Estimating Unpaid Claims Using the Case Outstanding Development Reserving Technique.

A Project as a Course requirement for

Bachelor of Science (Hons.) in Mathematics

K OMKAR REDDY

(Reg.No. 211202)

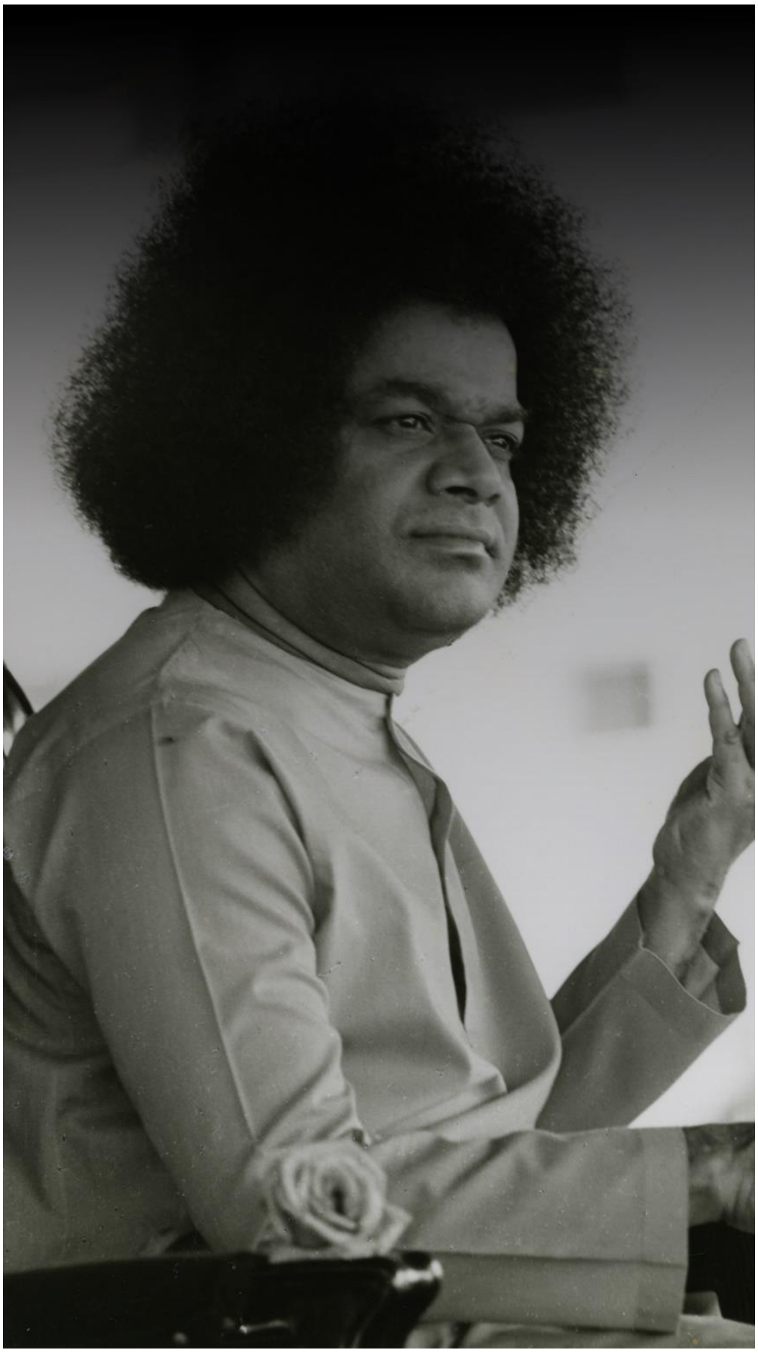
SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING

(Deemed to be University)

Department of Mathematics and Computer Science

Brindavan Campus

April 2024



I Dedicate This Work to Bhagawan Sri Sathya Sai Baba

SRI SATHYA SAI INSTITUTE OF HIGHER LEARNING
(Deemed to be University)

Dept. of Mathematics & Computer Science
Brindavan Campus

CERTIFICATE

This is to certify that this Project titled **"Estimating Unpaid Claims using the Case Outstanding Reserving Technique"** submitted by K OMKAR REDDY, 211202, Department of Mathematics and Computer Science, Brindavan Campus is a bonafide record of the original work done under my supervision as a Course requirement for the Degree of Bachelor of Science (Hons.) in Mathematics.

Darshan Gera

……..……………………..
Dr. Darshan Gera
Associate Head of Dept.

……..……………………..
Sri. Satya Sai Baba Mudigonda
Project Supervisor

Countersigned by

Place: Brindavan

…………………………………..
Dr (Ms.) Y Lakshmi Naidu
Head of the Department

Date: April 1, 2024

DECLARATION

The Project titled **"Estimating Unpaid Claims using the Case Outstanding Development Technique"** was carried out by me under the supervision of **Sri. Satya Sai Baba Mudigonda**, Department of Mathematics and Computer Science, Brindavan Campus as a Course requirement for the Degree of Bachelor of Science (Hons.) in Mathematics and has not formed the basis for the award of any degree, diploma or any other such title by this or any other University.

OmkarKatti

…………………………………..

Place: Bengaluru
Date: April 1, 2024

K OMKAR REDDY
Regd.No.: 211202
B.Sc. (Hons.) in Mathematics
Brindavan Campus

ACKNOWLEDGEMENTS:

I must confess that this work that I present below would not have been possible without expertise and contribution of many others. This work was only possible because I was fortunate enough to have been acquainted with some great people.

First and foremost, I would like to express my gratitude to the founding chancellor of the university, Bhagawan Sri Sathya Sai Baba, who has been the guiding light and inspiration for me and many others across the globe.

I would like to mention that this project was mentored by Sri. Satya Sai Baba Mudigonda, who guided me through this project and provided insights on areas to work on for the past one semester. It was his professional knowledge that helped me to complete this project and derive a huge amount of learning from the day this project began.

I would like to mention about the interest shown by Dr. Darshan Gera, associate HOD, Dept. of Mathematics and Computer Science, who motivated us to solve difficult problems being his students.

I would like to highlight about the consistent and persistent encouragement given by my parents throughout this project.

A special mention about the institute administration, Director of the SSSIHL Brindavan campus, Prof. Shiva Kumar; Warden of the Brindavan hostel, Dr Ravi Kumar; Assistant Warden, Dr. Malleswar; My wing teacher in the hostel, Sri M.G. Nanda Gopal.

I would like to say that this project was all possible due to the hard efforts of the teachers, who taught me since primary school.

I would say that it was my classmates and friends who spread the passion, zeal and collective vigor to push each other and move forward during the past one semester.

A huge mention to many other people who have shaped my life since it began, knowingly and unknowingly.

Contents:

Introduction:

Any company on this earth is interested to know the future costs it might incur as on today. Large Multinationals and other companies, thus, using their foresight and the expertise of their experts and talented employees set aside an amount, as a reserve or a provision to make “Good the loss, they might incur in the future”.

This is done by companies whose day-to-day activities may be trading, manufacturing and providing services to clients, these activities rather seem more certain.

However, the day-to-day operation of an insurance company lies in “Uncertainty and Risk”. The only major cost, Claims, that the insurance company might largely face lies in uncertainty. Thus, Reserving is even more pivotal for an insurance company.

