Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

Syed Danish Ali

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

Table of Contents

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

1) Introduction and overview

In this report we cover exploratory data analysis to review recent experience, uncover trends in data. We then proceed with basic ratemaking for sample medical health insurance using burning cost methodology. Advanced pricing is then implemented by modeling 6 Machine learning models. it is hoped that pricing for sample medical portfolio using basic burning cost and advanced machine learning have been able to open up data and business trends and produced valuable actionable insights for the management to implement and monitor over time as part of data driven decision making[1].

Accident and health also have index insurance products. For example, daily hospitalization cash product is fixed cash amount per day given to person upon hospital admission (usually higher cash amount if in ICU than normal hospital admission). A flat benefit is also given upon accidental death which usually serves as funeral policy since the fixed flat amounts are so low. Maternity fixed flat cash is paid upon delivery which is low sum too paid for nutrition and basic expenses. Microinsurance avails these benefits the most since index is simplifies operations and claims handling as well as customer dealing because these products are simple and easy to understand. This is a matter of building trust in population that has not historically relied upon insurance for protection. Make a low promise but deliver upon that and that is better than making a high promise of up to XYZ hospitalization or maternity coverage but then at time of delivery, playing hardball, debating, delaying, asking for too much documentation and nit picking over sub-limits to minimize coverage. And that too during an emotionally testing time for the insured population which is undergoing hospitalization. At that time, they don't have the time, efforts or emotional availability to bargain hardly with micro-managing insurers out to minimize their indemnity coverage. Such experiences leave scars and people are never able to trust insurance again. Hence, 10,000 that gets delivered without hassle is better than 25,000 gotten after tons of infighting and delays. Reliability and trust building matters.

Moreover, there is continuous innovation targeted in this accident and health segment for index insurance. People are realizing that key farmer becoming sick is as detrimental to agricultural earning and ability to repaying the agricultural micro-loans as having a natural catastrophe so bundling agricultural loans with daily hospitalization cash as simple easy and affordable index product is becoming more prevalent over time. Since this product acts as loss of income cover, other countries have made heat-wave index insurance as well where upon touching a set temperature with humidity (known as wet bulb phenomenon), small flat amount is paid directly to daily wage laborers and SMS sent to their mobile phones because it is assumed that they won't be able to work under such conditions and will likely be suffering from heat stroke and will lose their daily wage. During COVID19, a lot of insured population was disappointed that daily hospital cash should've functioned as loss of income cover for them where they recovered from COVID19 by staying at home instead of getting admitted in hospitals.

---

[1] Please note that "we" is just mentioned as Convention throughout this report. The sole author of this report is Syed Danish Ali.

Index for Accident and health is interrelated and simple; no need for complex indices modeling, satellite data, harvesting sample crops for yield for yield index insurance etc; helps people understand insurance better. And let's not talk about basis risk; people are used to it; anything is better than nothing; even the usual government support only covers a small proportion of their actual losses. Structure for indemnity does not exist outside the top cities of most countries so index is needed to create markets for insurance here. Even indemnity has huge basis risk called protection gap; for example, cancer average claim size is usually just 10% of the actual economic loss in private health insurance which doesn't make much sense because actual costs all well over 100 times this amount. So why is it so low when actual is so high? Many reasons for that such as indemnity health only covers up to limit and most of plans are for low paid staff which have low limits like 100,000 per annum hospitalization coverage. There are also many sub-limits restricting this further. Cancer is also not one episode of acute disease and one hospitalization but its treatment is spread over few years mostly and a lot of outpatient expenses are needed for it that person pays from his/her own pocket as health insurance only covers inpatient.

Whether it's normal indemnity coverage for earthquake or an index of earthquake parametric coverage, there is strong link with insurable interest to cover their losses and not provide gains and hence must be held as insurance (unlike cat bonds which have history of giving attractive returns to investors). It's not like just having the index means 0% link with insurable interest.

We also have to take the intention behind preferring index insurance to indemnity insurance in certain regions and parts of the world. The intention is to offer some sort of insurance risk mitigation in areas where it is unfeasible to launch indemnity insurance due to the many associates' challenges with it such as lack of infrastructure, high natural catastrophe risk, lack of pre-existing insurance coverage and many more. If that region develops enough over time, then of course we must transit from index to indemnity insurance products in that region. Thus, index insurance might not be the insurance that we deserve but is the one that we need right now for such regions in other to develop them further. Being blind to this intention behind index insurance and terming it as derivative fails to capture the essence behind the efforts in the first place because index is insurance by necessity and not an insurance of choice but intention behind it is the same; to offer some sort of insurance to previously uncharted markets. If there isn't any curative pill to COVID19, we don't say we have zero healthcare, we utilize vaccines which are not curative medicines of choice but part of healthcare system by necessity. Similar principle here for index insurance.

The other argument for index products to be categorized as insurance is that such products need to be regulated. Of-course index products on macro level with international agencies and governments can be immune to regulations because the products there will be developed on solid basis but there needs to be regulation for index products that are to be made locally by market players as well. Regulations force the maker of index product to first due all of it's due diligence and proper homework and provides protection to customers in case things go wrong. Both of these functions are essential otherwise some local poorly made products can benefit the makers to the detriment of consumers and lead to widespread trust issues over index insurance which can imperil any future efforts to launch index products as well. Banking and derivatives regulations are also not comprehensive as much as insurance regulations.

We can also quantify insurance risk and financial risk as proportions of total risk over number of years of historical data or simulated data. If insurance risk is average more than 10%, we can classify that index product as an insurance product; however, if a product has very high basis risk and insurance risk less than 10% on average, that index insurance is not insurance. Another point to note is that it doesn't have to be 100% specific to a party to the contract (it can be

50%-80% specific because in many index products, there isn't a single trigger; there are small triggers that ensure that small payments are made to reduce basis risk and they are increase the proportion of being specific to a party. Hence, products with multi-triggers (small losses, medium losses, high losses) that pays even when losses are small (even though payments are partial) can be considered insurance product. Hence, my conclusion is that index insurance needs to be looked separately from cat bonds on a case-to-case basis to know whether it's an insurance product or not rather than blankly stating that all index insurance is insurance or non-insurance. This is useful because in any regulations, we see existing debates on whether regulator should classify index insurance as insurance or a derivative.

A rate is an estimate of the expected value of future costs. A rate should be reasonable and not excessive, inadequate or unfairly discriminatory. The goal of ratemaking is to determine rates which will provide sufficient funds to pay expected losses and expenses; maintain an adequate margin for adverse deviation; and produce a reasonable return to shareholders.

Future Premiums = Future expected losses + Future expected expenses (both variable e.g., commissions & fixed costs) + Provision for Adverse Deviation + Profit Margin

Pricing first determines the Expected Losses and then adds loadings for expenses, profit & PADM to work out the final premium.

Expected losses = Expected No. of claims x Avg. Amount of claims

Burning cost = Frequency * Severity

Frequency = no. of claims / no. of lives insured

Severity = claims paid/ no. of claims

Loss ratio = claims paid/incurred claims

The base cost is determined from evaluating various risk features such as:

- Class of business (Hospitalization, major medical, index, critical illness, maternity, outpatient, pools)
- Age
- Gender
- Length of stay average
- Relationships (Son, daughter, mother, father, own self male, own self female, married
- female, husband and wife)
- Room Rent

- Inflation
- Expense loading
- Profit loading
- Margin for adverse deviation

It is important to realize the business portfolio composition of Group Health Vs individual health at an insurer. In retail space, data and products are usually more homogenous and if large portion of business is in individual retail, that lends itself to more credible and stable data analysis. However, in group health, the dynamics are different. There is significant concentration risk which means that for example, one jumbo group's adverse experience on parents for one year can lead to overall parents pricing factors increasing for overall portfolio for PGS. It also means that adverse experience in just 1-2 jumbo group can mean the company changing from small net profits to sizeable net losses. For example, one jumbo group's high losses on parents can mean that for year overall parents for shows high burning cost on total portfolio. Nowhere is this clearer than when comparing top diagnostics over years to evaluate inflation. The diagnostics over same hospitals over the years should show stable change which reflect the average medical inflation but due to jumbo groups entering and leaving over the years, this change can sometimes be 40% and sometimes be -10%.

Concentration risk also means that while the company might usually underwrite groups well within 300-400 or more groups in a Policy Year, the top 50 jumbo groups will likely account for about 70%-75% of the total Gross Premium for the company. Generally, for groups that are reasonably large like the top 100 groups, 100% credibility is given to client's own experience in experience rating and 0% weightage to generic book premium rates. For rest of the groups, price is blended with some credibility given to generic book rates and some weightage/credibility given to client's own experience. Where claims are too few as in small groups, full credibility is given to generic book rates since client's own experience can be very volatile/unreliable over time for small groups like 0% in one year but 200% in the next year just because of one employee contacting cancer in that small and medium (SME) company.

Given the practically limited role generic book rates plays in overall business results, it is important not to give over-importance to generic book rates and think that it is leading to huge impact on financial performance of the company. Moreover, generic book rates is only a model and a model cannot take a lot of unique factors that are on the ground realities for each jumbo client on a case-to-case basis. It is also an indicative premium rate and is not binding to be the final rate. The purpose of analytics is to aid in the sales decision-making process on quoting prices, and not to replace it. Management should not ask for generic book rates to take 30-50 factors into account and realize that generic book rates has no impact on majority 75%-85% of the premium.

Only by coincidence will pricing model align on their own with market/competitor premium rates. Every market player has their own pricing policy and expecting pricing model to align with it is an impossible expectation. Decision making cannot be delegated to a model or replaced by a model.

Therefore, it is important to realize that final premium prices quoted by insurer to customers are not just pricing model rates given automatically but goes through experience rating process to arrive at blended prices and that price itself is just the technical price from actuarial side. After that, technical discount is given to actuarial function. After that, the desired price might still not be arrived at especially for jumbo groups which have a lot of bargaining power and so commercial discounts given by Sales Director/COO/CEO becomes necessary to retain them. These multiple layers mean that technical risk considerations from actuarial function on client's actual experience is taken as well as commercial considerations by management of the company.

Different companies will have different 'break-even' best possible prices and that's why the rates by one insurer's pricing models shouldn't be expected to align highly to market average rates. Other competitors have more profitable business to cross-subsidize the highly competitive health insurance line of business with life insurance and non-life portfolio. Expenses can also be allocated more to profitable lines of business than health by other insurers. One insurer can under-quote to gain market share but then reject claims more and delay claim payments more to make up for it. The other insurer can under-quote initial policy but have higher prices on endorsements. Yet another insurer can price lower because it has a more profitable and higher top-line revenue on health insurance so it can afford to cross-subsidize loss making new business from profitable renewal business. Another insurer might have higher proportion of profitable individual health and micro-health segment so it would under-quote for group health. Another insurer might be new and so would have adequate capital which it will use to gain market share by underpricing aggressively.

Pricing models take only few major risk considerations whereas commercial considerations are also very important and taken into account by the management separately.

Example of commercial considerations are:

- Pricing based on elasticity of the customer; some customers might leave insurer if 10% price is increased but other customer might not leave at 10% price increase so that customer is given 10% price increase whereas the other person is not given that 10% price hike so as not to lose that customer.
- Cross-subsidizing loss-making business like done in health with other profitable lines like general and life insurance. Reputation of health as a loss leader to get one foot in the door is widely practiced. Then other profitable products are up-sold to the same customers.
- Cash flow underwriting. This is the practice where investment returns are relied upon to convert underwriting losses into net profits.
- Universal practice of price walking where new customers is gained on excessive under-pricing and then their prices are increased gradually upwards over time (this is called loyalty penalty).
- Underwriting loss-making clients because they are jumbo groups and they can be cross-subsidized with profit making groups and help the insurer meet their top line revenue targets. No matter the size of the customer, giving allowance to bargaining power should be up to a limit and not to unsustainable levels.
- Nit picking the contracts and fine prints in order to deny claims or delay claims and over charge hospital bills etc.
- Population that is more scattered and has lower awareness
- And many more

2) Exploratory Data analysis

From analyzing premium generally as % of sum insured over past 6 years, we usually see business underwriting cycles ups and downs over the years and that COVID19 seems to have triggered a shift from soft underwriting cycle to hard underwriting cycle in health insurance.

Loss-making onerous policies should be shown separately from profitable business to aid in pricing analysis. This is also required under IFRS17. The policy of recognizing onerosity is a policy to have above 100% combined ratio for each of top 50 groups. There are groups that have average of above 100% combined ratios but are still under-written to maintain top line revenue and is a universal practice.

The reason for the low onerosity of 2% as % of total GEP conventional (exc ASO) in FY2020 is because of COVID19 lockdowns, supply chain restrictions, deferred healthcare utilization from policyholders and increase in awareness of importance of health insurance leading to increase in premiums (increase in premium + decrease in claims = double effect).

There were abnormal profits in 2020 which is seen by rapid reduction in loss ratio in 2020. This was due to lockdowns and decrease in economic activities. 2020 premium did not fall proportionately to claims because by Dec 2019-Mid March 2020, most of the group business had renewed their business but COVID19 shocks only started after Mid-March 2020 and onwards.

The reason for increase in onerosity from 2% in 2020 to 8% in 2021 is because of high medical inflation, supply chain restrictions till half of 2021 third/fourth wave of COVID19, stagflation, pent up deferred healthcare getting utilized and pressure on premiums upon renewals due to low losses in 2020.

In 2021, we saw clients putting pressure on reinsurer that reinsurer made abnormal profits in 2020 and so the insurers in 2021 deserve cuts in prices. Thus 2021 pricing pressure, resumption of economic activities and inflation saw loss ratio increasing rapidly.

Demand for OP has been accelerating since the start of COVID19 to present as employees put pressure on employers to provide OP as part of employment benefit to reduce their out-of-pocket expenditure on health as COVID19 massively increased awareness and importance of maintaining adequate health in the population.

Health insurance is a highly competitive line of business which means that net profit % is usually very low at around break-even levels and it is a challenge to avoid net losses every year by focusing on underwriting and claim controls.

There are 2 types of Self-insured groups, 1) self-insured that have no in-house reimbursement for health claims and are shifting to professional insurer for the first time and 2) self-insured that have in-house reimbursement for health claims over a number of years and are now shifting to professional insurer for the first time. Both 1 and 2 self-insured groups are high risks but 2 is higher in risk because the level of existing awareness in employees is high. There's a difference that we should recognize though; in one instance, company itself might have no health coverage policy and for the first time they might come to us for insurance; in that case, awareness will take time to scale up; the second instance is when self-insured health coverage had been provided by the company itself for a number of years and then they come to us for health insurance. In that case, the awareness will already be high and there will be more chances of utilization and claims doubling.

Research shows that accidents are mostly caused by drivers' young males 18-25 years. That means that female claims for traffic accidents should be very low relative to male and especially young males.

82% of traffic accident claims are for Male and 18% for Female. This is same as what we would expect:

Gender	% by Number of claims	% by claims paid	Severity RF
Female	18%	18%	1.01
Male	82%	82%	1.00

Now we see results by gender and age:

Gender	Traffic Accidents only	% by Number of claims	% by claims paid	Severity RF
Female	01. 0 to 18	3%	3%	0.9
	02. 19 to 29	4%	3%	0.7
	03. 30 to 39	3%	3%	1.0
	04. 40 to 49	2%	2%	1.0
	05. 50 to 59	2%	3%	1.5
	06. 60 and above	4%	4%	1.2
Male	01. 0 to 18	10%	8%	0.8
	02. 19 to 29	23%	24%	1.0
	03. 30 to 39	14%	13%	0.9
	04. 40 to 49	11%	12%	1.0
	05. 50 to 59	8%	8%	1.0
	06. 60 and above	15%	18%	1.2

We can clearly see the age-gender effect here for traffic accidents. 19 to 29 young male have the highest total claims paid and number of claims. However, even for female we can see highest number of claims in 19-29 years of age. Why is that? Probing into data further, this is because females are there mostly as passengers and so when accidents increase for young male drivers, young female passengers also get injured in the same vein. Accidents then increase for male again for above 60 years which is to be expected as old age leads to higher accidents. Again, this same bump is repeated in female above 60 because they suffer accidents from male 60 years as passengers too. Older age claims have higher severity then claims at lower ages too but severity for female 50-59 years is higher than for male in 50-59 years age bracket.

when we filter for principal ownself only, the ratio of accidents for male increase further because passengers are not female as in all relationships 94% male 6% female.

overall without relationship filter, average claim size of female and male are equal over ages as aggregate; but with filter for ownself, male claim average is 24% higher than female.

Similarly, for COVID19 only claims, we can see that male form 70% of total claims paid and female only 30%:

COVID19 only	**% by number of claims**	**% by claims paid**	**Severity RF**
Female	31%	30%	0.95
Male	69%	70%	1.02

Age effect is also strong here that 32% of total claims paid is for above 60 years. However, there is one two-way effect also occurring which is that for Female, severity is high for above 50 years of age whereas for Male, severity is high for above 60 years of age.

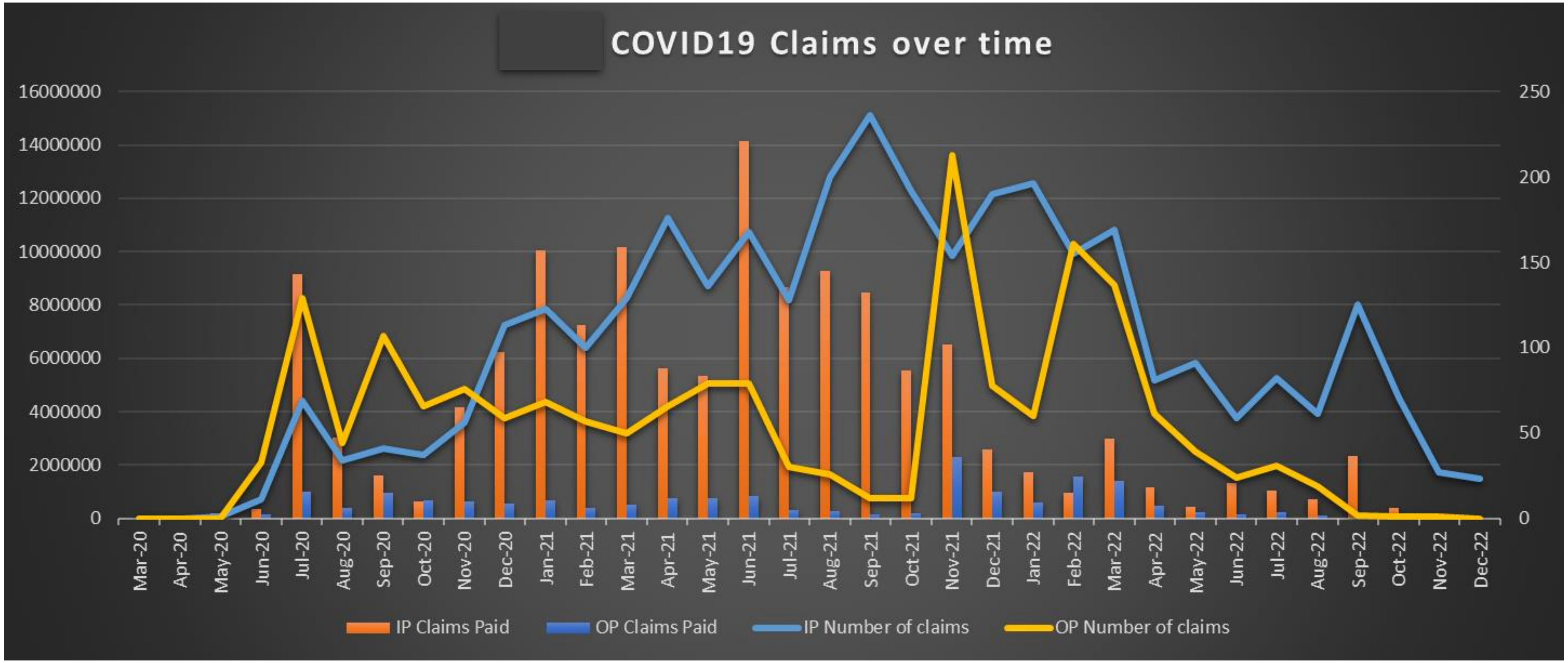
Gender	**Ages**	**% by number of claims**	**% by claims paid**	**Severity RF**
Female	01. 0 to 18	1%	1%	0.54
	02. 19 to 29	7%	2%	0.38
	03. 30 to 39	6%	4%	0.75
	04. 40 to 49	4%	3%	0.91
	05. 50 to 59	2%	3%	1.58
	06. 60 and above	11%	17%	1.83
Male	01. 0 to 18	2%	1%	0.85
	02. 19 to 29	13%	5%	0.46
	03. 30 to 39	13%	6%	0.53
	04. 40 to 49	12%	10%	0.96
	05. 50 to 59	8%	6%	0.99
	06. 60 and above	22%	42%	2.22

Plan wise effect is there; A and B are 74% of claims paid; rest are 26%; this is because plan A and B are more likely to seek hospitalization healthcare utilization upon COVID19 diagnosis while lower cadres will seek to recover more from their homes. male are more than female. Then plan wise, gender effect is there; 70% are male and 30% female.

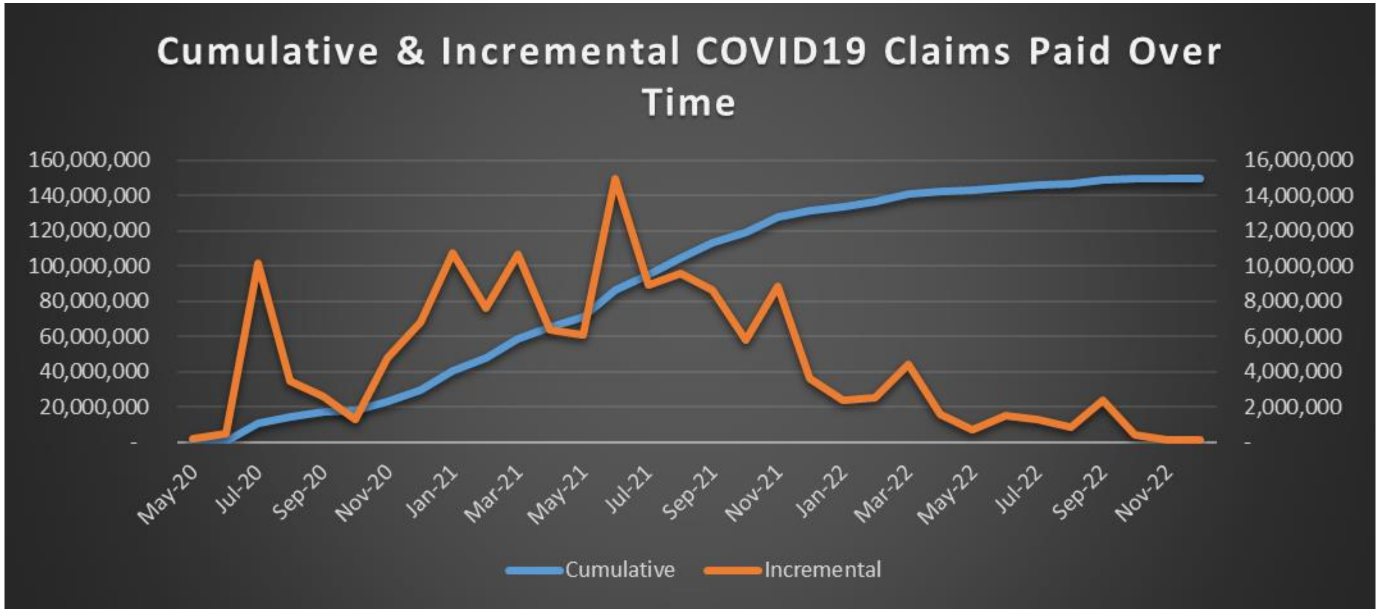
Plan Type	% claims paid Female	% claims paid Male	total claims paid
A	30%	70%	38%
B	33%	67%	35%
C	27%	73%	17%
D	31%	69%	6%
E	11%	89%	3%
F	22%	78%	0%
G,H,I,J,S,V	28%	72%	1%
Grand Total	**30%**	**70%**	**100%**

That means that higher the plan and sum insured, higher the chances of claims for covid19; plan and gender interaction is there because proportion of female and male across plans are very similar to each other; this is because if male gets hospitalized under covid19 in plan A then if his daughter or wife or mother gets covid19, they also get hospitalized under plan A and not let's say plan B or C.

If we see the cumulative claims movement over time to be able to assess how much COVID19 claims paid are as % of total claims paid than it shows an increasing trend over time from 2019-2021 and then decreasing trend from 2022:



Cumulative are total COVID19 claims paid so far and incremental are COVID19 claims paid in that particular month only. On average, cumulatively COVID19 forms on average 2.77% of total claims paid and incrementally it forms on average 2.69% of total claims paid which indicates a moderate impact. The increase in loss ratio in 2020 due to COVID19 claims was 2.19%, in 2021 it peaked at 5.88% and in 2022 it was a small 0.92%. COVID19 claims spiked in July 2020, Q1 2021, June 2021 and then started long term decrease from June 2021 to December 2022 due to high rates of vaccination and group herd immunity.



OP business has rapidly increased from Q3 2020 to present in response to employees demanding employers to pay for their COVID19 tests and vaccinations. Also, notion of seeping claims is relevant here; many small cases like for example, person had stomach flu, went to see a doctor, doctor prescribed tests worth 8 thousand; person said I have health card but I don't want to spend my own money; doctor says no problem; I will give IV antibiotics instead of oral and that will make it medically necessary to admit you in hospital for 1-3 days than it will be cashless coverage for you. Better to get your pre-existing illnesses looked at while you are at it too. Hence, these OP claims seep and become IP claims instead. After COVID19 March 2020, seeping claims possibility has reduced to a market extent due to pressure on healthcare supply (hospital rooms), lockdowns, fear of catching infection in hospital etc. so the claims that employees would make are reduced now hence greater pressure for OP. The seeping claims are quantitively evidenced as follows:

Length of Service	2017	2018	2019	2020	2021	2022	Grand Total
FEMALE	7.31	5.86	6.38	8.22	5.22	5.17	6.44
MALE	8.05	7.38	9.09	14.63	4.04	4.04	8.44
Grand Total	**7.65**	**6.54**	**7.54**	**11.08**	**4.78**	**4.76**	7.29

average confinement	2017	2018	2019	2020	2021	2022	Grand Total
FEMALE	1.31	1.37	1.51	1.50	1.49	1.45	1.43
MALE	1.24	1.30	1.46	1.49	1.46	1.39	1.37
Grand Total	**1.27**	**1.34**	**1.48**	**1.50**	**1.48**	**1.42**	**1.40**

Increase in Length of Service and average confinement in 2020 shows the clear aspect of lack of seeping claims there due to COVID19 in 2020; small claims came less and only serious cases came more so this increased (serious claims require more number of confinements on average than non-serious small claims). Seeping claims went back to normal levels in 2021 and onwards.

Investigating by relationship, we can see that COVID19 is much higher proportion of total claims paid of 57% for ownself male than for 36% of total claims paid for all diseases. Wife proportion is lower relatively in COVID19 than all diseases:

% of claims paid	COVID19 only	All Diseases
ownself male	57%	36%
wife	16%	21%
father	8%	15%
mother	6%	11%
married female	5%	5%
husband	4%	4%
ownself female	2%	4%
son	1%	2%
daughter	1%	2%

Diabetes is also another relevant example to explore.

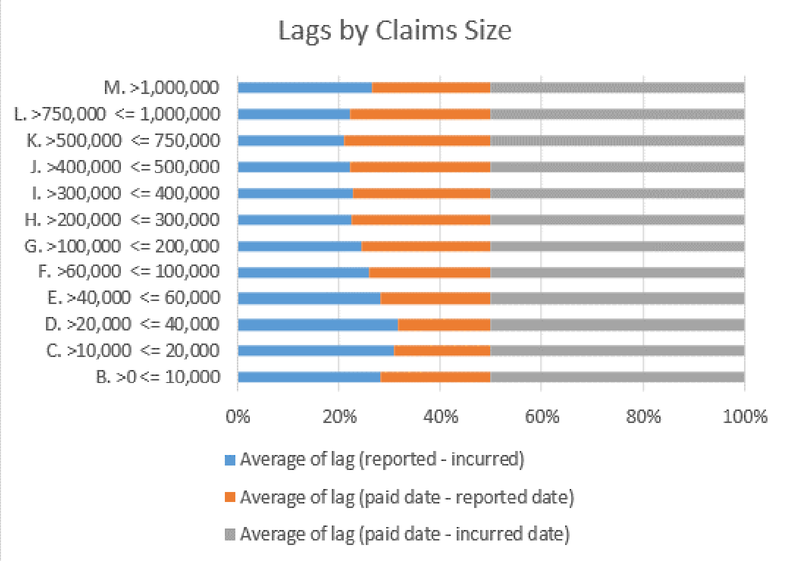
Diabetes	% by number of claims	% by claims paid	Severity RF
Type 2 diabetes mellitus without complications	88%	79%	0.90
Type 2 diabetes mellitus with complications	2%	5%	2.48
Gestational diabetes	2%	4%	2.81
Diabetic foot complications	1%	3%	4.84
Type 1 diabetes mellitus without complications	6%	3%	0.54
Diabetic Retinopathy	1%	3%	2.33
Type 1 diabetes mellitus with complications	1%	2%	2.82
Neonatal Diabetes Mellitus	0%	1%	2.32

Diabetes	% by number of claims	% by claims paid	Severity RF
Without Complications	94%	82%	1.00
With Complications	4%	12%	3.24
Maternity and newborn	2%	5%	3.09

94% claims are without complications of Diabetes. Average claim size of with complication is 3.24 times higher than with average claims for non-complications. 3.09 times higher for maternity and newborn.

