Term Project - Step 3

Tabbalat, Abed

2021-05-30

Introduction

The homeowners insurance market is going through some struggles right now after going through a hurricane season with high frequency in 2020 as well as the COVID-19 pandemic. It doesn't mean that it is the only reason. I have partnered up with a private insurance company to do a high-level dive into their data and analyze what is going on with their homeowners product and how we can view and analyze their data in different ways to give us a closer answer on how the product is performing.

Homeowners product consists of three types:

- **HO:** Homeowners Legacy Program
- **PHO:** New Homeowners Program
- **DP:** New Dwelling Fire Program

There are three datasets provided:

- **Claims Dataset:** Contains information about claims regarding the programs
- **Score Dataset:** Contains information about the reinsurance regarding the programs
- **Exposure Dataset:** Contains information about the premium details regarding the programs

The problem statement you addressed

The main question raised is how can we get the product more profitable? And what are the areas that are affecting the product's profitability? There were more questions that were suggested to investigate, but it seemed like those are the ones that fit most with the type of datasets provided. Since there are many comprehensive datapoints within the datasets, identifying the components needed will be key to solving our problem. In addition, there will be a metric to define the different views on how the product is performing starting with a combined ratio (a ratio that determines all components going into the product's performance).

How you addressed this problem statement

Addressing the problem starts with cleaning up the noise in the data that was determined to be not needed, as well as making sure that all three datasets have a common ID that will be the root of merging the needed components together. Calculating the combined ratio will involve 3 components. A fixed expenses ratio factored in at 30%. A loss ratio that will be calculated using the claims dataset. The calculation will be total incurred loss divided by the premiums. Finally, would be the CAT score ratio which is divided by the premium as well. Summing those three components will give us the combined ratio. Once we have that,

running a multiple linear regression model on both claims and CAT score to determine which has a stronger relationship to the covariance.

This may not solve the problem but will give good indications on where to dig deeper into this analysis and determining what is the root of a high combined ratio based on the metrics that will be used. Therefore, there will be 3 views derived from the datasets and summarized:

- Product/State view
- Construction Type view
- Customer Segmentation view

Analysis

Multiple Linear Regression on Losses Incurred

Looking at the regression models applied, below shows the results of multiple linear regression model for Losses Incurred:

```
##
## Call:
## lm(formula = `Total Incurred` ~ `Total Premium` + `Total Ceded Premium` +
##     Product + Property.State + Year.Built + Total.Square.Feet +
##     Building.Exposure + Other.Structures.Exposure + Contents.Exposure +
##     Loss.of.Use.Exposure + Roof.Age + Customer_Segment, data = df_exposure)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7244   -489   -295   -117 596473
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.005e+03  3.626e+03   0.553 0.580305
## `Total Premium`                -2.346e-02  5.075e-02  -0.462 0.643844
## `Total Ceded Premium`           1.548e-02  5.186e-02   0.298 0.765371
## ProductHO                       2.214e+02  2.241e+02   0.988 0.323110
## ProductPHO                      5.064e+02  2.026e+02   2.500 0.012432 *
## Property.StateLA               -6.511e+01  8.634e+01  -0.754 0.450817
## Property.StateNC                3.465e+02  2.399e+02   1.444 0.148732
## Property.StateSC               -6.821e-01  1.859e+02  -0.004 0.997073
## Property.StateTX                2.963e+02  1.187e+02   2.496 0.012566 *
## Year.Built                     -1.287e+00  1.803e+00  -0.714 0.475399
## Total.Square.Feet               1.955e-01  5.539e-02   3.530 0.000417 ***
## Building.Exposure               2.329e-03  5.988e-04   3.889 0.000101 ***
## Other.Structures.Exposure      -1.800e-03  2.547e-03  -0.707 0.479688
## Contents.Exposure              -2.243e-03  9.067e-04  -2.473 0.013385 *
## Loss.of.Use.Exposure           -5.901e-03  1.616e-03  -3.653 0.000260 ***
## Roof.Age                        3.015e+01  5.002e+00   6.028 1.68e-09 ***
## Customer_SegmentNon-Standard   -1.422e-01  2.726e+02  -0.001 0.999584
## Customer_SegmentPreferred      -2.044e+02  1.597e+02  -1.281 0.200361
## Customer_SegmentStandard       -5.245e+01  1.632e+02  -0.321 0.747916
## Customer_SegmentUltra Preferred -3.736e+02  1.624e+02  -2.300 0.021448 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5127 on 42569 degrees of freedom
##   (19017 observations deleted due to missingness)
## Multiple R-squared:  0.002988,   Adjusted R-squared:  0.002543
## F-statistic: 6.716 on 19 and 42569 DF,  p-value: < 2.2e-16
```

Looking at R2, it is showing an almost no relationship between the components used to the Incurred Loss indicating that losses have nothing to do with the structure of the home and the attributes that comes with it.

Multiple Linear Regression on CAT Score

```
##
## Call:
## lm(formula = `Total Ceded Premium` ~ `Total Premium` + Product +
##     Property.State + Year.Built + Total.Square.Feet + Building.Exposure +
##     Other.Structures.Exposure + Contents.Exposure + Loss.of.Use.Exposure,
##     data = df_exposure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4464.4  -194.5   -37.8   138.0  8873.3
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  8.509e+03  2.530e+02  33.637  < 2e-16 ***
## `Total Premium`              4.517e-01  2.846e-03 158.701  < 2e-16 ***
## ProductHO                   -3.566e+02  1.701e+01 -20.958  < 2e-16 ***
## ProductPHO                   3.425e+01  1.684e+01   2.033 0.042031 *
## Property.StateLA             9.033e+01  6.595e+00  13.698  < 2e-16 ***
## Property.StateNC             1.096e+02  1.154e+01   9.492  < 2e-16 ***
## Property.StateSC            -3.813e+01  1.017e+01  -3.750 0.000177 ***
## Property.StateTX            -1.901e+02  9.867e+00 -19.271  < 2e-16 ***
## Year.Built                  -4.215e+00  1.261e-01 -33.419  < 2e-16 ***
## Total.Square.Feet            5.703e-04  1.705e-04   3.345 0.000824 ***
## Building.Exposure            2.676e-04  3.786e-05   7.067  1.6e-12 ***
## Other.Structures.Exposure    2.334e-04  2.041e-04   1.143 0.252949
## Contents.Exposure           -8.800e-04  6.590e-05 -13.353  < 2e-16 ***
## Loss.of.Use.Exposure        -4.486e-04  1.340e-04  -3.348 0.000814 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 475 on 61592 degrees of freedom
## Multiple R-squared:  0.4862, Adjusted R-squared:  0.4861
## F-statistic:  4484 on 13 and 61592 DF,  p-value: < 2.2e-16
```