Reserving is a mundane operation; an insurance company must rather be interested to do. In the process, in a prospective basis, they ascertain an ultimate cost the company might face and this cost is divided in the form of “Premiums that the company will collect from the insured”.

It’s important for an insurance company to thus predict, present a correct and the most probable true cost they might incur.

An overestimation may lose them clients and business and an underestimation might lead the business to undergo liquidation in the future.

Well, there are many reserving techniques and each of them are best suitable for certain situations. This project is aimed at understanding the case outstanding development technique and applying this technique on a standard world problem.

This project thus will aim at:

1. Testing the suitability of the case outstanding technique on areas that it is meant to be used on.
2. Making some observations about when the predicted ultimate costs are close to true ultimate costs.
3. Trying to experiment the usage of Predictive modelling and find its utility, in case it may be found useful.
4. Automating the process, without which the above three objectives will be tedious to do.
5. To present an easy-convenient-one-click solution to calculate reserves, with better readable and understandable information using Graphs and other diagrams.
6. To Build a library in Python for other users to ease their work.

Prerequisite knowledge:

1. Microsoft Excel
2. Python language and in particular Pandas module in python
3. Basics of reserving and development triangle.

Motivation:

As mentioned in the introduction, the process of reserving is highly important for the stable existence of an insurance company. Only a true and best approximate reserve is one strong indicator for sustenance of the company. The closer the reserve amount is to the actual losses that will be reported in the future, the better it's for a company.

Thus, this entire process seems a lot fascinating as it's trying to look in to the future and trying to tell what future reported costs are going to be with full conviction, using the mathematical and statistical approaches already laid down.

A sense of satisfaction will surely be obtained in knowing how close the ultimate costs given by the technique used are to the actual ultimate costs. This is will by far improve the confidence a person has on a given technique and on its application on a given line of business.

The process of validating a reserving technique on a large number of companies' data calls forth for automation of the procedure. Solving a reserving technique on commonly used software like Microsoft Excel will involve a huge time, effort and is highly not productive. An involvement of computer programming will ease the process and will ensure the usability of the same program for future and wide use too.

Any alteration in the existing technique could also be made for the betterment of the predicted ultimate costs and improving the process of reserving to quite an extent. We live in a world that is very keen on and interested about machine learning and predictability.

Hence, an attempt to look at machine learning to improve the process by far also inspires to take up this project and experiment the use of machine learning in this field of actuarial reserving.

Over and above, this project will give an insight on how a particular reserving technique is built, the mathematical logic behind it and how to do reserving from start till the end.

General Idea of Reserving:

Reserving is the process done by an insurance company to anticipate the final ultimate costs the company will incur at the end of a specific policy or policies. Unlike any other businesses, in the insurance industry it is impossible to ascertain the total costs as they are not certain as of today.

Thus, Reserving is a process by which an actuary looks into the future and tries to present a true picture of the future costs to the best of his ability using the standard techniques based on mathematical and statistical approaches. The entire process is thus done with no bias and literally presents the most appropriate cost.

So, Actuarial professionals rely on historical data to ascertain the ultimate costs, whereby they get an idea of the reserves they have to maintain and fulfil the due claims on a future date.

Thus, several techniques have evolved to build reserves. Each technique is best suited for a certain type or line of business, based on a few factors.

Some of the famous reserving techniques include:

1. Chain Ladder method
2. Bornhuetter-Ferguson method
3. Cape cod method
4. Frequency Severity method
5. Case Outstanding technique.

The above techniques, each have a specific use and present the true picture when used in that line of business that they are meant for.

A Usage in the wrong line of business may lead to huge deviations and will lead to under or overestimation of the ultimate costs.

An insurance company may deal in mainly two types or lines of business:

1. Long tailed lines. (like Workers' compensation)
2. Short tailed lines. (like automobile industry)

Upon knowing the ultimate costs, an actuary can set aside a certain sum of money to pay for the losses that will be incurred in the future. These are known as reserves.

Ultimate costs = paid claims + case outstanding + IBNR (incurred but not reported)

The IBNR or the reserves are computed using an alteration of the above formula:

Ultimate costs – (paid claims + case outstanding) = IBNR (incurred but not reported)

Though we have to find the unpaid claims, the main process will be finding the ultimate costs and by the above alteration, we get the final unpaid claims.

Our procedures will also be the same, henceforth.

Abstract about the Case Outstanding Development Technique:

The Case outstanding development technique is a reserving technique. This method is a variation of the usual development method where losses are used to project ultimate costs. In this method however, we use case reserves to project future costs.

This technique is used in short tail lines of business.

Assumptions in this method:

1. Claims recorded to date will continue to develop in a similar fashion in the future.
2. Case outstanding give us relevant information on claims that are yet to be observed.
3. Throughout the policy period:
   a. there is consistent claims processing (claim settlement rates and case reserve adequacy).
   b. the mix of claim types is stable.
   c. policy limits (if any) are stable.
   d. reinsurance retention limits (if any) are stable.

The above are the same assumptions made in the chain-ladder technique as this technique too relies on historical data and trend to predict reserves. The Key assumption of this technique is that the relation between the case reserves and incremental paid claims will help us to predict the final costs.

It believes that the IBNR (Incurred but not recorded) claims are dependent on the case reserves. This is possible only in short lines and will not apply on long tail lines. Thus, the technique is not apt for long tail lines of business.

There are two variations used in the above discussed method.

The Casualty Actuarial Society names them: Approach 1 and Approach 2.

In the approach 1, We use historical data available to us to predict the ultimate costs of the future. This method is applicable only if we have sufficient data of the previous years to ascertain costs.

In the approach 2, We use the industry standard development factors for calculating the final ultimate cost. This is generally used when self sufficient information is not available to a given insurance company.

When is this method used?

Approach 1:

This method is best suited when majority of the claims are developed in the initial development period. In other words, it is best suited for short tail insurance policies. The best usage of this technique is in the automobile sector where the claims can easily be ascertained in the near future or the payments will not spread across a vast period of time.

Approach 2:

This method uses industry standard age to age factors and industry averages to develop the triangles. This method will be used for business that have just started and have no historical data to predict ultimate costs and build reserves.

Terminology and Formulae:

1. Reported Claims: Total amount of claims received by a company in a period of time after ascertainment done by TPA or Case examination.
   (paid claims + case outstanding)

2. Paid Claims: The amount of claims the company paid during a period of time.

3. Case Reserves: The amount of money to be set aside by an insurance company to pay for the losses that have been reported. This is interchangeably called case outstanding.

4. IBNR (Incurred but not reported claims): These are claim amounts of accidents that have already occurred but they have not yet been reported by the insured till date.

5. Remaining-in-case ratios: It's the ratio of the current year case outstanding compared to the prior (last year's) case outstanding.

6. Paid-on-case ratios: It's the ratio of current years incremental paid claims with the prior (last year's) case outstanding.

7. Ultimate costs: The sum or total costs the company will ultimately pay for a particular or a group of policy schemes in a given number of accident years.

Project Methodology:

Read Through the Jack Friedland's Book on Reserving.

Refer:Estimating Unpaid Claims using basic techniques

Solved a Case Outstanding Problem on Microsoft Excel.