This below shows that claims are getting reported more quickly as claim size increases; it also shows that it takes longer on average to pay the reported claims as claim size increases; This agrees with common practical sense. Larger claims are more complex, life-threatening and are given more priority by claimant as well as hospitals to be reported but then they take more time to get paid and settled since more complications have to be evaluated through more checks and balances as well as higher length of stay in complex claims.

Claim Size	Average of lag (reported - incurred)	Average of lag (paid date - reported date)	Average of lag (paid date - incurred date)
B. >0 <= 10,000	41	31	72
C. >10,000 <= 20,000	50	31	81
D. >20,000 <= 40,000	51	29	80
E. >40,000 <= 60,000	42	33	75
F. >60,000 <= 100,000	37	34	71
G. >100,000 <= 200,000	34	35	69
H. >200,000 <= 300,000	29	36	66
I. >300,000 <= 400,000	33	39	71
J. >400,000 <= 500,000	29	37	66
K. >500,000 <= 750,000	34	47	81
L. >750,000 <= 1,000,000	40	51	92
M. >1,000,000	33	29	62



Claim escalation and highlighting outliers (Both above average and below average) can be decided on automatic basis using statistics of percentiles. We can make model to see which claims are escalated or not and by how much automatically.

Low Trigger we have set limit as less than 30^{th} Percentile. Above average outlier limit has been set as above 75^{th} Percentile. No trigger is there for claims between 30^{th} -75^{th} Percentile. Grand total is addition of these 3 and is what data shows without any outlier detection.

Percentile is decided by arriving at mean and standard deviation assuming normal distribution for each procedure.

The key results are after seeing total of 5,541 procedures:

	above 75th percentile	less than 30th percentile	claim above 30, less than 75	Overall best estimate
	high trigger	**low trigger**	**no trigger**	**Grand Total**
Severity	2.23	0.46	0.82	1.00
no. of claims paid	20%	27%	53%	100%
total claims paid	44%	12%	44%	100%

This shows that severity is 2.23 times that of total severity for high trigger, 0.46 for low trigger and 0.82 for no trigger.

High trigger excludes 20% of number of claims but 44% of claims paid.

Low trigger excludes 27% of number of claims but 12% of claims paid

No trigger forms 53% of total number of claims and 44% of claims paid.

These metrices can be revised so that to make them more sensitive so that they exclude only the highest 10% claims and lowest 10% claims.

3) Machine Learning Methodology Key Notes

Machine Learning Predictive Models for Rate of Incidence/Severity/Burning Cost or Pure Premium can be built to accurately predict the rates. The usual range of models are generalized linear models, decision trees, random forests and boosted trees.

These models are built on the principle of bias-variance trade-off. The training dataset is used to train the models, and the prediction accuracy is validated on the test dataset. The training and test datasets are generally created by splitting the insured population data into 75% and 25%, respectively. Over-sampling technique is to be applied to the training data to balance the number of observations for the class of interest. Stratification of the training/test split is based on both datasets having rate of incidence equal to the entire population.

Training/Test Split and Bias-Variance Trade-off: To begin our model-building process, the insured population data is to be split into two sets, the training set and the test set. This is done to analyze the prediction accuracy and the prediction stability of our models. As the names suggests, the model/machine is trained/built on the training set and its prediction accuracy is validated on the test.

Models are selected on the principle of bias-variance tradeoff. In simple and easy terms, bias can be defined as the overall incorrect predictions for a model. Machine learned predictive models usually have a lower bias on the data they were trained upon and a higher bias on other data.

The model which is selected as the most accurate is one such that it has low bias on the data it was trained and the bias it produces on other validation datasets is not much different than the bias shown on the training set. The concept of selecting a model that produces minimal bias on the training set while also has low variation in bias on other datasets is known as the bias-variance trade-off.

Statistical Models for Rate of Incidence: The modelling methodologies for some of the most common predictive models are explained as follows:

1. ***Decision Trees:*** Decision trees provide a series of classification rules based on predictor variables that eventually lead to a final predicted value, or classification as in our case. Decision trees built for classification target variable are known as classification trees. In decision trees, recursive binary splitting criteria is adopted. Each spilt created is based off just one predictor variable. Further, each split is selected such that it produces the smallest impurity in the data node. Impurity can be defined as the difference between the actual classification versus the predicted classification of the split. This way the most important predictor variables based on most important values come automatically to the top in a decision tree while the lesser important ones are filtered out. A decision tree is a tree like collection of nodes intended to create a decision on values affiliation to a class or an estimate of a numerical target value. Each node represents a splitting rule for one specific Attribute. For classification this rule separates values belonging to different classes, for regression it separates them in order to reduce the error in an optimal way for the selected parameter criterion.

Maximal depth is the depth of a tree that varies depending upon the size and characteristics of the data. This parameter is used to restrict the depth of the decision tree.

2. ***Random Forests:*** Random forests are based of many decision trees. The algorithm for each tree is based on the same criteria as explained above, however, the datasets considered for each tree are created by random bootstrap sampling. This means that each observation that is selected to be included in dataset replaced in the original data for re-selection. Random forest requires far more computational than a single tree and are usually much more accurate as well.
3. ***Gradient Boosted Trees:*** Boosted tree work on the principle of sequential learning. They are different from random forests in their approach. The boosting approach builds a tree based on a previous decision tree minimizing residual (bias) of the previous tree. The iteration continues and new tree are built until a certain level of accuracy is reached. Like Random Forests, boosted trees also require far more computational than a single tree and are usually much more accurate as well. Boosting is a flexible nonlinear regression procedure that helps improving the accuracy of trees. By sequentially applying weak classification algorithms to the incrementally changed data, a series of decision trees are created that produce an ensemble of weak prediction models.
4. ***Deep Learning:*** Deep Learning is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, dropout and L1 or L2 regularization enable high predictive accuracy. Each compute node trains a copy of the global model parameters on its local data with multi-threading (asynchronously), and contributes periodically to the global model via model averaging across the network. The activation function (non-linearity) to be used by the neurons in the hidden layers.
    a. Tanh: Hyperbolic tangent function (same as scaled and shifted sigmoid).
    b. Rectifier: Rectifier Linear Unit: Chooses the maximum of (0, x) where x is the input value.
    c. Maxout: Choose the maximum coordinate of the input vector.
    d. ExpRectifier: Exponential Rectifier Linear Unit function. Epochs means how many times the dataset should be iterated or streamed. The implemented Adaptive Learning Rate Algorithm (ADADELTA) automatically combines the benefits of learning rate annealing and momentum training to avoid slow convergence. Specification of only two parameters (rho and epsilon) simplifies hyper parameter search.
5. Support Vector Machine (SVM): a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mapping used by the SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel

function K(x,y) selected to suit the problem. The hyperplanes in the higher dimensional space are defined as the set of points whose inner product with a vector in that space is constant.

6. ***Generalized Linear Models:*** Generalized linear models or GLMs belong to the linear model family. They provide a simple equation connecting the target variable to the predictor variables via a link function. Each predictor variable is assigned a coefficient that explains the magnitude and direction of the effect of that predictor on our target. To account for non-linear relationships, we can factorize the variables. Collinearity between numeric variables is to be also checked. Generalized linear models (GLMs) are an extension of traditional linear models. This algorithm fits generalized linear models to the data by maximizing the log-likelihood. The elastic net penalty can be used for parameter regularization. Gaussian family was used in our case as BC is a numeric label. Iteratively Reweighted Least Squares Method (IRLSM) is the solver on GLM regression method by default. Other solver methods are L_BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm), Coordinate descent, coordinate descent naïve, gradient descent likelihood and gradient descent squared error.

Deep learning for regression is a type of machine learning that uses artificial neural networks to predict a continuous output value (i.e. a real number) given a set of input features. The neural network architecture used in this task is usually a feedforward neural network, which consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the input features, and each hidden layer applies a non-linear transformation to the input using a set of learnable parameters (weights and biases). The output layer produces the predicted output value.

A perceptron is a simple type of artificial neural network that is used for binary classification tasks. It consists of a single layer of input nodes connected to a single output node. The output of the perceptron is determined by applying a step function to the weighted sum of the inputs. Perceptrons can also be used as building blocks in more complex neural networks, such as multi-layer perceptrons (MLPs). In deep learning, a perceptron is a type of a single layer neural network. In deep learning for regression, the goal is to train the neural network to minimize the difference between the predicted output value and the true output value. This is typically done by using a loss function, such as mean squared error, which measures the difference between the predicted output and the true output. The network is trained by adjusting the weights and biases so as to minimize the loss.

The basic idea behind SVM for regression is to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest data points from each class. The margin is maximized by finding the support vectors, which are the data points that are closest to the hyperplane. These support vectors determine the position of the hyperplane.

The SVM algorithm for regression uses a different loss function than deep learning, called epsilon-insensitive loss function, which allows some tolerance for the difference between the predicted and actual values. This is useful in scenarios where the errors may be large, but the number of errors is small.

5 types of errors were evaluated as part of model diagnostics:

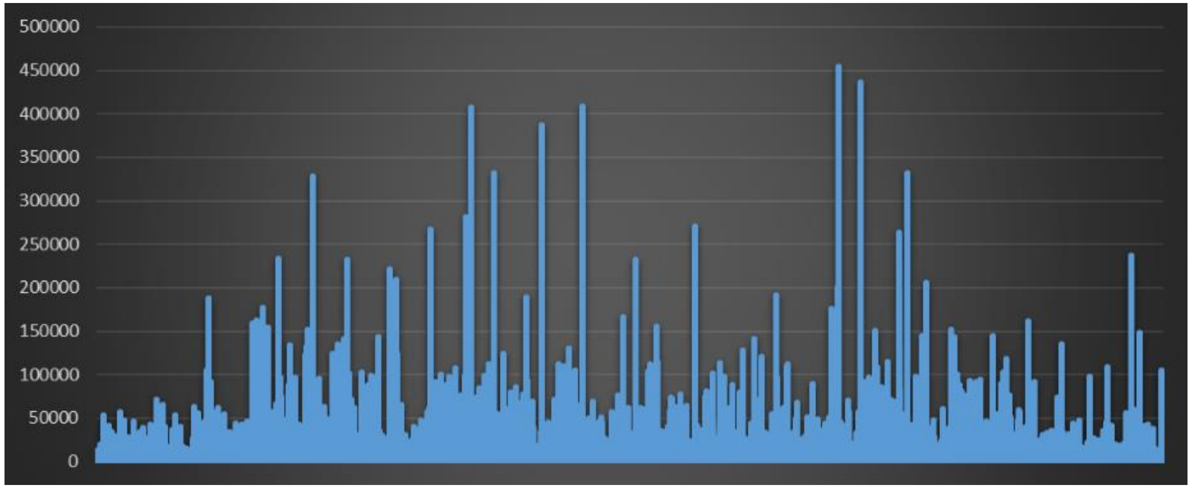
- The average relative error is the average of the absolute deviation of the prediction from the actual value divided by actual value.

- Root mean squared error is the averaged root-relative-squared error.
- Absolute error is the average absolute deviation of the prediction from the actual value.
- Squared error is the averaged squared error.
- the correlation coefficient estimates the correlation between the label and prediction attributes.

Each of the 6 models are optimized before being put to the ranking list. Multi-objective evolutionary algorithm is applied for finding the best feature sets. Each feature set is pareto-optimal with respect to complexity vs. model error. It is inspired by the process of natural selection, where the fittest individuals in a population are selected to reproduce and pass on their genetic traits to the next generation. In the context of model optimization, a genetic algorithm begins by randomly generating a set of model parameters, or "chromosomes," which are evaluated based on a predefined "fitness" function. The chromosomes with the highest fitness are then selected to "breed" and produce a new generation of chromosomes, with some degree of random variation introduced through mutation. This process is repeated for multiple generations until a satisfactory solution is found.

4) Machine Learning Modeling Key Results

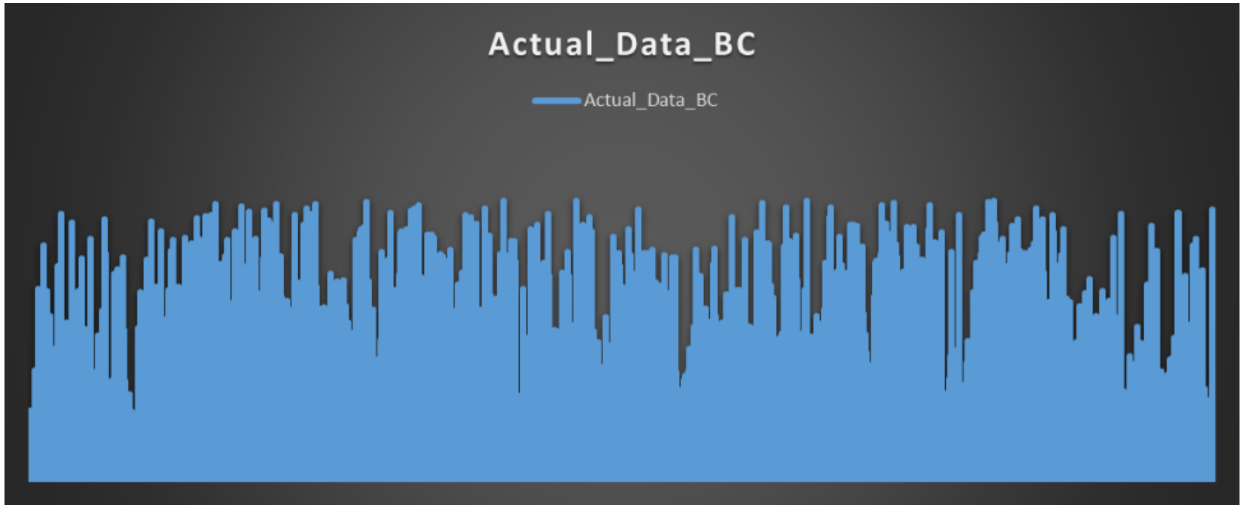
The BC actual data has volatility outliers in it:



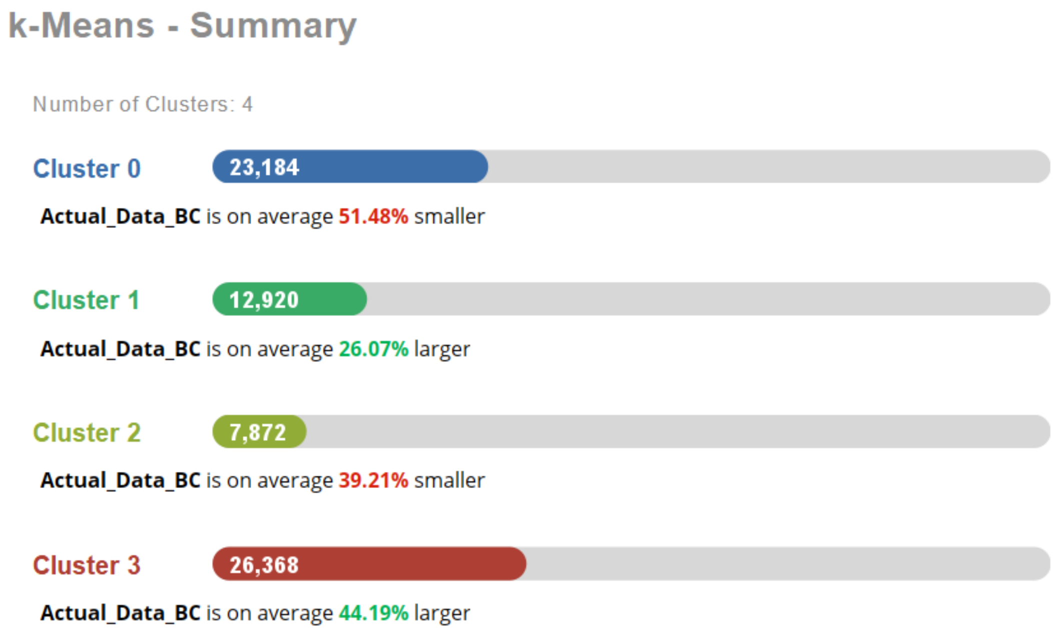
To handle this, we can do 2 things; first is outlier detection in ML and then removing them. Then actual BC data will be more smooth and predicted datasets from different ML models will also be closer to the actual BC data.

Once BC has been determined, loading for outliers as adverse deviation, cat can be done to transform BC to Gross Technical premium.

50,000 is selected as outlier and if >=50,000 of BC is there, 50k is replaced from actual data figure. 370 rows were >=50k and so were removed from the data. 70,714 rows now 70,344. The graph of actual data trimmed now is much smoother than previously:



Now we can do clustering on it using k-means clustering; we have selected to reach 4 clusters.



The clusters are:

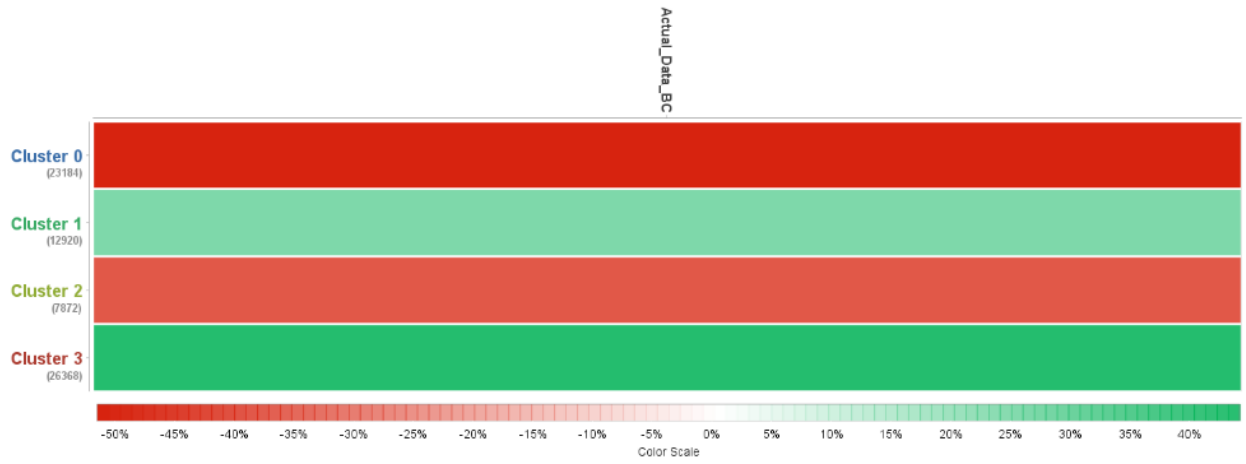
Gender	Age	Nationality	network_type	Relation	clustering	Average of Sum of Actual_Data_BC
F	Band 1 (0-18)	AMERICAN	Platinum	CHILD	cluster_0	119,810
M	Band 2 (19-35)	BRITISH	Bronze			
	Band 3 (36-50)	EGYPTIAN	Gold			
	Band 4 (50+)	EMIRATI	Silver			
		INDIAN	Others			
		JORDANIAN				
		Others				
		PAKISTANI				
		PALESTINIAN				
		PHILIPPINE				
		SUDANESE				
		SYRIAN				

Gender	Age	Nationality	network_type	Relation	clustering	Average of Sum of Actual_Data_BC
F	Band 1 (0-18)	BRITISH	Platinum	CHILD	cluster_1	213,918
	Band 2 (19-35)	EGYPTIAN	Gold	SPOUSE		
	Band 3 (36-50)	EMIRATI	Bronze	MEMBER		
	Band 4 (50+)	INDIAN	Silver			
		PAKISTANI	Others			
		AMERICAN				
		JORDANIAN				
		Others				
		PALESTINIAN				
		PHILIPPINE				
		SUDANESE				
		SYRIAN				

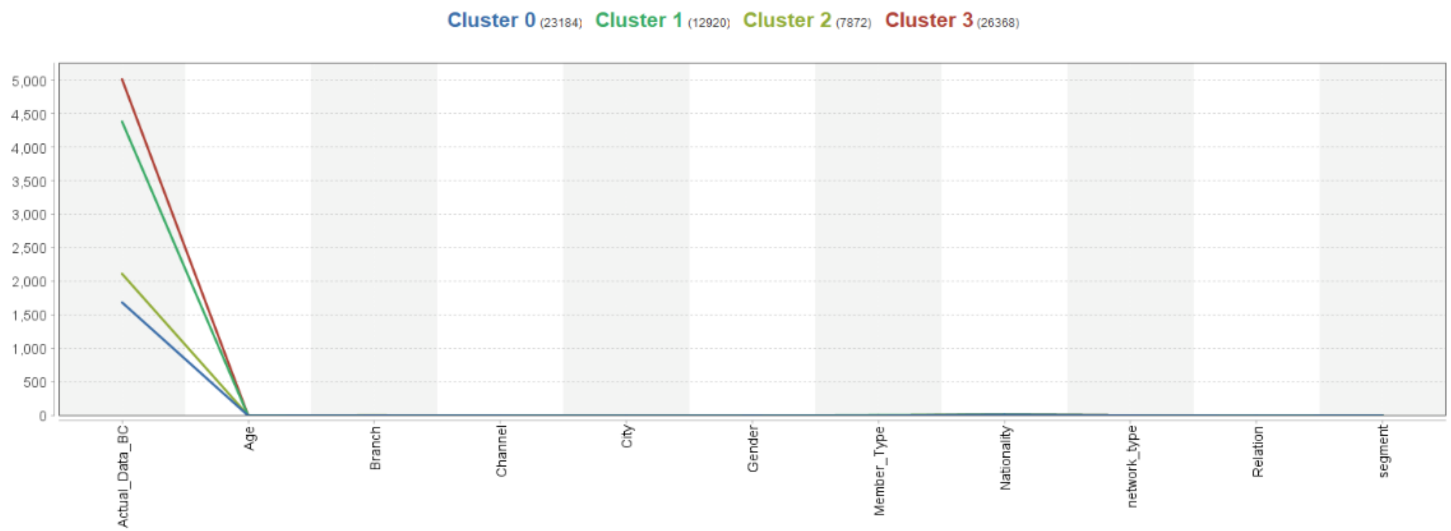
Gender	Age	Nationality	network_type	Relation	clustering	Average of Sum of Actual_Data_BC
F	Band 1 (0-18)	BRITISH	Platinum	CHILD	cluster_2	35,574
M	Band 2 (19-35)	EGYPTIAN	Gold	SPOUSE		
	Band 3 (36-50)	JORDANIAN	Bronze	MEMBER		
	Band 4 (50+)	Others	Silver			
		PALESTINIAN	Others			
		AMERICAN				
		EMIRATI				
		INDIAN				
		PAKISTANI				
		PHILIPPINE				
		SUDANESE				
		SYRIAN				

Gender	Age	Nationality	network_type	Relation	clustering	Average of Sum of Actual_Data_BC
F	Band 1 (0-18)	BRITISH	Platinum	CHILD	cluster_3	192,987
M	Band 2 (19-35)	EGYPTIAN	Gold	MEMBER		
	Band 3 (36-50)	JORDANIAN	Silver			
	Band 4 (50+)	Others	Bronze			
		PALESTINIAN	Others			
		AMERICAN				
		EMIRATI				
		INDIAN				
		PAKISTANI				
		PHILIPPINE				
		SUDANESE				
		SYRIAN				

k-Means - Heat Map



k-Means - Centroid Chart

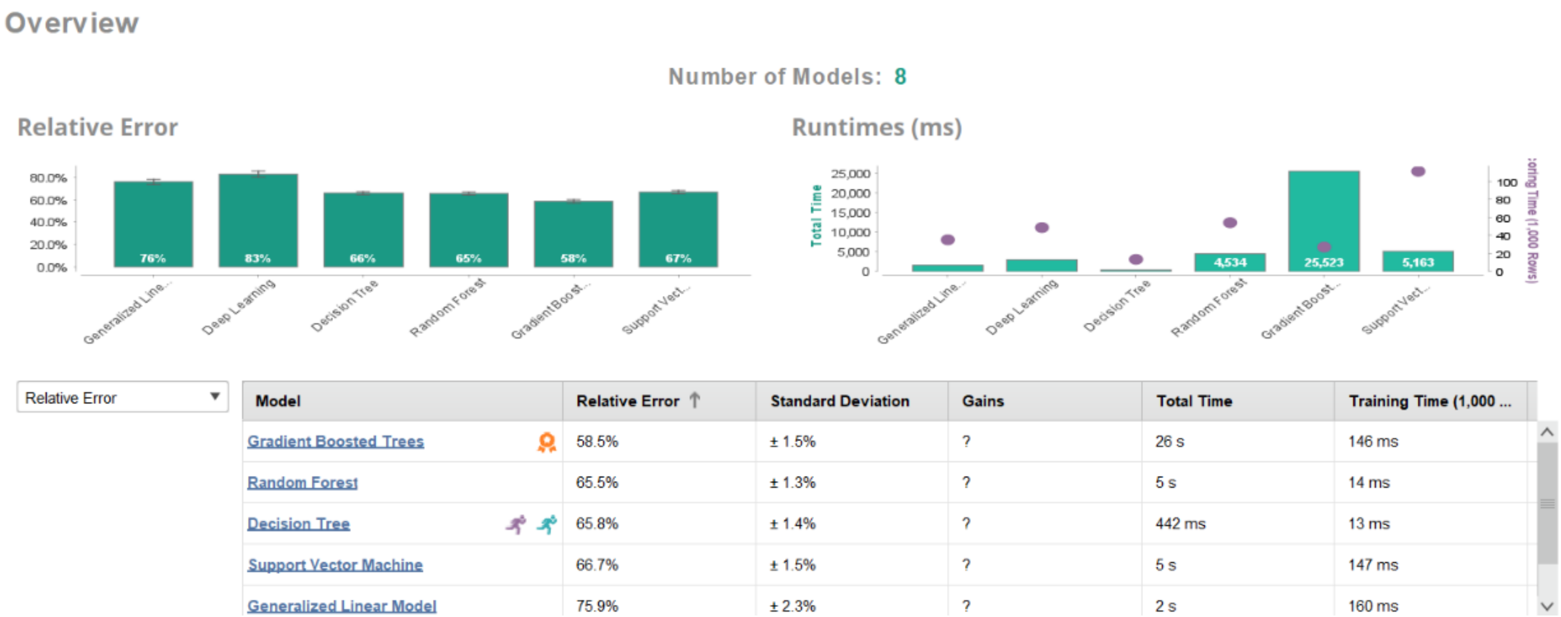


k-Means - Centroid Table

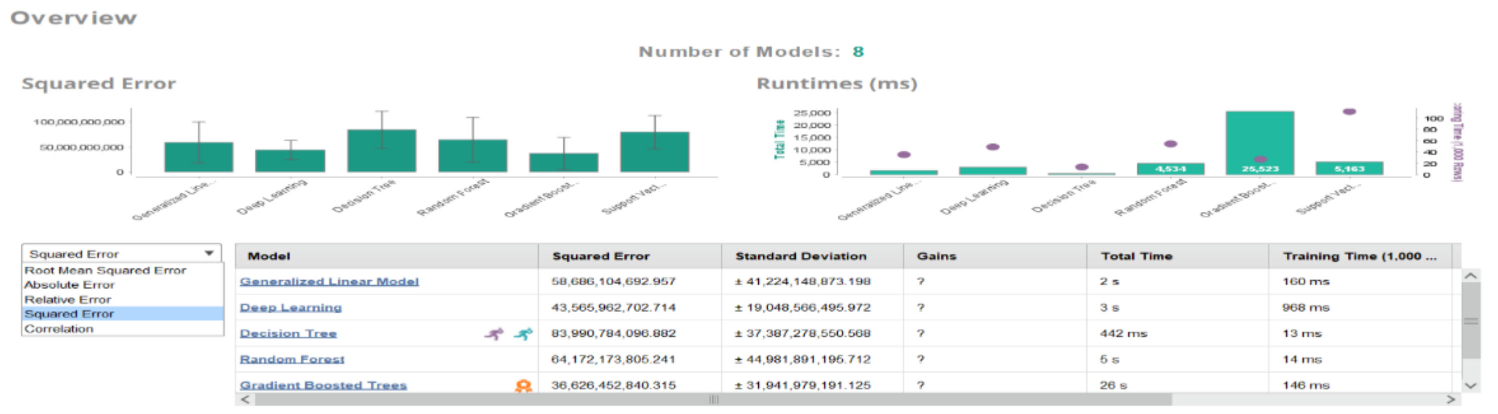
Cluster	Actual_Data...	Age	Branch	Channel	City	Gender	Member_Ty...	Nationality	network_type	Relation	segment
Cluster 0	1685.706	0	3	1	0	1	0	12	3	0	2
Cluster 1	4379.801	1	3	1	0	0	7	12	3	2	2
Cluster 2	2111.725	2	5	1	2	1	2	12	3	1	1
Cluster 3	5009.072	1	3	1	0	1	2	12	3	1	2

The results of running 6 optimized AutoML models is that in the scoreboard, GBM has the lowest relative error and deep learning the highest error.

