files the statutory filing.
Assets	Total assets at the end of the accident year.
Liabilities	Total liabilities at the end of the accident year.
DPW	Direct premiums written during the accident year.
NPW	Total net premiums written during the accident year.
NPE	Total net premiums earned during the accident year.
External Environment Variables	
Inflation	Personal consumption expenditure growth at state level in the previous year
GDP	GDP level at state level in the previous year
GdpChg	GDP growth at state level in the previous year
Estimation Error Analyses	
ManagerError	Calculated as the managers' initial estimation for the losses in year t minus the true losses, divided by the total assets.
ModelError	Calculated as the model estimation (model without managers' estimates) for the losses in year t minus the true losses, divided by the total assets.

Table 10 (continued)

TaxShield	Calculated as the sum of net income and true reserve, divided by total assets
Smooth	Calculated as the average return on assets over the previous three years before year t.
SmallProfit	Set to 1 if the insurer's earnings fall into the lowest 5% of the positive earnings distribution and zero otherwise.
Insolvency	Set to 1 if the risk capital ratio is smaller than two and zero otherwise.
Violation	Set to 1 if the insurer has at least one ratio violations and zero otherwise.
Liab	Total liabilities divided by total assets.
Crisis	Dummy variable equals to 1 if the observation is for the accident year 2008 and 0 otherwise.
Size	Total assets of the firm in thousands, taking logarithm.
SmallLoss	Set to 1 if the insurer's earnings fall into the highest 5% of the negative earnings distribution and zero otherwise.
Profit	Set to 1 if the insurer's earnings fall into the top 90% of the positive earnings distribution and zero otherwise.
Loss	Set to 1 if the insurer's earnings fall into the bottom 90% of the negative earnings distribution and zero otherwise.
LineSize	Calculated as the total premiums written in this line divided by the total premium written by the insurer for all lines combined
Reinsurance	Calculated as the premiums reinsured divided by total premiums written
Public	Set to 1 if the company is publicly traded and zero otherwise.
Mutual	Set to 1 if the company has a mutual organization structure and zero otherwise.
Group	Set to 1 if the company is a member of a group and zero otherwise

Table 11 Cross-validation results of machine learning models

Business line	Training/ Validation Sample	Machine learning without manager estimates								Machine learning with manager estimates							
		Random forest		Artificial neural networks		Gradient boosting machine		Linear regression		Random forest		Artificial neural networks		Gradient boosting machine		Linear regression	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Private Passenger Auto Liability	1996–2005	8213	34,687	12,663	31,984	7602	33,483	12,627	35,154	7758	36,071	11,903	30,652	6665	25,494	10,562	31,780
	1996–2006	7848	34,547	13,013	35,454	7500	29,786	13,173	37,857	7220	30,305	12,502	33,185	8447	32,637	11,006	33,500
	1996–2007	7869	35,047	13,330	33,554	7524	30,620	13,559	40,615	6902	30,220	12,571	38,788	8261	32,253	5208	18,047
Commercial Auto Liability	1996–2005	3565	14,051	5384	16,632	3583	13,797	5864	18,765	3446	13,555	5192	16,345	3469	13,456	5593	18,262
	1996–2006	3520	13,881	5221	15,090	3770	13,143	5657	17,466	3266	13,583	4855	14,598	3216	12,590	5357	17,559
	1996–2007	3575	13,671	5269	15,036	3610	12,816	5741	17,480	3322	13,121	4531	15,283	3259	12,300	5741	17,480
Workers' Compensation	1996–2005	7518	29,418	10,864	32,549	7398	29,170	12,272	37,910	7144	28,629	11,371	32,327	6966	27,674	12,282	38,926
	1996–2006	7434	29,387	10,793	30,622	7095	28,136	12,728	38,606	6988	26,888	10,981	32,549	6659	25,527	12,527	36,483
	1996–2007	7298	29,468	11,053	32,779	7412	28,603	12,923	36,799	6861	26,574	10,806	30,921	6945	26,847	12,716	36,833
Commercial Multi-Peril	1996–2005	5103	22,060	6719	25,646	5250	23,334	7820	27,533	4854	22,062	6694	25,765	4948	21,780	7255	27,250
	1996–2006	5151	23,404	6459	25,735	5067	23,018	8022	35,718	4968	22,308	6889	24,612	4969	20,965	7469	34,569
	1996–2007	4963	22,556	6807	24,917	5555	22,499	6876	29,755	4534	21,265	7696	26,464	4953	20,529	7380	28,584
Homeowner/Farmowner	1996–2005	5401	35,313	7084	23,375	4844	27,593	5674	22,069	4620	30,677	5944	21,404	4620	30,314	4402	16,359
	1996–2006	5228	34,615	7251	22,557	5293	27,254	5687	21,070	4791	31,937	6187	19,070	5416	39,621	4203	16,201
	1996–2007	5121	33,550	7008	21,237	5314	27,187	5548	21,269	4590	31,305	6166	19,291	4349	26,346	4321	16,674