Refer:Reference excel

Found an Automobile Claims dataset on CAS website.

Refer:Commercial Auto dataset

Performed Data Pre-processing on the Data set and divided into separate companies

Built a Class and methods to automate the entire process.

Ran the Processed data on the defined functions.

Displayed the solved steps in an excel for each company.

Tried to use Predictive modelling in predicting the average age to age factors.

Compared the Real ultimate costs, the ultimate cost given by the technique, costs given by predictive modelling

Final interpretations and observations were deducted or inferred from the procedures.

The project was started by reading the chapter 12: Case Outstanding technique in the CAS book on reserving by: Jacqueline Friedland.

An exercise problem was solved in excel to understand the procedure about how the reserves could be predicted using this method.

A basic program in python was made and the same problem was solved using pandas and python. Both the procedures or technologies produced the same result.

A pre-processing of the given data that had 100+ companies, had to be done and companies that had cleaner data were chosen. This ensured that sufficient and clean data was available to test the code.

Separate CSV files were created for each of the company that could be further used or called to test the program.

The same functions were converted into methods and a python class **CaseOutstandingReserving** was created in python.

Better visual representation to understand the highly numerical ratios and ultimate costs for each of the accident year were made, so that the user can use the technology and make interpretations with utmost ease.

The final results were then put in an excel format and saved, that clearly explains the entire procedure, for each of the companies

An attempt was made to predict the age-to-age factors by using a predictive modelling technique - the generalised linear modelled gamma distribution.

The results were then compared.

A comparison was made among the actual ultimate costs, technique predicted ultimate costs and the cost s predicted by predictive modelling.

Dataset and its Features:

The dataset was taken from the CAS website Research resources section: Commercial_Auto_dataset.

The dataset had 15800 rows and 12 columns.

The dataset comprises of collated data of 157 companies.

The columns in the dataset are:

a. GRCODE NIAC: The company Code.
b. GRNAME NAIC: The name of the company.
c. AccidentYear: The year in which the accident happened.
d. DevelopmentYear: The year in which the accident was reported to the company.
e. DevelopmentLag: The difference between the Development year and the accident year
f. IncurLoss_: incurred losses and allocated expenses at the year end.
g. CumPaidLoss_: The total sum of losses paid across the year.
h. BulkLoss_: Bulk and IBNR reserves on net losses and DCC expenses.
i. PostedReserve97_: Posted reserves in year 1997.
j. EarnedPremDIR_: Premium earned at incurral year-direct and assumed.
k. EarnedPremCeded_: premiums earned at incurral year -ceded
l. EarnedPremNet_: premiums earned at incurral year -net

The relevant features or columns useful in our technique are:

a. GRCODE NIAC: The company Code.
b. GRNAME NAIC: The name of the company.
c. AccidentYear: The year in which the accident happened.
d. DevelopmentYear: The year in which the accident was reported to the company.
e. DevelopmentLag: The difference between the Development year and the accident year
f. IncurLoss_: incurred losses and allocated expenses at the year end.
g. CumPaidLoss_: The total sum of losses paid across the year.

Despite having a huge dataset of 157 companies, not all the companies' data could be used for computation and calculating reserves. A few companies had insufficient historical data that was not enough to calculate reserves using the technique.

Hence, during the process of pre-processing the companies' data were considered and divided under two types or classes:

a. Good_data_sets: datasets that have more than 50 rows of data.
b. Ok_data_sets: datasets that have more than 20 rows of data.

The number of good datasets were: 116 in number.

The number of ok datasets were: 25 in number.

Thus, the total number of companies were used in our process or computation are: 141.

A key note about our dataset is that it contains the data of accidents that happen between 1988 and 1997 and were reported in about 10 years from when they have incurred.

Methodology in finding the ULTIMATE COSTS:

(using the case outstanding technique)

The dataset has several columns. The important columns are Accident year, Development Year, Incurred losses, Cumulative Paid losses and the case reserves.

However, we are not provided with the case reserves column. This however can be computed using the formula:

Case Reserves=Incurred losses – Paid losses.

Thus, it's possible for us to obtain the Case reserves from this formula.

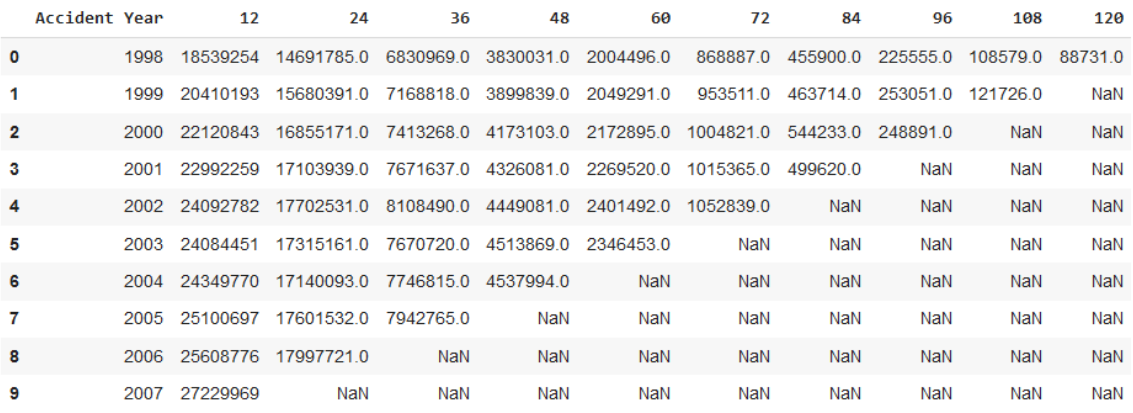
Building Triangles:

The next step once we have the Cumulative Incurred losses, Cumulative paid losses and Case Reserves is to build development triangles from the given data.

There are two triangles that are useful for us in particular:

1. Incremental Paid Claims.
2. Case Reserves.

A Development triangle looks like this:



	Accident Year	12	24	36	48	60	72	84	96	108	120
0	1998	18539254	14691785.0	6830969.0	3830031.0	2004496.0	868887.0	455900.0	225555.0	108579.0	88731.0
1	1999	20410193	15680391.0	7168818.0	3899839.0	2049291.0	953511.0	463714.0	253051.0	121726.0	NaN
2	2000	22120843	16855171.0	7413268.0	4173103.0	2172895.0	1004821.0	544233.0	248891.0	NaN	NaN
3	2001	22992259	17103939.0	7671637.0	4326081.0	2269520.0	1015365.0	499620.0	NaN	NaN	NaN
4	2002	24092782	17702531.0	8108490.0	4449081.0	2401492.0	1052839.0	NaN	NaN	NaN	NaN
5	2003	24084451	17315161.0	7670720.0	4513869.0	2346453.0	NaN	NaN	NaN	NaN	NaN
6	2004	24349770	17140093.0	7746815.0	4537994.0	NaN	NaN	NaN	NaN	NaN	NaN
7	2005	25100697	17601532.0	7942765.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	2006	25608776	17997721.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	2007	27229969	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(An example of incremental paid claims triangles)

It consists of the accident year and number of months that are computed from the difference between when they occurred and when they were reported. The Nan in the table indicate the values that are not known as of today, i.e., the future values that need to be predicted.

Any diagonal traversal on this table will give you the values that were reported in a particular year irrespective of which accident year they happened.