Looking at R2, 0.486 shows a closer relationship between premiums and its attributes to the CAT score meaning that, CAT score can be impacted by the attributes related to the premium which means if we want to show a better CAT score we need to visit the premium price and the attributes factored in it.

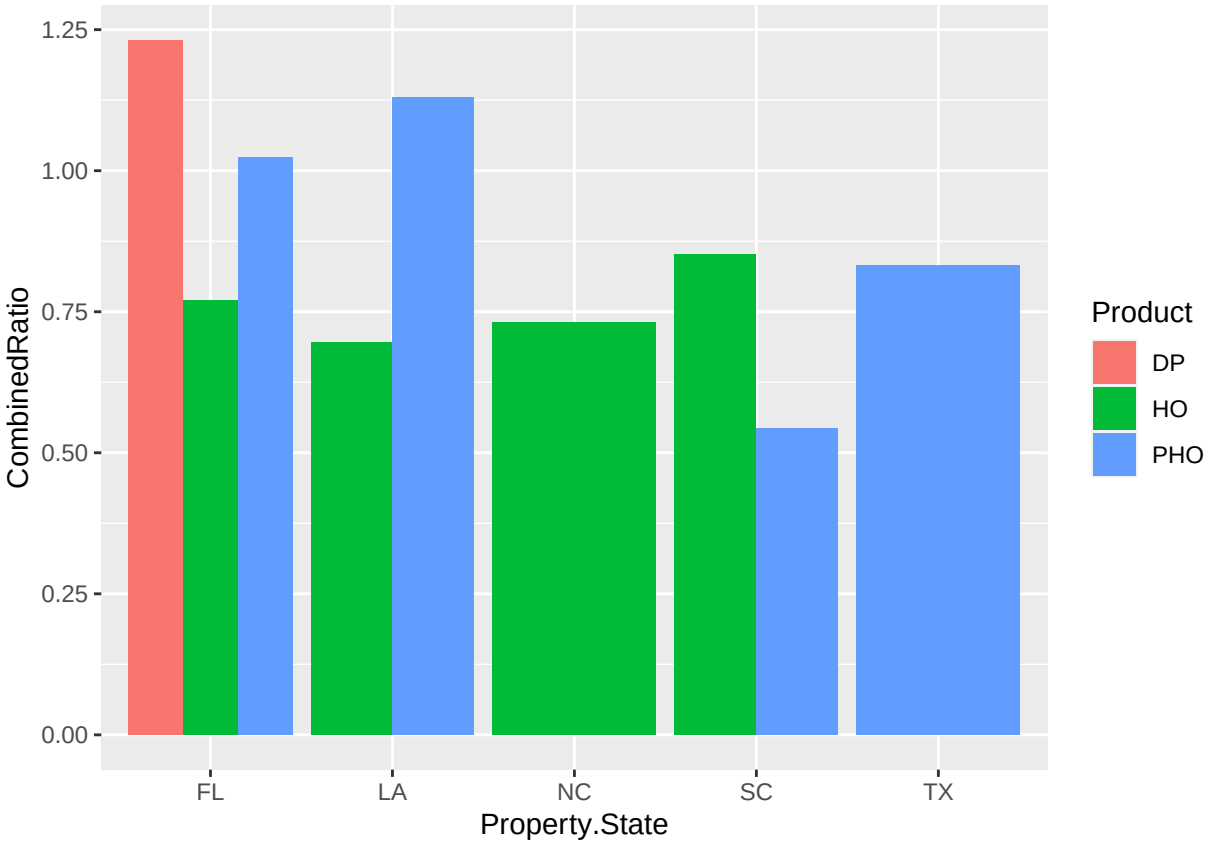
Once we have looked over the relationships between the Incurred Loss vs Premium and CAT score vs Premiums, we can now see the results of the different views that we need to analyze to see what could cause a high combined ratio and what can we do to solve it.

Product State View

The Product State view will contain components that differentiate the products and the states they reside in. Those components will show a Loss Ratio, CAT ratio, and combined ratio.

