Development and Evaluation of Models for Claims Reserves

Lillith Chute and Paul LeBorgne

Description:

The worker's compensation industry deals with a tremendous amount of data from images to video to various forms of text files, to standard relational databases. Most small to medium (likely many processes within large firms) deal with the processing of this data with a high degree of human touch. This significantly slows aspects of the workflow creating high wait times for the processing of claims to underwriting policies.

One such area of interest for the insurance industry is in calculating the reserves necessary to cover claims. In much of the industry, particularly with small to mid-sized insurers, this is accomplished via actuaries or claims adjuster experience. In many cases, regarding actuaries, they must create features by hand and include them in any of their models. This can be time-consuming and tedious. A significant advantage of recent machine learning models simplifies the previous drawback: recent models learn nonlinear transformations and interactions between variables from the data without manually specifying them. This is performed implicitly with tree-based models and explicitly with neural networks.

The problem we would like to address is attempting to accurately predict the reserve amount based on features such as class code, claim type (indemnity or medical only), body part(s), nature of injury, injury cause, and total amount required once the claim is closed.

Currently at my company, the way the reserves are set is based primarily on some basic business rules and adjuster experience. Using a large amount of historical data to generate a machine learning model as a decision aid would significantly bolster the accuracy of the initial reserve amount. Having this ability would maximize the amount of money that the company would have as investment while minimizing the penalty of under-reserving.

State Space:

The set of possible states or values that a model can have, is our state space. Regarding insurance claims reserving that means all the different possible values for the parameters and variables the model considers. For this reserving model, those features would be Claimant Type (whether it is medical or indemnity), Injury Cause, Body Part, Nature of Injury, Class Code, Injury State, Jurisdiction State, Claimant Age, Wages, Occupation Description, and various Co-morbidity flags. These features by the number of rows of data, represent the state space of the reserve model.

State Transition:

The state space for a reserve is stochastic in nature, due to the nature of not having all pieces of information immediately available. Some of this information comes in over the course of the claim, however, the reserve amount is initially set and then adjusted over time. For instance, we may not know what the wages are for the claimant until sometime after the claim is filed.

Problem Representation:

The issue of creating a model to predict the amount of money to reserve when a claim is initially entered could be characterized along the parameters as follows:

1. **Fully observable**: This is because all the relevant information about the claim is available at the time of claim entry.

2. **Single Agent**: There is only one actor that needs to decide about the reserve amount of a claim in this system.

3. **Deterministic**: This problem can be considered deterministic since the reserve amount being predicted will be determined by the values of the features (for instance body part injured, claim type, wages, etc.) and the parameters of the model.

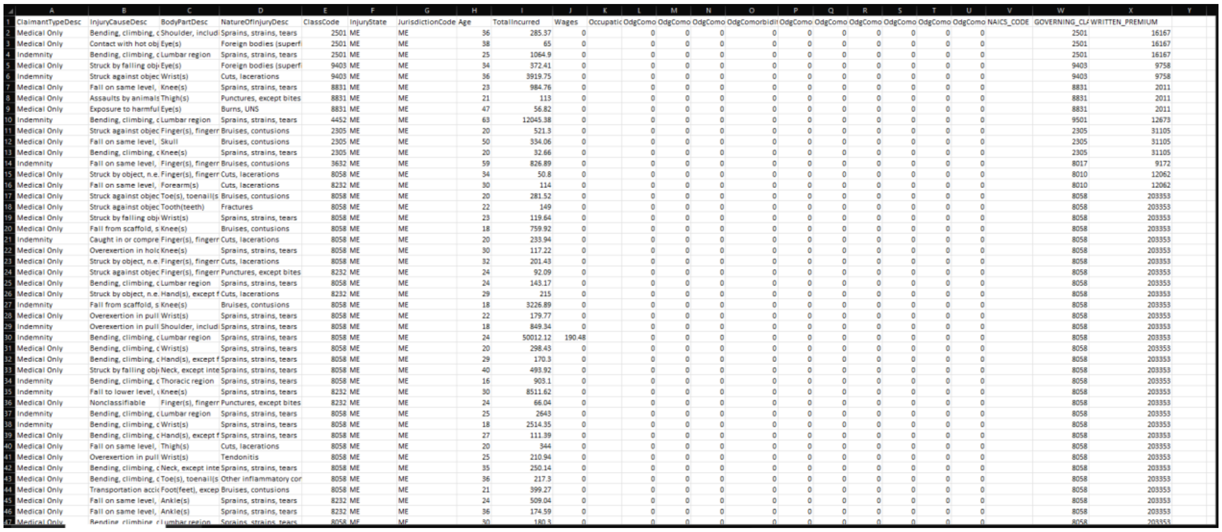
4. **Sequential:** The predictions of each claim are made one after another. Additionally, the information from each prediction can be used to make future predictions. Therefore, the problem is sequential.

