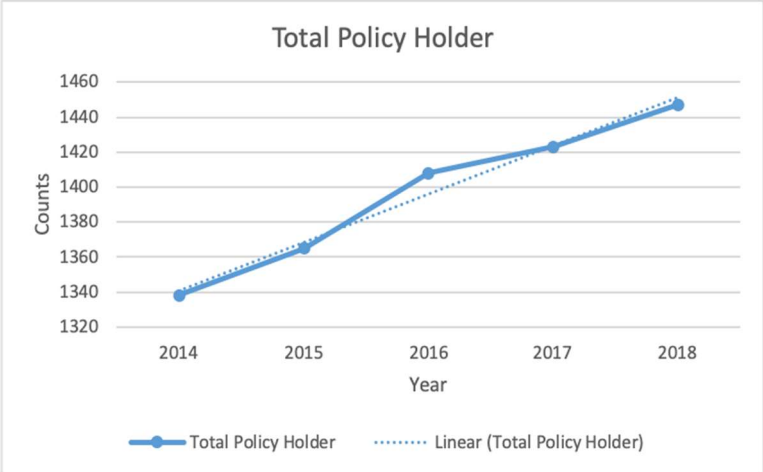
Report: Model Review for Unicorn Insurance

Executive summary:
Unicorn Insurance provides approximately 15 million customers with health insurance services in the Asia-Pacific. As a part of the internal review program for this year, the review on the pricing team's claims model for Cranberry city has been assigned to me, in order to supervise the company's risk management policy. Specific tasks are as follows: building an independent model through analyzing historical data, making comparison between own model and the model from the pricing team and putting forward reasonable suggestions for improvement. Generalized Linear Model has been taken to carry out analysis on historical data and construct an independent model to perform the comparison with the model of pricing team. The conclusion is that the model of pricing team is reasonably good, although there are still some improvements should be taken. Recommendations discussed include: 1. Enriching the database. 2. Increasing relative variables. 3. Enhancing supervisory methods.

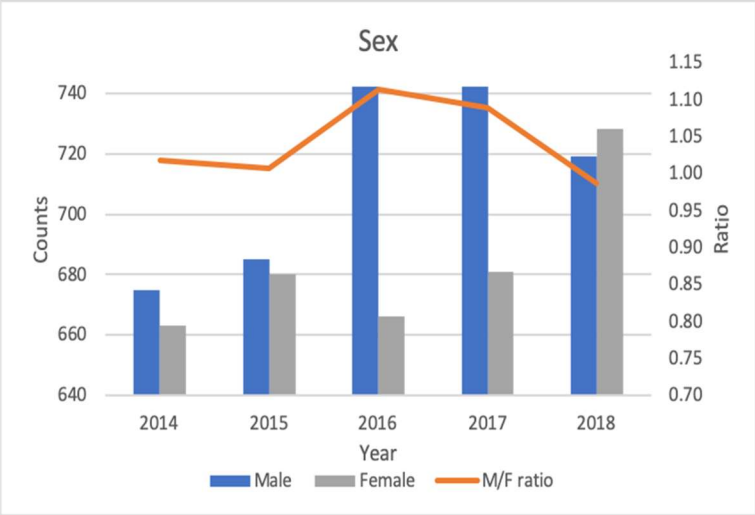
1.The analysis of demographic composition of policyholders

Total policyholders change (2014~2018)



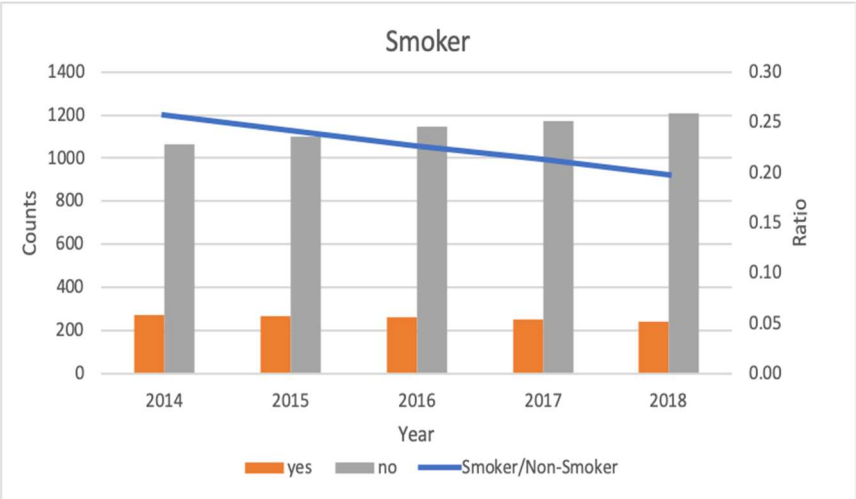
The statistics from year 2014 to 2018 have reflected a near linearly growth of the number of total policyholders, although there is a relatively higher increase between 2015 and 2016.