Model	Relative Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Gradient Boosted Trees	0.585	0.0		25523.0	145.7	27.2
Random Forest	0.655	0.0		4534.0	14.1	54.3
Decision Tree	0.658	0.0		442.0	13.0	13.6
Support Vector Machine	0.667	0.0		5163.0	146.7	111.4
Generalized Linear Model	0.759	0.0		1629.0	159.8	35.3
Deep Learning	0.827	0.0		3042.0	968.5	48.9



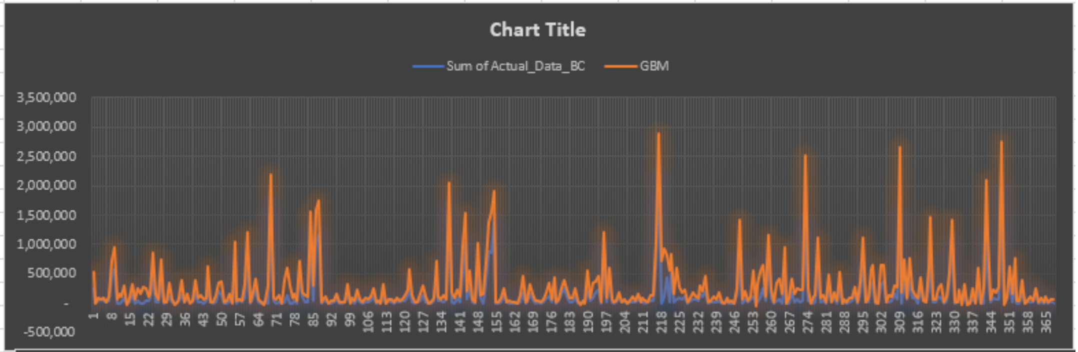
Can show by other error types too:

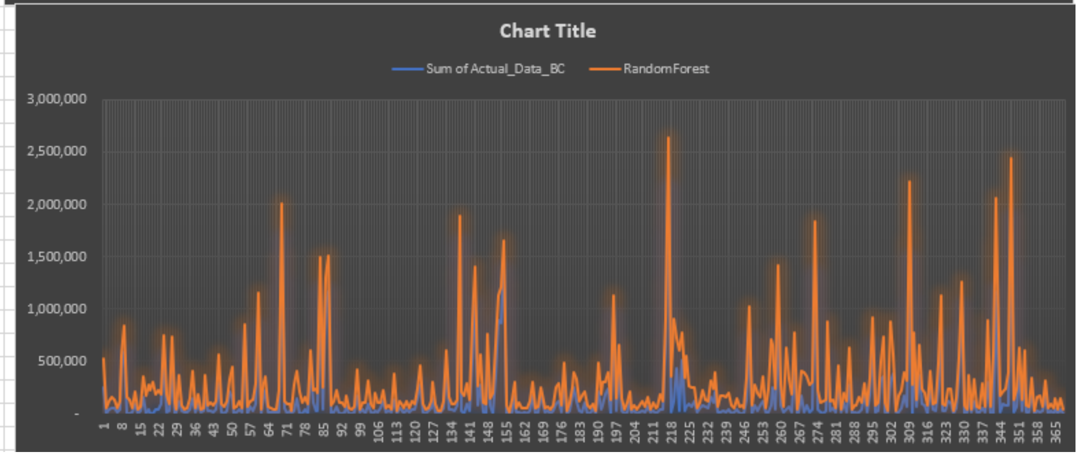


Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

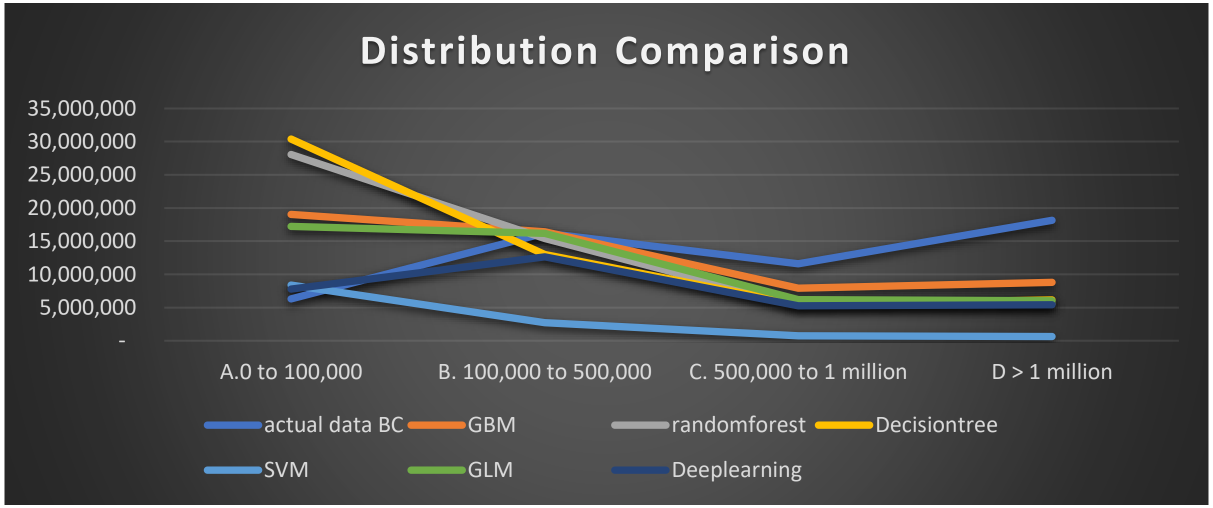
Now actual data BC with predictions for BC based on different models should be compared. Prediction BC by different ML models shouldn't be 100% same as actual data BC because that would be over-fitting that captures both the trend and the noise in the actual data. It shouldn't also be too different because then there would be under fitting. The goal is parsimonious fit to the data that can be used to predict just trends into the future:

network_type	Relation	clustering	Sum of Actual_Data_BC	GBM	RandomForest	DecisionTree	SVM	GLM	DeepLearning
Platinum	CHILD	cluster_0	251,854	279,966	269,733	116,549	36,789	166,390	170,277
Bronze	CHILD	cluster_0	2,478	3,532	45,210	54,614	20,994	- 192,361	- 95,271
Gold	CHILD	cluster_0	37,106	53,928	77,736	25,088	29,197	- 16,511	17,004
Platinum	CHILD	cluster_1	55,233	11,387	98,942	116,075	34,651	- 17,948	- 63,418
Platinum	CHILD	cluster_3	44,050	34,046	80,499	116,075	43,006	- 84,414	- 69,438
Silver	CHILD	cluster_0	7,763	15,275	43,139	4,550	23,292	- 139,519	- 47,061
Bronze	CHILD	cluster_0	40,640	80,806	69,264	34,345	23,187	- 42,366	2,765
Gold	CHILD	cluster_0	428,053	288,988	122,690	74,217	31,743	133,483	103,596
Platinum	CHILD	cluster_0	584,592	358,929	256,106	22,788	39,771	243,939	218,507
Platinum	MEMBER	cluster_3	2,651	50,727	148,864	344,846	51,510	381,059	172,201
Platinum	SPOUSE	cluster_1	2,299	144,991	127,623	252,692	38,601	387,432	229,948
Bronze	CHILD	cluster_0	7,516	92,066	48,018	2,652	25,987	- 12,888	46,980
Gold	CHILD	cluster_0	92,828	195,996	112,580	84,158	34,576	162,961	149,940
Gold	SPOUSE	cluster_1	1,219	- 20,316	39,800	46,549	32,865	306,454	176,983
Others	CHILD	cluster_0	18,824	76,862	47,042	27,650	28,961	- 95,391	- 49,492
Bronze	CHILD	cluster_0	148,399	180,852	199,199	88,366	32,850	126,978	123,001
Bronze	SPOUSE	cluster_1	2,385	64,909	171,020	512,290	31,370	270,471	181,996
Gold	CHILD	cluster_1	34,295	206,881	233,635	23,726	38,078	190,934	77,508
Others	CHILD	cluster_0	851	154,581	214,623	284,548	35,545	44,475	28,963
Platinum	SPOUSE	cluster_1	3,370	277,081	300,552	512,290	45,727	556,776	398,950
Silver	CHILD	cluster_1	33,230	208,305	146,317	23,726	33,075	67,926	20,507
Platinum	CHILD	cluster_3	36,802	74,950	186,389	331,503	44,767	53,213	39,610
Silver	CHILD	cluster_0	92,726	64,578	84,647	25,933	24,960	- 1,893	30,331
Gold	CHILD	cluster_0	529,305	333,822	211,932	252,766	35,808	233,308	203,805
Gold	MEMBER	cluster_3	10,861	154,638	165,806	742,873	47,038	370,428	190,304
Bronze	CHILD	cluster_0	22,648	37,654	63,846	174,136	21,879	- 101,446	- 20,544
Platinum	CHILD	cluster_0	421,729	331,766	306,521	174,136	38,370	184,859	208,617
Silver	SPOUSE	cluster_1	2,156	14,684	104,138	179,193	21,695	94,888	46,567
Bronze	CHILD	cluster_0	10,922	10,394	27,713	20,556	20,129	- 166,658	- 52,091
Platinum	CHILD	cluster_0	114,131	233,847	246,698	95,841	36,001	119,647	136,411
Silver	CHILD	cluster_0	25,339	31,264	44,032	8,365	22,312	- 113,817	- 32,756
Silver	MEMBER	cluster_3	4,148	- 32,149	29,954	29,744	32,429	23,303	- 53,298
Bronze	CHILD	cluster_0	9,775	2,371	42,534	86,773	21,947	- 111,686	- 15,952
Gold	CHILD	cluster_0	34,349	64,281	105,436	86,773	30,354	64,163	77,189





The comparison to evaluate fitting results by distributions is shown:



The BC as per actual data and relativity factors as per actual data is compared with that BC and RF arrived at from different models.

Different models ordered in relevance to relative error so that GBM is first because it has least error.

Ensemble can sumproduct RF of different models by weightage.

As we can see, some results of models are closely similar to actual BC but also different as well for other cases:

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

Gender	BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
	F	127,064	115,920	116,999	125,143	31,638	81,895	47,386
	M	162,523	176,962	189,791	182,956	37,148	181,234	135,118
	Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Gender	Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning	Ensemble
	F	0.89	0.82	0.79	0.84	0.93	0.66	0.56	0.80
	M	1.14	1.25	1.28	1.22	1.09	1.46	1.60	1.27
			50%	25%	10%	5%	5%	5%	

Age	BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
	Band 1 (0-18)	132,350	163,985	164,044	176,128	33,438	118,794	86,074
	Band 2 (19-35)	120,711	133,738	144,775	141,201	34,225	126,924	82,727
	Band 3 (36-50)	183,316	160,326	142,995	133,721	33,737	160,564	121,992
	Band 4 (50+)	137,695	114,353	145,393	161,397	34,277	78,585	40,772
	Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Age	Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
	Band 1 (0-18)	0.93	1.16	1.11	1.18	0.98	0.96	1.02
	Band 2 (19-35)	0.85	0.94	0.98	0.94	1.01	1.02	0.98
	Band 3 (36-50)	1.29	1.13	0.97	0.89	0.99	1.29	1.44
	Band 4 (50+)	0.97	0.81	0.98	1.08	1.01	0.63	0.48

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

Nationality	BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
	AMERICAN	153,614	141,113	135,837	119,381	35,038	174,322	107,029
	BRITISH	53,489	59,316	81,825	62,626	32,221	13,023	- 11,522
	EGYPTIAN	175,851	148,075	142,352	124,473	33,467	152,627	77,655
	EMIRATI	187,936	149,561	146,210	119,880	36,204	137,748	96,450
	INDIAN	184,252	270,402	301,124	421,403	39,417	259,839	215,064
	JORDANIAN	121,550	137,451	145,348	161,763	33,049	142,385	88,613
	Others	407,828	315,187	263,243	324,307	40,619	281,196	234,804
	PAKISTANI	124,334	130,584	145,388	112,499	32,377	99,255	82,794
	PALESTINIAN	49,486	53,058	72,692	46,437	31,557	32,209	10,141
	PHILIPPINE	77,454	102,976	117,701	84,223	30,013	52,070	33,594
	SUDANESE	34,624	50,176	69,697	52,867	29,108	31,659	7,733
	SYRIAN	64,076	75,856	89,275	63,302	31,826	57,715	17,601
	Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
AMERICAN	1.08	1.00	0.92	0.80	1.03	1.41	1.27
BRITISH	0.38	0.42	0.55	0.42	0.95	0.11	- 0.14
EGYPTIAN	1.24	1.04	0.96	0.83	0.99	1.23	0.92
EMIRATI	1.32	1.05	0.99	0.80	1.07	1.11	1.14
INDIAN	1.30	1.91	2.04	2.82	1.16	2.10	2.54
JORDANIAN	0.86	0.97	0.98	1.08	0.97	1.15	1.05
Others	2.87	2.22	1.78	2.17	1.20	2.27	2.78
PAKISTANI	0.88	0.92	0.98	0.75	0.95	0.80	0.98
PALESTINIAN	0.35	0.37	0.49	0.31	0.93	0.26	0.12
PHILIPPINE	0.55	0.73	0.80	0.56	0.88	0.42	0.40
SUDANESE	0.24	0.35	0.47	0.35	0.86	0.26	0.09
SYRIAN	0.45	0.53	0.60	0.42	0.94	0.47	0.21

BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
Bronze	47,851	66,233	113,081	112,789	28,902	36,945	27,853
Gold	142,808	148,622	143,544	139,464	34,098	156,590	104,391
Others	10,229	53,287	122,083	169,936	33,156	10,002	- 18,183
Platinum	267,453	230,154	197,415	200,497	40,249	217,853	162,584
Silver	68,110	88,385	114,391	109,687	28,767	47,873	19,206
Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
Bronze	0.34	0.47	0.76	0.75	0.85	0.30	0.33
Silver	0.48	0.62	0.77	0.73	0.85	0.39	0.23
Gold	1.01	1.05	0.97	0.93	1.00	1.26	1.23
Platinum	1.88	1.62	1.34	1.34	1.18	1.76	1.92

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
cluster_0	119,810	121,105	126,913	119,401	31,423	99,155	69,401
cluster_1	213,918	163,334	142,318	139,442	30,569	152,430	101,143
cluster_2	35,574	60,091	92,350	87,622	28,418	28,147	5,294
cluster_3	192,987	204,839	207,749	223,031	42,091	196,222	145,858
Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
cluster_0	0.84	0.85	0.86	0.80	0.92	0.80	0.82
cluster_1	1.51	1.15	0.96	0.93	0.90	1.23	1.20
cluster_2	0.25	0.42	0.62	0.59	0.84	0.23	0.06
cluster_3	1.36	1.44	1.41	1.49	1.24	1.58	1.72

BC	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
CHILD	92,786	130,707	147,952	153,428	33,190	73,257	44,521
MEMBER	133,535	136,433	149,707	151,351	35,781	135,139	93,489
SPOUSE	237,334	172,891	142,511	139,416	30,017	165,900	117,288
Grand Total	**142,096**	**141,796**	**147,856**	**149,651**	**33,974**	**124,006**	**84,576**

Relativity Factor	Average of Sum of Actual_Data_BC	Average of GBM	Average of RandomForest	Average of DecisionTree	Average of SVM	Average of GLM	Average of DeepLearning
CHILD	0.65	0.92	1.00	1.03	0.98	0.59	0.53
MEMBER	0.94	0.96	1.01	1.01	1.05	1.09	1.11
SPOUSE	1.67	1.22	0.96	0.93	0.88	1.34	1.39

As we can see from these results:

- Male are more expensive than female. This is because male have higher hospitalization for 0-18 years, higher accidents, higher overall severity but lower frequency, earlier mortality relative to female and more cardiovascular diseases which account for more deaths than all of cancer deaths combined on global average.
- Female have lower severity, lower accidents, lower hospitalization for 0-18 years but still their costs are not that low because of maternity related claims.
- Band 4 50+ has lower cost than Band 3 36-50 years because no maternity claims are there for ages 50+. Band 2 19-35 years has lower cost than Band 1 0-18 years because of higher hospitalizations in age 0-1 years and high accidents in ages 16-18 years.
- Some nationalities have above 1 RF are more costly and some have below 1 RF are less costly.
- It is not surprising that BC increases as plans increase from low benefits to high benefits (Bronze to Silver to Gold to Platinum). Although BC increases from plans, loss ratios and premium as % of sum insured might have different experience across plans like Platinum might have lower loss ratio and lower premium as % of sum insured than let's say Gold plan.
- Cluster 2 has the lowest BC whereas Cluster 0 has second lowest BC. Cluster 1 and 3 are close-by in BC which means instead of asking the model to make 4 clusters, we could've asked it for 3 clusters as each cluster should have distinct trends. Cluster 0 is least costly because it consists mostly of Children. Cluster 1 is next least costly because it has Female only. Cluster 2 and 3 are similar in constitution.
- Child has lowest BC whereas member has around 1 RF. Spouse is expensive because most spouses are female and eligible for maternity benefit.

Tables and graphs for the 6 ML models for first sample Medical Pricing

GLM model

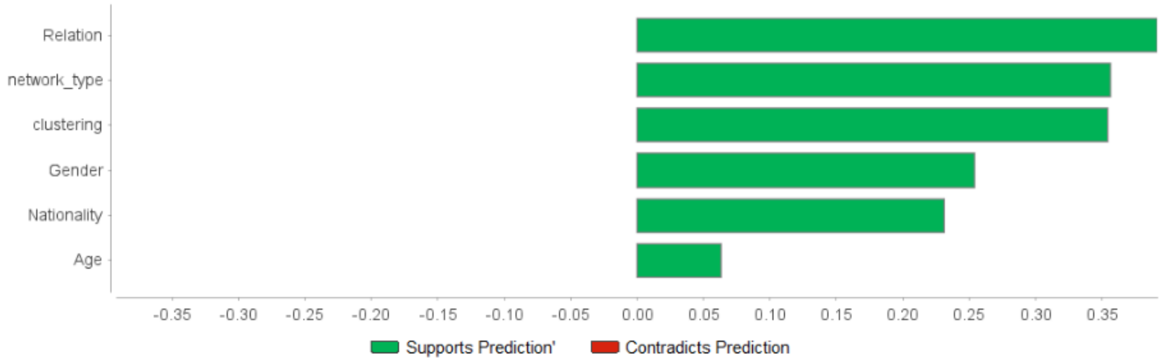
Generalized Linear Model - Model

Attribute	Coefficient	Std. Coefficient	Std. Error	z-Value	p-Value
Nationality.AMERICAN	-14636.340	-14636.340	?	?	?
Nationality.BRITISH	-87081.089	-87081.089	?	?	?
Nationality.EGYPTIAN	62913.109	62913.109	?	?	?
Nationality.EMIRATI	92391.295	92391.295	?	?	?
Nationality.INDIAN	232257.171	232257.171	?	?	?
Nationality.JORDANIAN	50545.713	50545.713	?	?	?
Nationality.Others	162737.999	162737.999	?	?	?
Nationality.PAKISTANI	3833.173	3833.173	?	?	?
Nationality.PALESTINIAN	-61378.980	-61378.980	?	?	?
Nationality.PHILIPPINE	-6406.939	-6406.939	?	?	?
Nationality.SUDANESE	-61389.747	-61389.747	?	?	?
Nationality.SYRIAN	-50398.991	-50398.991	?	?	?
Nationality.MISSING	0	0	?	?	?

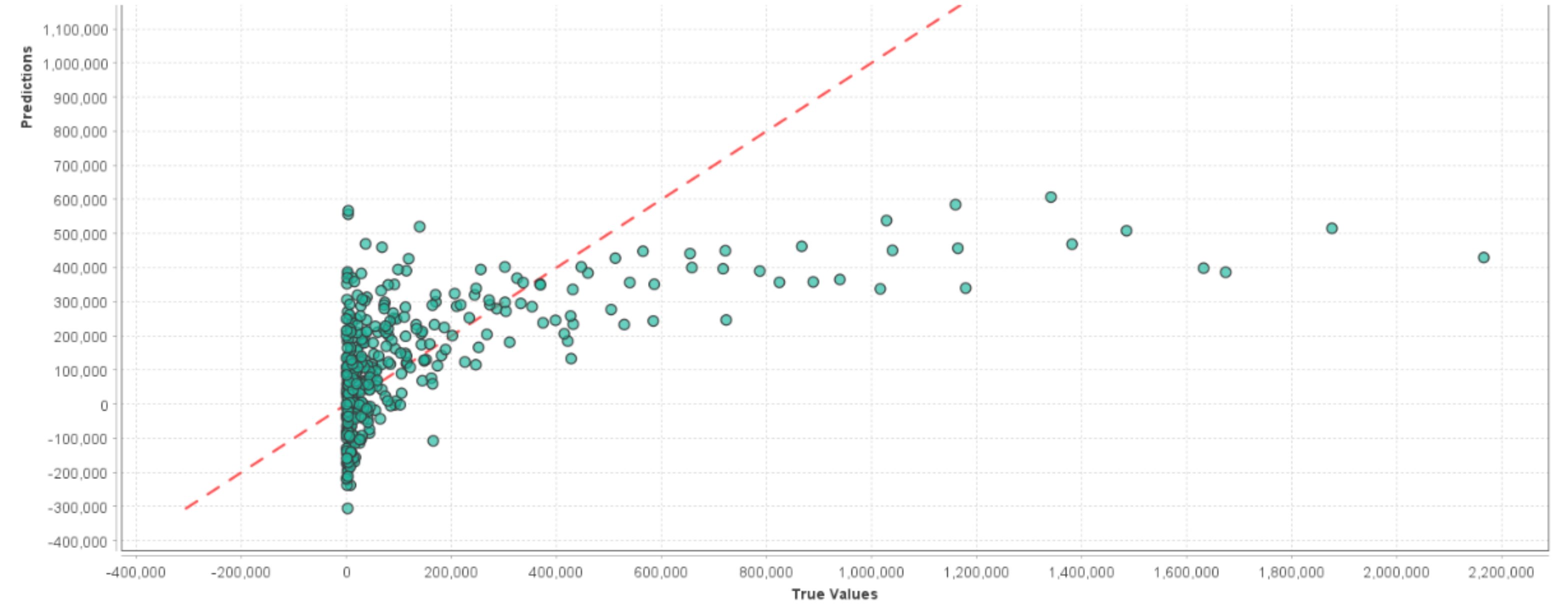
Generalized Linear Model - Weights

Attribute	Weight
Relation	0.280
network_type	0.127
Gender	0.119
clustering	0.085
Nationality	0.043
Age	0.038

Important Factors for Prediction



Generalized Linear Model - Predictions Chart



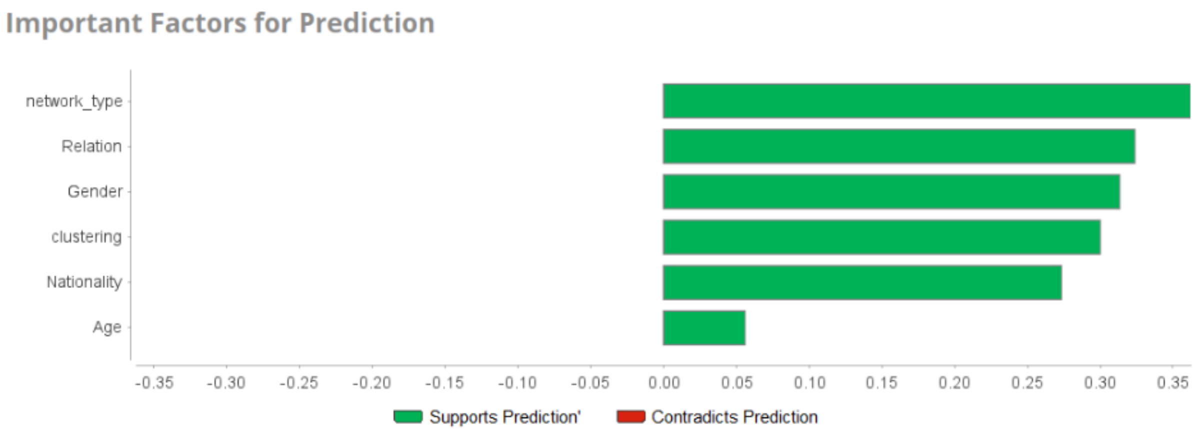
Deep Learning

Deep Learning Model

```
Model Metrics Type: Regression
 Description: Metrics reported on full training frame
 model id: rm-h2o-model-model-5
 frame id: rm-h2o-frame-model-5
 MSE: 5.1916296E10
 RMSE: 227851.48
 R^2: 0.426673
 mean residual deviance: 5.1916296E10
 mean absolute error: 127549.88
 root mean squared log error: NaN
Status of Neuron Layers (predicting Sum of Actual_Data_BC, regression, gaussian distribution, Quadratic loss, 4,751
weights/biases, 60.7 KB, 5,520 training samples, mini-batch size 1):
 Layer Units      Type Dropout       L1       L2 Mean Rate Rate RMS Momentum Mean Weight Weight RMS Mean Bias Bias
RMS
     1    42     Input  0.00 %
     2    50 Rectifier       0 0.000010 0.000000  0.289578 0.447053 0.000000    0.001297   0.149062  0.487253 0.039506
     3    50 Rectifier       0 0.000010 0.000000  0.007302 0.021922 0.000000   -0.004594   0.142105  0.991515 0.017857
     4     1    Linear         0.000010 0.000000  0.000255 0.000194 0.000000    0.025569   0.208087 -0.005756 0.000000
Scoring History:
           Timestamp   Duration Training Speed   Epochs Iterations     Samples Training RMSE Training Deviance Training
MAE Training r2
 2022-12-11 01:33:12  0.000 sec                   0.00000          0    0.000000           NaN                 NaN
NaN         NaN
 2022-12-11 01:33:12   0.359 sec    2936 obs/sec   1.00000           1   552.000000   291660.30813 85065735338.04150
219433.29406     0.06059
 2022-12-11 01:33:13   0.447 sec    4295 obs/sec   2.00000           2 1104.000000   258889.70566 67023879696.61295
135334.31547     0.25984
 2022-12-11 01:33:13   0.523 sec    5191 obs/sec   3.00000           3 1656.000000   256161.65858 65618795327.63316
135367.41780     0.27535
 2022-12-11 01:33:13   0.584 sec    6016 obs/sec   4.00000           4 2208.000000   243456.20191 59270922248.91696
138115.65280     0.34545
 2022-12-11 01:33:13   0.640 sec    6666 obs/sec   5.00000           5 2760.000000   241281.16491 58216600539.83927
152823.45810     0.35710
 2022-12-11 01:33:13   0.696 sec    7184 obs/sec   6.00000           6 3312.000000   235705.65608 55557156306.36876
145892.98340     0.38647
```

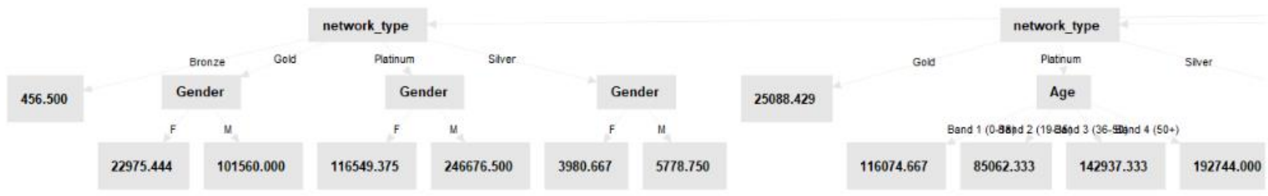
```
 2022-12-11 01:33:13  0.751 sec    7606 obs/sec  7.00000            7 3864.000000  236816.52831 56082068079.64516
137516.93623     0.38067
 2022-12-11 01:33:13  0.803 sec    8029 obs/sec  8.00000            8 4416.000000  229760.38970 52789836677.04465
135949.58882     0.41703
 2022-12-11 01:33:13  0.867 sec    8184 obs/sec  9.00000            9 4968.000000  239097.93903 57167824447.90935
143545.39730     0.36868
 2022-12-11 01:33:13  0.917 sec    8492 obs/sec 10.00000           10 5520.000000  227851.47746 51916295782.39195
127549.88325     0.42667

H2O version: 3.30.0.1-rm9.8.1
```