5. **Dynamic:** This is because the reserve amount for each claim can change over time as added information becomes available or as each claim is settled.

6. **Continuous:** The underlying values of each feature, for example, age or wages are continuous, hence the problem itself is continuous.

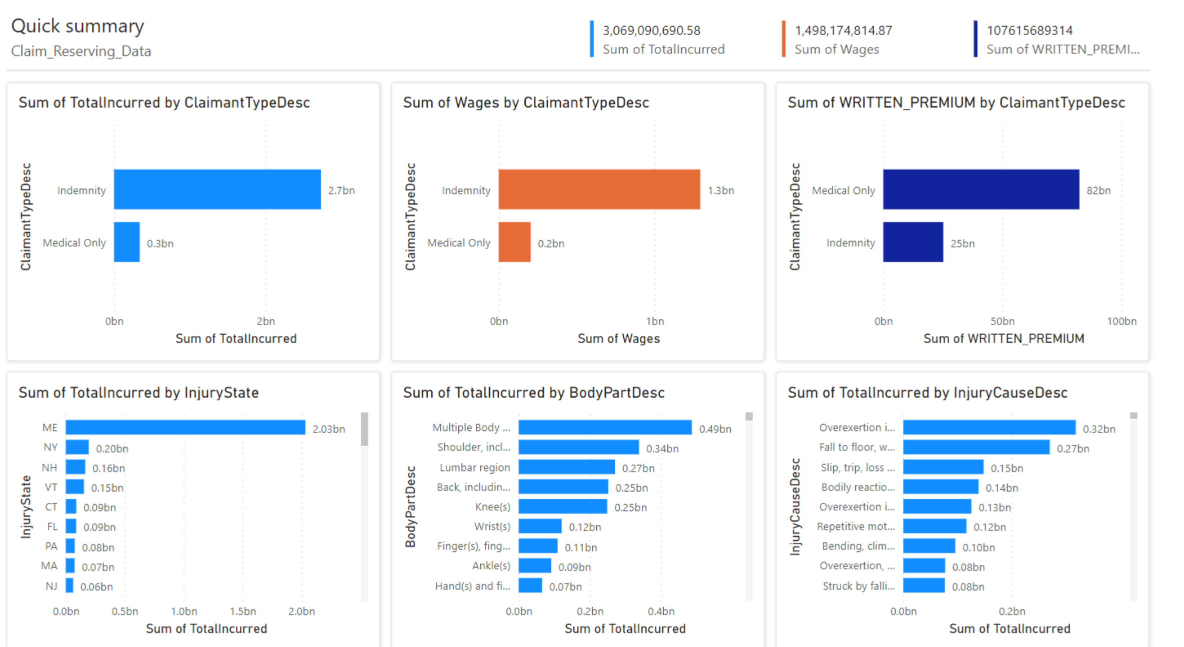
Datasets:

This is a real-world project. The data we are going to be using is a current set of worker's compensation claims data for the company that Lillith works for. The data has been de-identified and approved by her company. It will consist of approximately 474,552+ rows. A sample of what we are working with is shown below.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	ClaimantTypeDesc	InjuryCauseDesc	BodyPartDesc	NatureOfInjuryDesc	ClassCode	InjuryState	JurisdictionCode	Age	TotalIncurred	Wages	Occupatic	OdgComo	OdgComo	OdgComo	OdgComorbidit	OdgComo	OdgComo	OdgComo	OdgComo	OdgComo	OdgComo	NAICS_CODE	GOVERNING_CL/	WRITTEN_PREMIUM
2	Medical Only	Bending, climbing, c	Shoulder, includi	Sprains, strains, tears	2501	ME	ME	36	285.37	0		0	0	0	0	0	0	0	0	0	0		2501	16167
3	Medical Only	Contact with hot obj	Eye(s)	Foreign bodies (superfi	2501	ME	ME	38	65	0		0	0	0	0	0	0	0	0	0	0		2501	16167
4	Indemnity	Bending, climbing, c	Lumbar region	Sprains, strains, tears	2501	ME	ME	25	1064.9	0		0	0	0	0	0	0	0	0	0	0		2501	16167
5	Medical Only	Struck by falling obj	Eye(s)	Foreign bodies (superfi	9403	ME	ME	34	372.41	0		0	0	0	0	0	0	0	0	0	0		9403	9758
6	Indemnity	Struck against objec	Wrist(s)	Cuts, lacerations	9403	ME	ME	36	3919.75	0		0	0	0	0	0	0	0	0	0	0		9403	9758
7	Medical Only	Fall on same level,	Knee(s)	Sprains, strains, tears	8831	ME	ME	23	984.76	0		0	0	0	0	0	0	0	0	0	0		8831	2011
8	Medical Only	Assaults by animals	Thigh(s)	Punctures, except bites	8831	ME	ME	21	113	0		0	0	0	0	0	0	0	0	0	0		8831	2011
9	Medical Only	Exposure to harmful	Eye(s)	Burns, UNS	8831	ME	ME	47	56.82	0		0	0	0	0	0	0	0	0	0	0		8831	2011
10	Indemnity	Bending, climbing, c	Lumbar region	Sprains, strains, tears	4452	ME	ME	63	12045.38	0		0	0	0	0	0	0	0	0	0	0		9501	12673
11	Medical Only	Struck against objec	Finger(s), fingerr	Bruises, contusions	2305	ME	ME	20	521.3	0		0	0	0	0	0	0	0	0	0	0		2305	31105
12	Medical Only	Fall on same level,	Skull	Bruises, contusions	2305	ME	ME	50	334.06	0		0	0	0	0	0	0	0	0	0	0		2305	31105
13	Medical Only	Bending, climbing, c	Knee(s)	Sprains, strains, tears	2305	ME	ME	20	32.66	0		0	0	0	0	0	0	0	0	0	0		2305	31105
14	Indemnity	Fall on same level,	Finger(s), fingerr	Bruises, contusions	3632	ME	ME	59	826.89	0		0	0	0	0	0	0	0	0	0	0		8017	9172
15	Medical Only	Struck by object, n.e.	Finger(s), fingerr	Cuts, lacerations	8058	ME	ME	34	50.8	0		0	0	0	0	0	0	0	0	0	0		8010	12062
16	Medical Only	Fall on same level,	Forearm(s)	Cuts, lacerations	8232	ME	ME	30	114	0		0	0	0	0	0	0	0	0	0	0		8010	12062
17	Medical Only	Struck against objec	Toe(s), toenail(s	Bruises, contusions	8058	ME	ME	20	281.52	0		0	0	0	0	0	0	0	0	0	0		8058	203353
18	Medical Only	Struck against objec	Tooth(teeth)	Fractures	8058	ME	ME	22	149	0		0	0	0	0	0	0	0	0	0	0		8058	203353
19	Medical Only	Struck by falling obj	Wrist(s)	Sprains, strains, tears	8058	ME	ME	23	119.64	0		0	0	0	0	0	0	0	0	0	0		8058	203353
20	Medical Only	Fall from scaffold, s	Knee(s)	Bruises, contusions	8058	ME	ME	18	759.92	0		0	0	0	0	0	0	0	0	0	0		8058	203353
21	Indemnity	Caught in or compre	Finger(s), fingerr	Cuts, lacerations	8058	ME	ME	20	233.94	0		0	0	0	0	0	0	0	0	0	0		8058	203353
22	Medical Only	Overexertion in holc	Knee(s)	Sprains, strains, tears	8058	ME	ME	30	117.22	0		0	0	0	0	0	0	0	0	0	0		8058	203353