Table 12 Holdout test results of machine learning models

Business line	Holdout Sample	Machine learning without manager estimates								Machine learning with manager estimates							
		Random forest		Artificial neural networks		Gradient boosting machine		Linear regression		Random forest		Artificial neural networks		Gradient boosting machine		Linear regression	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Private Passenger Auto Liability	2006	8434	35,271	14,030	56,955	6622	23,086	13,432	39,100	8083	36,037	15,581	43,738	6564	22,781	11,538	37,290
	2007	8390	40,116	14,639	52,220	10,289	58,980	15,813	59,098	7670	35,205	13,760	45,449	10,557	51,826	12,021	37,903
	2008	8507	41,440	17,640	40,571	8000	37,251	14,347	45,600	8664	39,732	13,816	42,966	9450	38,189	13,049	56,944
Commercial Auto Liability	2006	3679	14,527	6143	15,658	3680	13,389	5940	16,399	3475	14,468	5454	15,511	3312	18,336	5503	15,866
	2007	3056	10,413	5856	14,268	3865	11,930	5304	14,677	2912	10,228	5289	18,205	3738	14,594	4937	13,565
	2008	3216	9638	6173	14,515	3318	9656	5561	15,185	3268	11,353	3910	10,812	3139	11,526	5561	15,185
Workers' Compensation	2006	7675	28,922	13,455	54,781	6957	24,741	13,920	36,752	6981	25,458	15,031	47,215	6053	21,130	14,069	39,753
	2007	6477	24,704	15,590	99,623	6684	22,606	13,859	37,107	6237	26,793	10,607	29,750	7080	28,250	13,380	41,593
	2008	9081	36,131	15,651	38,397	8595	32,830	15,124	44,808	8841	37,460	15,074	47,788	8529	35,409	14,753	42,760
Commercial Multi-Peril	2006	4543	14,943	7149	21,334	5722	20,705	7230	17,687	3889	14,430	7864	20,603	3872	14,142	6716	18,734
	2007	4783	19,821	6996	31,622	4042	12,352	9077	40,515	4567	19,633	7427	26,317	4021	14,364	7939	34,973
	2008	8400	39,315	11,738	42,790	9152	42,002	11,366	40,373	7475	36,082	8939	31,997	7642	35,677	10,247	36,373
Homeowner/ Farmowner	2006	5079	31,213	8495	21,218	4678	26,112	5219	14,457	3731	23,916	6316	16,639	4061	25,751	3413	11,225
	2007	5006	31,099	5475	12,992	5553	34,897	5202	15,297	4509	25,861	8105	20,640	5116	41,415	4064	13,748
	2008	11,787	119,233	15,285	48,258	11,302	117,113	7968	23,580	10,728	119,615	9484	28,268	11,433	118,083	5628	20,881

References

A. M. Best Company. (1994). *Best's aggregates and averages: Property-casualty edition*. Oldwick: A. M. Best Company.

Anderson, D. R. (1971). Effects of under and overevaluations in loss reserves. *Journal of Risk and Insurance*, 585–600.

Anderson, D. R. (1973). Effects of loss reserve evaluation upon policyholders' surplus. Madison, Wisconsin: Bureau of Business Research and Service, University of Wisconsin, monograph, 6.

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research, 58*(1), 199–235.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417.

Beaver, W. H., & McNichols, M. F. (1998). The characteristics and valuation of loss reserves of property casualty insurers. *Review of Accounting Studies, 3*(1–2), 73–95.

Beaver, W. H., McNichols, M. F., & Nelson, K. K. (2003). Management of the loss reserve accrual and the distribution of earnings in the property-casualty insurance industry. *Journal of Accounting and Economics, 35*(3), 347–376.

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2020). Using machine learning to detect misstatements. Review of Accounting Studies, forthcoming.

Bierens, H. J., & Bradford, D. F. (2005). Are property-casualty insurance reserves biased? A Non-Standard Random Effects Panel Data Analysis 1.