Decision Tree

Decision Tree - Model

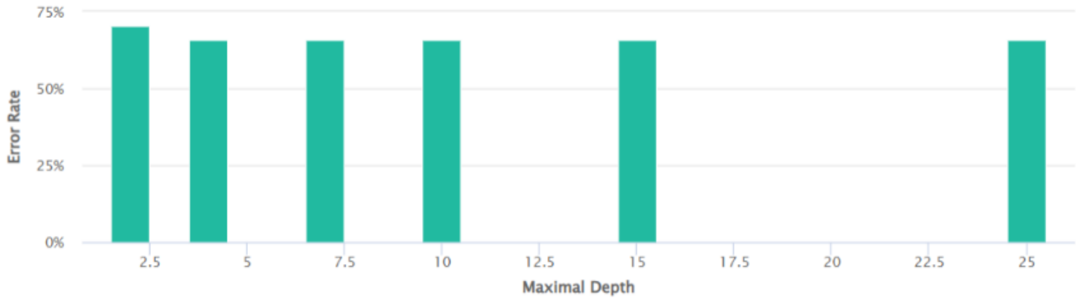


Decision Tree - Optimal Parameters

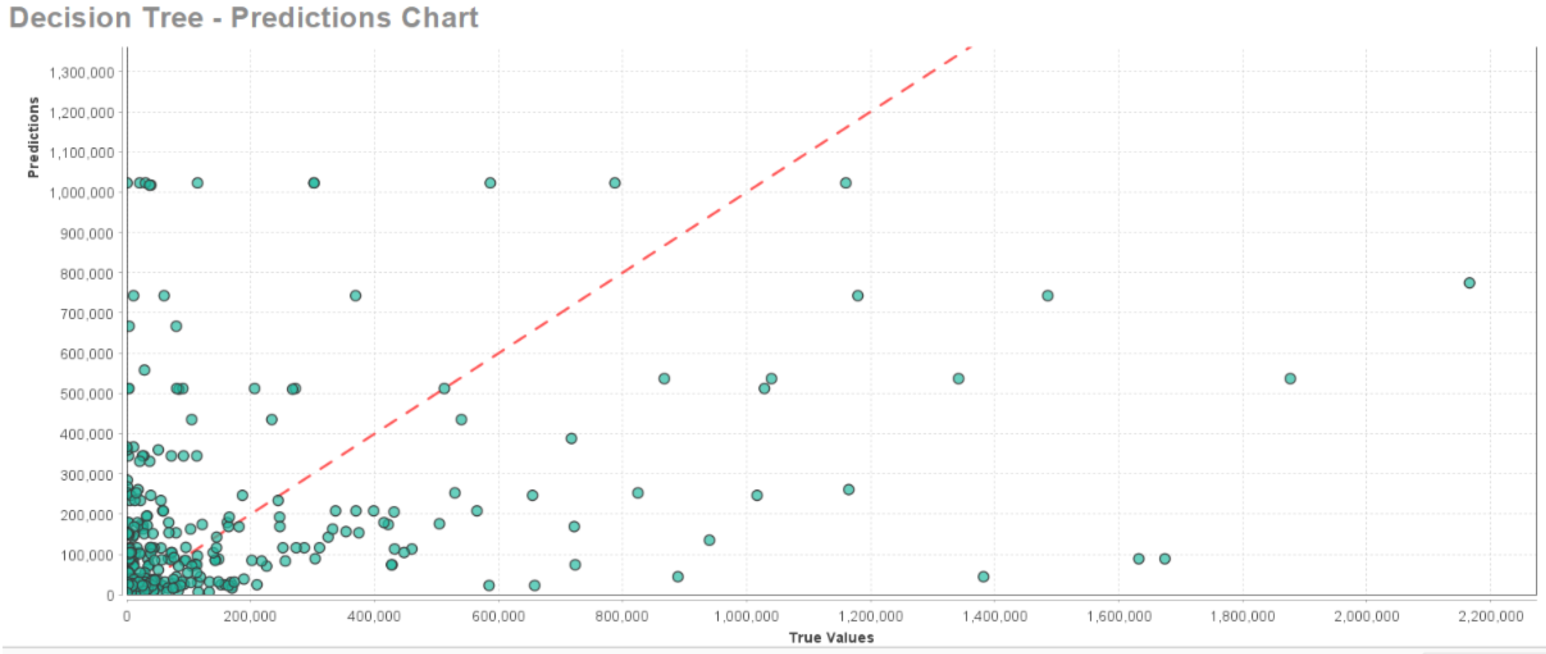
Optimal Parameters

Maximal Depth: 4

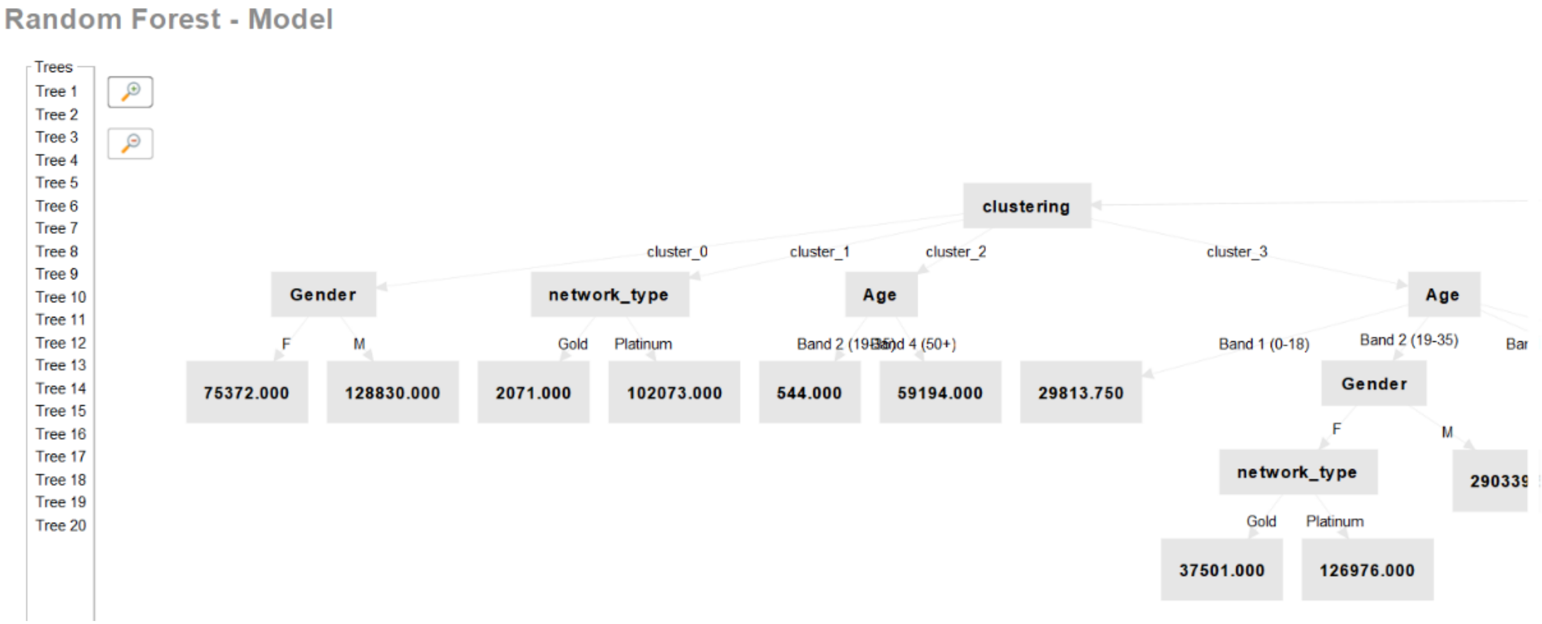
Error Rates for Parameters



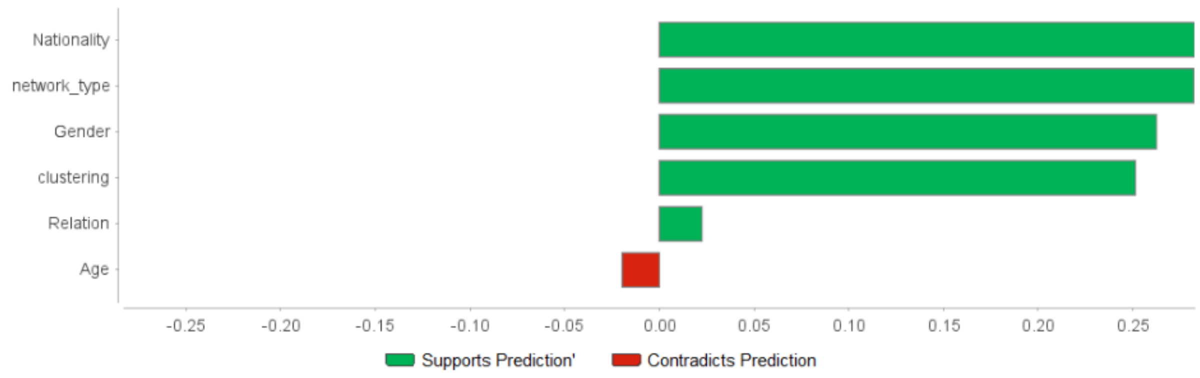
Maximal Depth	Error Rate
2	70.5%
4	65.9%
7	65.9%


Decision Tree - Predictions Chart
Predictions
True Values
0
100,000
200,000
300,000
400,000
500,000
600,000
700,000
800,000
900,000
1,000,000
1,100,000
1,200,000
1,300,000
1,400,000
1,600,000
1,800,000
2,000,000
2,200,000

Random forest



Important Factors for Prediction



Random Forest - Optimal Parameters

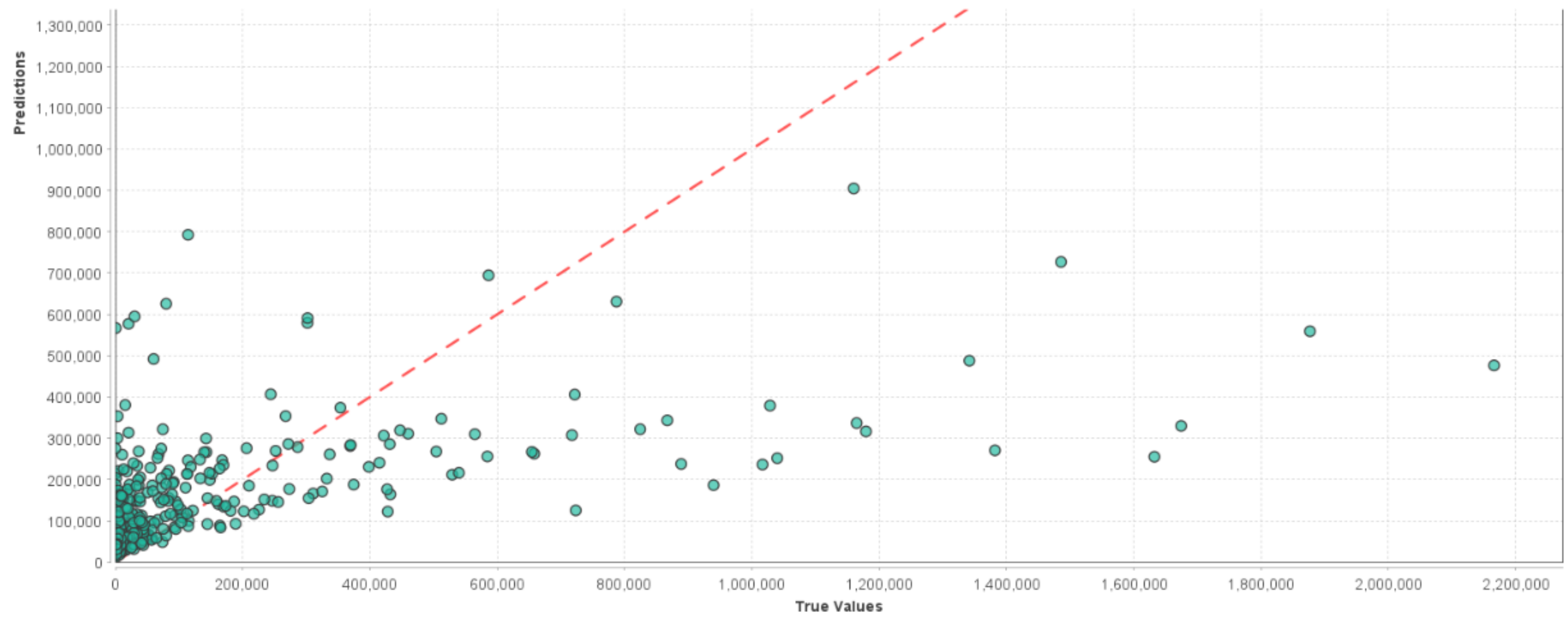
Optimal Parameters

Number Of Trees: 20
Maximal Depth: 7

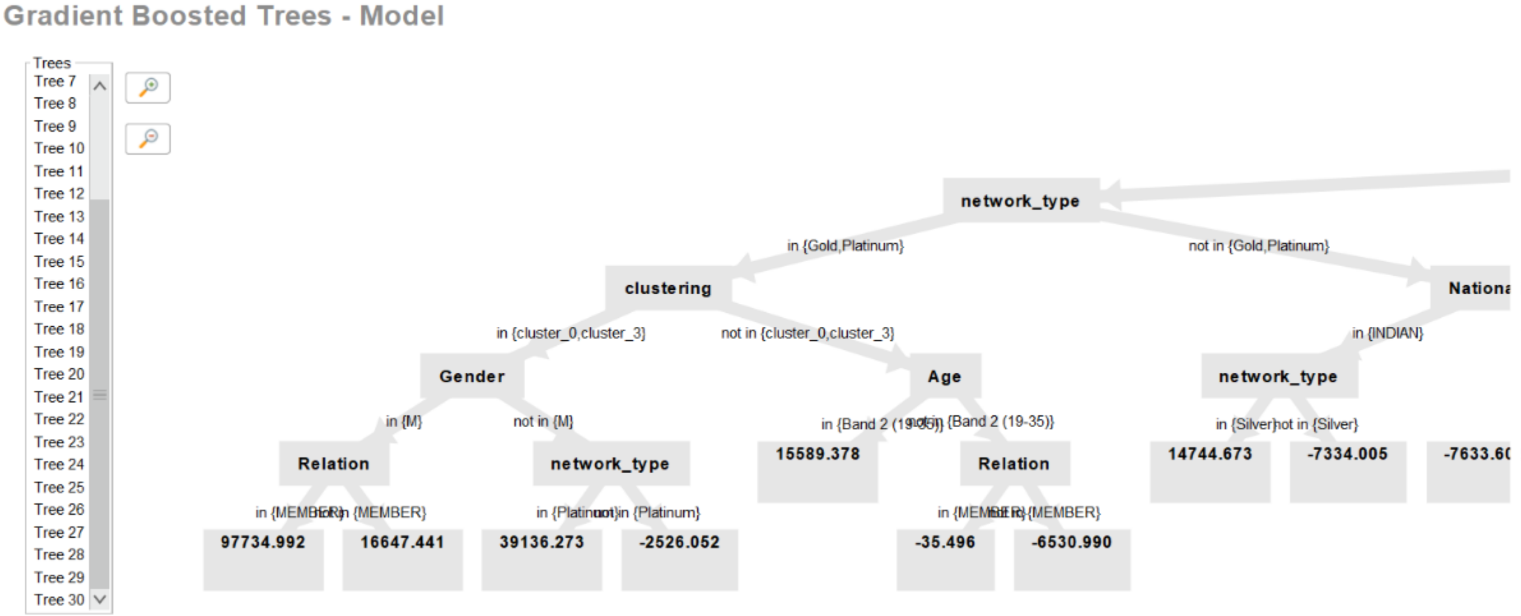
Error Rates for Parameters

Number of Trees	Maximal Depth	Error Rate
20	2	70.1%
60	2	70.1%
100	2	70.1%

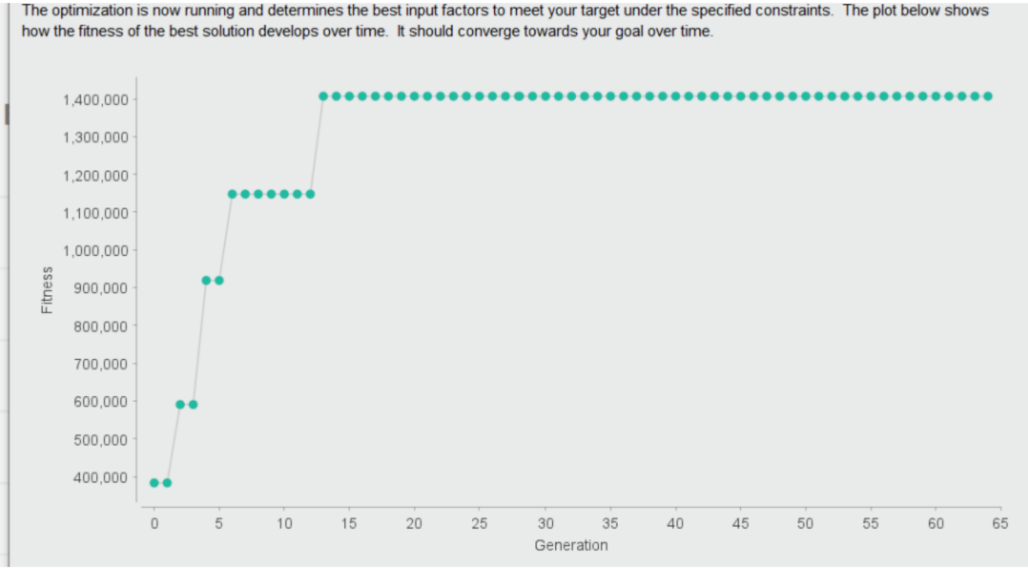
Random Forest - Predictions Chart



GBM


Gradient Boosted Trees - Model
Trees
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
Tree 24
Tree 25
Tree 26
Tree 27
Tree 28
Tree 29
Tree 30
network_type
in {Gold,Platinum}
not in {Gold,Platinum}
clustering
in {cluster_0,cluster_3}
not in {cluster_0,cluster_3}
Gender
in {M}
not in {M}
Relation
network_type
in {Platinum}
not in {Platinum}
Age
Relation
in {MEMBER}
not in {MEMBER}
97734.992
16647.441
39136.273
-2526.052
15589.378
-35.496
-6530.990
in {INDIAN}
network_type
in {Silver}
not in {Silver}
14744.673
-7334.005

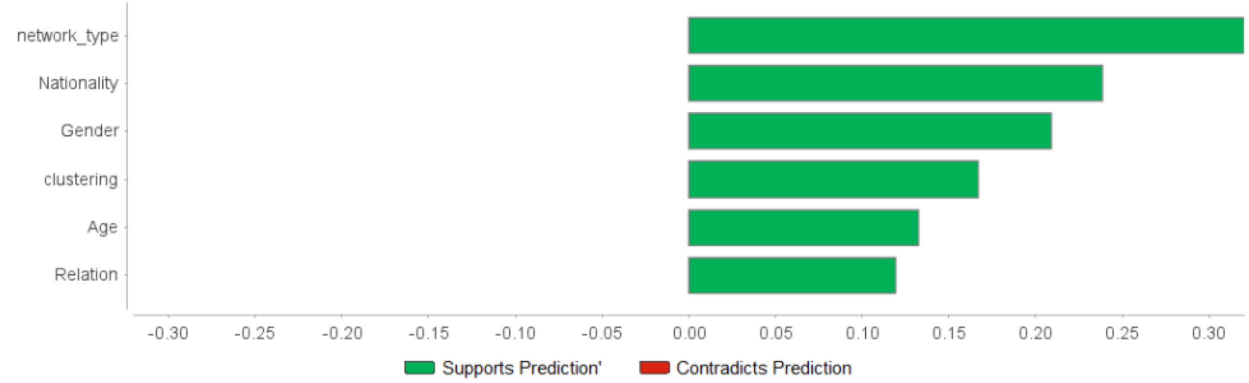
Genetic algorithm that optimizes all machine learning parameters using evolutionary selection method.



Gradient Boosted Trees - Weights

Attribute	Weight
network_type	0.141
Relation	0.106
Gender	0.104
clustering	0.077
Nationality	0.047
Age	0.039

Important Factors for Prediction

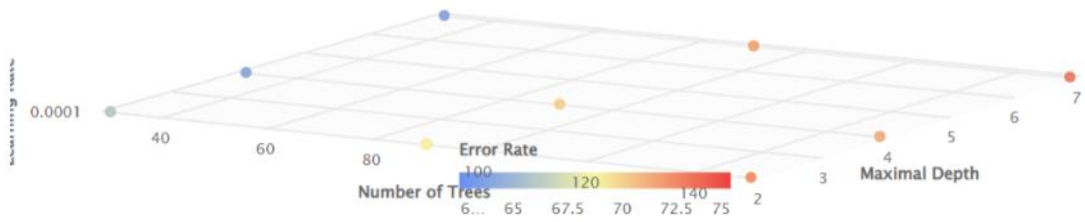


Gradient Boosted Trees - Optimal Parameters

Optimal Parameters

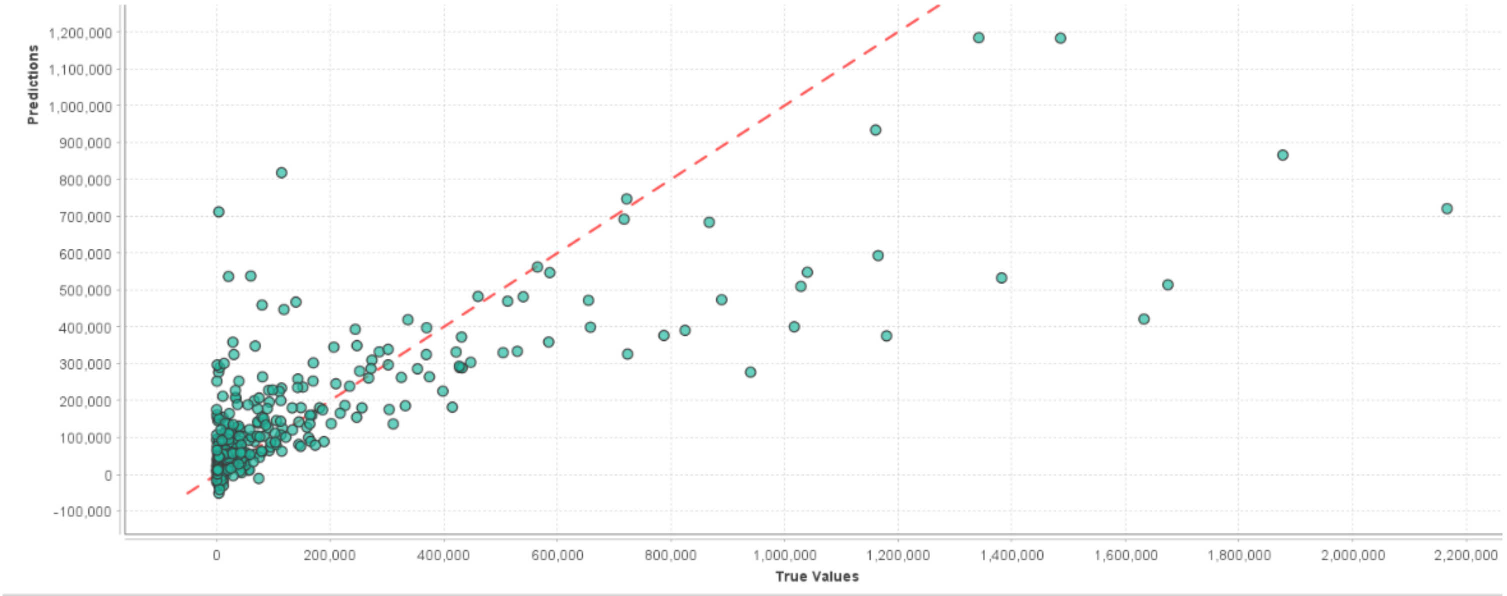
Number Of Trees: 30
Maximal Depth: 7
Learning Rate: 0.100

Error Rates for Parameters

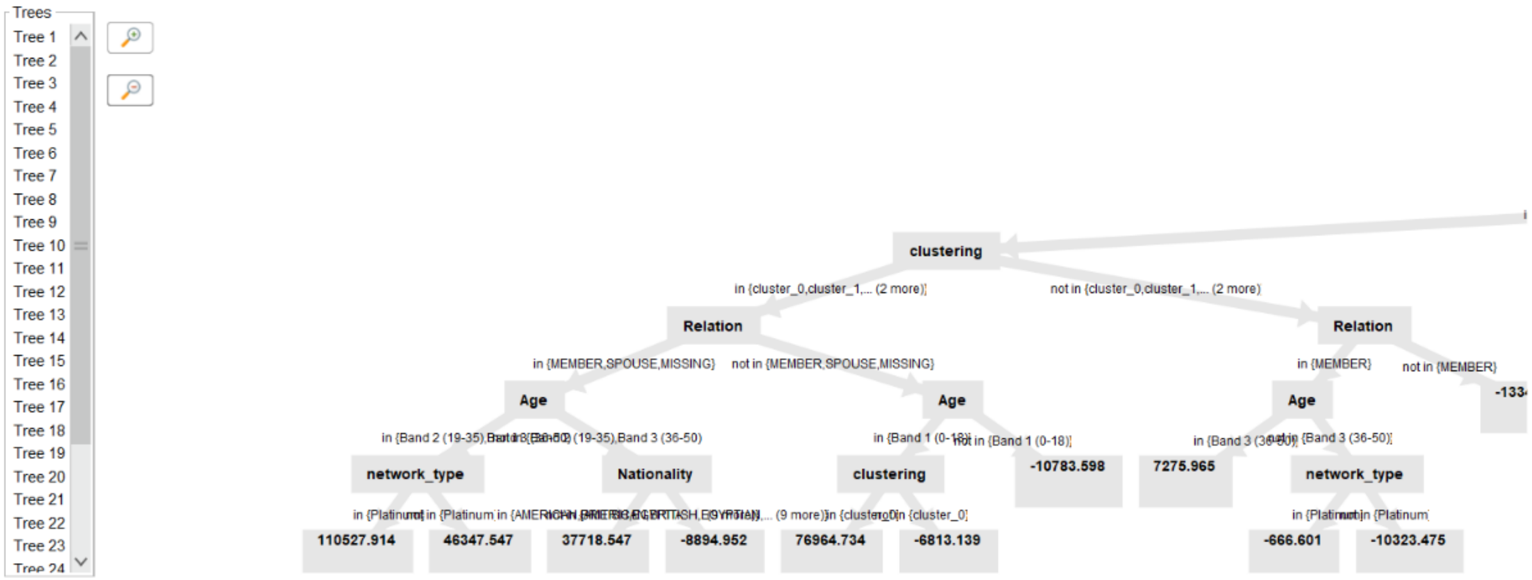


Number of Trees	Maximal Depth	Learning Rate	Error Rate
30	2	0.001	71.2%
90	2	0.001	71.2%
150	2	0.001	71.2%

Gradient Boosted Trees - Predictions Chart



Gradient Boosted Trees - Production Model



Support Vector Machines SVM

Kernel Model

```
Total number of Support Vectors: 920
Bias (offset): 34772.785

w[Relation] = 33497.030
w[clustering] = 149193.676
w[Gender] = 63518.061
w[network_type] = 166154.495
w[Age] = 16690.667
w[Nationality] = 86409.647
```

Support Vector Machine - Weights

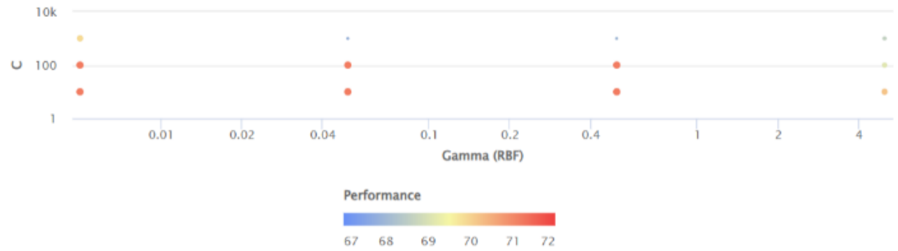
Attribute	Weight
network_type	0.666
clustering	0.523
Nationality	0.521
Gender	0.295
Relation	0.176
Age	0.032