23	Medical Only	Struck by object, n.e.	Finger(s), fingerr	Cuts, lacerations	8058	ME	ME	32	201.43	0		0	0	0	0	0	0	0	0	0	0		8058	203353
24	Medical Only	Struck against objec	Finger(s), fingerr	Punctures, except bites	8232	ME	ME	24	92.09	0		0	0	0	0	0	0	0	0	0	0		8058	203353
25	Medical Only	Bending, climbing, c	Lumbar region	Sprains, strains, tears	8058	ME	ME	24	143.17	0		0	0	0	0	0	0	0	0	0	0		8058	203353
26	Medical Only	Struck by object, n.e.	Hand(s), except f	Cuts, lacerations	8232	ME	ME	29	215	0		0	0	0	0	0	0	0	0	0	0		8058	203353
27	Indemnity	Fall from scaffold, s	Knee(s)	Bruises, contusions	8058	ME	ME	18	3226.89	0		0	0	0	0	0	0	0	0	0	0		8058	203353
28	Medical Only	Overexertion in pull	Wrist(s)	Sprains, strains, tears	8058	ME	ME	22	179.77	0		0	0	0	0	0	0	0	0	0	0		8058	203353
29	Indemnity	Overexertion in pull	Shoulder, includi	Sprains, strains, tears	8058	ME	ME	18	849.34	0		0	0	0	0	0	0	0	0	0	0		8058	203353
30	Indemnity	Bending, climbing, c	Lumbar region	Sprains, strains, tears	8058	ME	ME	24	50012.12	190.48		0	0	0	0	0	0	0	0	0	0		8058	203353
31	Medical Only	Bending, climbing, c	Wrist(s)	Sprains, strains, tears	8058	ME	ME	20	298.43	0		0	0	0	0	0	0	0	0	0	0		8058	203353
32	Medical Only	Bending, climbing, c	Hand(s), except f	Sprains, strains, tears	8058	ME	ME	29	170.3	0		0	0	0	0	0	0	0	0	0	0		8058	203353
33	Medical Only	Struck by falling obj	Neck, except inte	Sprains, strains, tears	8058	ME	ME	40	493.92	0		0	0	0	0	0	0	0	0	0	0		8058	203353
34	Indemnity	Bending, climbing, c	Thoracic region	Sprains, strains, tears	8058	ME	ME	16	903.1	0		0	0	0	0	0	0	0	0	0	0		8058	203353
35	Indemnity	Fall to lower level, u	Knee(s)	Sprains, strains, tears	8232	ME	ME	30	8511.62	0		0	0	0	0	0	0	0	0	0	0		8058	203353
36	Medical Only	Nonclassifiable	Finger(s), fingerr	Punctures, except bites	8232	ME	ME	24	66.04	0		0	0	0	0	0	0	0	0	0	0		8058	203353
37	Indemnity	Bending, climbing, c	Lumbar region	Sprains, strains, tears	8058	ME	ME	25	2643	0		0	0	0	0	0	0	0	0	0	0		8058	203353
38	Indemnity	Bending, climbing, c	Wrist(s)	Sprains, strains, tears	8058	ME	ME	18	2514.35	0		0	0	0	0	0	0	0	0	0	0		8058	203353
39	Medical Only	Bending, climbing, c	Hand(s), except f	Sprains, strains, tears	8058	ME	ME	27	111.39	0		0	0	0	0	0	0	0	0	0	0		8058	203353
40	Medical Only	Fall on same level,	Thigh(s)	Cuts, lacerations	8058	ME	ME	20	344	0		0	0	0	0	0	0	0	0	0	0		8058	203353
41	Medical Only	Overexertion in pull	Wrist(s)	Tendonitis	8058	ME	ME	25	210.94	0		0	0	0	0	0	0	0	0	0	0		8058	203353
42	Medical Only	Bending, climbing, c	Neck, except inte	Sprains, strains, tears	8058	ME	ME	35	250.14	0		0	0	0	0	0	0	0	0	0	0		8058	203353
43	Medical Only	Bending, climbing, c	Toe(s), toenail(s	Other inflammatory cor	8058	ME	ME	36	217.3	0		0	0	0	0	0	0	0	0	0	0		8058	203353
44	Medical Only	Transportation acci	Foot(feet), excep	Bruises, contusions	8058	ME	ME	21	399.27	0		0	0	0	0	0	0	0	0	0	0		8058	203353
45	Medical Only	Fall on same level,	Ankle(s)	Sprains, strains, tears	8232	ME	ME	24	509.04	0		0	0	0	0	0	0	0	0	0	0		8058	203353
46	Medical Only	Fall on same level,	Ankle(s)	Sprains, strains, tears	8232	ME	ME	36	174.59	0		0	0	0	0	0	0	0	0	0	0		8058	203353
47	Medical Only	Bending, climbing, c	Lumbar region	Sprains, strains, tears	8058	ME	ME	30	180.3	0		0	0	0	0	0	0	0	0	0	0		8058	203353