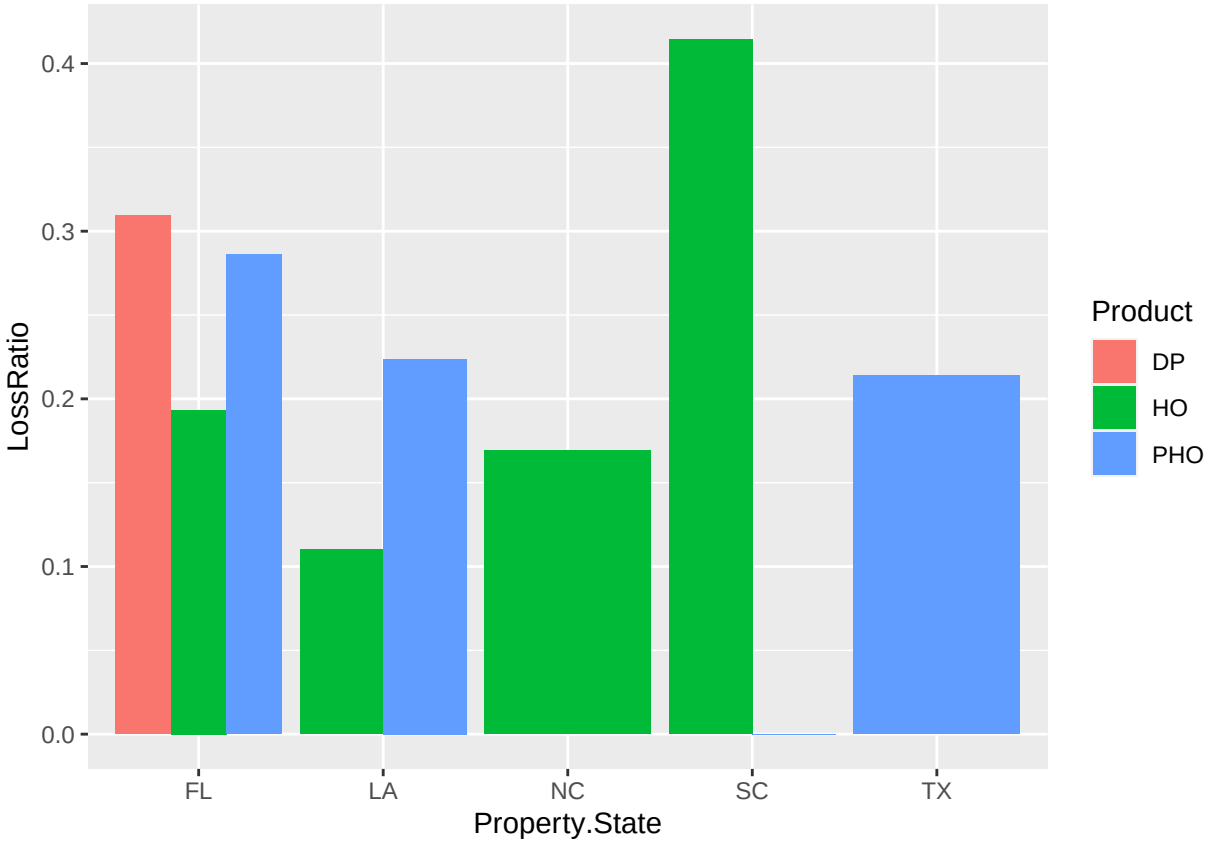
Below shows the results for the combined ratio:

```
## `summarise()` has grouped output by 'Product'. You can override using the `.groups` argument.
```



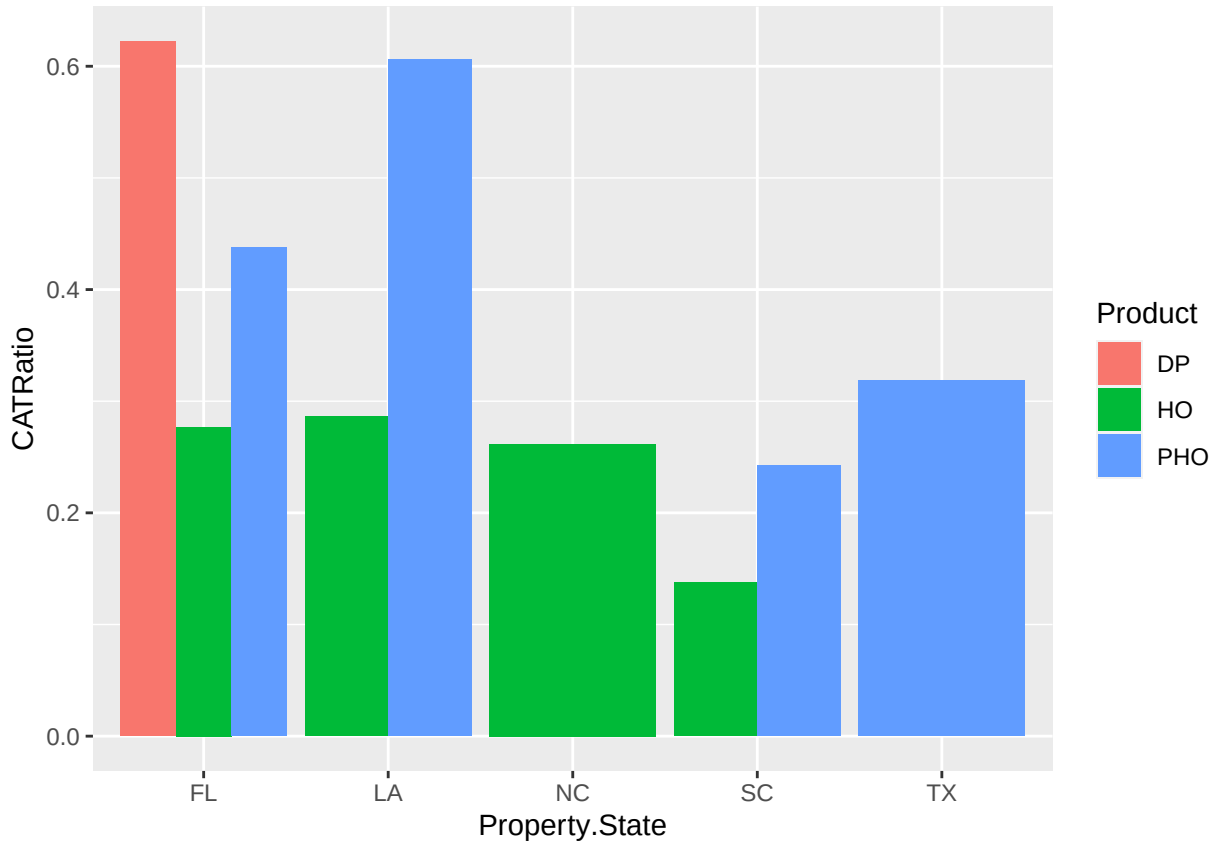
We can determine from this chart that DP product in the state of FL has the highest combined ratio, however we can count that as a false metric as the volume of the premium is not high enough to be compared to the rest of the product's ratios therefore we will skip analyzing DP and mention that it will be put on hold until volume and time add up to be able to analyze. As an expectation, I thought that the legacy product would be an answer to the profitability as there must be a valid reason for the product to be restructured and produced. However, in this scenario, it is clear that the PHO product has a higher combined ratio than legacy in all states except for SC.