A horizontal traversal will tell you the values corresponding to one accident year across different months they were reported.

A vertical traversal on the table will tell you values that were reported corresponding to one given time period in months across different accident years.

Our method requires us to build such triangles for incremental paid claims and cumulative paid claims. After finding or building the development triangles, we have to find the age-to-age factors which will help us to build the triangle completely.

Age-to-age factors:

The age-to-age factors will help us to find the ultimate costs. There are two ratios that are important for using this technique:

a. Remaining-in-case ratios.
b. Paid-on-case ratios.

The formula for Remaining-in-case ratios is:

$$\text{Remaining-in-Case Ratio} = \frac{\text{Current Case Outstanding}}{\text{Prior Case Outstanding}}$$

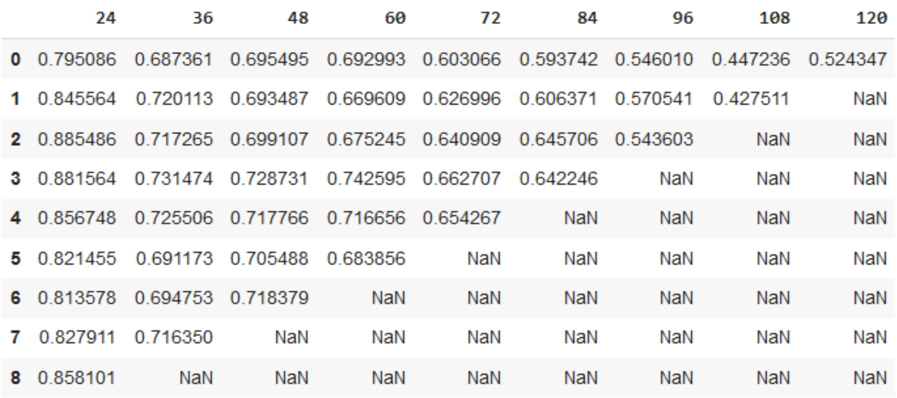
Thus, Remaining-in-case ratio × prior case outstanding will give you current case outstanding.

Similarly, the formula for paid-on-case ratio is:

$$\text{Paid-on-Case Ratio} = \frac{\text{Incremental Paid Claims}}{\text{Prior Case Outstanding}}$$

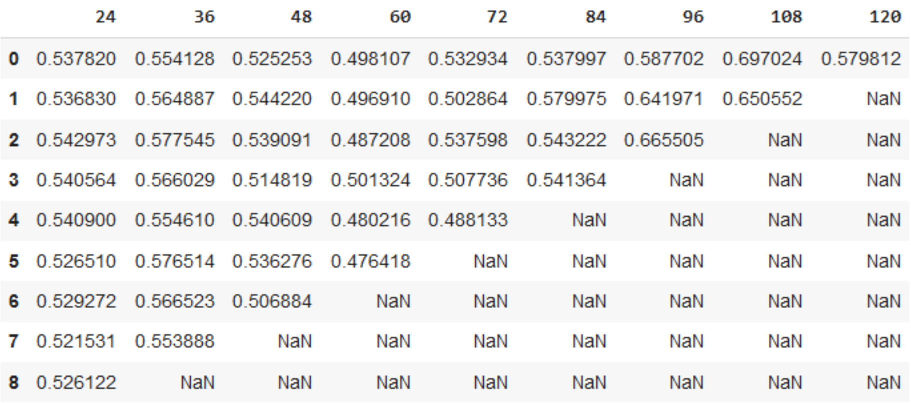
Similarly, Paid-on-Case Ratio × prior case outstanding will give you incremental paid claims for the current year. These results are important for us to fill the triangle.

We build a triangle consisting of all the age-to-age factors for both the ratios.



	24	36	48	60	72	84	96	108	120
0	0.795086	0.687361	0.695495	0.692993	0.603066	0.593742	0.546010	0.447236	0.524347
1	0.845564	0.720113	0.693487	0.669609	0.626996	0.606371	0.570541	0.427511	NaN
2	0.885486	0.717265	0.699107	0.675245	0.640909	0.645706	0.543603	NaN	NaN
3	0.881564	0.731474	0.728731	0.742595	0.662707	0.642246	NaN	NaN	NaN
4	0.856748	0.725506	0.717766	0.716656	0.654267	NaN	NaN	NaN	NaN
5	0.821455	0.691173	0.705488	0.683856	NaN	NaN	NaN	NaN	NaN
6	0.813578	0.694753	0.718379	NaN	NaN	NaN	NaN	NaN	NaN
7	0.827911	0.716350	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	0.858101	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(An example of paid-on-case ratio triangle)



	24	36	48	60	72	84	96	108	120
0	0.537820	0.554128	0.525253	0.498107	0.532934	0.537997	0.587702	0.697024	0.579812
1	0.536830	0.564887	0.544220	0.496910	0.502864	0.579975	0.641971	0.650552	NaN
2	0.542973	0.577545	0.539091	0.487208	0.537598	0.543222	0.665505	NaN	NaN
3	0.540564	0.566029	0.514819	0.501324	0.507736	0.541364	NaN	NaN	NaN
4	0.540900	0.554610	0.540609	0.480216	0.488133	NaN	NaN	NaN	NaN
5	0.526510	0.576514	0.536276	0.476418	NaN	NaN	NaN	NaN	NaN
6	0.529272	0.566523	0.506884	NaN	NaN	NaN	NaN	NaN	NaN
7	0.521531	0.553888	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	0.526122	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(An example of remaining in case ratio triangle)

Finding the average age-to-age factors:

We continue our process by finding the average age-to-age factors. We can observe that this triangle is also just half filled. We compute average age-to-age factors for each of the 12th month by considering an average of all vertical values of the 12th -monthly columns.

The average can be arithmetic mean, median, geometric mean etc. Once we find the average age-to-age factor for each column, the next step is to fill the case-outstanding and paid incremental triangles.

Filling the Case-Outstanding triangle:

Since we already know the average age-to-age factors, we use them to build the Case-Outstanding triangle. For each of the unknown values in the Case-Outstanding triangle, we get the predicted value by our discussed formula: Remaining-in-case ratio × prior case-outstanding = current case-outstanding.

Since, we do not have the Remaining-in-case ratio for any of the missing values, as it is actually a value of the future, we substitute it with the average age-to-age factor corresponding with that 12th monthly figure. Doing the same for each of the unknown value in the triangle, we end by filling this triangle completely.

Filling the Incremental-Paid Claims Triangle:

The Penultimate step in our process is to fill the incremental paid claims triangle. We already know the paid-on case ratios and computed the average age-to-age factors for the same.

The prior discussed formula is critical at this step, that is, paid-on-case ratio × prior case outstanding = incremental paid claims.

In the previous step, we have already computed and filled the case outstanding triangle. Thus, we have all the predicted values for all the case reserves.

By multiplying the average age-to-age factors of the paid-on-case ratios to the case outstanding or case reserves of the previous year, we will obtain the incremental paid claims for the current year. Repeating this process, we will obtain a filled incremental paid claims triangle.

The Final Step:

The final step in our process is to ascertain the ultimate claims the company will likely pay in the future. This is done by summing up the values corresponding to a given accident year, i.e., across each row. Thus, the last column in the cumulated triangle will give us the sum or the ultimate claims corresponding to one accident year.