Policyholder gender distribution (2014~2018)



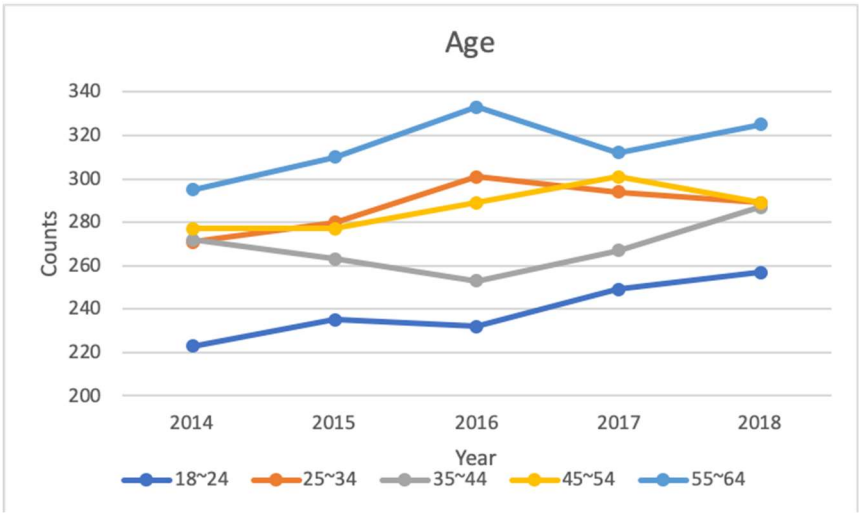
The Male/Female ratio of policyholders fluctuates only slightly between 1.0 and 1.1 from 2014 to 2018. The number of male policyholders increases from 2014, then it has a peak in 2016, afterwards it shows a decreasing tendency. In addition, the number of female policyholders has a fluctuation between the years of 2014 and 2016, after which it has a significant growth until the year of 2018.

Policyholder smoker distribution (2014~2018)



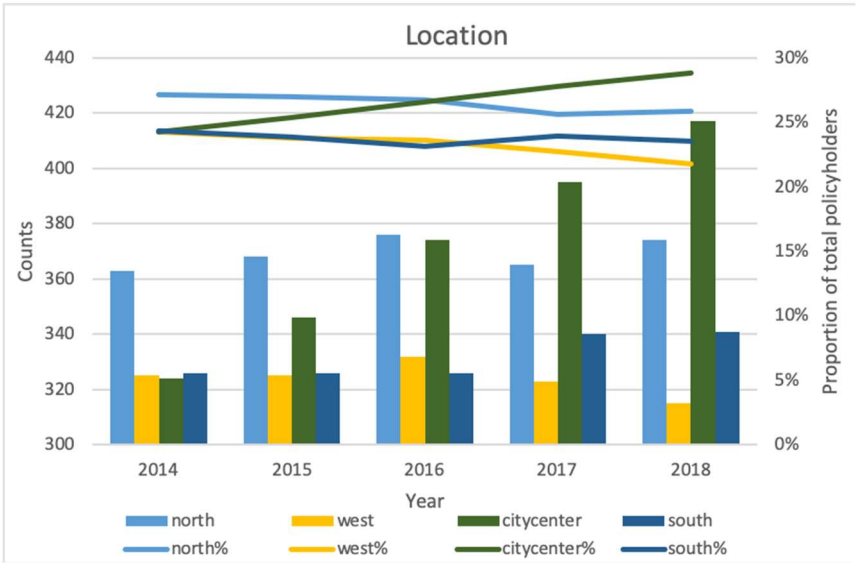
Considering smoking status of the policyholders, the majority of policyholders are non-smokers, while the number of smokers approximately remains unchanged over the 5 years. In addition, the ratio between smoker and non-smoker is around 0.25 in the year of 2014, then it begins dropping to 0.20 until 2018.