Below shows the results for the Loss Ratio:



For PHO, we can see a higher loss ratio relative to HO. Except for the state of SC, a loss ratio that exceeds 40%. From this chart we can recommend that the company needs to address their product in the state of SC or even exit. In addition, the state of FL has a higher Loss Ratio than the remainder of the states. Those can be areas to investigate and attempt to improve by collaberating with the claims department and see what can be done to improve the Loss Ratio.

Below shows the results for the CAT Ratio:



This was a surprise to me, the state of LA and FL have an extremely high CAT Ratio which is the main reason that the product is getting a significantly high combined ratio. In order to lower the CAT Ratio and have our total combined lower than 1, the company needs to increase the rate of their premiums to offset the increase in CAT premium since CAT premium is controlled by the reinsurer and the company would have no choice but to increase it's rates to level set.

The table below shows the summary that puts all those charts together for Product/State:

Table 1: Summary Calculations By Product & State

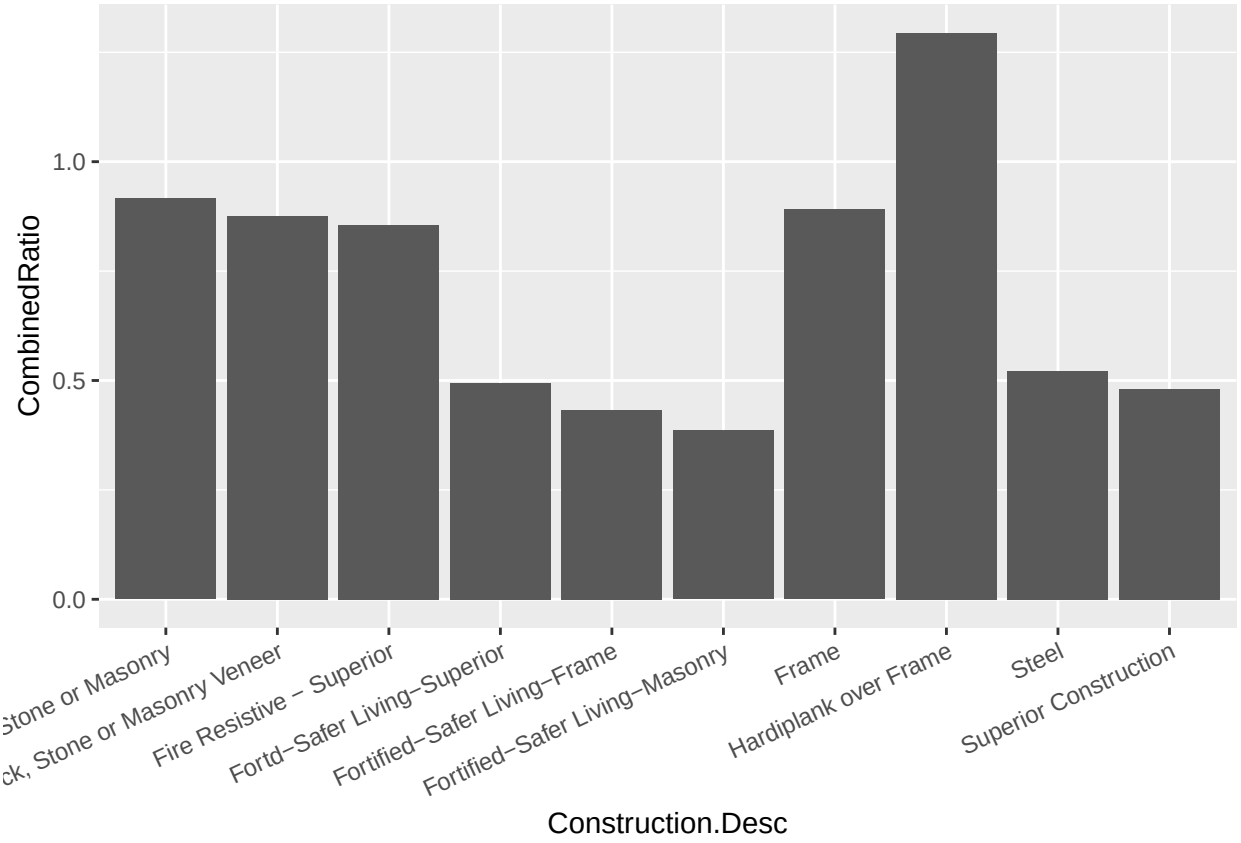
Product	Property.State	Total Premium	Total Incurred	Total Ceded Premium	LossRatio	CATRatio	CombinedRatio
DP	FL	662476	204892.1	412264.01	0.3092823	0.6223078	1.2315901
HO	FL	18085866	3498277.0	5008000.92	0.1934260	0.2769014	0.7703274
HO	LA	7869043	865112.8	2254177.56	0.1099388	0.2864615	0.6964002
HO	NC	2618547	443266.0	685011.66	0.1692794	0.2615999	0.7308793
HO	SC	3116780	1291671.2	428250.32	0.4144249	0.1374015	0.8518264
PHO	FL	30507967	8722226.3	13344938.61	0.2859000	0.4374247	1.0233247
PHO	LA	9329294	2087277.9	5653167.72	0.2237337	0.6059588	1.1296925
PHO	SC	67492	0.0	16359.81	0.0000000	0.2423963	0.5423963
PHO	TX	10034275	2145257.1	3197426.04	0.2137929	0.3186504	0.8324434