Additionally, a few quick points of analysis to get a sense of which category of claim might have the highest impact.



Approach to Solution:

1. Do some data analysis to determine how to handle any data imbalance and determine the most correlative features to the label we are trying to predict. This may reduce the number of features or point to the fact that there may be features we need to engineer or track down.
2. **Preprocess data:** Our data is real life data that may require significant work to clean up.

- Convert any text data. In categorical data such as body part, if that proves useful, we may need to merge or discard categories.
- Perform typical pre-processing tasks dealing with issues such as
    - Null columns
    - Null rows
    - Encoding categorical data
    - Data splitting

3. **Algorithms**: Initially, I think working with gradient boosting and neural network models would be most fruitful.
    - Gradient Boosting Machine.
    - Neural Network.
4. **Evaluate:** Use statistical evaluation techniques to determine the efficacy of the models. Such techniques are discussed in the next section. Additionally, compare the two models to determine the most effective.

Evaluation:

If we consider that we are using a form of regression model, then there are a couple of evaluation functions that we can use to determine the effectiveness of the model. The first that we can use is the mean squared error. This will measure the average squared difference between the predictions made and the actual value. Additionally, we will consider the mean absolute error. This measures the average absolute difference between the predictions and actual values. These measures assume the target values are continuous and that we have a normal distribution.

Using both provides a way that we can quantify the difference between the predications made by the model and the actual target values. We can use this metric to compare different models and determine the best overall performing model.

Deliverable:

At the conclusion of this project, we plan to deliver models that, given a claim with the requisite features, will produce a reserve amount for that claim. We also will produce an ablation table comparing the two models' performance and accuracy. If given enough time, it is possible that we could provide some sort of interface where different values could be entered, and the model run to see what sort of prediction it would produce.

Responsibilities:

There will be a significant amount of time and work needed to ensure the dataset has been thoroughly cleaned and prepared for use in the project. As such this portion of the project is currently planned to be done collaboratively, with both team members working to get it completed. Once that is done, each team member will work to implement a machine learning algorithm and generate their data models.

Related Research:

1. Blier-Wong C, Cossette H, Lamontagne L, Marceau E. Machine Learning in P&C Insurance: A Review for Pricing and Reserving. Risks. 2021; 9(1):4. https://doi.org/10.3390/risks9010004

2. Carrato, Alessandro, and Michele Visintin. 2019. From the chain ladder to individual claims reserving using machine learning techniques. Paper presented at ASTIN Colloquium, Cape Town, South Africa, April 2–5; vol. 1, pp. 1–19.

3. Crèvecoeur, Jonas, and Katrien Antonio. Methods for Claim Reserving in Non-Life Insurance: Modeling the Occurrence, Reporting and Development of Individual Claims. 2020.

4. Härkönen, V. On Claims Reserving with Machine Learning Techniques. In Mathematical Statistics; Stockholms Universitet: Stockholm, Sweden, 2021; p. 1 - 71.