Bishop, C. M. (2006). *Pattern recognition and machine learning*: Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L. (2002). Using models to infer mechanisms. *IMS Wald Lecture, 2*, 59–71.

Browne, M. J., Ma, Y.-L., & Wang, P. (2009). Stock-based executive compensation and reserve errors in the property and casualty insurance industry. *Journal of Insurance Regulation, 27*(4).

Brownlee, J. (2016). Linear regression for machine learning. Machine learning mastery. https://machinelearningmastery.com/linear-regression-for-machine-learning/, accessed the last time on 22 June 2019.

Brownlee, J. (2018). A gentle introduction to k-fold cross-validation, may 2018. Available in https://machinelearningmastery.com/k-fold-cross-validation/, accessed the last time on 22 June 2019.

Eckles, D. L., & Halek, M. (2010). Insurer reserve error and executive compensation. *Journal of Risk and Insurance, 77*(2), 329–346.

Fischer, P. E., & Verrecchia, R. E. (2000). Reporting bias. *The Accounting Review, 75*(2), 229–245.

Forbes, S. W. (1970). Loss reserving performance within the regulatory framework. *Journal of Risk and Insurance*, 527–538.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Gaver, J. J., & Paterson, J. S. (2004). Do insurers manipulate loss reserves to mask solvency problems? *Journal of Accounting and Economics, 37*(3), 393–416.

Grace, E. V. (1990). Property-liability insurer reserve errors: A theoretical and empirical analysis. *Journal of Risk and Insurance*, 28–46.

Grace, M. F., & Leverty, J. T. (2012). Property–liability insurer reserve error: Motive, manipulation, or mistake. *Journal of Risk and Insurance, 79*(2), 351–380.

Guttman, I., & Marinovic, I. (2018). Debt contracts in the presence of performance manipulation. *Review of Accounting Studies, 23*(3), 1005–1041.

Hoyt, R. E., & McCullough, K. A. (2010). Managerial discretion and the impact of risk-based capital requirements on property-liability insurer reserving practices. *Journal of Insurance Regulation, 29*(2).

Lambert, R. A. (1984). Income smoothing as rational equilibrium behavior. *Accounting Review*, 604–618.

Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. R. (2000) Boosting algorithms as gradient descent. In *Advances in neural information processing systems,* (pp. 512–518).

Nelson, K. K. (2000). Rate regulation, competition, and loss reserve discounting by property-casualty insurers. *The Accounting Review, 75*(1), 115–138.

Perols, J. (2011). Financial statement fraud detection: An analysis of statistical and machine learning algorithms. *Auditing: A Journal of Practice & Theory, 30*(2), 19–50.

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding needles in a haystack: Using data analytics to improve fraud prediction. *The Accounting Review, 92*(2), 221–245.

Petroni, K., & Beasley, M. (1996). Errors in accounting estimates and their relation to audit firm type. *Journal of Accounting Research, 34*(1), 151–171.

Petroni, K. R. (1992). Optimistic reporting in the property-casualty insurance industry. *Journal of Accounting and Economics, 15*(4), 485–508.

Ramon, J. 2013. Responses to question "how to determine the number of trees to be generated in random Forest algorithm?". **ResearchGate**. https://www.researchgate.net/post/How_to_determine_the_number_of_trees_to_be_generated_in_Random_Forest_algorithm

Samuels, D., Taylor, D. J., & Verrecchia, R. E. (2018). Financial misreporting: Hiding in the shadows or in plain sight? *Available at SSRN, 3157222*.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227.

Smith, B. D. (1980). An analysis of auto liability loss reserves and underwriting results. *Journal of Risk and Insurance*, 305–320.

Sun, T., & Vasarhelyi, M. A. (2017). Deep learning and the future of auditing: How an evolving technology could transform analysis and improve judgment. *CPA Journal, 87*(6).

Trueman, B., & Titman, S. (1988). An explanation for accounting income smoothing. *Journal of Accounting Research*, 127–139.

Weiss, M. (1985). A multivariate analysis of loss reserving estimates in property-liability insurers. *Journal of Risk and Insurance*, 199–221.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. 3rd Edition, Burlington: Morgan Kaufmann Publishers.

Zhang, C., & Browne, M. J. (2013) Loss reserve errors, income smoothing and firm risk of property and casualty insurance companies. In *Annual Meeting of the American Risk and Insurance Association, Working Pa,* (pp. 1–55).