Support Vector Machine - Optimal Parameters

Optimal Parameters

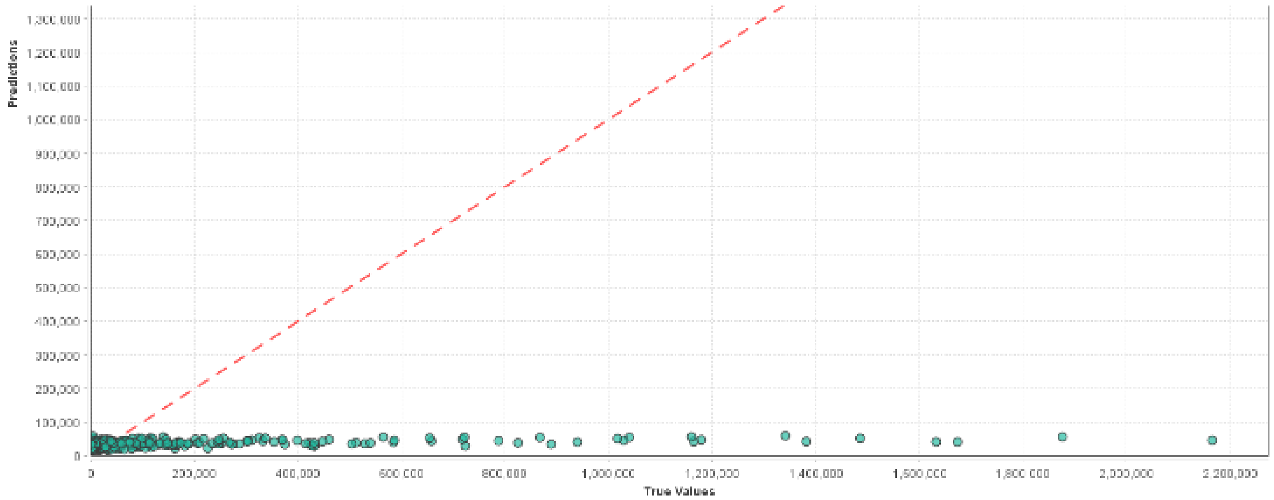
Kernel Gamma: 0.050
C: 1000

Error Rates for Parameters

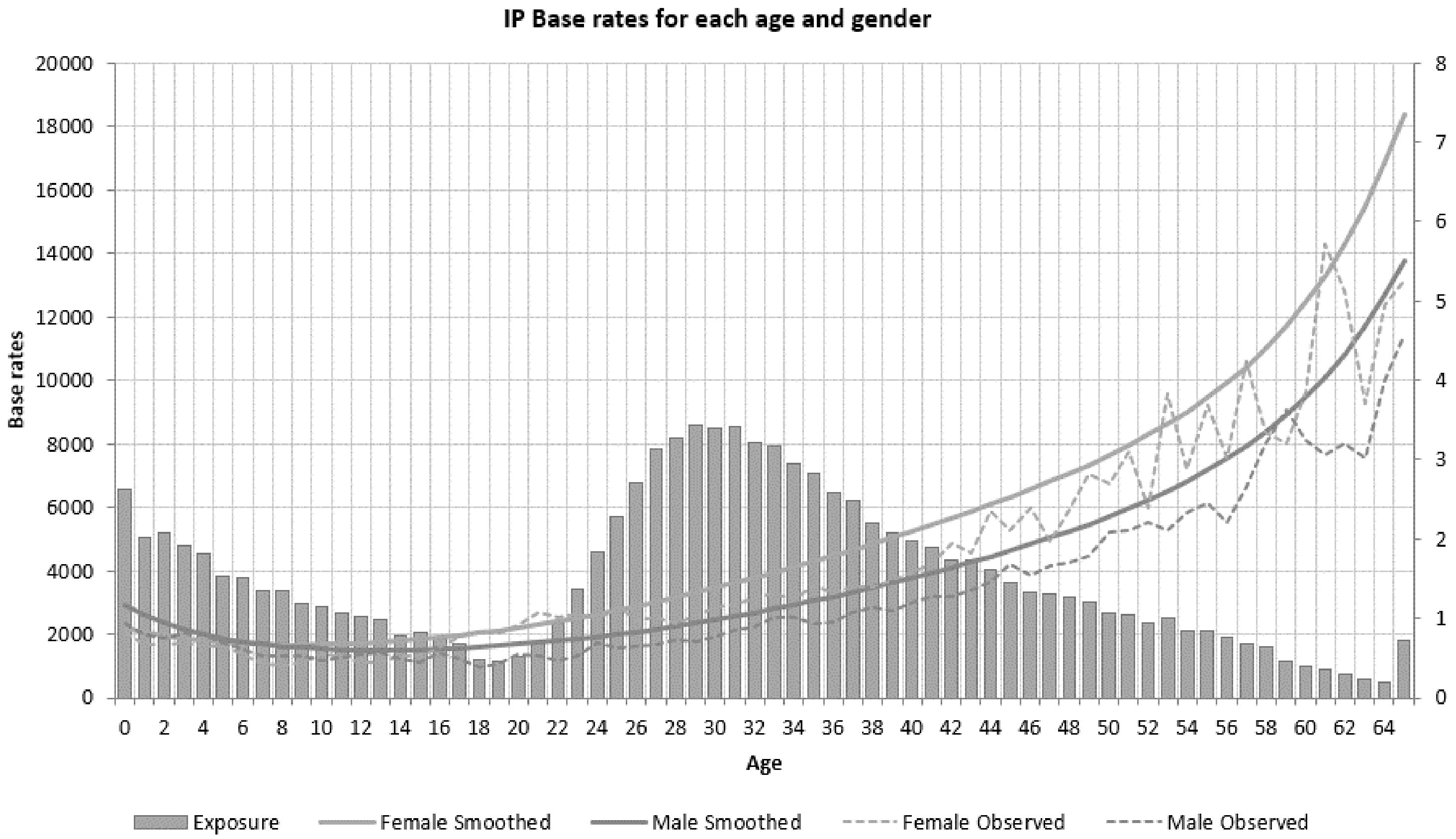


Gamma (RBF)	C	Performance
0.005	10	71.2%
0.050	10	71.2%
0.500	10	71.2%

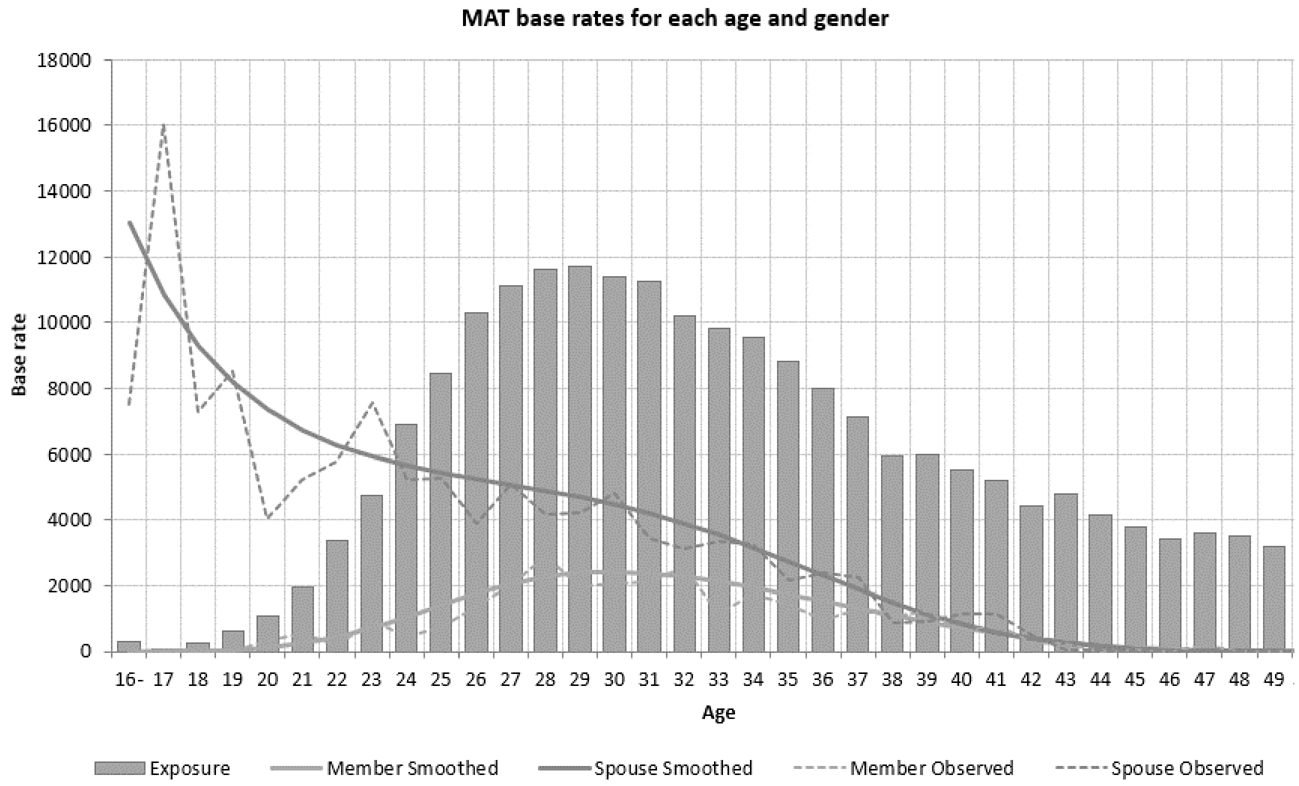
Support Vector Machine - Predictions Chart



In-Patient : Base rates trend



Maternity : Base rates trend



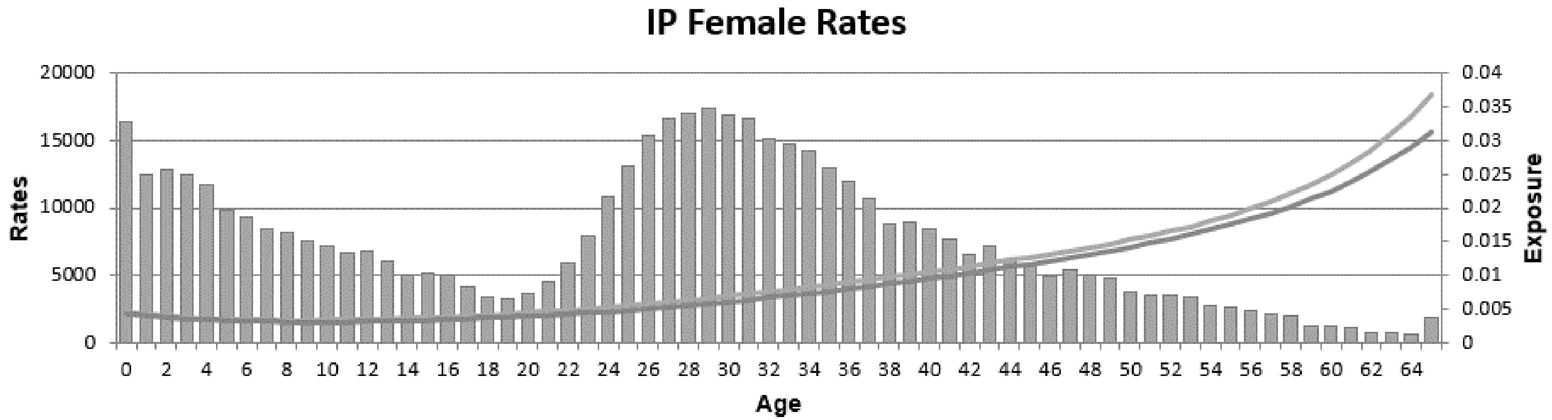
The riders can be various but some common riders are:

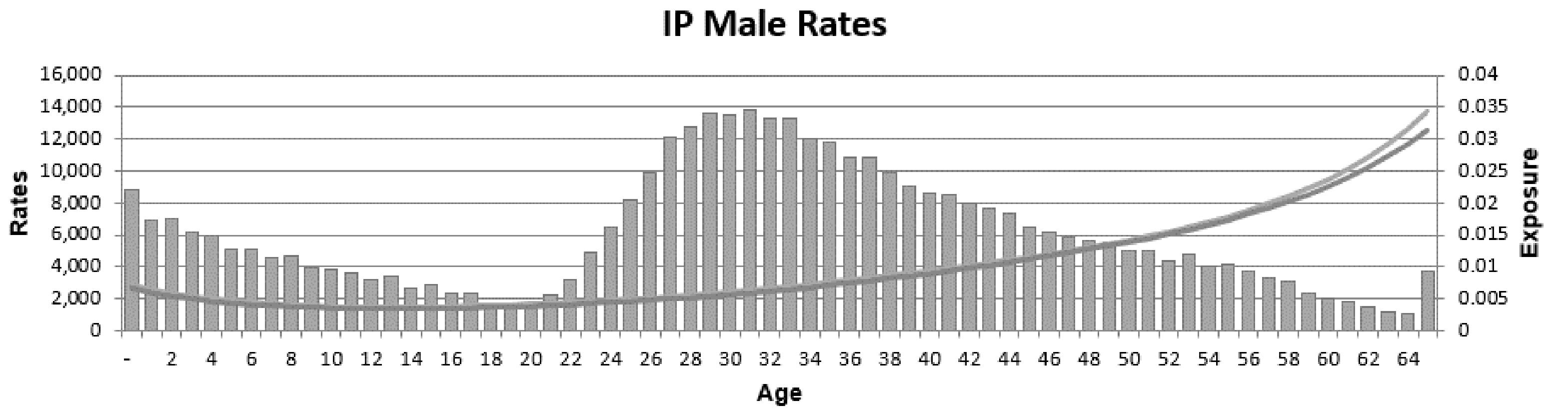
- Dental
- Optical
- Repatriation and Evacuation
- Alternative medicine
- Annual screening
- Routine health checkup
- Psychiatric treatment
- Vaccination
- Medical assistance and appliances
- Hepatitis
- Renal Dialysis
- Rehabilitation
- Toiletries

Additional loadings are applied to the base rates to determine the total burning cost. The actual loading factors that are generally used depend on the following rating factors:

- Segment
- Network
- Geographical limit
- Co-insurance
- Deductible
- Annual limit
- Sub limits
- Room and boarding

IP Base rates – 2021 (blue) vs 2022 (red)

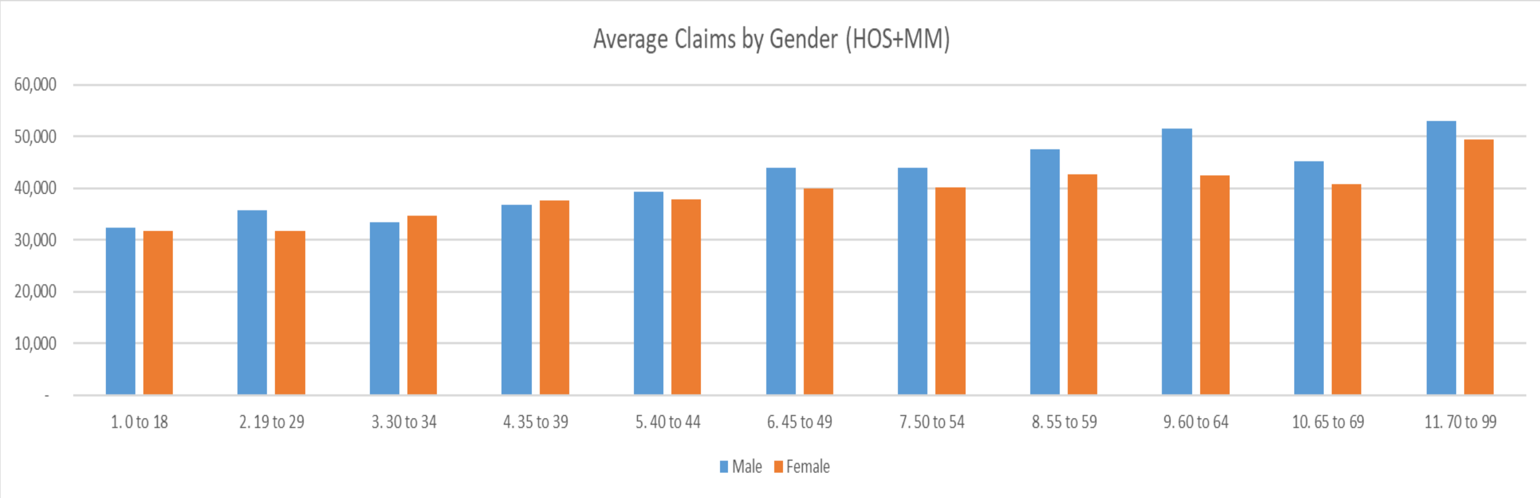




5) Key Results of Basic Burning Cost Pricing Analytics

Age and gender both have impact on the amount of premium for a healthcare product. The cost of healthcare service increases with age and is on average greater for females compared to males especially for certain age groups. The following graphs show the composition of the exposure by age groups and by gender:

IP excluding maternity	Total Earned Lives			Total number of claims		
Age	**Male**	**Female**	**Total**	**Male**	**Female**	**Total**
01. 0 to 18	21%	23%	22%	28%	26%	27%
02. 19 to 29	22%	22%	22%	12%	13%	13%
03. 30 to 34	14%	14%	14%	11%	11%	11%
04. 35 to 39	12%	12%	12%	10%	11%	11%
05. 40 to 44	10%	10%	10%	9%	10%	9%
06. 45 to 49	7%	7%	7%	8%	9%	8%
07. 50 to 54	6%	6%	6%	7%	7%	7%
08. 55 to 59	4%	4%	4%	6%	5%	6%
09. 60 to 64	2%	1%	1%	4%	4%	4%
10. 65 to 69	1%	1%	1%	2%	3%	3%
11. 70 to 99	1%	1%	1%	2%	2%	2%
Grand Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**



From the above graphs we can see that most of the exposure lies in the younger age bands.

Around 58% of the exposure pertains to ages up to 34 years (but this is decreasing over time. 6 years ago, this proportion was 75%). Similarly, males constitute around 54% of the total exposure while females constitute around 46% (this was 60% Male and 40% Female six years ago). Increase in female proportion means increase in claim costs/Burning Cost because while Female are cheaper than Male when it comes to HOS+MM, after inclusion of MAT into it HOS+MM+MAT female are more expensive than male).

By looking at Burning Cost for ages for HOS+MM only, we can quantify the impact of increase in average age proportion (58% portfolio having ages up to 34 years whereas till 6 years ago, this proportion was 75%):

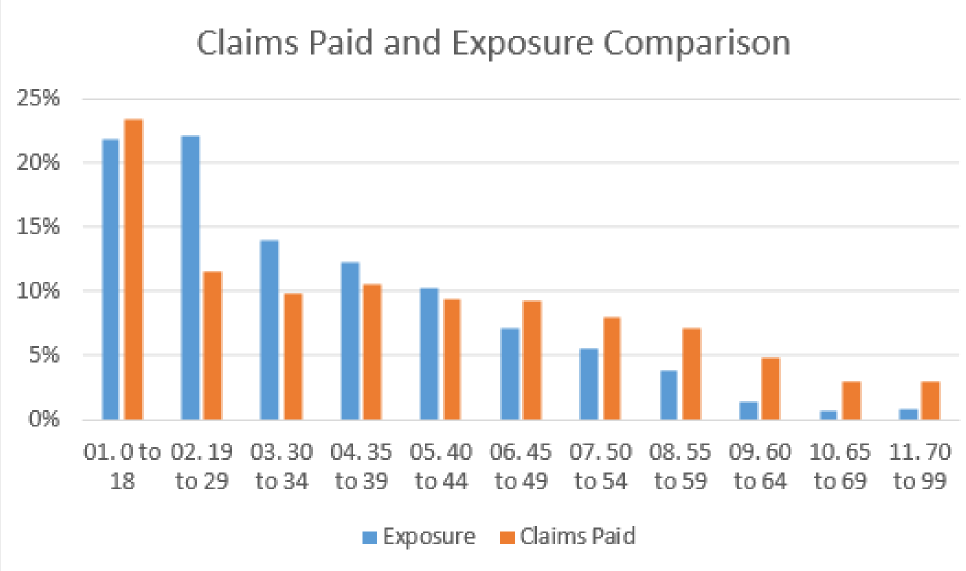
average BC 0 to 34 years	% population 0 to 34 years	Average BC 35 to 99 years	% population 35 to 99 years	Weighted Average BC	BC RF
1,032	58%	2,936	42%	1,831.95	1.21
1,032	75%	2,936	25%	1,508.30	

Age (IP excluding maternity)	Frequency RF	Severity RF	BC RF
01. 0 to 18	1.24	0.86	1.07
02. 19 to 29	0.57	0.91	0.52
03. 30 to 34	0.78	0.91	0.71
04. 35 to 39	0.86	0.99	0.86
05. 40 to 44	0.89	1.04	0.92
06. 45 to 49	1.14	1.13	1.29
07. 50 to 54	1.26	1.14	1.43
08. 55 to 59	1.55	1.22	1.89
09. 60 to 64	2.59	1.27	3.30
10. 65 to 69	3.58	1.16	4.15
11. 70 to 99	2.61	1.38	3.60

This shows that burning cost has increase by 21% due to effect of proportion of population within ages 34 years changing from 75% to 58% over 6 years' time period. This proportion is expected to decrease over time, and average age is expected to increase over time because 1) Country average age is increasing over time as mortality rates decrease and birth rates decrease and 2) insurer does not capture or target segments that have young ages or focuses on products that target young ages such as school, student policies, individual retail policies to mitigate this long-term trend. Cost curtailment features to reduce claims over time such as focus on prevention through wellness is immaterial as well. The only recourse hence is to continue increasing prices every year as higher premium gets charged upon increase in age.

Burning Cost is high for 0 to 18 years which is to be expected given high frequency of claims from babies and teenagers; within teenagers, accidents incidence for male increases from 15-25 years due to risk social behavior. BC falls than from 19 years and then keeps increasing progressively. For further look into these trends:

Age (IP excluding Maternity)	Exposure	Claims Paid	BC RF
01. 0 to 18	22%	23%	1.07
02. 19 to 29	22%	11%	0.52
03. 30 to 34	14%	10%	0.71
04. 35 to 39	12%	11%	0.86
05. 40 to 44	10%	9%	0.92
06. 45 to 49	7%	9%	1.29
07. 50 to 54	6%	8%	1.43
08. 55 to 59	4%	7%	1.89
09. 60 to 64	1%	5%	3.30
10. 65 to 69	1%	3%	4.15
11. 70 to 99	1%	3%	3.60



Here we can see that BC increases rapidly from 45 years and onwards for HOS+MM. 45 to 49 years increase is still moderate at yellow but 60 to 99 (99 is practically 75 years) increase is very high at red zone. Thus, we can categorize claim cost into 3 classes for ages for HOS+MM:

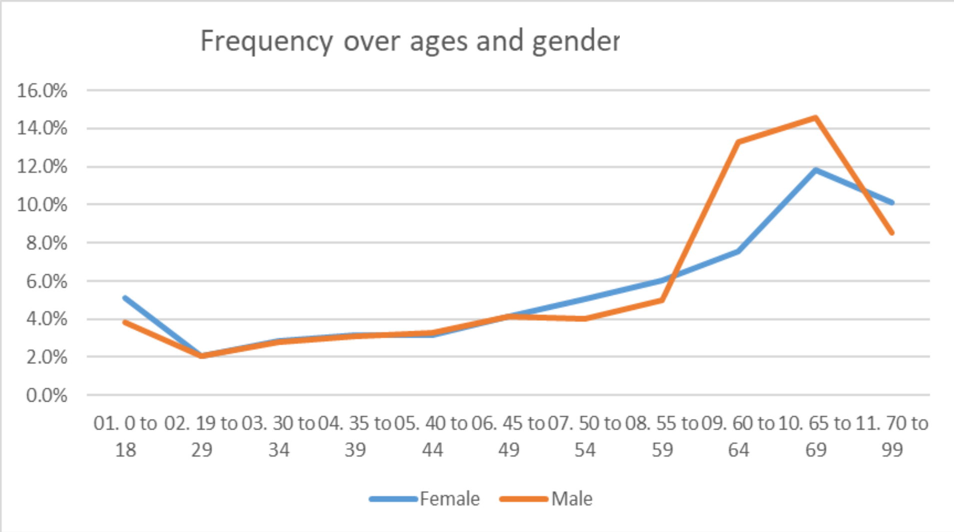
- green zone = 19 years to 44 years
- yellow zone = 0 to 18 years + 45 to 59 years,
- red zone = 60 to 75 years
- From above graph, we can also see that exposure to ages 40 years and above is 30% but claims paid are 45%. 6 years ago, exposure to ages 40 years and above was 17% but was 42% of claims paid. This shows that the skewness of distribution of age has reduced over the years which means that more people are growing older now but their claims burden is less extreme now than before.

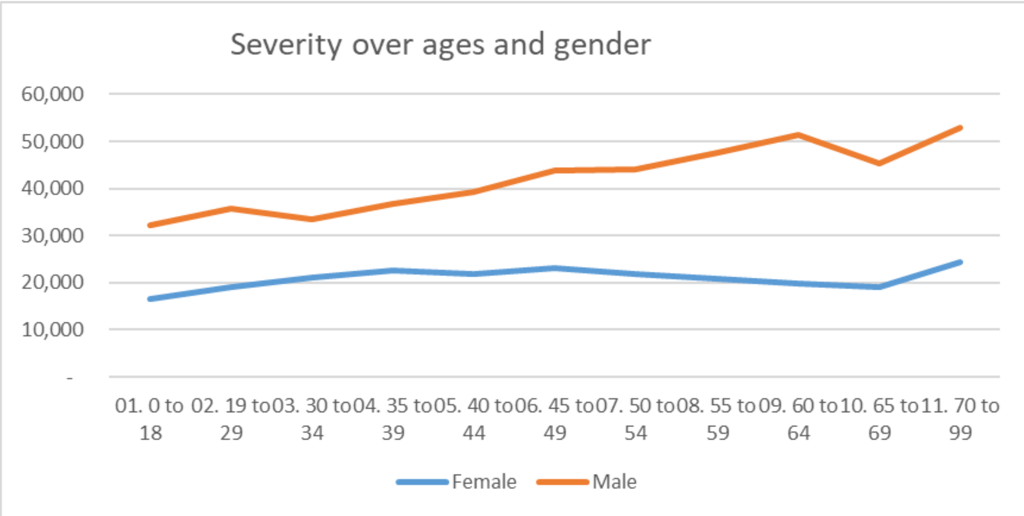
While under group health which is majority of insurer's portfolio usually, we cannot select which age comes to us under group client's population census, we should aim to increase our focus and target green zone and minimize the red zone. This can mean that we minimize underwriting parents in the policies where possible as parents have high ages. However, age has interaction effect with relationship and gender as well as seen here:

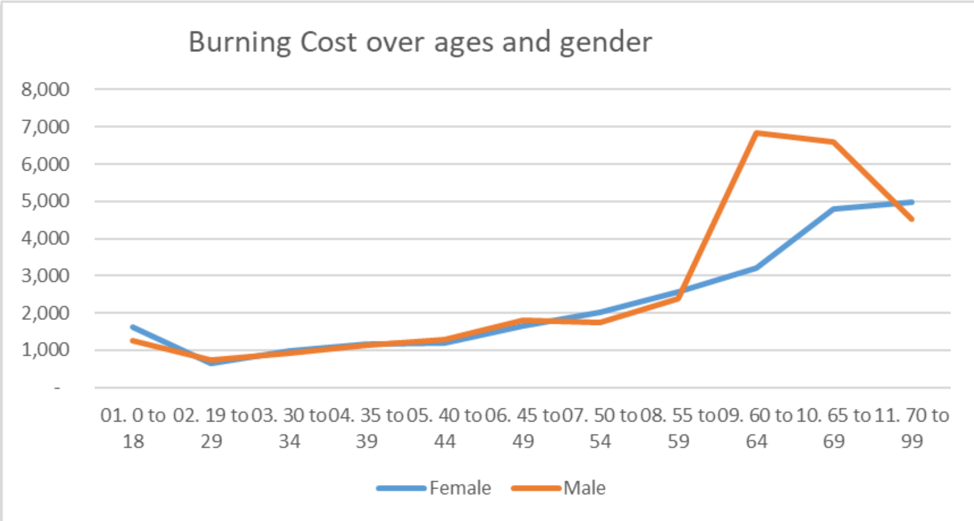
Gender	Relationship	BC RF
Female	Mother	1.01
Male	Father	1.17
Female	Daughter	0.87
Male	Son	0.89
Female	Wife	1.02
Male	Husband	1.19
Female	Ownself - Female	0.91
Male	Ownself - Male	1.07
Female	Married Female	1.02

Thus, we can see here that daughter and son have low costs due to lower ages and parents have higher costs due to high ages but also that this high cost is largely in father male parent rather than for female mother. Male is higher than female across mother father, daughter son, wife husband and ownself female+married female Vs Ownself male. Thus, our practical focus should be where possible to incentivize group clients to include children in their policies too or increase their coverage and minimize parents' coverage and increase rates for male on HOS+MM rather than on female.

The graphs shows the frequency, severity and burning cost for age-gender for Inpatient (excluding maternity):







From here we can see that frequency is higher overall for female than male. Severity is higher for male then female but frequency for female is high enough to push burning cost for female to be higher than for male.

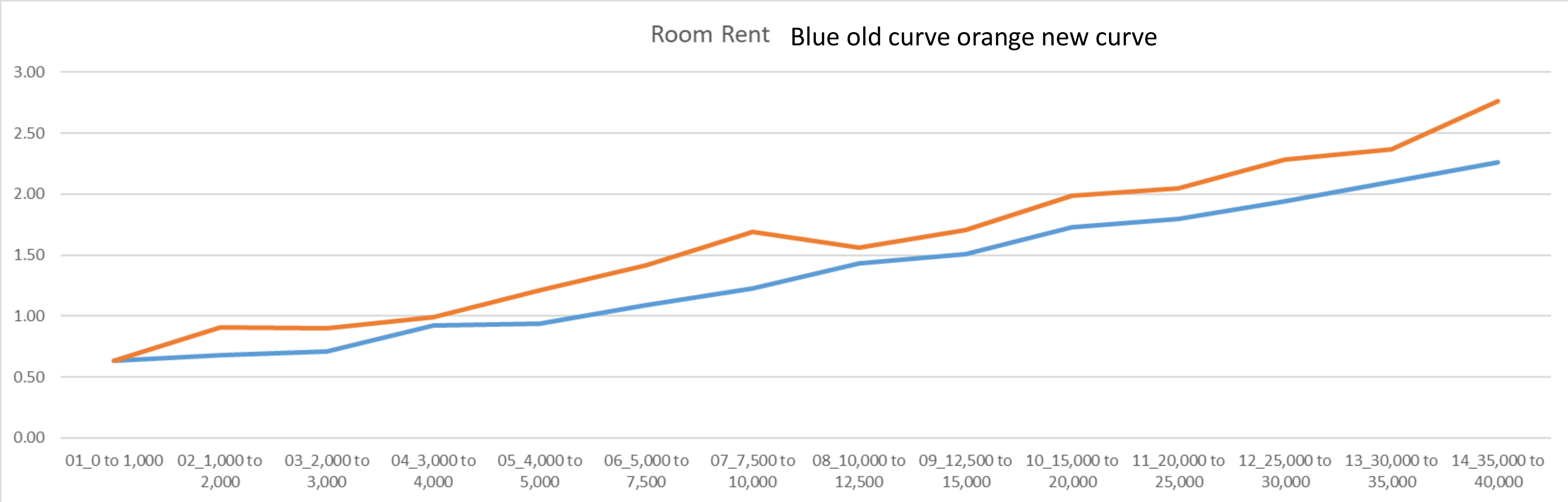
This is a universal trend that frequency is more careful/conscious about their health and so utilize healthcare for minor issues even whereas male avoid healthcare utilization due to culture that deters male from being conscious about their health. However, when male do get hospitalized, their claims are on average more severe than that of female then.

The life expectancy is also at play here. Burning Cost is high for Male at 60-69 years whereas it is high for female at 65-75 years. This is because female have 5 years higher life expectancy than male and this gap (65 years female – 60 years for male) reflects that.

Maternity Incidence rates are usually quite high between 7%-26% for maternity paying eligible married female only. Severity is also about 2 times higher than Inpatient excluding maternity claims. Average frequency is stable over time and slightly decreasing as fertility rates are falling over the years due to various social reasons but average severity is increasing due to higher C-Sections and medical inflation. C-section forms 60% by number of claims but 76% by claims paid; because its average claim is 211% times than that of normal meaning 100:211

A private health insurer's C-Section ratio is usually 2.6-3 times higher than the National average due to various reasons. Higher limits for C-section so greater financial incentive, doctors seek surgical experience so prefer C-Section, educated people in top 3 cities prefer C-section as it is less risky and time consuming. That is why Universal Social insurance programs put in same limits for C-Section or Normal maternity so as to remove any financial incentive from doing C-Section surgeries and do it only when medically necessary.

As we can see in the graph ahead, differentiated inflation over the years is at play here. Differentiated means that hospitals do not increase prices the same way across all room types; general wards are more price sensitive so prices are increased there to a lesser degree than in private rooms or VIP rooms where people have higher purchasing powers and so can meet the higher medical inflation rates:



Different results can come if we compare BC with premium as % of sum insured and loss ratio. For example, maternity can have high BC but have low loss ratio and high premium as % of sum insured because sum insured limits are low on average and premium charged for them is high. Major medical has lower frequency than hospitalization but higher severity. It can have high limits and high loss ratio as premium charged for major medical might be very competitive. OP has high frequency but very low severity. C-Section maternity has higher premium as % of sum insured than normal maternity. 60% of maternities are C-Section and 40% normal deliveries. These deliveries are 80% of total maternity claims because the rest 20% relates to complications of maternity where it leads to abortions and failed pregnancies.

