Shuai(Fiona) Jiang
Professor Yajima
MA678 Midterm Project Proposal
7 November, 2019

Travel insurance with the target attribute of claim status

Travel has become a more common event for many people because of the improvement of transportation systems and other technologies. People gain more exposure to the risks that could occur during their trips. Travel insurance may seem not as well-known as those insurances everybody owns that cover the risks caused by critical illness or a crashed car. However, it interests me because the factors contributing to the overall travel risks are changing dynamically just like other insurances' factors. Therefore, from actuarial science perspectives, there are some patterns and facts presented by real world's behaviors and figures needed to be observed. That is the exact work I would like to apply in my future career as an actuarial practitioner -- using predictive analysis tools to determine a fair price amount for insurance products.

I found a fun dataset provided by a Singaporean third-party travel insurance servicing company. The dataset contains standard information for each insurance policy and decent amount of observations (63326 records). Claim status is indicated to be target, and other information is assumed to be influential to claim status. The following table is listing the names of 11 attributes and their description.

1	Agency	Name of agency
2	Agency.Type	Type of travel insurance agencies
3	Distribution.Channel	Distribution channel of travel insurance agencies
4	Product.Name	Name of the travel insurance products
5	Duration	Duration of travel
6	Destination	Destination of travel
7	Net.Sales	Amount of sales of travel insurance policies
8	Commission	Commission received for travel insurance agency
9	Gender	Gender of insured
10	Age	Age of insured
11	Claim.Status	Claim Status

Shuai(Fiona) Jiang
Professor Yajima
MA678 Midterm Project Proposal
7 November, 2019

I wanted to explore if there exist some patterns between predictive variables and claim status. Also, I was wondering about the concrete relationship between them. If so, it would be feasible to make some predictive analysis about future claim status, which is helpful for actuarial risk classification application.

The potential problems I foresee include multilevel relationships. For instance, the claim status, which is the outcome, could vary by different destinations or durations. If that's true, then it is not reasonable to simply run the data together with a logistic regression model. Also, there are many destinations, I need to figure out if I am supposed to group the destinations by certain criteria, same for the durations. Another problem is that there are some missing values within each column. So far, the decisions for dealing with missing values cannot be made until further EDA and other statistical analysis.

I plan to start the initial exploratory data analysis in this week (the week of Nov 4th), then come up with the solutions for missing values considering the practical applications behind the data. By the end of the week of Nov 11th, the work mentioned above is expected to be completed. Then I will be able to work on figuring out the multilevel relationships. Further work such as making prediction models by groups, analyzing their statistical inferences and justifying the interpretability of the models by using data visualization is expected to start in the week of Nov 18th. The claim status prediction will be roughly completed after thanksgiving week.

As for future direction, trying more data and more variables is motivating. Besides, I hope I will obtain some general ideas about how to generate claim prediction's models, not only for the travel insurance. Also, I will read the material for SOA PA exam and see if there's anything helpful to add into this project.