From this the reserve amount can be easily calculated, the formula:

Ultimate Claims – Reported Claims = IBNR (or reserves)

Thus, we have finished understanding the complete procedure of the Case Outstanding technique approach 1. The next step consists of validating this approach by running this technique on large number of datasets pertaining to various companies and comparing the results with actual values.

Validating the Technique:

The next phase of the project discusses about validation of the usage of the case outstanding technique on the individual companies' datasets.

Additional information about the datasets:

There is a unique aspect about the dataset. It has information about various companies with accident years ranging from 1988 to 1997 and 10 reported years for each of the accident year. In other words, using this data will give us completely filled triangles.

But if we stop at 1997 as the current year, we will get only a semi-filled triangle on which this technique can be run. This will help us to validate this technique.

Data Pre-processing:

The dataset comprises of the data of various companies put together to form one consolidated database. We obtain two types of data pertaining to each of the company:

1. Initial dataset: This dataset of a company comprises of data up to development year 1997.This is the dataset on which we will run our technique.

2. Final dataset: This dataset comprises of all the data about a given company. This is used to validate the results we get from the initial dataset by comparing it with the original values.

Building a Python framework to solve the problem:

This project called for an automation procedure given the scale of data to be processed. Thus, a python framework with a python class CaseOutstanding_Reserving was made with methods or functions to automate the entire process of reserving.

The Class attributes are the raw csv file as a dataframe, names of accident years, reported years, reported losses, paid losses and case outstanding columns. The reported and paid losses, here refer to the cumulated ones.

The program is built only for cumulative figures (reported loss and paid loss) and not the incremental ones.

The Class methods include:

1. **d_triangle(parameter):** will build a development triangle given the parameter.
2. **incremental_triangles** (): will build the incremental paid claims triangles.
3. **paid_tr_on_case_ratio** (): will build up the paid-on case ratio triangle.
4. **Rem_in_case_ratio** (): will build up the remaining in case ratio triangle.
5. **Calc_measures(ratio_triangle):** will give us a data frame with measures like mean and median pertaining to the ratio triangle given.

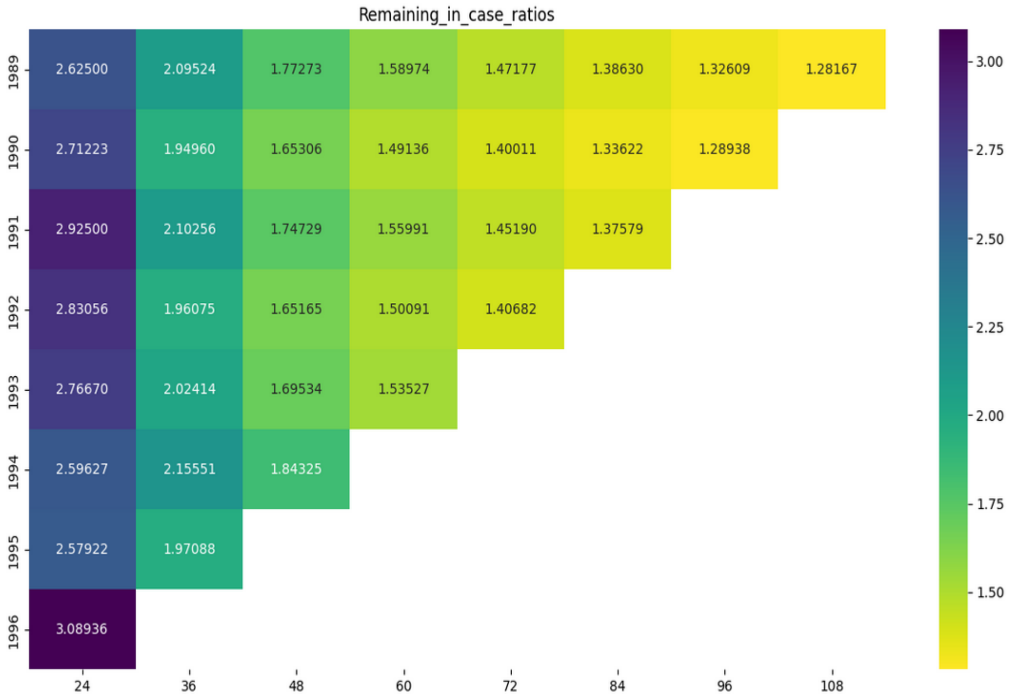
6. **fill_triangle_case_outstanding (measure, method):** will fill the case reserves triangle based on the measure data frame and a method like mean in it.
7. **fill_triangles_paid_triangle (case_out, paid_inc, measure, method):** will fill the incremental paid claims triangle based on the filled case outstanding triangle, paid incremental claims triangle, measure data frame and methods like mean in measures data frame.

Better readability and understanding the technique:

The main aspect in the procedure which needs actuarial judgement is the selection of the average age-to-age factors. Well, it's not humanly possible for us to look at a vast set of numbers, compare and observe the trend in the data.

The only way to do this is the usage of pictorial representations for the same. Thus, heatmaps were generated for each of the ratio triangle that will help in understanding the trend in data and selection of average age-to-age factors.

Here is a sample of a heatmap generated by the program:



(A heatmap generated to visualise the trend in case reserves.)

Bar graphs were also made to understand the ultimate costs of various accident years and identify the trends across them.

"A Picture truly speaks more than thousand numbers do." Thus, visualisations of data and trends were created to improve the understandability.

Predicting the age-to-age ratios using GLM:

In the prior process, we used the standard methods like mean, median and geometric mean to fill in the missing data. However, if there is a trend in the age-to-age ratios, averages might not give the best picture of the results.

Hence, there was an attempt to fit a statistical model to predict the age-to-age factors and see if the results can be improved for the method.

Several statistical and machine learning models were run and tried to fit the ratios. The best fit was obtained by Gamma distribution. Further work can be done to make sure that the predicted values converge to the actual true values, in case there exists better fits.

Datasets which have values that the model can understand and analyse, meaning no zero values in the them, were fed to model and results were obtained.

Training Data:

The training data consisted of all data rows where the development year was less than or equal to 1997 and the columns:

1. Accident year
2. Development lag

were chosen as the features for the model to predict the remaining-in-case ratios and paid-on-case ratios.

The target variables for predictions were remaining-in-case ratio and paid-on-case ratio.

Testing Data:

The testing data comprised of all data rows where the development year was more than 1997, where the same features were chosen as those in the train data set.

The model was made to predict the remaining-in-case ratio and paid-on-case ratio.

Given the ratios by the model, ratio triangles were filled with the predicted values given by the model, and the case outstanding triangle and incremental triangles were subsequently filled.

The actual ultimate costs and ultimate costs obtained by using these statistical tools were analysed.

A total of about 105 companies were trained by the model and the ultimate costs were ascertained using the predicted age-to-age factors.

Observations and topics for discussion:

Related to the usual case outstanding development technique:

A Total of 141 companies were tested using the code and these companies' data as mentioned earlier had been classified as good data sets and ok data sets.