The blanket statement that we should encourage customers to shift to network hospitals instead for healthcare utilization due to better controls and coordination is not 100% correct because network claims on average are 2.35 times higher than non-network claims too and so more network claims would increase our losses at no additional premium. That is why over time, market practice has become to waive any deductions to non-panel reimbursement claims so that losses are controlled for insurers. Another reason why network claims as % of total claims are decreasing over time is that people's purchasing power has increased over the years and so they prefer convenience and trust to utilize healthcare from providers that they have used historically and that have their trust. Going to network hospitals is that questions start first and treatments start later which micro-managing is a huge inconvenience to customers even though network offers cashless coverage. In non-network they have flexibility, trust, past historical experience and convenience so they pay first and claim from insurer later. Another determinant is different claim practices at different insurers. One insurer might compare charges with most expensive hospital in the city or even region for reimbursement whereas the other insurer might handle claims by comparing to reasonable and customary charges instead of most expensive leading to lower claims. One insurer might get higher volume discounts from Hospitals, providers and TPAs due to higher business volumes than other smaller insurers. One insurer might have better fraud wastage and abuse controls than other insurer leading to lower loss ratio or more profitable business segments.

There is also a Plan Matrix that has been analyzed. As you can see below, in Blue, plans in premium data and claims data are both the same. However, in green, the plans are higher in premium data but lower in claims data and vice versa for orange. This is a structural phenomenon and part of business as usual and is not an outlier or unusual thing.

Plans in Premium data	**A**	**B**	**C**	**D**	**E**	**F**	**G,H,I,J,S,V**
A	82%	19%	15%	6%	0%	0%	0%
B	10%	61%	34%	8%	1%	0%	0%
C	5%	14%	34%	16%	9%	3%	0%
D	2%	6%	15%	62%	51%	30%	21%
E	0%	0%	1%	3%	12%	25%	10%
F	0%	0%	0%	3%	13%	15%	4%
G,H,I,J,S,V	0%	0%	0%	3%	14%	27%	65%
Grand Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

	claims paid	**number of claims**	**Severity RF**
claims paid same plans in premium and claims	61%	66%	0.93
claims paid higher in premium lower in claims	22%	23%	0.96
claims paid lower in premium higher in claims	18%	12%	1.49

The main usual reasons for blue not being 100% same across policy and claims dataset are:

1) in premiums, person gets registered in B but after few months he claims and employer says he has been promoted to plan A so his claims get paid under plan A
2) ad-hoc special considerations given by medical claims team upon request of client;
3) third reason during year company says economy is bad we are putting most of people from plan B to plan C to reduce our premium burden of group health; So, plan A was registered but claim occurred before plans could be downgraded.
4) emergency; whatever plan is available gets allocated in times of urgency and emergency to be decided later onwards. Payment is given by insurer approved as matter is time sensitive and group client pays back any amounts paid over the limits to the insurer later onwards.
5) If a treatment or procedure is not available at a provider of adherent class then the insured may be referred to a provider from a higher provider class
6) The proportion of business where premium is lower plan but claims is higher should be aimed to be minimized where possible as it is a moderate cause of higher-than-otherwise losses.

82% of total claims paid are in Plans A, B and C. 74% of total number of claims are in Plans A, B and C only.

Room Type	% of total claims paid	% of total no. of claims
Private Room	53%	51%
Semi Private	27%	29%
General Ward	19%	20%

Plan Type % of total claims paid	Private Room	Semi Private	General Ward
A	28%	21%	22%
B	32%	32%	27%
C	22%	28%	31%

Top 3 cities form 79% of the total claims paid. Top 10 cities form 88% of the number of claims.

91% of billed is paid. Claim rejection ratio is evaluated based on two criteria. First is that whatever claim amount is reported, zero amount is paid because claim is rejected all together. Hence, here claim numbers are seen rather than amounts. Based on this, the claims acceptance ratio first method is 95% which means that only 5% of claims are rejected altogether due to reasons like being ineligible for pre-existing coverage etc. The second method is to see where claims are paid partially than reported so claim amounts for those selected claims is seen. Partial payments can be due to sub-limit deductions, other various deductions like room-rent, higher plan, limit bursting and other reasons. As per the second method, the claims acceptance ratio is 90% which means that 10% deductions are there as deductions for partial from reported claims. Total acceptance ratio is then 85% where 5% claims are rejected altogether and 10% claims deductions are there.

Claim development table shows the important feature that claims get developed faster on loss/accident year than on policy/risk/underwriting year basis. Timely development is important for up-to-date business decision making. This is going to be an issue as IFRS17 calls for underwriting year whereas currently IFRS4 focuses on accident development which is faster.

Claim Development	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Claims Paid in Loss/Accident Year as % of Claims Paid in Calendar/Financial Year	102%	100%	106%	100%	99%	84%
Claims Paid in Risk Year as % of Claims Paid in Calendar/Financial Year	111%	104%	97%	113%	96%	62%

Regarding type of claim charge, room and board is not significant expense as it is just 8% of claims paid on average. pre and post hospitalization is also not significant claims as it is just 1% of claims paid; ex gratia is 0%. medicines is important cost 24% of total claims paid; analysis is still incomplete because main category of 67% of claims paid is "Others" which we need to open up. what does "Others" represent? Surgeries? Day care procedures? Tests?

Deductions are 15% on claims paid. Reasons for claim deductions need to be opened up because “Other Reasons” form 96% of the deductions and reasons like out-of-network deduction, pre-existing deduction, room rent difference deduction and limit sum insured bursting deduction combined form only 4% of deductions. The selection of room which will make overall difference upon the charges, higher the room more will be the package charges for a particular procedure. Hence the overall deduction shall be applicable considering the actual room type or limit awarded against the actual utilization. Usually consultation, surgery, Anesthesia and OT Charges varies in each room category. More the cost of room, higher would be these charges. In case an individual opted for room of more than his limit (Private room instead of semi-private then he has to pay the difference of not only the room charges but difference of consultation, surgery, anesthesia and OT charges also. even if he has sufficient limit to accommodate the charges of higher room. Even in case of package pricing, he has to pay the difference of package. Suppose his room limit is 5000 (Semi-Private) and he opted for 10,000 private) the difference would be of package in case of package pricing and in case of open billing it would be difference in rates of two rooms, surgery, anesthesia and OT charges, even if his limit is sufficient.

Key results across diagnostic codes reveals a wealth of information on burden of diseases in the population covered by the company. Like Nationally, there is double burden of disease because infections are rising over time (due to climate change, antibiotic/antimicrobial resistance, worsening pollution, COVID19, rising catastrophes, heatwaves, floods, mosquitos, water crisis) and increase in non-communicable diseases (NCD) over time as population ages and NCDs increase due to lifestyle factors. Heart attacks, gall bladder removal surgeries, diabetes, hypertension, appendix issues, mental health issues, cancers, C-sections in maternity, are increasing claims. Increasing burden of diseases means long-term increasing claims which can be countered by increasing premiums over time, focusing on preventive measures such as awareness campaigns, executive checkups, provision of nets and first aid, making wellness products, digitization to streamline processes and reduce expenses and so on.

Sample Medical Actuarial Pricing Basic Burning Cost and Advanced Machine Learning Report

Diagnostic Codes Description	**% proportion of paid claims**	**% of total no of claims**	**Severity RF**
Pregnancy, childbirth and the puerperium	23%	11%	2.0
Diseases of the digestive system	12%	10%	1.2
Injury, poisoning and certain other consequences of external causes	8%	12%	0.7
Diseases of the circulatory system	7%	5%	1.3
Diseases of the genitourinary system	7%	5%	1.3
Diseases of the respiratory system	7%	10%	0.6
Codes for special purposes	6%	13%	0.4
Certain infectious and parasitic diseases	5%	4%	1.2
Neoplasms	4%	2%	2.4
Diseases of the eye and adnexa	4%	4%	1.1
Diseases of the musculoskeletal system and connective tissue	3%	5%	0.7
Endocrine, nutritional and metabolic diseases	3%	6%	0.6
Certain conditions originating in the perinatal period	3%	2%	1.9
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	3%	6%	0.5
Congenital malformations, deformations and chromosomal abnormalities	1%	0%	3.4
Diseases of the nervous system	1%	1%	1.6
Diseases of the skin and subcutaneous tissue	1%	2%	0.6
Factors influencing health status and contact with health services	1%	2%	0.5
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	0%	0%	1.0
Diseases of the ear and mastoid process	0%	0%	1.1
Mental and behavioural disorders	0%	0%	0.5

This shows that maternity forms the largest proportion by claims paid, then digestive system issues and then injury poisoning and external causes which includes COVID19. The top 3 by number of claims are codes for special purposes, injury and poisoning and external causes and pregnancy. Severity is high for those above 1 Severity RF and lower than the average for those with lower than 1 Severity RF. Congenital is expensive then neoplasms cancers then pregnancy.

Diagnostic Codes Description	Severity RF
Congenital malformations, deformations and chromosomal abnormalities	3.4
Neoplasms	2.4
Pregnancy, childbirth and the puerperium	2.0
Certain conditions originating in the perinatal period	1.9
Diseases of the nervous system	1.6
Diseases of the genitourinary system	1.3
Diseases of the circulatory system	1.3
Certain infectious and parasitic diseases	1.2
Diseases of the digestive system	1.2
Diseases of the ear and mastoid process	1.1
Diseases of the eye and adnexa	1.1
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	1.0
Diseases of the musculoskeletal system and connective tissue	0.7
Injury, poisoning and certain other consequences of external causes	0.7
Diseases of the respiratory system	0.6
Endocrine, nutritional and metabolic diseases	0.6
Diseases of the skin and subcutaneous tissue	0.6
Mental and behavioural disorders	0.5
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	0.5
Factors influencing health status and contact with health services	0.5
Codes for special purposes	0.4

We can also see by sub-codes instead of by main codes only as follows. The top 30 sub-codes form 77% of total claims paid and 72% of total number of claims:

main codes description	sub codes description	% proportion of paid claims	% of total no of claims	Severity RF
Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	10%	4%	2.4
Pregnancy, childbirth and the puerperium	Other maternal disorders predominantly related to pregnancy	9%	5%	1.6
Diseases of the circulatory system	Ischaemic heart diseases	4%	1%	3.0
Diseases of the digestive system	Noninfective enteritis and colitis	3%	4%	0.8
Codes for special purposes	Persons encountering health services in circumstances related to reproduction	3%	7%	0.4
Diseases of the eye and adnexa	Disorders of lens	3%	2%	1.2
Diseases of the respiratory system	Pneumonia in diseases classified elsewhere	3%	2%	1.6
Diseases of the respiratory system	Acute upper respiratory infections	3%	6%	0.4
Diseases of the genitourinary system	Renal tubulo-interstitial disorders in diseases classified elsewhere	3%	2%	1.3
Diseases of the digestive system	Liver disorders in diseases classified elsewhere	2%	1%	3.2
Endocrine, nutritional and metabolic diseases	Diabetes mellitus	2%	4%	0.6
Injury, poisoning and certain other consequences of external causes	Sequelae of injuries, of poisoning and of other consequences of external causes	2%	1%	1.9
Diseases of the genitourinary system	Vulvovaginal ulceration and inflammation in diseases classified elsewhere	2%	1%	1.4
Certain infectious and parasitic diseases	Intestinal infectious diseases	2%	2%	1.1
Pregnancy, childbirth and the puerperium	Pregnancy with abortive outcome	2%	1%	1.9
Certain conditions originating in the perinatal period	Disorders related to length of gestation and fetal growth	2%	1%	2.0
Factors influencing health status and contact with health services	Persons encountering health services for examination and investigation	2%	5%	0.3
Pregnancy, childbirth and the puerperium	Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium	2%	1%	1.8
Certain infectious and parasitic diseases	Viral infections of the central nervous system	2%	1%	1.3
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	General symptoms and signs	2%	4%	0.4
Diseases of the circulatory system	Hypertensive diseases	2%	4%	0.4
Diseases of the digestive system	Hernia	2%	1%	2.8
Diseases of the genitourinary system	Urethral disorders in diseases classified elsewhere	2%	2%	1.0
Diseases of the musculoskeletal system and connective tissue	Dorsopathies M50-M54 Other dorsopathies	2%	2%	0.7
Diseases of the digestive system	Other diseases of intestines	2%	1%	1.4
Diseases of the respiratory system	Pleural effusion in conditions classified elsewhere	1%	3%	0.5
Diseases of the digestive system	Diseases of appendix	1%	0%	2.3
Diseases of the digestive system	Diseases of oral cavity, salivary glands and jaws	1%	2%	0.5
Neoplasms	Benign neoplasms	1%	0%	2.8
Diseases of the digestive system	Diseases of oesophagus, stomach and duodenum	1%	2%	0.5

Severity is high for those above 1 Severity RF and lower than the average for those with lower than 1 Severity RF. Liver disorders are expensive then ischemic heart diseases then benign neoplasms and hernia.

main codes description	sub codes description	Severity RF
Diseases of the digestive system	Liver disorders in diseases classified elsewhere	3.2
Diseases of the circulatory system	Ischaemic heart diseases	3.0
Neoplasms	Benign neoplasms	2.8
Diseases of the digestive system	Hernia	2.8
Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	2.4
Diseases of the digestive system	Diseases of appendix	2.3
Certain conditions originating in the perinatal period	Disorders related to length of gestation and fetal growth	2.0
Pregnancy, childbirth and the puerperium	Pregnancy with abortive outcome	1.9
Injury, poisoning and certain other consequences of external causes	Sequelae of injuries, of poisoning and of other consequences of external causes	1.9
Pregnancy, childbirth and the puerperium	Oedema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium	1.8
Pregnancy, childbirth and the puerperium	Other maternal disorders predominantly related to pregnancy	1.6
Diseases of the respiratory system	Pneumonia in diseases classified elsewhere	1.6
Diseases of the genitourinary system	Vulvovaginal ulceration and inflammation in diseases classified elsewhere	1.4
Diseases of the digestive system	Other diseases of intestines	1.4
Certain infectious and parasitic diseases	Viral infections of the central nervous system	1.3
Diseases of the genitourinary system	Renal tubulo-interstitial disorders in diseases classified elsewhere	1.3
Diseases of the eye and adnexa	Disorders of lens	1.2
Certain infectious and parasitic diseases	Intestinal infectious diseases	1.1
Diseases of the genitourinary system	Urethral disorders in diseases classified elsewhere	1.0
Diseases of the digestive system	Noninfective enteritis and colitis	0.8
Diseases of the musculoskeletal system and connective tissue	Dorsopathies M50-M54 Other dorsopathies	0.7
Endocrine, nutritional and metabolic diseases	Diabetes mellitus	0.6
Diseases of the digestive system	Diseases of oral cavity, salivary glands and jaws	0.5
Diseases of the digestive system	Diseases of oesophagus, stomach and duodenum	0.5
Diseases of the respiratory system	Pleural effusion in conditions classified elsewhere	0.5
Codes for special purposes	Persons encountering health services in circumstances related to reproduction	0.4
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	General symptoms and signs	0.4
Diseases of the circulatory system	Hypertensive diseases	0.4
Diseases of the respiratory system	Acute upper respiratory infections	0.4
Factors influencing health status and contact with health services	Persons encountering health services for examination and investigation	0.3

Lastly, top 35 primary diagnosis description form 63% of total claims paid and 69% of total number of claims and are shown below:

Ranking	maincodes description	sub codes description	primary diagnosis description	% proportion of total claims paid of top 100	% proportion of total no of claims of top 100	Severity RF
1	Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	Maternal care for unspecified type scar from previous cesarean delivery	8%	2%	3.36
2	Diseases of the eye and adnexa	Disorders of lens	Other specified cataract	3%	1%	2.45
3	Endocrine, nutritional and metabolic diseases	Diabetes mellitus	Type 2 diabetes mellitus without complications	3%	4%	0.72
4	Pregnancy, childbirth and the puerperium	Other maternal disorders predominantly related to pregnancy	Malnutrition in childbirth	3%	1%	2.55
5	Injury, poisoning and certain other consequences of external causes	Sequelae of injuries, of poisoning and of other consequences of external causes	Covid - 19, virus identified	3%	1%	3.75
6	Codes for special purposes	Persons encountering health services for examination and investigation	Encounter for other specified special examinations	3%	4%	0.53
7	Diseases of the digestive system	Noninfective enteritis and colitis	Noninfective gastroenteritis and colitis, unspecified	2%	2%	0.85
8	Codes for special purposes	Persons encountering health services in circumstances related to reproduction	Encounter for antenatal screening, unspecified	2%	4%	0.50
9	Diseases of the circulatory system	Hypertensive diseases	Essential (primary) hypertension	2%	3%	0.60
10	Certain infectious and parasitic diseases	Intestinal infectious diseases	Typhoid fever, unspecified	2%	1%	1.97
11	Certain infectious and parasitic diseases	Viral infections of the central nervous system	Dengue fever [classical dengue]	2%	1%	2.23
12	Diseases of the digestive system	Liver disorders in diseases classified elsewhere	Calculus of gallbladder and bile duct without cholecystitis without obstruction	2%	0%	4.85
13	Diseases of the respiratory system	Pleural effusion in conditions classified elsewhere	Other specified respiratory disorders	2%	3%	0.44
14	External causes of morbidity and mortality	Ill-defined and unknown causes of mortality	UNKNOWN AND UNSPECIFIED CAUSE	1%	20%	0.06
15	External causes of morbidity and mortality	Transport accidents	Person injured in unspecified motor-vehicle accident, traffic, initial encounter	1%	1%	2.32
16	Diseases of the circulatory system	Ischaemic heart diseases	Atherosclerotic heart disease of native coronary artery without angina pectoris	1%	0%	7.55
17	Diseases of the respiratory system	Acute upper respiratory infections	LRTI /RTI/CHEST INFECTIONS	1%	4%	0.34
18	Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	Maternal distress during labor and delivery	1%	0%	3.08
19	Codes for special purposes	Persons encountering health services in circumstances related to reproduction	Encounter for care and examination of mother immediately after delivery	1%	2%	0.50
20	Injury, poisoning and certain other consequences of external causes	Transport accidents	ROAD TRAFFIC ACCIDENT	1%	1%	1.97
21	Pregnancy, childbirth and the puerperium	Pregnancy with abortive outcome	MISSED ABORTION	1%	1%	1.38
22	Diseases of the genitourinary system	Renal tubulo-interstitial disorders in diseases classified elsewhere	CALCULUS OF URETER	1%	0%	3.37
23	Codes for special purposes	Persons encountering health services in circumstances related to reproduction	Encounter for supervision of other normal pregnancy, unspecified trimester	1%	1%	1.03
24	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	General symptoms and signs	Fever presenting with conditions classified elsewhere	1%	2%	0.57
25	Diseases of the digestive system	Noninfective enteritis and colitis	GASTROENTERITIS (acute, catarrhal, chronic, congestive, hemorrhagic,non-infectious)	1%	2%	0.59
26	Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	Other uterine inertia	1%	0%	3.22
27	Factors influencing health status and contact with health services	Persons encountering health services for examination and investigation	Encounter for general adult medical examination without abnormal findings	1%	2%	0.53
28	Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	CURRENT CONDITION EFFECTING PREG. DEL	1%	0%	2.48
29	Pregnancy, childbirth and the puerperium	Maternal care related to the fetus and amniotic cavity and possible delivery problems	NON PROGRESS OF LABOR/UNSPECIFIED UTERINE INERTIA DEL	1%	0%	2.99
30	Neoplasms	Benign neoplasms	Leiomyoma of uterus, unspecified	1%	0%	4.33
31	Injury, poisoning and certain other consequences of external causes	Persons without reported diagnosis encountered during examination and investigation c	PRE/POST HOSPITALIZATION CONSULTATION/MEDICINES	1%	2%	0.38
32	Diseases of the circulatory system	Ischaemic heart diseases	Chronic ischemic heart disease, unspecified	1%	0%	1.81
33	Endocrine, nutritional and metabolic diseases	Disorders of other endocrine glands	Type 2 diabetes mellitus without complications	1%	2%	0.51
34	Diseases of the genitourinary system	Renal tubulo-interstitial disorders in diseases classified elsewhere	CALCULUS OF KIDNEY	1%	0%	2.44
35	Diseases of the digestive system	Liver disorders in diseases classified elsewhere	Calculus of gallbladder without cholecystitis without obstruction	1%	0%	4.07

A lot of useful insights can be derived from this. For example, gallbladder removal surgery is a top hospitalization reason due to lifestyle factors. Infections, digestive issues are very common and so is COVID19 since 2019 till 2022. Diabetes and hypertension issues are very common and increasing although only the complications part is expensive. This suggests need to make specialized product for diabetes managed care and inpatient hospitalization for complications of diabetes. Similarly, we can see congenital issues having high costs, knee replacement as well as heart attacks and cancers. Traffic accidents need a separate product given how common they are. Also, we can make specialized tropical diseases cover that covers dengue, malaria, zika, chikungunya and enchiphalitis as well as product for WaSH diseases (Water, sanitation hygiene issues as most digestive issues are from WaSH). Kindley liver like hepatitis also need to seen in detail especially hepatitis B and C which should show decrease in both frequency and severity since modern treatments have become affordable and widespread over the past decade.

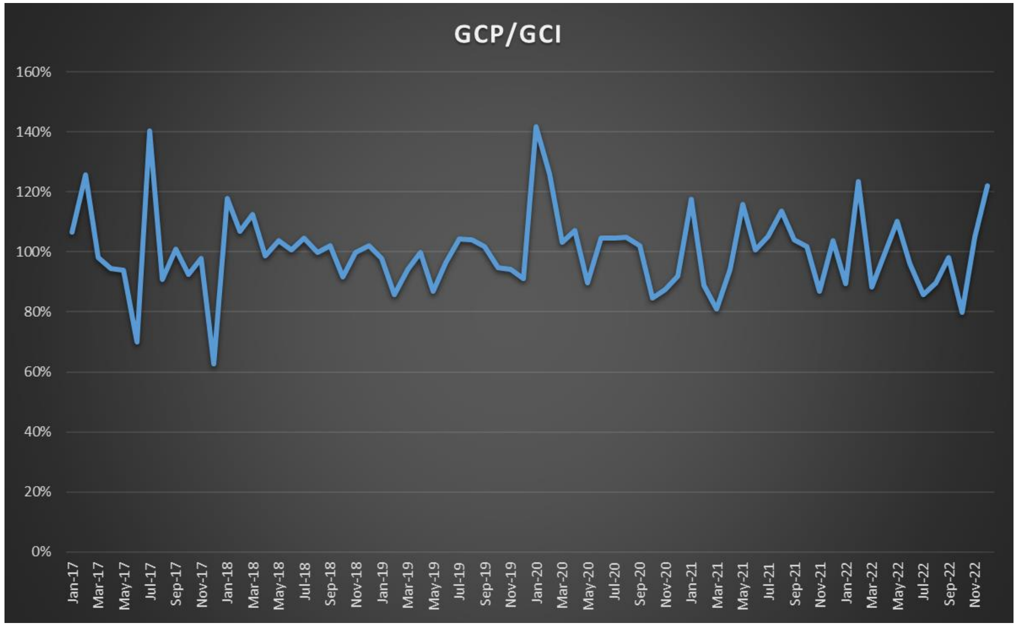
The results by top 20 procedures:

Top 20 First Procedures Description	sum of payable	Count	Average
Extraction of Products of Conception Low Open Approach	1,269,885,266	21,929	57,909
Low Cervical Cesarean Section	652,561,057	12,316	52,985
Resection of Gallbladder Percutaneous Endoscopic Approach	181,910,555	6,367	28,571
SPONTANEOUS VAGINAL DELIVERY	170,329,801	2,163	78,747
Delivery of Products of Conception External Approach	158,321,015	5,927	26,712
Division of Female Perineum External Approach	131,935,517	2,290	57,614
X-RAY OF SPECIFIED SITE	128,696,147	1,888	68,165
Ultrasonography of Right and Left Heart	113,318,995	3,224	35,149
Extraction of Right Lens Percutaneous Approach	110,406,357	554	199,289
ROUTINE CHEST X-RAY	95,770,099	1,202	79,676
PHACOEMELCIFICATION	94,239,730	205	459,706
ANGIOPLASTY OF CORONARY VESSEL(PTCA)	92,972,198	348	267,161
Introduction of Other Therapeutic Substance into Respiratory Tract Via Natural or Artificial Opening	91,447,103	1,503	60,843
LAPAROSCOPIC CHOLECYSTECTOMY	91,124,358	939	97,044
Introduction of Other Antineoplastic into Peripheral Vein Percutaneous Approach	69,004,858	2,001	34,485
Dilation of Coronary Artery One Artery Percutaneous Approach	66,377,104	2,851	23,282
Magnetic Resonance Imaging (MRI) of Brain	61,075,004	1,308	46,693
Extraction of Products of Conception Retained Via Natural or Artificial Opening	55,887,498	214	261,157
Performance of Urinary Filtration Intermittent Less than 6 Hours Per Day	54,962,097	199	276,191

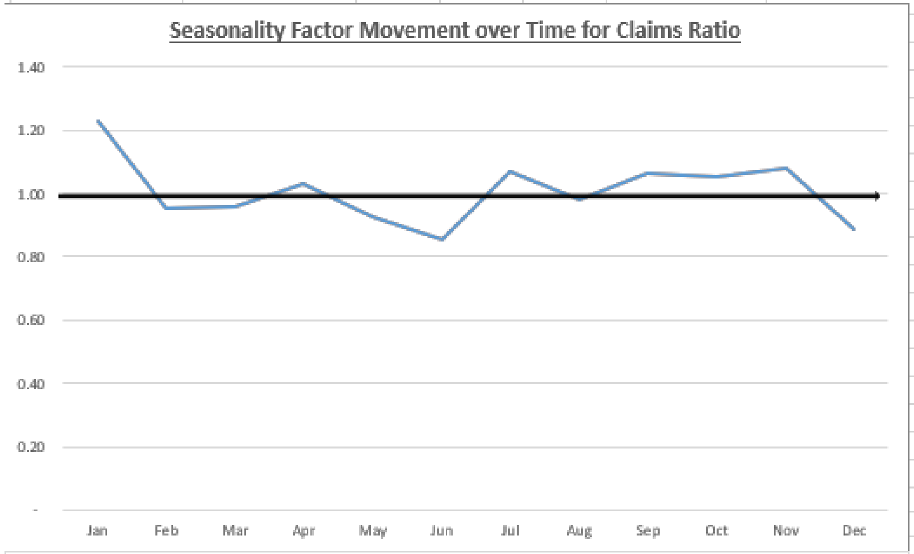
Analyzing results by top 30 hospital providers, we can see that top 30 hospitals account for 57% of total claims paid and 63% of total number of claims: Well-known big hospital brands and general practitioners clubbed by city form top 30 hospitals and healthcare providers. The top hospitals are on network and average claim size is quite high as these are on network plus because more complicated cases are handled by them and because their fees and quality are much higher than average.

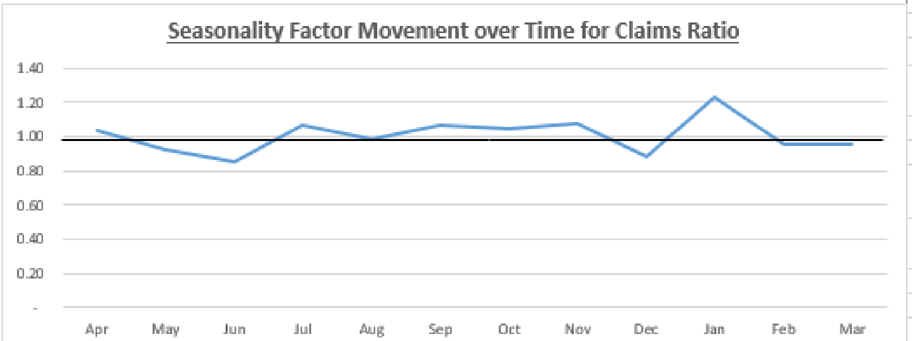
There would be some method to categorize the hospitals by tiers of brand, quality, fees; Analysis by making tiers of hospital providers can increase our analysis and lead to insights that are not observed right now. Medical claims team would be the best resource to categorize healthcare providers into Tiers.

