Term Project - Step 1

Tabbalat, Abed

2021-05-14

Introduction

With my insurance background I have decided for the term project to be insurance related. I have met with an insurance company, and they agreed to provide data that I can analyze their homeowner's product in relation to premiums and losses. Premiums are the revenue collected from the policy holder to activate their policy. Usually different products have different policy time periods, and for this specific project, I am focusing on homeowners which determines that each policy lasts for a full year period. Losses are the claim amounts that happens when someone's home gets damaged that is covered by a homeowner's insurance policy. The loss amounts are dependent on the type of damage that needs to be repaired.
The company provided historical loss data that goes back to 5 years and current in force exposure for its current policies. The problem that is being identified is the company performance is not performing as expected and have requested to analyze their data and ask the right question to determine specific areas that can be improved to enhance its performance. The data includes in force premiums, loss amounts, and CAT Scores.

Research Questions

- Based on the historical trend, would a predictive model show improvement in the upcoming years?
- What would be the combined ratio for each segment in the homeowner's product?
- Is there any specific characteristics that can swing the losses or the CAT score?
- Where are the concentrations by geographical locations in losses and CAT score?
- Is the company non-renewing the right policies?

Approach

Economics is everywhere, the insurance company are required to pay claims when they come through, making it the highest expense coming through the company. The second highest expense will be the reinsurance that is purchased to protect the company from insolvency, for example when hurricane Irma landed on Florida, the industry lost more than 5 billion dollars. In this situation, the reinsurance kicks in that will cede the bulk of claims coming through since it happens all at once, and therefore it is needed to insure from a regulators stand point and making sure all claims have been covered.
Fixed expenses will be assumed as flat and therefore it will be factored out of the analysis. From this point.

How your approach addresses (fully or partially) the problem.

(Exposure File) contains the earnings of the company, expenses are claims, reinsurance purchase, and fixed cost expense. Earnings should be greater than expenses in any segment to show profit and that is the ideal

case scenario. This project is defined to identify characteristics in each segment such as County, Roof Age, Year Built, etc.... Calculating the combined ratio will be the first step in observing the characteristics of the product. From that point, a correlation exercise will be performed to determine what variable has the most impact on the loss ratio. Once it is determined a regression analysis will be applied to predict how the product is going to perform in the future. In addition, Once we know the dominating variables and finding out outliers, a recommendation can be given to pinpoint the areas that need improvement in order to lower the loss ratio and have more profitability.

Data

File Name: *Set 1_Claim Details*
Source: *Provided by the company via FTP, file is available in repository*

This dataset contains all the information about the claims that have been coming through for the past 5 years (2016 – 2021), the main use of this dataset is determine the loss ratio and outliers that comes with it.

File Name: *Set 2_Exposure File*
Source: *Provided by the company via FTP, file is available in repository*

This dataset contains all the property characteristics and earning that are part of the premiums that are received from the insured.

File Name: *Set 3_CAT Score*
Source: *Provided by the company via FTP, file is available in repository*

This dataset contains the detail behind the reinsurance that covers the policies which is a large expense variable that impacts the economics of the product.

Required Packages

- `library(plyr)`
- `library(dplyr)`
- `library(ggplot2)`
- `library(tidyverse)`
- `library(readxl)`
- `library(lubridate)`
- `library(outliers)`
- `library(lmtest)`
- `library(car)`

Plots and Table Needs

The approach will be cleaning all 3 datasets and determining which variables that are most critical to the economics to be a part of the calculation which can be combined into 1 data frame to perform a combined ratio calculation, correlation, regression, and outliers.

The plots that will be presented are:

1. Scatter
2. Line
3. Histograms
4. Box Plot

Questions for future steps

The question that I have for future steps are:

- Are all 3 data sets compatible in executing a regression model?
- Is there any other data needs that are needed to amend it?
- Will the results give insight on how to improve the product results?
- Will I need more packages to solve some of the problems other than the listed?