Construction Type View

The Construction Type view will contain components that differentiate the type of materials the homes have been built with. Similar to the Product view, the metrics used will be Loss Ratio, CAT ratio, and combined

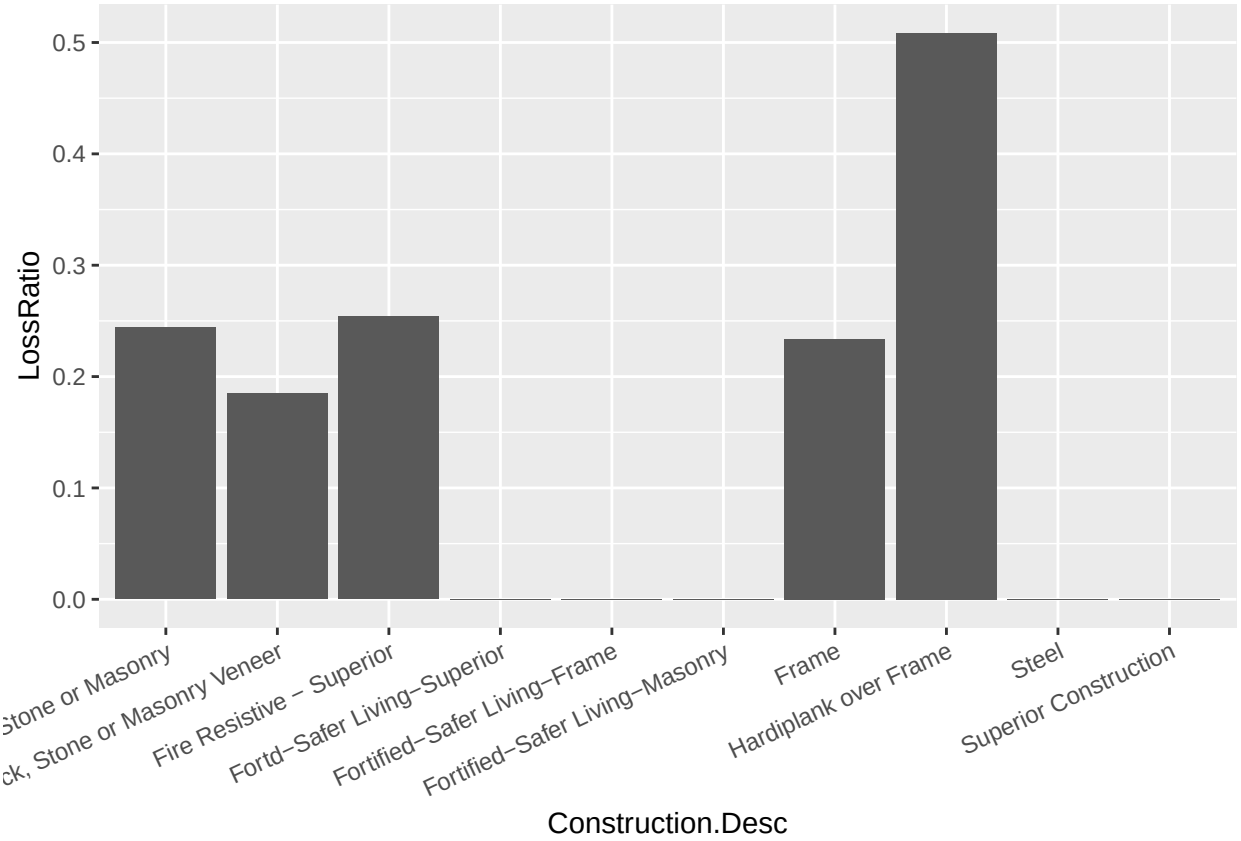
ratio.

Below shows the results for the combined ratio:



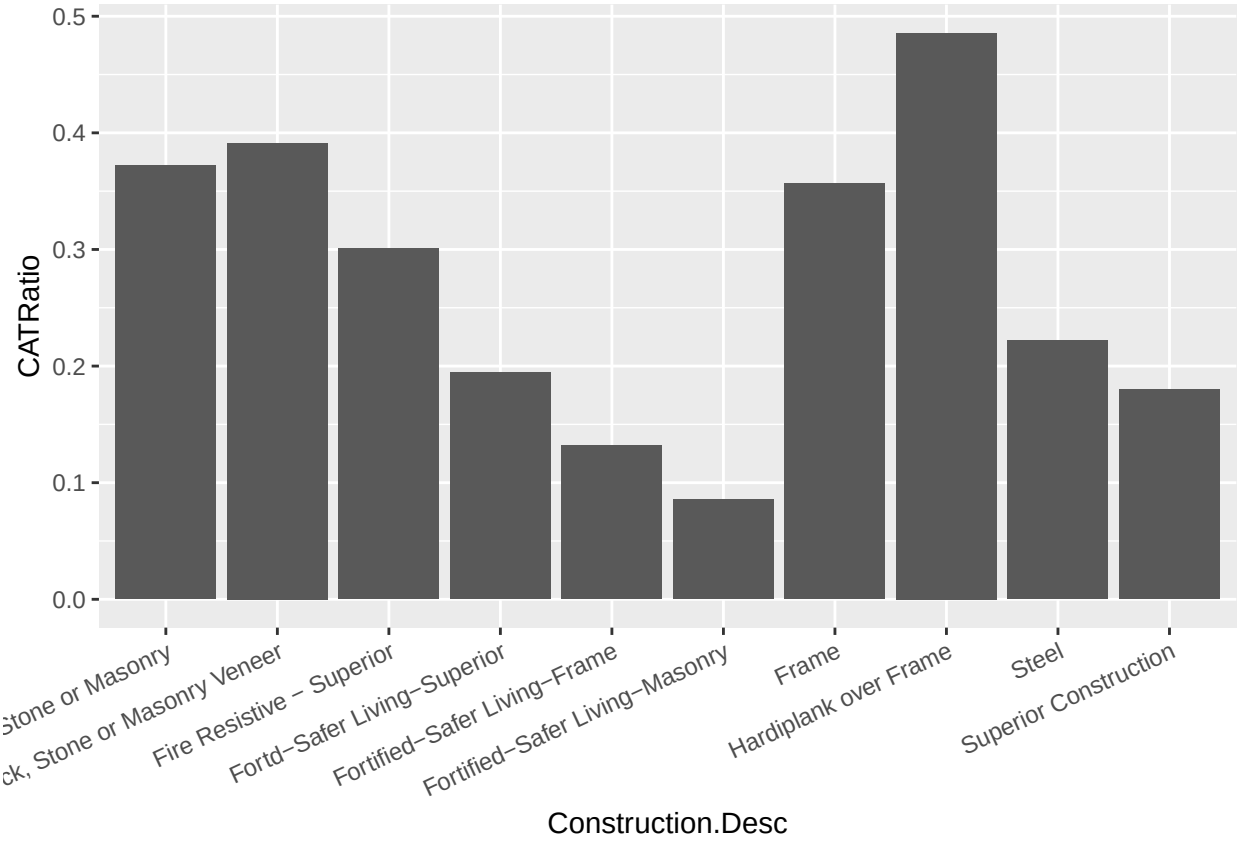
Looking at the Construction Type, this seems to be stragitforward. There is only one component that has a combined ratio higher than 1 (meaning unprofitable), and that is the `Hardiplan Over frame`. This is a component that will need to be addressed by the underwriting department. Before asking them to investigate we need to dive into the breakdown of this ratio, as it will be the focus in analyzing the Construction Type view.

Below shows the results for the Loss Ratio:



It is loud and clear that `Hardiplank Over Frame` is the main driver that, if addressed, can significantly lower the Loss Ratio across the products. A more detailed analysis in comparing which products are writing policies that are covering `Hardiplank Over Frame` can be a starting point to understand why it is a Loss Ratio driver component.

Below shows the results for the CAT Ratio:



The chart above gives us the remainder answer to our question in regards to Construction Type. `Hardiplank Over Frame` also has the highest CAT Ratio relative to the other components, which tells me that the reinsurers count it as the highest risk of a claim. Revisiting the business models on the type of construction could be a significant help in lower the combined ratio for the products and improve the results.

The table below shows the summary that puts all those charts together for Construction Type:

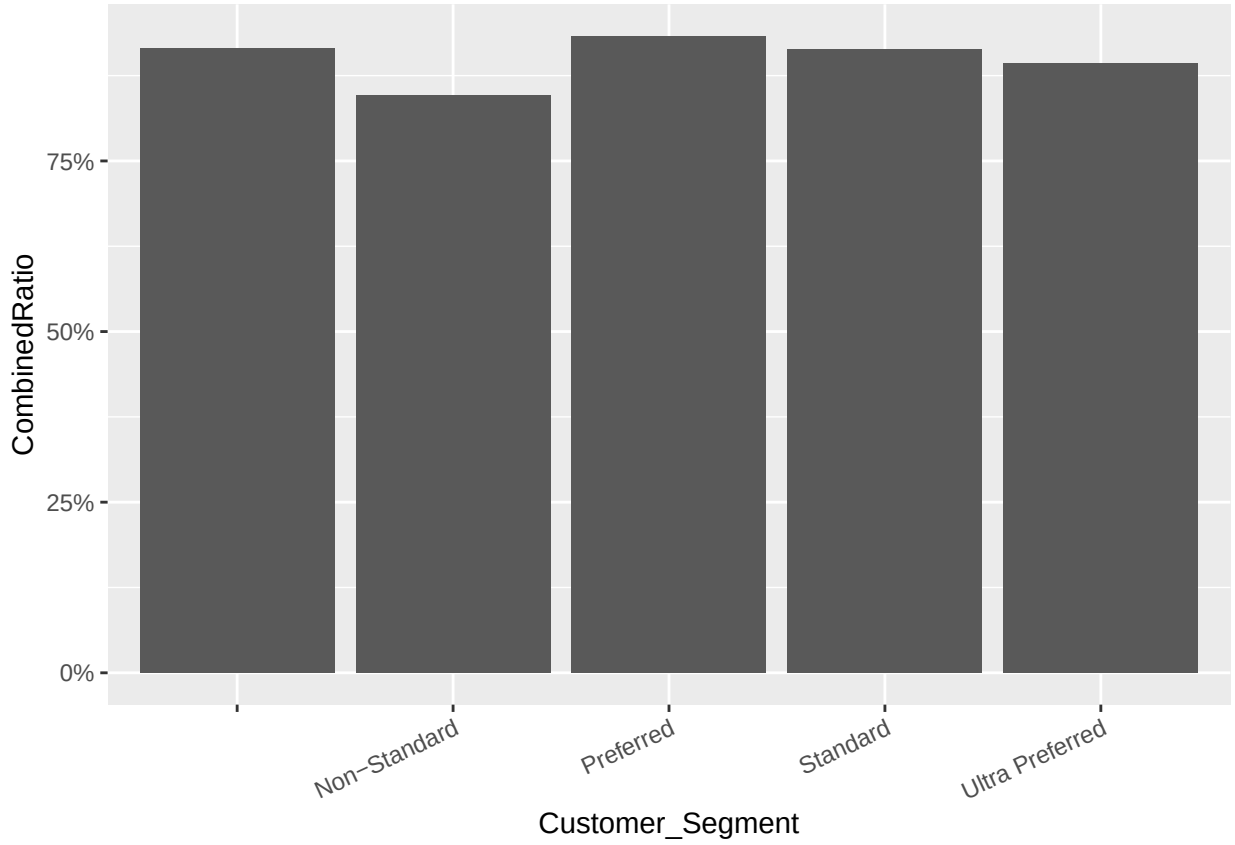
Table 2: Summary Calculations By Construction Type