We also investigated seasonality. Lockdowns, summer, winter, dengue season, smog, monsoon floods, heatwaves are the usual seasonality factors. But the most important seasonality factor is renewal season of Q4 Dec-Jan and Q2 June-July. Dec claim ratio goes down and shoots up in Jan; June claim ratio goes down but shoots up in July. One way to monitor claim settlement speed is to take ratio of Gross Claims Paid (GCP) to Gross Claims Incurred (GC):



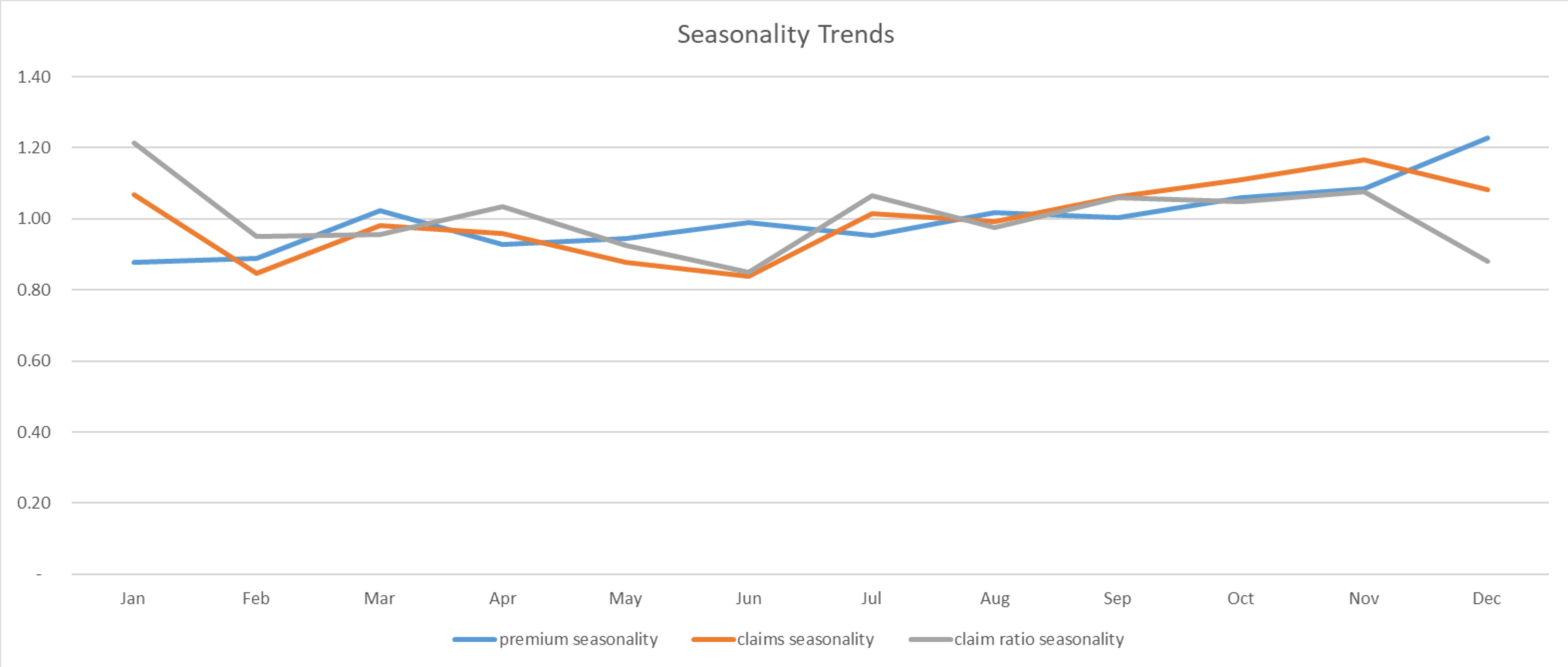
This shows how claims paid decrease in months of Ramadan and speed up after Eid. These also show any reduction in claims paid due to let's say year-end's additional workloads. It can also show how factors like for example claim settlement speed picked up at year end 2020 (Dec and Jan 2020). This is important as this effects the level at which Outstanding claims are held which in turn impacts IBNR.





But since claims ratio takes into account both claims and premiums, is this decrease in Dec and increase in Jan claims ratio and decrease in May-June and increase in July claims ratio being caused due to premium movements or claim movements? To dissect this, we look at this table and graph (above 1 means over the average and below 1 means lower than the average):

Months	premium seasonality	claims seasonality	claim ratio seasonality
Jan	0.88	1.07	1.21
Feb	0.89	0.85	0.95
Mar	1.02	0.98	0.96
Apr	0.93	0.96	1.04
May	0.95	0.88	0.93
Jun	0.99	0.84	0.85
Jul	0.95	1.02	1.07
Aug	1.02	0.99	0.98
Sep	1.00	1.06	1.06
Oct	1.06	1.11	1.05
Nov	1.08	1.17	1.08
Dec	1.23	1.08	0.88



Here we can see that Dec decrease in claims ratio is both due to increase in premium above the average and claims although premium seasonality is higher at 1.23 than for claims side 1.08. This makes sense because in renewal season, a substantial portion of clients are renewed in December as well as substantial amount of premium is earned over month of Dec as many policies are 1st Jan to 31st Dec.

The substantial seasonal increase in claims ratio in Jan is because of below the average premium in Jan and above average claims in Jan. Premiums drop sharply in Jan as much of renewal business gets covered in December and few business gets renewed in Jan. However, since claims factor is 1.08 in Dec and 1.07 in Jan this can mean higher than above claims due to winter (winter increases healthcare problems and utilization) and it can also mean that in renewal times, client hold on to claims and under-report them to get better renewal prices and report them in one go pent up once renewal has been completed.

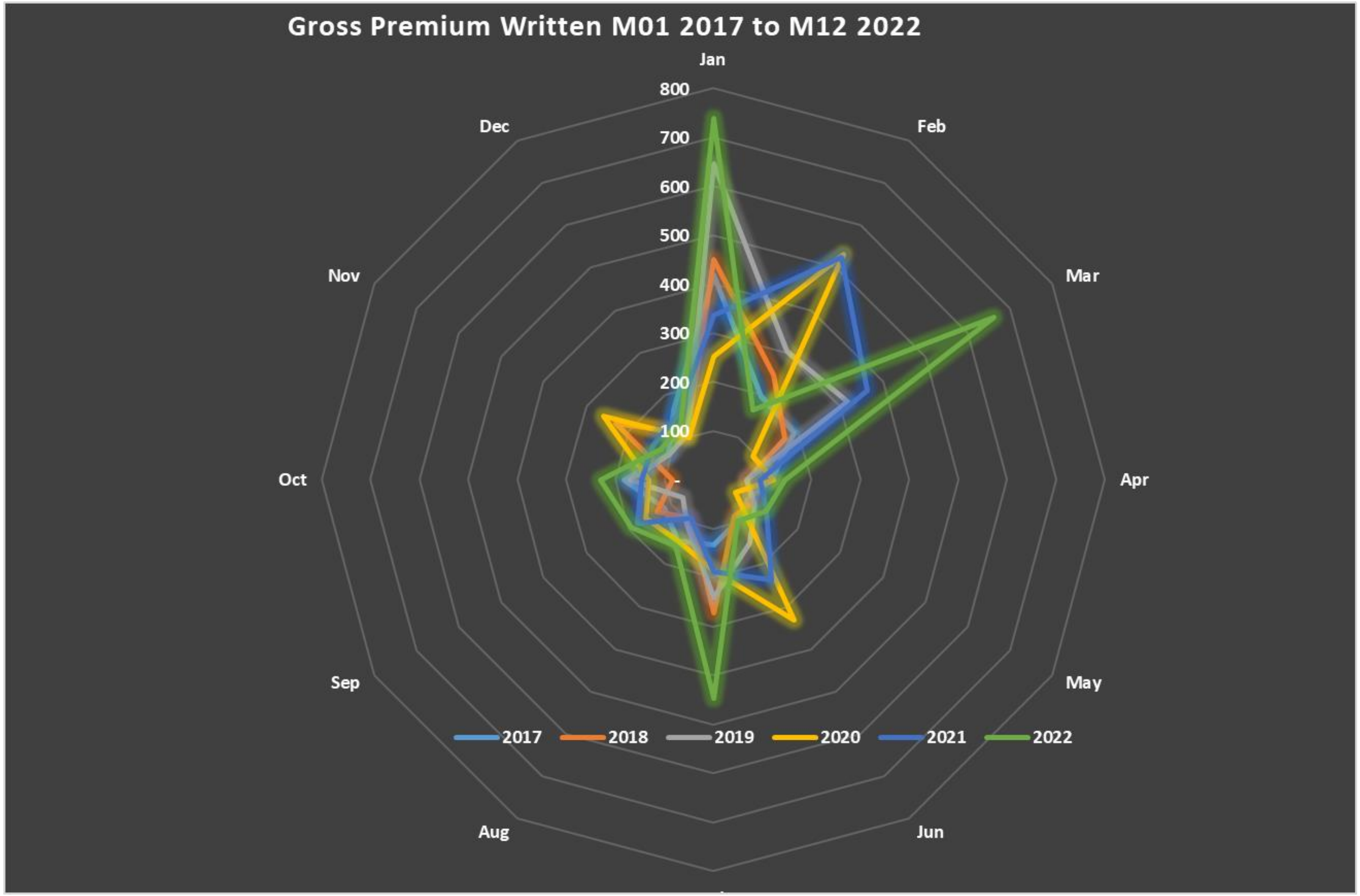
The high claims in Oct-Nov of 1.11-1.17 are related to higher-than-average maternity claims (marriage season is in winter (then add 9-10 months to it). Oct Nov also have higher claims due to higher illnesses during winter season changing and smog and dengue season.

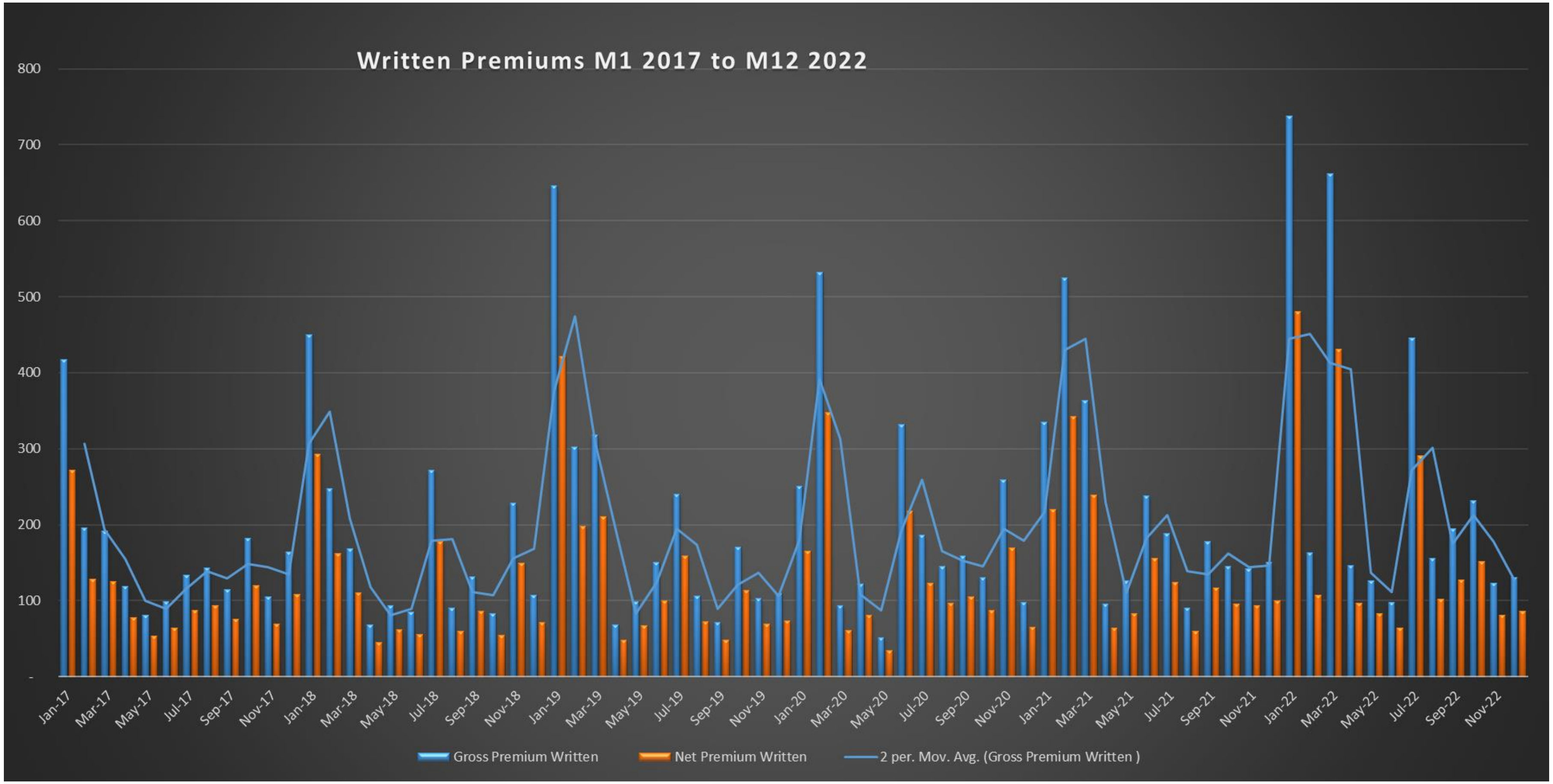
The second major renewal season Q2 shows that claims ratio decrease in May June and increase in July. Here premium seasonality is near to 1 in all 3 months so fluctuation is coming from claim side. This can be due to lower claims during summer months but it can also be under-reporting of claims by clients in order to get best renewal rates and then pent-up reporting those claims in July to get paid once renewal has been done in June. This emphasizes to us that we should not be overly focused on just the latest year's claim ratio as clients demand to act as benchmark for renewal premium price but recognize that we have to be prudent since there is element of under-reporting here as well. Since over seasonality trend is average over 6 years 2017-2022, April, May, June claims are also decreasing because people defer elective procedures during Ramadan months which feel during these 3 months over these 6 years.

Another factor is that insurers treat new business and renewal differently. New business has very competitive rates (price walking practice which is nearly universal) but rates are made more sustainable over time and/or claim approval is made more stringent over time. Thus, important to evaluate results through this segment of new business and renewal business separately too. As we can see, loss ratio is on average 33% higher for new business than renewal business but it doesn't increase the overall loss ratio by that much because new business is on average 20% of total earned premium and 25% of total claims paid:

Loss Ratio	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
New Business	94%	93%	111%	129%	120%	110%
Renewal Business	73%	78%	83%	70%	80%	78%
Difference	21%	15%	28%	59%	40%	32%
Total overall	76%	80%	92%	78%	88%	85%
% of earned premium	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
New Business	16%	15%	33%	13%	20%	22%
Renewal Business	84%	85%	67%	87%	80%	78%
% of claims paid	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
New Business	20%	18%	39%	22%	25%	26%
Renewal Business	80%	82%	61%	78%	75%	74%

Premiums are usually spiked around quarterly time periods with main at Jan and June. In 2020, the highest figures of Jan shifted to Feb instead. In 2021, premium movements have been spread over monthly basis more smoothly.


Gross Premium Written M01 2017 to M12 2022
Jan
Feb
Mar
Apr
May
Jun
Aug
Sep
Oct
Nov
Dec
800
700
600
500
400
300
200
100
-
2017
2018
2019
2020
2021
2022

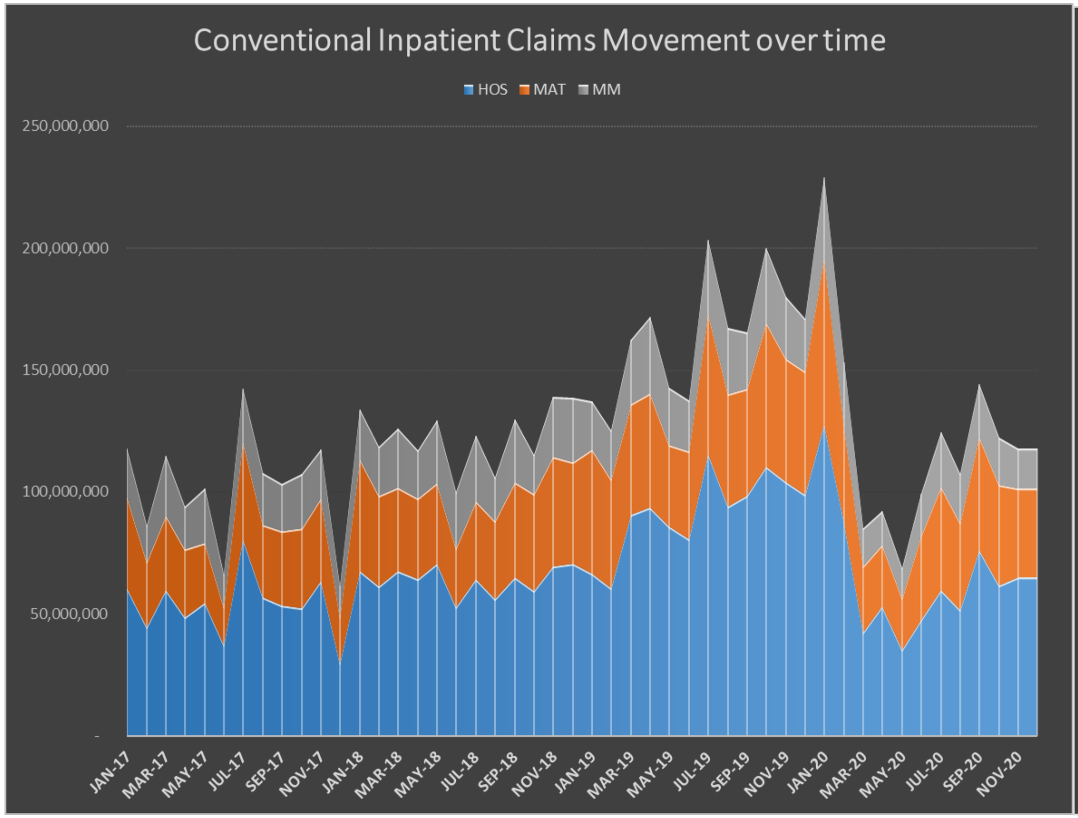


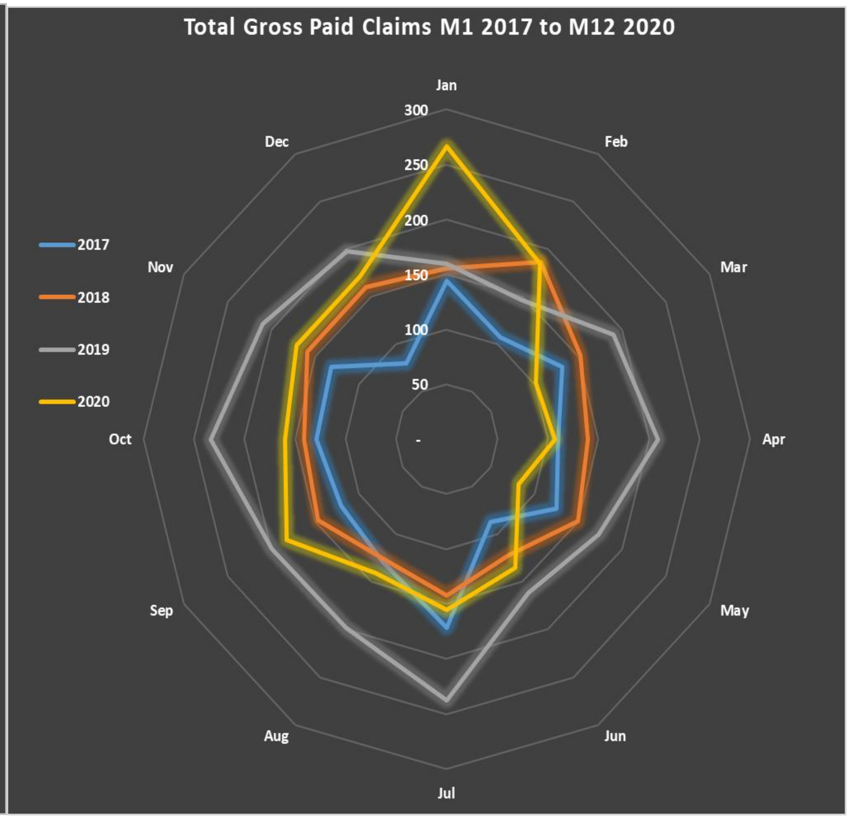
ultimate loss ratio of any ASO pool is 100% as 115% price is charged with 100% going for claims and 15% going for admin; if claim ratio is less than 100%, that amount is reimbursed back to the client at the end of the policy duration because that is the client's money; when 100% pool is burst within policy duration, client is notified for enhancement. Of-course as we are seeing only claims ratio (GCP/GWP) on Calendar year, the ratio is bound to come up differently than a static 100%. If we want to see how loss ratio in financials would look excluding ASO Pools, we can see that ASO Pools increase the loss ratio by 9%-4% on average historically.

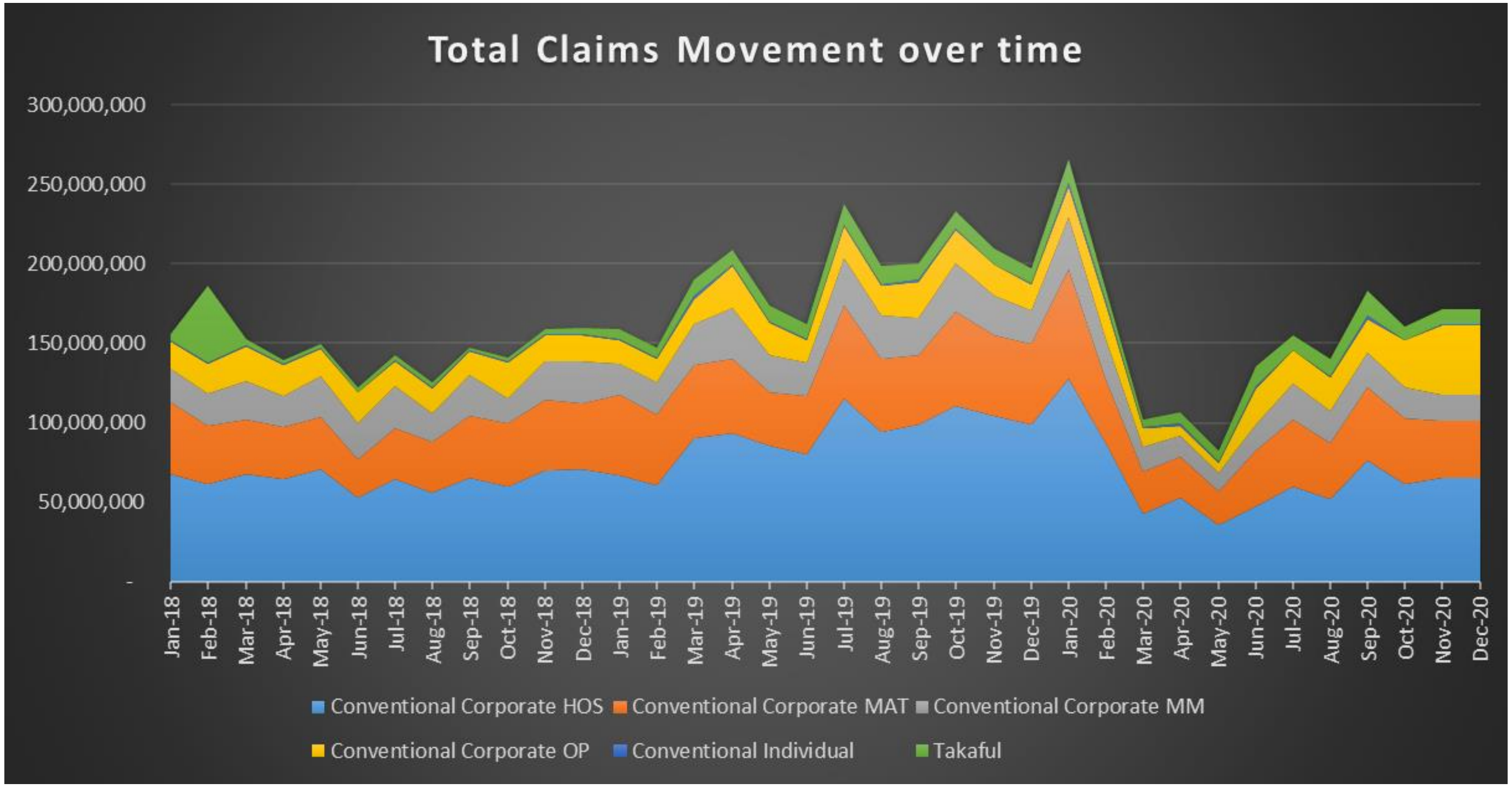
ASO Pool is important to see in IFRS17 as once it gets implemented, the ASO pools will be remove from premiums, claims as non-insurance risk and only insurance risk will be shown in the Statement of Income. We can see OP increasing as proportion of business over time. OP business has rapidly increased from Q3 2020 to present in response to employees demanding employers to pay for their COVID19 tests and vaccinations. Also notion of seeping claims is relevant here; many small cases like for example, person had stomach flu, went to see a doctor, doctor prescribed tests worth 8 thousand; person said I have health card but I don't want to spend my own money; doctor says no problem; I will give IV antibiotics instead of oral and that will make it medically necessary to admit you in hospital for 1-3 days than it will be cashless coverage for you. Better to get your pre-existing illnesses looked at while you are at it too. Hence, these OP claims seep and become IP claims instead. After COVID19 March 2020, seeping claims possibility has reduced to a market extent due to pressure on healthcare supply (hospital rooms), lockdowns, fear of catching infection in hospital etc. so the claims that employees would make are reduced now hence greater pressure for OP.

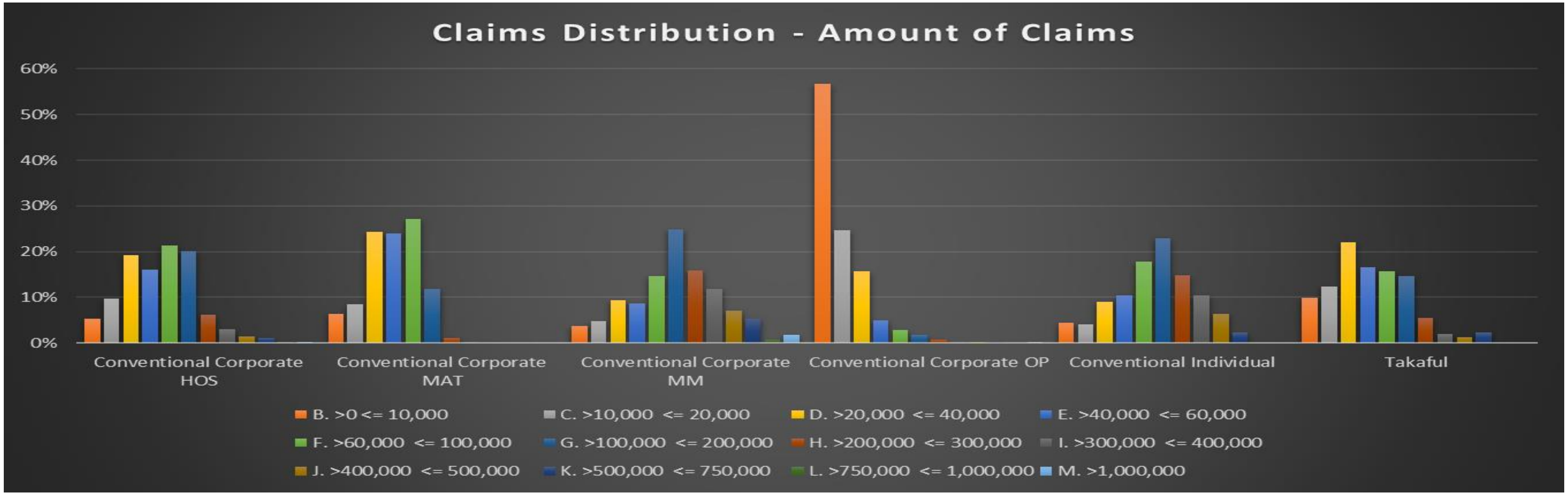
If we assign risk on basis of age of policyholders covered, we can see that 25% of our portfolio has greater than 40 years of age. 48% of our insured lives are less than 30 years of age. This is lower than Nationally having 65% population below 30 years because of higher than average ages of employees in large companies.

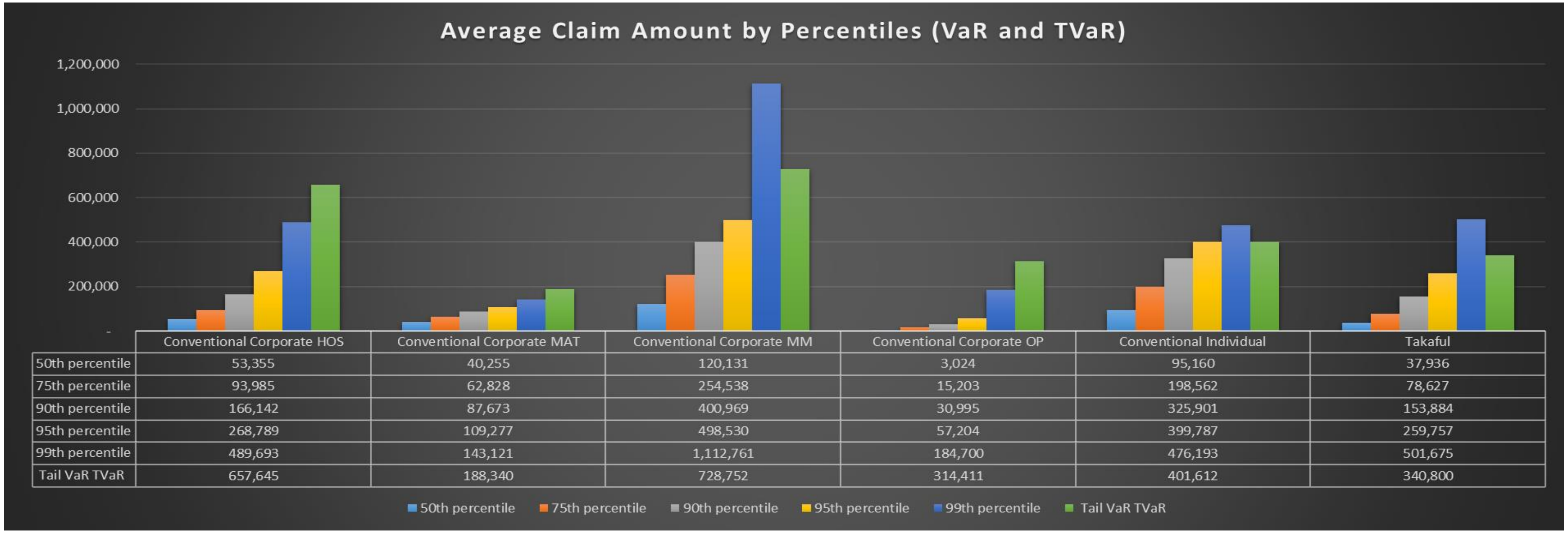
Claim movements are more spread out over time unlike premium which has more spikes and concentration in few months. The Q2 2020 drop in claims due to COVID19 are evident. Claims average amount severity is also analyzed percentile wise and as per TVaR.