Policyholder age distribution (2014~2018)



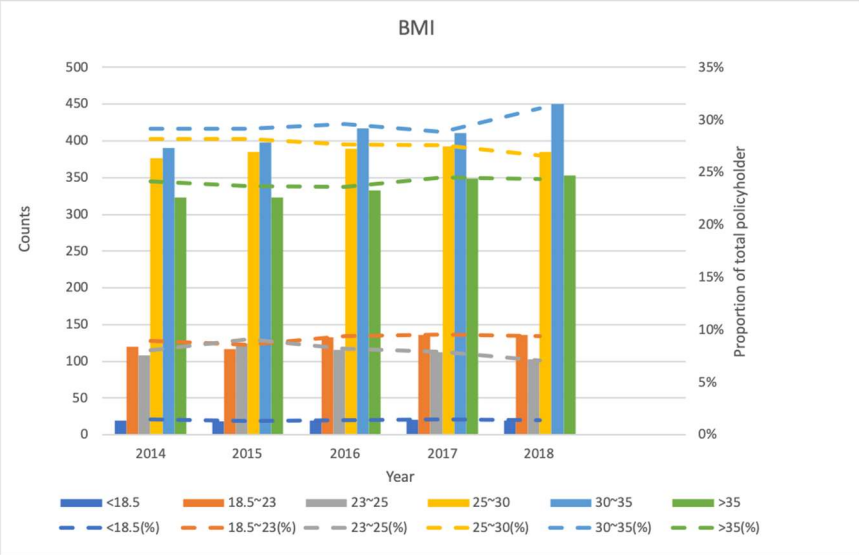
Policyholders have been separated into 5 age groups: 18-24, 25-34, 35-44, 45-54 and 55-64. The number of youngest group and elderly group respectively take up the least and largest proportion of total policyholders, and both of these age groups show a similar increasing trend. The policyholders aged 45-54 have an increase and achieve around 300 in 2017, then it decreases gradually. Likewise, people between the ages of 25 and 34 grow and have a peak of approximate 300 in 2016, after that it starts declining steadily. It is found that the policyholders aged 35-44 has an opposite trend compared with 25-34 age group.

Policyholder location distribution (2014~2018)



The policyholders living in the north keep a high and steady proportion from 2014 to 2018. The number of people living in the west is relatively small and shows a decline from 2016 and the proportion of southern policyholders rises slowly during the same period. It is noticeable that there is a significant growth of city center people over the 5 years, from 5% to 24% approximately.

Policyholder BMI distribution



We separate policyholders into 5 groups based on their BMI levels. The proportions of policyholders with different levels remain stable from 2014 to 2018. The percentage of people with BMI under 18.5 is the lowest, people who have BMI over 25 account for the largest proportion. Additionally, the normal people (i.e. people who have BMI between 18.5 and 25) have a moderate size of number.

2.Experience analysis of claims

Considering discussing the relationship between features and claim costs, we use Generalized Linear Model to assist this analysis. Since some features described by strings, we need to convert strings data into numerical values by the following rules: "male"⟶ 1, "female"⟶2, "yes"⟶ 1, "no"⟶0, "citycenter"⟶1, "north"⟶2, "south"⟶ 3, "west"⟶ 4. We construct Generalized Linear Model, in which the response variable is claim costs that fists the data of age, sex, bmi, smoker, location and year by using least squares. The result table is shown below:

Variable	Estimate	Standard Error	t-Statistic	P-Value
age	253.936	5.15604	49.2503	0.
sex	77.4428	144.751	0.535009	0.592661
bmi	301.038	11.767	25.5833	0.
smoker	23783.	186.78	127.331	0.
location	-92.9337	65.0743	-1.42812	0.153303
year	-5.52766	0.2456	-22.5067	0.
$R^2 = 0.88$				

Through comparing P-Value with a significance level of 0.05, we can conclude that changes in these variables (i.e. age, bmi, smoker and year) are correlated with changes in claim costs. However, the variables (sex and location) are insignificant correlated with claims, and hence we remove them and perform the modelling again. Here is the adjusted result:

Variable	Estimate	Standard Error	t-Statistic	P-Value
age	253.972	5.1558	49.2594	0.
bmi	301.795	11.7549	25.674	0.
smoker	23797.7	186.536	127.577	0.
year	-5.59633	0.201847	-27.7257	0.
$R^2 = 0.88$				

From the result table, it is found that the variables of age, bmi and smoker are positively correlated with claim costs while the variable year has negatively correlated with claims, and the coefficient of determination is 0.88 by using Generalized Linear Model, which means that the model is reasonably good to fit the data.