Construction.Desc	Total Premium	Total Incurred	Total Ceded Premium	LossRatio	CATRatio	CombinedRatio
Brick, Stone or Masonry	32251415	7875725.8	1.200173e+07	0.2441978	0.3721302	0.9163281
Brick, Stone or Masonry Veneer	23348602	4312732.9	9.140370e+06	0.1847105	0.3914740	0.8761845
Fire Resistive - Superior	758773	192717.6	2.283384e+05	0.2539858	0.3009311	0.8549170
Fortd-Safer Living-Superior	233	0.0	4.531782e+01	0.0000000	0.1944971	0.4944971
Fortified-Safer Living-Frame	576	0.0	7.606951e+01	0.0000000	0.1320651	0.4320651
Fortified-Safer Living-Masonry	1350	0.0	1.156452e+02	0.0000000	0.0856631	0.3856631
Frame	22940786	5362619.7	8.180381e+06	0.2337592	0.3565868	0.8903460
Hardiplank over Frame	2977182	1514184.3	1.445979e+06	0.5085965	0.4856870	1.2942835
Steel	6114	0.0	1.356560e+03	0.0000000	0.2218776	0.5218776
Superior Construction	6709	0.0	1.207374e+03	0.0000000	0.1799634	0.4799634

Customer Segmentation

So far we analyzed by Product/State, and construction type. Now we are analyzing what type of customer is holding our policy and how can it impact the Combined Ratio. The metrics used will be Loss Ratio, CAT ratio, and combined ratio.

Below shows the results for the combined ratio:



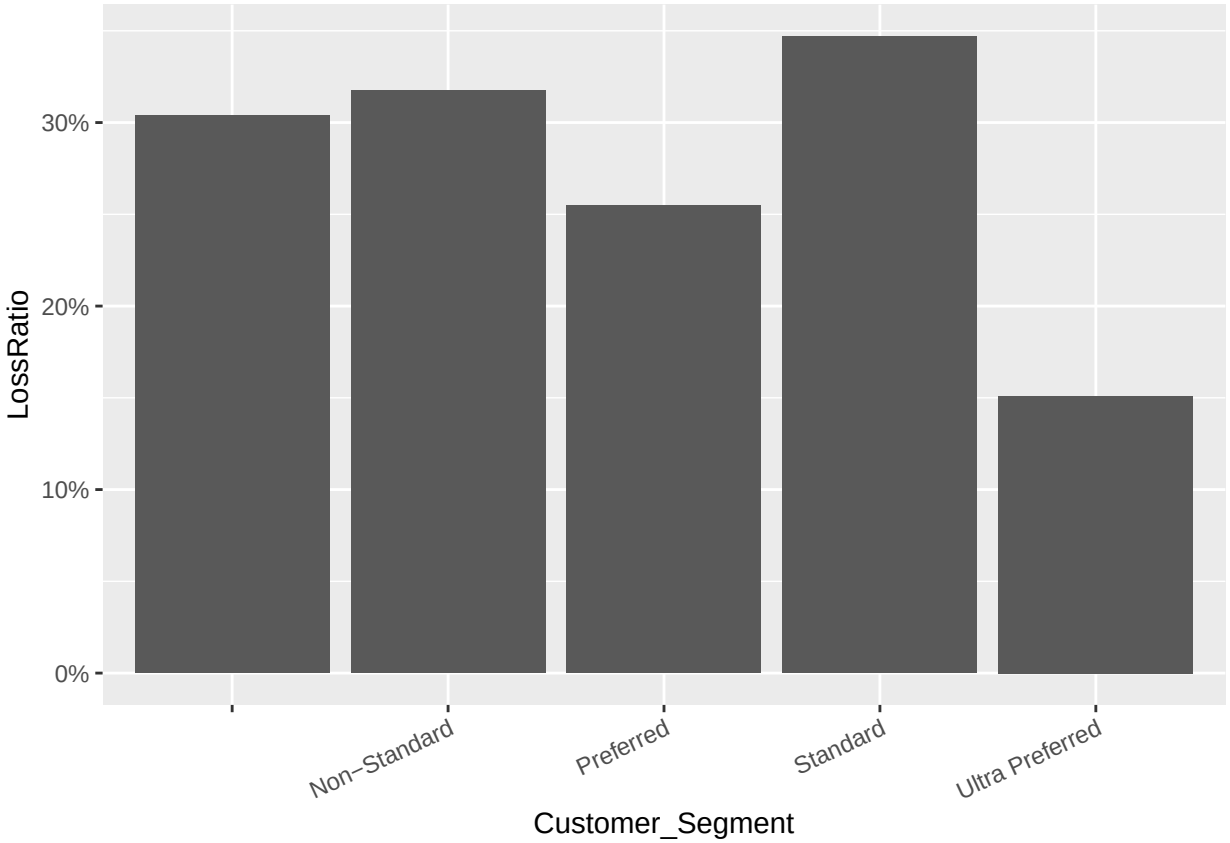
The customer segmentation is broken out into 4 components:

- `Non-Standard`
- `Standard`
- `Preferred`
- `Ultra Preferred`

Looking at the chart, there is an additional blank component, which means the dataset contains policies that are missing the mapping or haven't been mapped. This needs to be reported back to the company explaining the importance to have their data updated and fixed in order to provide a better analysis.