Conventional Inpatient Claims Movement over time
HOS MAT MM
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
-
JAN-17 MAR-17 MAY-17 JUL-17 SEP-17 NOV-17 JAN-18 MAR-18 MAY-18 JUL-18 SEP-18 NOV-18 JAN-19 MAR-19 MAY-19 JUL-19 SEP-19 NOV-19 JAN-20 MAR-20 MAY-20 JUL-20 SEP-20 NOV-20


Total Gross Paid Claims M1 2017 to M12 2020
2017
2018
2019
2020
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
300
250
200
150
100
50
-


Total Claims Movement over time
300,000,000
250,000,000
200,000,000
150,000,000
100,000,000
50,000,000
-
Jan-18 Feb-18 Mar-18 Apr-18 May-18 Jun-18 Jul-18 Aug-18 Sep-18 Oct-18 Nov-18 Dec-18 Jan-19 Feb-19 Mar-19 Apr-19 May-19 Jun-19 Jul-19 Aug-19 Sep-19 Oct-19 Nov-19 Dec-19 Jan-20 Feb-20 Mar-20 Apr-20 May-20 Jun-20 Jul-20 Aug-20 Sep-20 Oct-20 Nov-20 Dec-20
Conventional Corporate HOS
Conventional Corporate MAT
Conventional Corporate MM
Conventional Corporate OP
Conventional Individual
Takaful


Claims Distribution - Amount of Claims
60%
50%
40%
30%
20%
10%
0%
Conventional Corporate HOS
Conventional Corporate MAT
Conventional Corporate MM
Conventional Corporate OP
Conventional Individual
Takaful
B. >0 <= 10,000
C. >10,000 <= 20,000
D. >20,000 <= 40,000
E. >40,000 <= 60,000
F. >60,000 <= 100,000
G. >100,000 <= 200,000
H. >200,000 <= 300,000
I. >300,000 <= 400,000
J. >400,000 <= 500,000
K. >500,000 <= 750,000
L. >750,000 <= 1,000,000
M. >1,000,000



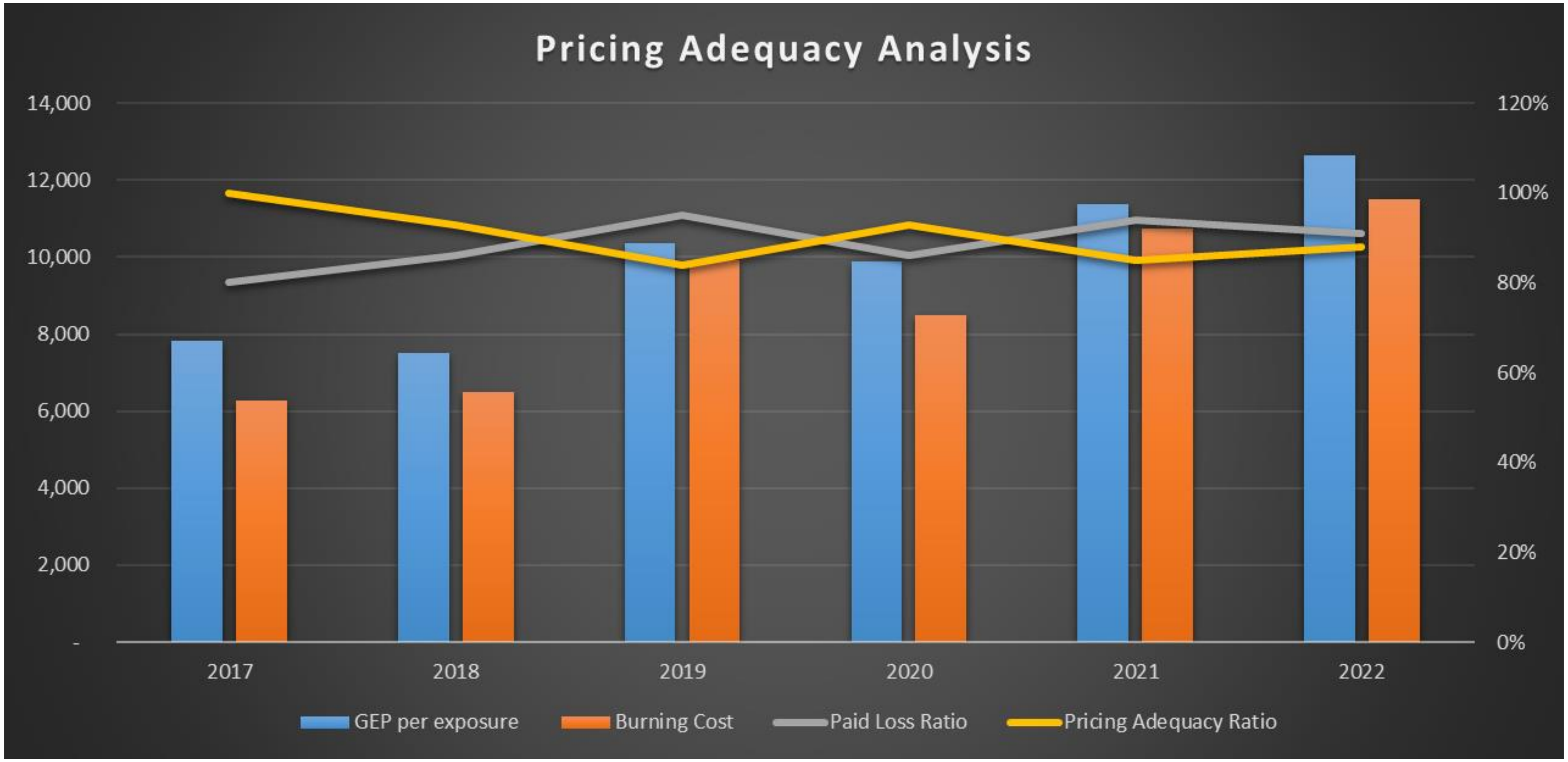
	Conventional Corporate HOS	Conventional Corporate MAT	Conventional Corporate MM	Conventional Corporate OP	Conventional Individual	Takaful
50th percentile	53,355	40,255	120,131	3,024	95,160	37,936
75th percentile	93,985	62,828	254,538	15,203	198,562	78,627
90th percentile	166,142	87,673	400,969	30,995	325,901	153,884
95th percentile	268,789	109,277	498,530	57,204	399,787	259,757
99th percentile	489,693	143,121	1,112,761	184,700	476,193	501,675
Tail VaR TVaR	657,645	188,340	728,752	314,411	401,612	340,800

Pricing Adequacy Analysis

Pricing Adequacy Ratio (PAR) = actual premium sold/technical risk premium. Technical risk premium will be simply burning cost*(1+loadings) = technical risk premium from the actuary that only takes risk factors into consideration. Actual premium sold will be actual premium that policy was sold at that considered risk as well as commercial considerations. PAR below 80% indicates underpricing and so does combined ratio above 100% or loss ratio above 85% as well as presence of any premium deficiency reserve PDR. If there is underpricing, PAR should be at around 100% and prices increased gradually every year to make it 100% over 2-3 or 5 years.

We can also arrive at pricing adequacy analysis where we compare gross earned premium per exposure Vs Burning Cost. If BC is above GEP per exposure, that indicates above 100% loss ratio and vice versa. GEP per exposure is gross premium actual whereas loaded BC is Gross Premium Technical. Pricing Adequacy Ratio (PAR) is then simply Gross Premium Actual/Gross Premium Technical. Below 100% indicates potential underpricing and price increase recommended is 1-PAR when PAR is below 100%. The PAR is 93% overall which is stable and near to 100%. But segments within overall PAR can still reveal inadequacies in pricing as some might have PAR above 100% and some below 100%.

As can be seen, gap between GEP per exposure and BC decreases when loss ratio increases and vice versa. PAR decreases as loss ratio increases and vice versa. Overall, the PAR is 90% average over 6 years which means 10% slight moderate increase in prices is required.



We need to compare gross earned premium with gross claims paid for each benefit type over the years (HOS = hospitalization, MM = major medical, MAT =maternity and OP = Outpatient):

Earned premium	2017	2018	2019	2020	2021	2022	Total
HOS	48%	49%	49%	50%	49%	51%	49%
MM	17%	17%	15%	14%	13%	12%	14%
MAT	24%	24%	26%	22%	19%	20%	22%
OP	11%	11%	11%	14%	19%	17%	14%

Claims Paid	2017	2018	2019	2020	2021	2022	Total
HOS	45%	46%	51%	47%	47%	48%	47%
MM	24%	26%	26%	27%	22%	23%	24%
MAT	16%	16%	13%	13%	13%	11%	13%
OP	16%	12%	10%	14%	18%	18%	15%

Earned premium proportion for HOS and OP is on the increase over the years and MM is on the decrease as well as MAT. However, MM proportion of claims paid is stable and high over the years than % for earned premium. This shows MM has underpricing. One reason could also be the factor individual underwriting loading upon members due to miscellaneous factors like pec, overage, penalty loading premium like delayed intimation of addition of a member, all such loadings are applied under HOS class but not on MM so it remains underpriced.

No of lives	2017	2018	2019	2020	2021	2022	Total
HOS	37%	37%	49%	47%	50%	54%	47%
MM	27%	24%	23%	16%	16%	14%	19%
MAT	10%	9%	8%	5%	5%	6%	7%
OP	26%	30%	21%	31%	30%	26%	27%

No. of Claims	2017	2018	2019	2020	2021	2022	Total
HOS	18%	22%	33%	30%	28%	27%	26%
MM	5%	6%	6%	6%	6%	5%	5%
MAT	11%	15%	21%	22%	18%	17%	17%
OP	66%	57%	40%	42%	48%	51%	52%

Number of lives % is increasing for HOS over the years and decreasing for MM over the years. MAT increase in no. of claims while decrease in number of lives for Maternity over time indicates that due to slight increase in average age portfolio over time, the average proportion of married female spouses as % of total population is increasing and/or average number of children per married family is increasing over time. Older married couples also have more children on average than younger couples.

This may be due to the facts that HOS being the compulsory core class whereas MM either remains optional in some policies or some of the insured members are restricted till HOS while individual underwriting due to miscellaneous factors like overage and pre-existing. On the other hand, MAT may be carrying the element anti selection claims are mostly forecasted considering the population covered falling under a certain age criteria.

Correlation matrices have been made for sub-segments within data:

0%	No Correlation
From 0% to 30%	Weak Correlation
From 30% to 60%	Reasonable Correlation
From 60% to 90%	Strong Correlation

Reinsurance Pricing: Practical Issues & Considerations, 8th Sep 2006; 2006 GIRO Reinsurance Matters! Working Party, Mark Flower et al.

https://www.casact.org/sites/default/files/2021-02/working-paper-ali1-2017-08.pdf

Correlation Matrix	1) Hospitalization	2) Major Medical	3) Maternity	4) Outpatient
1) Hospitalization	100%	35%	25%	75%
2) Major Medical	35%	100%	52%	33%
3) Maternity	25%	52%	100%	48%
4) Outpatient	75%	33%	48%	100%

Correlation Matrix	1) Hospitalization	2) Major Medical	3) Maternity	4) Outpatient
1) Hospitalization	N/A	Reasonable Correlation	Weak Correlation	Strong Correlation
2) Major Medical	Reasonable Correlation	N/A	Reasonable Correlation	Reasonable Correlation
3) Maternity	Weak Correlation	Reasonable Correlation	N/A	Reasonable Correlation
4) Outpatient	Strong Correlation	Reasonable Correlation	Reasonable Correlation	N/A

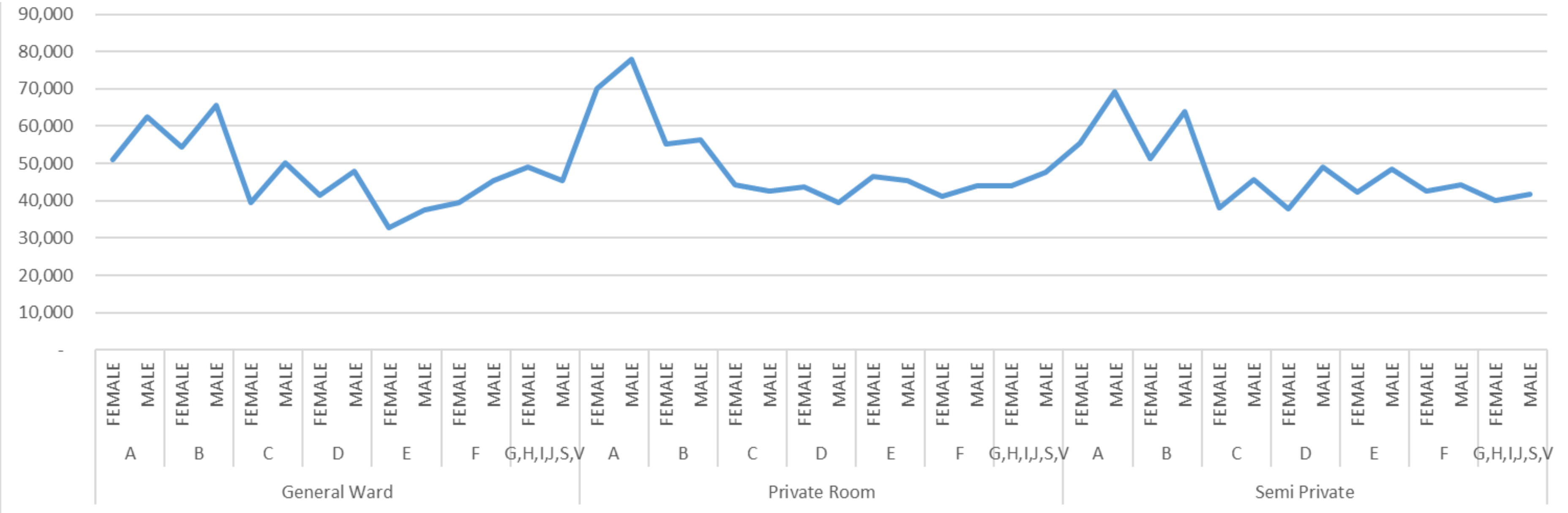
The relationship proportion of total claims paid shows own self and wife forming the majority average 73% of total claims paid:

Relationship	2,017	2,018	2,019	2,020	2,021	2,022	Grand Total
Own Self	34%	34%	33%	38%	42%	39%	37%
Wife	38%	38%	39%	37%	34%	34%	36%
Son	11%	11%	11%	10%	9%	10%	10%
Daughter	8%	7%	8%	7%	6%	8%	7%
Mother	5%	4%	5%	4%	5%	5%	5%
Father	3%	3%	4%	3%	3%	3%	3%
Husband	1%	2%	1%	1%	2%	2%	1%
Grand Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

The relationship proportion of total number of claims shows own self and wife forming the majority average 72% of total number of claims:

Relationship	2,017	2,018	2,019	2,020	2,021	2,022	Grand Total
Own Self	44%	43%	36%	39%	42%	40%	41%
Wife	29%	31%	35%	34%	31%	30%	31%
Son	11%	11%	11%	10%	9%	11%	10%
Daughter	9%	9%	8%	7%	7%	8%	8%
Mother	4%	4%	6%	6%	6%	7%	5%
Father	2%	2%	3%	3%	4%	4%	3%
Husband	1%	1%	1%	1%	1%	1%	1%
Grand Total	100%	100%	100%	100%	100%	100%	100%

Male is consistently above female and plans are on increasing trend from A to F and below with some outliers. There is hardly any difference between general ward and semi-private which needs to be investigated as semi-private should be higher than general ward. Private room cost is higher than other room types.



The frequency RF shows that RF is consistently high for above 60 years ages and that 19 to 59 years frequency has decreased over the years; 60 to 80 years and 0 to 18 years frequency has increased over the years. This is because of cohort effects. Since most business 80% is renewal and policies are renewing for an average 6 years before shifting to other insurers, the previous generation is becoming older over the years and shifting from middle age to old age and increasing their frequency of claiming. New generation of young and middle-aged employees are coming up over the years which so far have lower claims frequency but as time passes, their frequency will increase too as they grow older. 0 to 18 years increase in frequency needs to be investigated more thoroughly. It can be because more proportion of married people are now there in the data so and older couples have more children than younger couple so more children means higher frequency of claims. It can also be due to increase in congenital diseases or increase in underwriting terms and policies with children where previously policies might be more focused on employees only and not include children and so on.

Frequency RF								
Age	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**Total**	**Trendlines**
01. 0 to 18	0.82	0.89	1.15	1.50	1.36	1.76	1.24	
02. 19 to 29	0.62	0.58	0.49	0.60	0.56	0.58	0.57	
03. 30 to 34	0.84	0.81	0.76	0.79	0.76	0.75	0.78	
04. 35 to 39	1.04	0.98	0.87	0.84	0.87	0.81	0.86	
05. 40 to 44	1.32	1.27	0.96	0.81	0.82	0.82	0.89	
06. 45 to 49	1.73	1.61	1.40	1.00	1.11	0.87	1.14	
07. 50 to 54	2.00	2.21	1.57	1.07	1.12	0.98	1.26	
08. 55 to 59	2.43	2.30	1.97	1.28	1.51	1.21	1.55	
09. 60 to 64	2.13	2.07	3.11	2.26	3.01	2.80	2.59	
10. 65 to 69	2.48	2.47	4.25	3.66	4.22	4.26	3.58	
11. 70 to 80	1.94	1.63	3.60	2.71	3.14	2.60	2.61	
Grand Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
Change in frequency yearly		-8%	2%	-48%	20%	-9%		

The BC has same trend as frequency:

BC RF								
Age	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**Total**	**Trendlines**
01. 0 to 18	0.75	0.83	1.08	1.43	1.18	1.62	1.13	
02. 19 to 29	0.53	0.49	0.47	0.58	0.54	0.58	0.52	
03. 30 to 34	0.80	0.76	0.71	0.75	0.71	0.72	0.73	
04. 35 to 39	1.02	0.98	0.85	0.85	0.89	0.80	0.87	
05. 40 to 44	1.29	1.24	0.99	0.85	0.82	0.87	0.91	
06. 45 to 49	1.88	1.78	1.52	1.13	1.39	1.00	1.30	
07. 50 to 54	2.72	3.01	1.67	1.12	1.25	1.10	1.50	
08. 55 to 59	3.57	3.34	2.55	1.48	1.74	1.33	1.97	
09. 60 to 64	3.05	3.10	3.57	2.46	3.48	2.99	3.20	
10. 65 to 69	3.34	3.11	4.64	3.52	3.99	3.81	3.88	
11. 70 to 80	3.59	2.50	5.03	2.84	3.48	2.47	3.35	
Grand Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
change in BC yearly		19%	55%	-42%	17%	-5%		

Severity trends are more volatile and increasing and then decreasing. 0 to 39 years is stable over time, 40 to 49 is increasing trend and 50 to 80 years is increasing trend over time:

Severity RF								
Age	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**Total**	**Trendlines**
01. 0 to 18	0.91	0.93	0.94	0.95	0.87	0.92	0.91	
02. 19 to 29	0.86	0.85	0.95	0.97	0.95	1.00	0.91	
03. 30 to 34	0.95	0.94	0.93	0.95	0.94	0.95	0.94	
04. 35 to 39	0.98	1.00	0.98	1.01	1.02	1.00	1.01	
05. 40 to 44	0.97	0.97	1.03	1.04	1.00	1.07	1.02	
06. 45 to 49	1.09	1.11	1.09	1.13	1.25	1.15	1.14	
07. 50 to 54	1.36	1.37	1.07	1.05	1.12	1.13	1.19	
08. 55 to 59	1.47	1.45	1.29	1.15	1.15	1.10	1.27	
09. 60 to 64	1.43	1.50	1.15	1.09	1.15	1.07	1.23	
10. 65 to 69	1.35	1.26	1.09	0.96	0.94	0.89	1.08	
11. 70 to 80	1.85	1.54	1.40	1.05	1.11	0.95	1.29	
Grand Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00	
change in severity yearly		29%	52%	12%	-2%	5%		

6) Overall Conclusions and Recommendations

The Health insurer should make plans for phased enhancement of analytics so that pricing can be surgical and more comprehensive to lead to better profitability.

- The first phase will be to capture data items not captured currently. As well as capturing, refining and extracting what data is currently available. For example, number of lives and size of groups not captured. Capturing this can lead to different groups being binned such as jumbo, large, medium, SME and see pricing results of how burning cost differs as each segment based on size of the group. Health Insurer needs to be able to capture reinsurance for non-quota on transactional level as well. Receivables/payables, expenses and counterparties also need to be captured at unit of account level.
- The second phase will be to pilot and carry out the analytics broadly and finalize the most useful analytics. For example, how does C-Section ratio of the company compare to national ratio? ICD10 and other coding claim analytics to reveal difference in burden of disease by various sub-segments such as age, gender, relationship status, occupation, cities, rural/urban and others.
- Third will be to codify the analytics from IT to a Business Intelligence (BI) layer. That will allow Health insurer to better bargain using analytics and these can be shown to brokers, group clients, hospitals to increase bargaining power of Health insurer in manner that is quick and automated. For example, one year most patients went to the most expensive hospital in the city leading to high losses, another time client should be told that due to cardiac episodes, this year the group has been loss-making. Health insurer can also show that high losses were due to parents and higher than average frequency for maternity etc.

The insurer should develop these functions, on-site forensics fraud verifier employee, marketing/business development function, medical coding, epidemiology/biostatistics function.

Loss ratios of “additions” versus “initial” cohort should be monitored to test for anti-selection.

BI layer for bargaining with top jumbo clients, brokers and hospitals; performance management for agents

- Jumbo client results Vs Insurer results Vs National results
- broker results Vs Insurer results Vs National results
- hospitals results Vs Insurer results Vs National results
- Agents results Vs Insurer results Vs National results

What can analytics tell us about product development? OP is increasing in demand so individual retail OP product should be launched. Few expensive claim procedures should be excluded like knee replacement etc. and launched. Instead of comprehensive, specific products should be launched; for example, product that targets only dengue, Chikunya, malaria by analyzing claim details on mosquito borne diseases, same for one product for waterborne sanitation diseases; potential is immense to analyze data and launch diabetes insurance product like in India, cover mental health like in India, launch male specific cancer, female specific cancer after analyzing claims to price it in detail. offer ways to increase preventive health and so reduce claim burden of disease, parents only policy, pets policy, Chinese only policies etc.

Some usually lacking things in health insurance analytics are; 1) epidemiology 2) forensic fraud handling officer 3) marketing officer 4) medical auditor for ICD coding. FWA 20% of health claims.

Improve data capturing and MIS. For e.g., be able to identify between hospitalization and day care procedure claims, capture occupation etc.

Example of with jumbo clients; helps in bargaining at times of renewal objectively. Some sample examples:

- Frequency for maternity for one jumbo client was twice that of our portfolio average
- One client's policyholders claimed from expensive hospital for 80% of total claims paid leading to loss
- One jumbo client faced spike in health claims in one policy year due to spike in cardiac arrest claims
- One company faced spike in infections one year. They looked at their distribution procedures, identified the root cause and solved it
- One policy was loss-making because of parents
- Historically above 100% loss ratio because families can be added and deleted as and when required under their policy historically. Controls are that family once insured cannot be removed for next 3 years; comparing enrolment of families cross-verified with National family registration certificate and all children born during policy to be enrolled. Part of the premium is paid by employees for family health insurance and the employer does not pay the whole premium.
- Many employees in few jumbo groups didn't pay premium to cover health insurance because they are covered under Universal Social insurance. However, plenty of them came back as Universal Social insurance has its own problems such as very low limits, denial of coverage issues and so on. This is an evolving situation and it remains to be same to what extent will Universal Social insurance cannibalize the private health sector.

One major pricing challenge is using a group product underwriting model with a flat rate while allowing for voluntary individual choices. This is demonstrated by the disconnect between the insurer's use of experience data to set prices versus the results of empirical pricing due to changes in policyholders' health care behavior after purchasing insurance. As experience does not represent future reality, the contradiction between the two means we must look for a better way to set pricing and predict customer behavior.

In order to reduce anti-selection insurer should seek definition of the plans from the client (e.g., which cadres are covered in each plans).

Insurer should strictly monitor practices at our network hospitals and we should try to negotiate rates periodically (say every two years). During negotiations insurer should mainly focus on the volume of business being given to our network hospitals.

Burning Cost is high for 0 to 18 years which is to be expected given high frequency of claims from babies and teenagers; within teenagers, accidents incidence for male increases from 15-25 years due to risk social behavior. BC falls than from 19 years and then keeps increasing progressively.

BC increases rapidly from 45 years and onwards for HOS+MM. 45 to 49 years increase is still moderate at yellow but 60 to 99 (99 is practically 75 years) increase is very high at red zone. Thus, we can categorize claim cost into 3 classes for ages for HOS+MM:

- green zone = 19 years to 44 years
- yellow zone = 0 to 18 years + 45 to 59 years,
- red zone = 60 to 75 years

While under group health is usually majority of an insurer's health portfolio, the insurer cannot select which age comes to us under group client's population census, insurer should aim to increase our focus and target green zone and minimize the red zone. This can mean that insurer minimize underwriting parents in the policies where possible as parents have high ages. However, age has interaction effect with relationship and gender as well as seen here:

Gender	Relationship	BC RF
Female	Mother	1.01
Male	Father	1.17
Female	Daughter	0.87
Male	Son	0.89
Female	Wife	1.02
Male	Husband	1.19
Female	Ownself - Female	0.91
Male	Ownself - Male	1.07
Female	Married Female	1.02

Thus, we can see here that daughter and son have low costs due to lower ages and parents have higher costs due to high ages but also that this high cost is largely in father male parent rather than for female mother. Male is higher than female across mother father, daughter son, wife husband and ownself female+married female Vs Ownself male. Thus, the practical focus should be where possible to incentivize group clients to include children in their policies too or increase their coverge and minimize parents coverage and increase rates for male on HOS+MM rather than on female.

The skewness of distribution of age has reduced over the years which means that more people are growing older now but their claims burden is less extreme now than before.

Frequency is higher overall for female under Inpatient excluding maternity than male. Severity is higher for male then female. This is a universal trend that frequency are more careful/conscious about their health and so utilize healthcare for minor issues even whereas male avoid healthcare utilization due to culture that deters male from being conscious about their health. However, when male do get hospitalized, their claims are on average more severe than that of female then.

The life expectancy is also at play here. Burning Cost is high for Male at 60-69 years whereas it is high for female at 65-75 years. This is because female have 5 years higher life expectancy than male and this gap (65 years female – 60 years for male) reflects that.

Maternity Incidence rates suggest that frequency for maternity is moderately increasing over time. Age effect that age of people in cohorts increasing so older married couples have more children than younger couples on average.

National C-section ratio till 5 years ago was 22%; but private health insurers usually have C-Section ratio as 60% which is 2.6 times higher due to various reasons. Higher limits for C-section so greater financial incentive, doctors seek surgical experience so prefer C-Section, educated people in top 3 cities prefer C-section as it is less risky and time consuming. That is why Universal Social Insurance has put in same limits for C-Section or Normal maternity so as to remove any financial incentive from doing C-Section surgeries and do it only when medically necessary.

Room rents show differentiated inflation. Differentiated means that hospitals do not increase prices the same way across all room types; general wards are more price sensitive so prices are increased there to a lesser degree than in private rooms or VIP rooms where people have higher purchasing powers and so can meet the higher medical inflation rates.

IBNR for each of top 50 groups should be compared AvE.

Regarding type of claim charge, room and board is not significant expense as it is just 8% of claims paid on average. pre and post hospitalization is also not significant claims as it is just 1% of claims paid; ex gratia is 0%. medicines is important cost 24% of total claims paid; analysis is still incomplete because main category of 67% of claims paid is "Others" which we need to open up. what does "Others" represent? Surgeries? Day care procedures? Tests? charges can be bifurcated into for example, consultation fee, diagnostics, ICU, medicines, bed charges, surgeon fees.

Deductions are 18% on claims paid. Reasons for claim deductions need to be opened up because “Other Reasons” form 96% of the deductions and reasons like out-of-network deduction, pre-existing deduction, room rent difference deduction and limit sum insured bursting deduction combined form only 4% of deductions.

Key results across diagnostic codes reveals a wealth of information on burden of diseases in the population covered by the company. Like Nationally, there is double burden of disease because infections are rising over time (due to climate change, antibiotic/antimicrobial resistance, worsening pollution, COVID19, rising catastrophes, heatwaves, floods, mosquitos, water crisis) and increase in non-communicable diseases (NCD) over time as population ages and NCDs increase due to lifestyle factors. Heart attacks, gall bladder removal surgeries, diabetes, hypertension, appendix issues, mental health issues, cancers, C-sections in maternity, are increasing claims. Increasing burden of diseases means long-term increasing claims which can be countered by increasing premiums over time, focusing on preventive measures such as awareness campaigns, executive checkups, provision of nets and first aid, making wellness products, digitization to streamline processes and reduce expenses and so on.

We had access to first procedures description but could not find a way to bin that data into main codes or sub-codes like we had available for primary disease description. That is why we cannot summarize results by procedures undertaken because then top 300 procedures start forming more than 50% of total claims paid then. We cannot evaluate two-way analysis of disease description with procedures due to same lacking.

There should be some method to categorize the hospitals by tiers of brand, quality, fees; Analysis by making tiers of hospital providers can increase our analysis and lead to insights that are not observed right now. We should also aim to differentiate provider by type of provider so hospital, clinic, laboratory and pharmacy.

We should discuss protection gap or basis risk and see what we can do there. Basis risk or protection gap is difference between the insured loss and the actual loss. So, for example, how does IP insurance reduce out-of-pocket expenditures for policyholders? Cancer average claim size is usually hardly 10% of the actual economic loss which cannot be thought of as a realistic figure for such an expensive and chronic life-threatening disease like Cancer which runs into millions so how much basis risk is there?

In conclusion, it is hoped that pricing for sample medical portfolio using basic burning cost and advanced machine learning have been able to open up data and business trends and produced valuable actionable insights for the management to implement and monitor over time as part of data driven decision making.