Let's analyse the results based on our classification:

a. Good data sets:

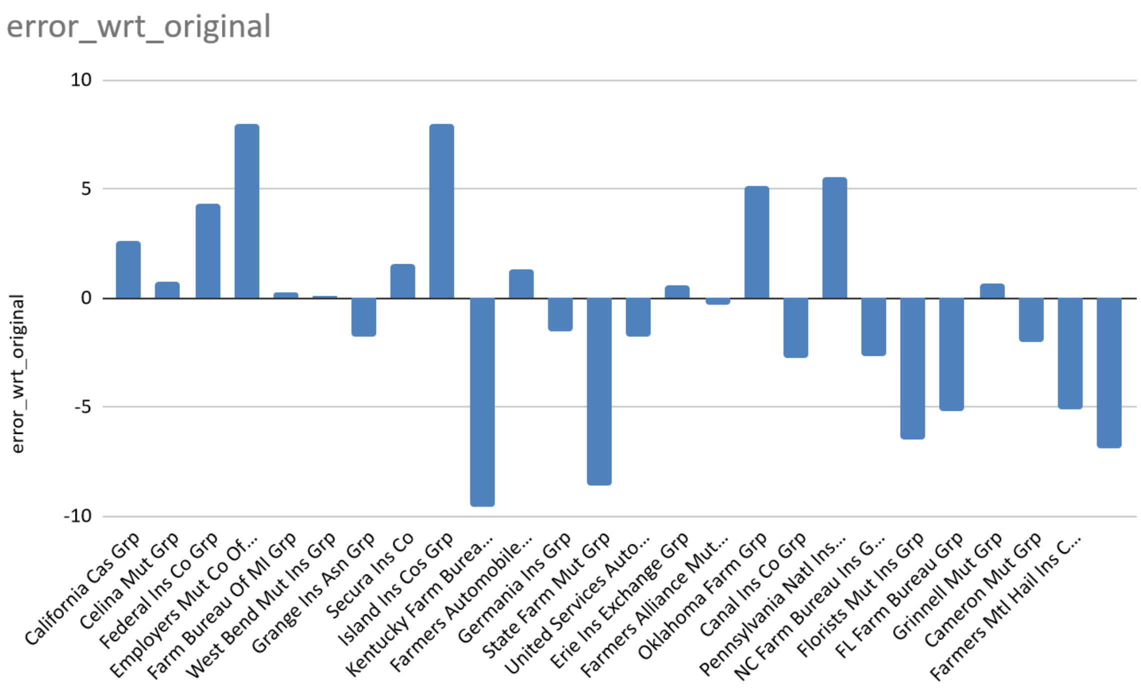
Regarding the good data sets, the error between the method generated ultimate costs and the true ultimate costs is very less and minute for several companies.

After applying standard normalisation on the true ultimate costs and method generated ultimate costs for all the companies, the following are results:

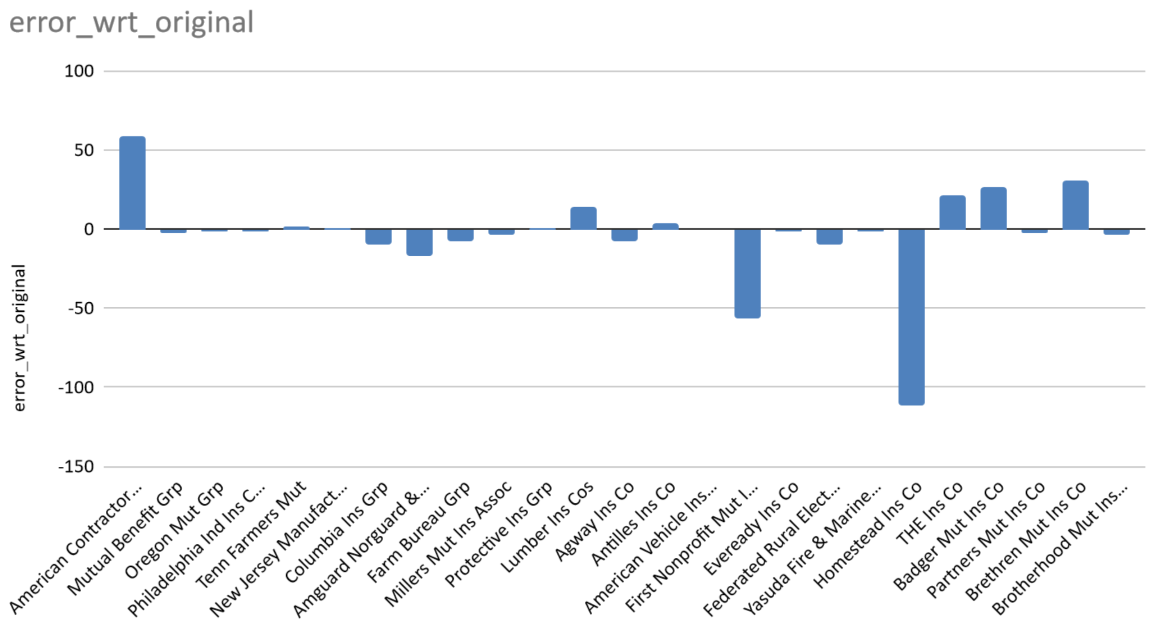
1. Mean_absolute_error: 0.011594325854566394
2. Mean_squared_error: 0.001032960239277601

Almost all companies have an error of less than 10%. 86 of 116 companies have an error less than 10% in their projections. About 75% companies have a less than 10% error. 47 out of 116 companies have an error of less than 1%.

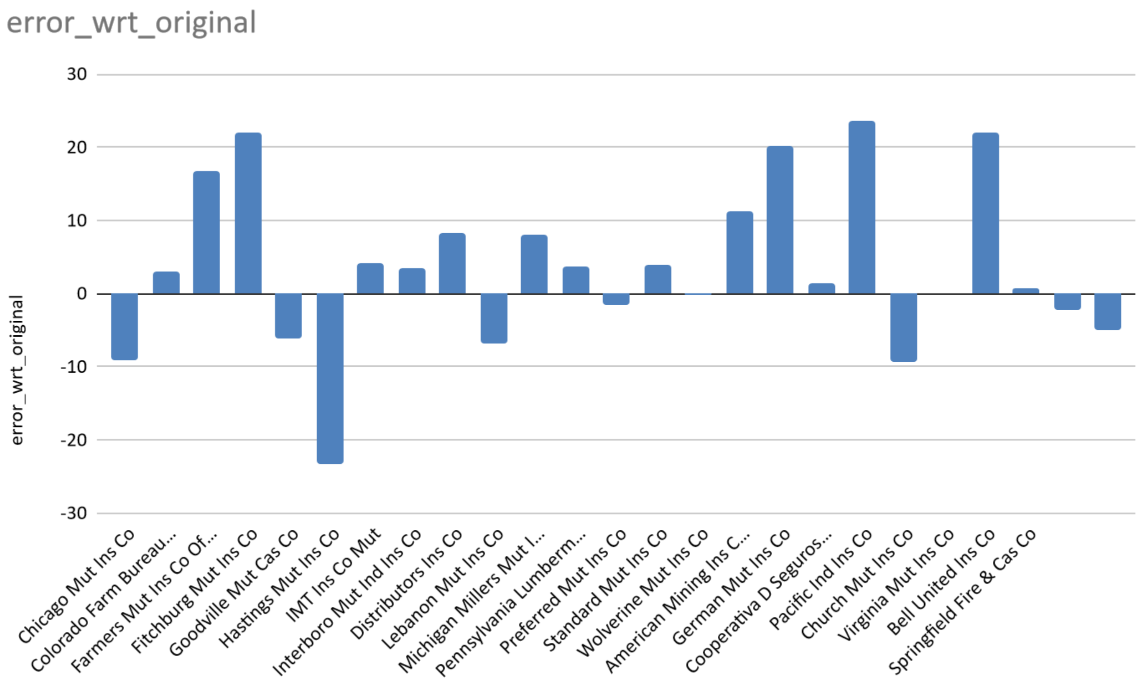
Error Representation for the first 25 companies in percentage terms:



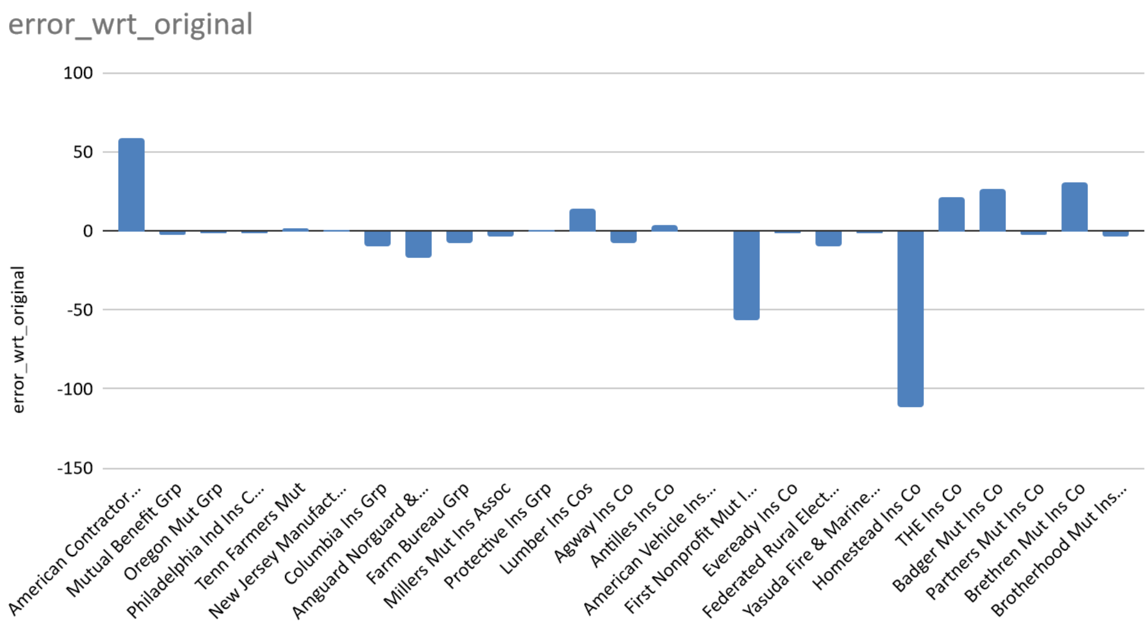
Error Representation for the companies 25-50 in percentage terms:



Error Representation for the companies 50-75 in percentage terms



Error Representation for the companies 76-116 in percentage terms



b. OK data sets:

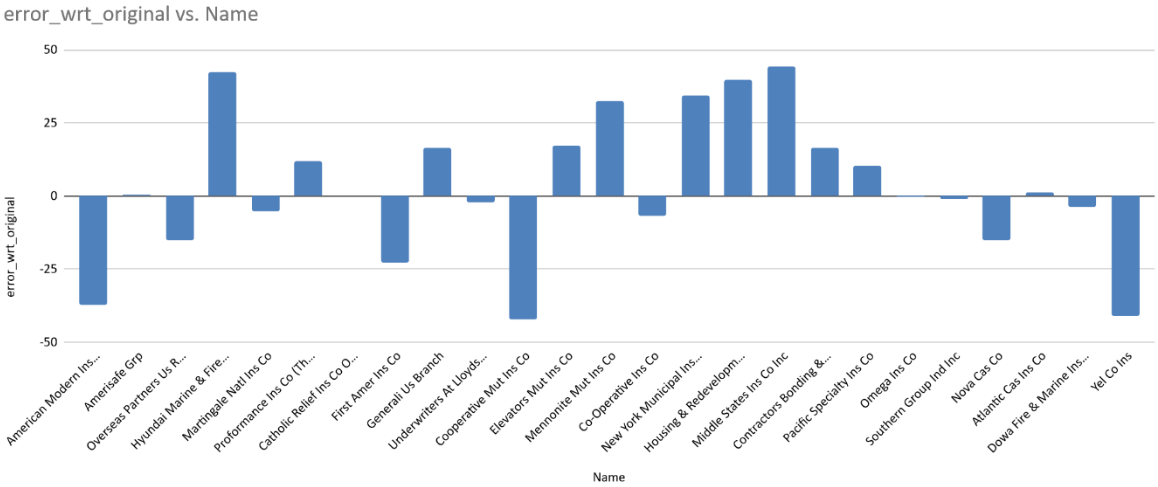
For the Ok data sets, the error between the method generated ultimate costs and the true ultimate costs is also comparatively less.
After applying standard normalisation on the true ultimate costs and method generated ultimate costs for all the companies, the following are the results:

1.Mean_absolute_error: 0.12462294193949244
2.Mean_squared_error: 0.06449797915890235

The number of companies which have a less than 10% error is 9 out 25. The number of companies that have a less than 10% error is 36%.

The number of companies which have a less that 1% error is 4 out 25, which is about 16% of the companies.

Error Representation for the companies 1-25 in percentage terms



Comparisons between Good datasets and Ok datasets:

1. After seeing the above results of the good datasets and ok datasets, we can observe that the predictability of good datasets is more than the predictability of the ok datasets.

2. The reasons why the good datasets beat the ok datasets is because the good datasets have more historical data than the ok data sets. That was how the data was pre-processed.

3. This leads us to an important observation in the process: More data will lead to better predictability.

4. There are some company outliers in both datasets whose error is more than the true ultimate cost.

Related to the usage of predictive modelling in finding the average age-to-age factors:

The main driving force to experiment with predictive modelling was to find more accurate age-to-age factors. The traditional way of trying to use mean, median or geometric mean might not give the best way to look at the true picture of future ratios.

Thus, there was an attempt to use the generalised linear model (GLM) to fit the age-to-age factors. The gamma distribution was used for fitting and predicting the ratios.

The general way of doing reserving is to directly fit the incremental paid claims with the model and find the ultimate costs. But my concern was to use GLM in this particular model to better represent the age-to-age factors.

However, the results could not beat the results obtained by using the technique as it is.

Let's look at a few metrics:

1. Mean_absolute_error: 0.04795088330122375
2. Mean_squared_error: 0.03214898506509198

Potential reasons why predictive modelling did not work out well while fitting the ratios:

1. The traditional way of splitting the dataset generally is a random split. But in our approach the training data set consists of splitting the data into training data where the development year is lesser than or equal to 1997 and the testing data where the development year is above 1997.

2. This implies that there is no trace of development years of the test data set in the train dataset. This makes it difficult for the model to predict ratios accurately as the test and the train dataset are a bit apart and do not have a homogeneous representation.