3. Modelling

In order to model the future claims that need to be paid by policy insurer, we can use regression analysis to perform the modelling. Through using historical data, we obtain a general equation of the model for claims:

$$\text{Claims} = \text{c}_1 \times age + c_2 \times bmi + c_3 \times smoker + c_4 \times year + c_5$$

Before modelling, there are several assumptions should be built based on the experience analysis above:

1. Since the amount of the claim of historical data is inflation-adjusted to 2019 AUD values, we assume the expected costs for 2019 are also inflation adjusted, which keeps the consistency of modeling.
2. Since there is no "year" variable consideration in the 400 policyholder profiles, we assume the coefficient of year in the equation above: $c_4 = 0$.
3. We do not consider "sex" and "location" due to the following reasons:
   Qualitative: during the 5 years of historical data, the gender ratio almost stays 1, thus the effect of gender is relatively small. In addition, due to the short study period, the impact of population mobility cannot be reflected within a short time-frame.
   Quantitative: through the use of Generalized Linear Model, we can obtain that the P-Value of "sex" and "location" are 0.592661 and 0.153303 respectively, both of them are larger than predetermined significance level (i.e. 0.05).
4. The data of "age" and "bmi" in 400 policyholder profiles only contains specific values (i.e. ages are 20, 30, 40, 50 and 60. bmi are 18.5, 23, 25, 30 and 35), thus we consider it is assumed that the "age" and "bmi" have been rounded.
5. The proportion of policyholders who have bmi less than 18.5 is extremely low, therefore we do not consider such policyholders.
6. The variables: "age", "bmi", "smoker" and "year" are statistically independent.
7. The variables: "age", "bmi" and "smoker" are positively correlated with claim costs. For these reasons:
   Qualitative: these three variables are closely related with health status, and hence we assume that the policyholders with elder age or with a higher bmi and the smoker policyholders tend to have high claim costs.

Quantitative: through regression analysis, we can clearly find that the coefficient of these three variables are larger than 0, which means that claim costs will increase with rising of these three variables' values.

Using the same method above, here is the result table:

Variable	Estimate	Standard Error	t-Statistic	P-Value
age	192.851	4.91044	39.2737	0.
bmi	30.8221	6.88132	4.47909	0.
smoker	23316.2	195.677	119.157	0.
$R^2 = 0.87$				

From the table above, we can find that the three variables are positively correlated with claim costs, and the P-Value of them are extremely small and the coefficient of determination is 0.87, thus we can say that this model is reasonable. Here is the model:

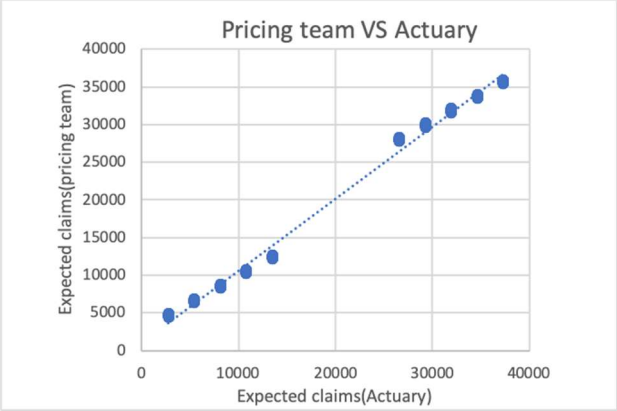
$$Claims = 192.851 \times age + 30.8221 \times bmi + 23316.2 \times smoker$$

4.Comparison and Recommendation

Using the model and the data of 400 policyholder profiles to calculate expected claims, then make comparison between the outputs of pricing team model and actuary model:

Similarity:

1. The claim costs are positively correlated with two variables: "age" and "smoker".
2. We can compare the claim outputs between pricing team and actuary through the plot below:



It is found that the expected claim outputs between models of pricing team and actuary display an approximately linear relation.

Differences:

1. For pricing team model: a) the expected claims are only positively correlated with two variables: age and smoker. b) this model can be only applied to some specific values in terms of age and bmi (i.e. age:20, 30, 40, 50 and 60; bmi: 18.5, 23, 25, 30 and 35)
2. For actuary model: a) the expected claims are positively correlated with three variables: age, bmi and smoker b) this model can be used more general (i.e. age can be more integer numbers. Bmi can be more values)

5. Recommendations

1. For existing variables, a broader range of values should be considered into the model to improve the accuracy of claims predictive.
2. Since there are only two variables considered in this model, combining more factors will optimize the current model.
3. More valid historical data should be selected and collected for testing model in order to modify models effectively.
4. Conduct claims leakage analysis, since it focuses on some claims leakage problems which can not be explained statistically by the patterns in historical data. This approach will provide a well-rounded improvement on the claim costs.
5. Claims quality self-assessment can be implemented by claim managers to perform a periodic test in order to ensure that the claim operations can be performed smoothly.