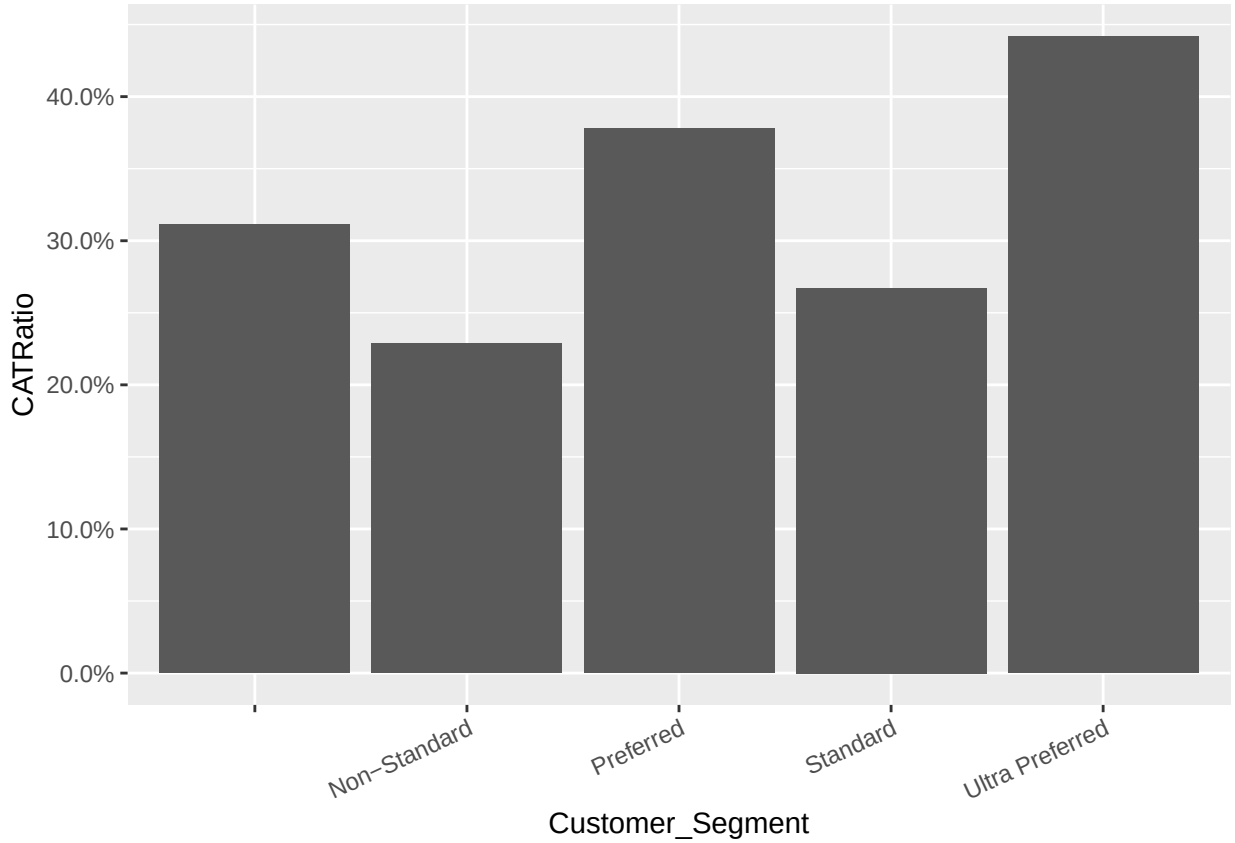
If we skip the blanks, looking at the Combined Ratio from a Customer Segmentation perspective, we can determine that the combined ratio is fairly close to each other. With `Preferred` to be the highest combined ratio and `Non-Standard` to be the lowest. Ideally, from the naming mechanism used by the company, I would expect for `Ultra-Preferred` to be the lowest combined ratio and the `Non-Standard` to be the highest.

Below shows the results for the Loss Ratio:



The chart above makes sense to what the expectation would be, `Ultra Preferred` has a significant low Loss Ratio, and Standard has the highest. Even though the `Non-Standard` would be expected higher than `Standard` the volume of `Standard` policies is much higher making it more exposed and having a higher loss occurrence chances.

Below shows the results for the CAT Ratio:



The CAT Ratio has been the main reason behind the increased Combined Ratio in `Preferred` and `Ultra Preferred` segmentation. This goes back to our original answer by saying that the company will need to increase its rates to level set the premium. The combined ratio should be lower for those two relative to the other and the only way to do so is by increasing its rates as CAT premium is an uncontrollable factor.

The table below shows the summary that puts all those charts together for Customer Segmentation:

Table 3: Summary Calculations By Customer Segmentation

Customer_Segment	Total Premium	Total Incurred	Total Ceded Premium	LossRatio	CATRatio	CombinedRatio
	4587213	1393887.9	1427094.3	0.3038638	0.3111027	0.9149664
Non-Standard	1790371	568637.2	409565.4	0.3176086	0.2287601	0.8463687
Preferred	28625928	7292984.7	10822341.7	0.2547685	0.3780608	0.9328293
Standard	14602452	5066008.9	3901271.6	0.3469286	0.2671655	0.9140942
Ultra Preferred	32685776	4936461.7	14439323.6	0.1510278	0.4417617	0.8927895

Conclusion

After slicing and dicing the data, analyzing it, we can conclude that there are areas of improvement that the company can act upon to help improve its product's results. Loss Ratio can be improved by the company by initiating a deep dive analysis into the states mentioned above that needs to be revised, as well as the construction type and customer segmentation. However, CAT Score is and will always be an unknown variable as it is determined by external large reinsurance companies. The market's performance as well contains an impact. Even though this analysis does not give a final answer to what the company has asked

for, yet it has certainly identified where the issues are coming from in order to act and improve its results.