3. Since our method is interested in finding the age-to-age factors, even a small deviation in the prediction of the age-to-age factors will lead to a huge difference in the final prediction of the ultimate cost.

4. There are only a few features in the train dataset, which is insufficient for fitting the target.

However, there might be better ways to fit the model and get closer results to the actual value, since there is a multitude of regression models available in various libraries to get better results.

This is truly an area of future research where we have an advantage of easily and efficiently validating the results generated by machine learning and reserving techniques.

Conclusions:

The Usual Case_Outstanding_Development Technique:

The case outstanding technique has really performed very well on the commercial auto insurance line. It can be seen from the results that the case outstanding technique is apt for short tail lines of business.

With sufficient data, there may be situations where this technique might give an error of less than 1%. This validates the conviction on this method.

The results however may not be so accurate given insufficient historical data. In these situations, the second approach may be followed which involves using the industry standard age to age factors.

For further faith, the reader may use the excels attached at the end to validate and be convinced about the findings.

Using Predictive modelling on the method:

Upon using the generalised linear model with the gamma distribution, it may be noted that, the results could not beat the results of the case outstanding technique on a macro-level. However, they are companies where the GLM has given more accurate results.

It must be suggested that a better statistical measure or regression model can be found to get better results.

These areas are of future work and experimentation, and require more in depth understanding of data science regression and statistical models and must get the wherewithal to use them in the course of action.

As in the project, the gamma generalised linear model could not beat the traditional method.

For further examination and details, the reader is advised to read the attached excel at the end to see the findings.

The GLM was not applied to OK datasets as they have very less data for fitting and predicting. There is no hope for the GLM to work on these datasets.

Summary of the findings:(after standard scaling the parameters)

Metrics	Method on good datasets	Method on OK datasets	GLM Results:
Mean absolute error	0.011594325854566394	0.12462294193949244	0.04795088330122375
Mean squared error	0.001032960239277601	0.06449797915890235	0.03214898506509198

Off-Shoots of the Project:

Due to the automation and validation of the case outstanding reserving technique, these were the following off-shoots:

1. A package in python was made for calculating the reserves using the case outstanding technique, given that the data is consistent and the reported and paid claims are cumulative in nature.

2. Given the link of the csv file which contains the raw data, the program is capable of generating excel file consisting of all the steps that are done to compute the final ultimate costs and reserves.

3. Predicting the ultimate costs using a statistical distribution which is a new way to obtain the missing age-to-age factors.

4. Generation of graphs for better understanding of the numbers in triangles and trends about how the claims move.

Future Work:

1. To develop better statistical and machine learning models to fit our data and beat the traditional models.

2. In reality there are huge disturbances in some datasets, where these companies are outliers for all our tests. There must be ways to reduce their deviations, though they are minimal in number.

Learnings From the Project:

This project has helped in a dual way.

1. It has helped me to understand how reserving actually happens, and in particular how the case outstanding technique actually works.

2. It has also helped me get hands-on programming experience using classes in python to automate the mundane tasks generally done on excel. This project has scaled my learnings as a programmer in python.

Bibliography:

1.The Reserving manual by Jacqueline Friedland

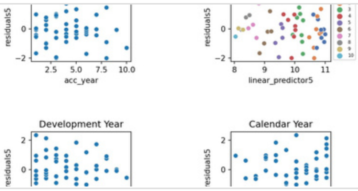
https://www.casact.org/sites/default/files/database/studynotes_friedland_estimating.pdf

2. Blog about machine learning and reserving.

Machine Learning in Reserving Working Party ≣ Reserving with GLMs in Python

We revisit our previous example on reserving with GLMs in R ≣ this time we use Python.

https://institute-and-faculty-of-actuaries.github.io/mlr-blog/post/foundations/python-glm s/index.html#chain-ladder-model



3. Link to dataset (commercial auto)

Loss Reserving Data Pulled from NAIC Schedule P | Casualty Actuarial Society

Glenn G. Meyers, PhD, FCAS Peng Shi, PhD, ASA Please direct all comments to Glenn Meyers at ggmeyers@metrocast.net

https://www.casact.org/publications-research/research/research-resources/loss-reserving-data-pulled-naic-schedule-p

4. An article on medium by Nicholas Misawo

Calculating claims reserve using Python

In this article, I will walk you through a step by step guide on how you can save tremendous time in calculating claims reserve using the...

https://medium.com/@nmisawo/calculating-claims-reserve-using-python-6720e5bf7b50

5. Excels given by the reserving manual of reserving written Jacqueline Friedland.

6. A blog on machine learning and reserving.

Machine Learning in Reserving Working Party ≣ My Machine Learning in Reserving Journey

An ML beginner recounts his first steps into using Machine Learning for reserving

https://institute-and-facul t y-of-actuar ies.github.io/mlr-blog/post/foundations/ml-jour ney/

Links for Code and Resources:

1. First example automated:

Google Colaboratory

https://colab.research.google.com/drive/1eaBT30×0_hmdUjrBcf8sLzyIA3Bf2RWT?usp=sharing

2. Code for data Pre_Processing:

Google Colaboratory

https://colab.research.google.com/drive/16pGbKdzJsak9EeuVgx4xk2xneAtMln5z?usp=sharing

3. Code that generated Excel files for the good datasets:

Google Colaboratory

https://colab.research.google.com/drive/1dzPU21XCBrbApamUUzOhsPvD_XK1peql?usp=drive_link

4. Code that generated Excel files for OK datasets:

Google Colaboratory

https://colab.research.google.com/drive/1n1nCBgDGJG3-sBDex_6phrRaGCBbrWuM?usp=drive_link

5. Links to the library or package created:

Google Colaboratory

https://colab.research.google.com/drive/1iKvBbXx6_XzJAyyrV7OiT3EjJGvkP78X?usp=sharing

6. Code that generated results of machine learning:

Google Colaboratory

https://colab.research.google.com/drive/1rGclu9wkVv97Q6Hhc0E-Snr-aluthmCu?usp=sharing

7. The Excel files of companies generated by code:

   https://drive.google.com/drive/folders/1nRNbaGVV6AdahNM-8Rv75vrBai4j70kN?usp=drive_link

8. Processed data files:

   a. final data_sets for each company

      https://drive.google.com/drive/folders/1DqenTbjgFdeQ6Mvorjg6TWNXTSaUYVPb?usp=drive_link

   b. initial data_sets:(where dev year ≤ 1997)

      https://drive.google.com/drive/folders/1_xoCprXKB1wkDumLWcuC89_FuDsZil7r?usp=drive_link

9. Results:

   a. Results of good datasets:

      https://docs.google.com/spreadsheets/d/16vPtTpKIst8Fd0cIh39r7TsFmxz0hIY1/edit?usp=drive_link&ouid=102915954087936884603&rtpof=true&sd=true

   b. Results of Ok datasets:

      https://docs.google.com/spreadsheets/d/1-J3MLQn1RR-yhRHZWeSLuHT2HPwH0Q4N/edit#gid=975200194

   c. Results using predictive modelling:

      https://docs.google.com/spreadsheets/d/1iMUqBqxrQhSxA3oCsYV6RxKVB8n_I3s9/edit#gid=1966708612