Insurance Claim Prediction

Jiali Lu
jialilu.physics@gmail.com

1. Introduction

In addition to make more accurate estimations and predictions, we also turn to machine learning for new perspective and fresh insight. In this project, I show how to make sense of the patterns learned by a model and come up with new ideas to improve the operation.

Actuarial loss prediction dataset [1] is used in this project, since we all have some common sense about insurance yet the problem is complex enough to provoke some thinking. This particular dataset is on workers' compensation, a type of insurance for work related injury and covers medical costs, lost wage as well as benefit to dependents in case of death. The objective is to predict the total payout (or ultimate loss) for each incident. In addition to the usual demographic information (such as age and marital status) and work related information (such as weekly wages and working hours), it also includes a free text description on the incident.

2. Result

A gradient boosting machine is used to model the data, after cleaning up invalid records and designing new features. Tow important classes of new features are inflation adjusted quantities such as real wage, and bag-of-word features of the word stems extracted from the text description.

The main focus of this project is on the interpretation. Below I show the effect of each factor using a technique called partial dependency analysis, which essentially measures the effect of the features by changing one feature at a time while leaving other features fixed. It provides clear graphical illustration of the average effect over the whole population.

2.1. Effect of Initial Loss Estimate

Initial loss estimate is the estimation made by the insurance company when the claim is filed, likely driven by domain knowledge. It is the most effective predictor of the ultimate loss, and is responsible for 86% of the error explained by the model. Each percent of increase in initial loss is associated with 0.72% increase in ultimate loss. This is shown in Fig 1.

1. https://www.kaggle.com/competitions/actuarial-loss-estimation/

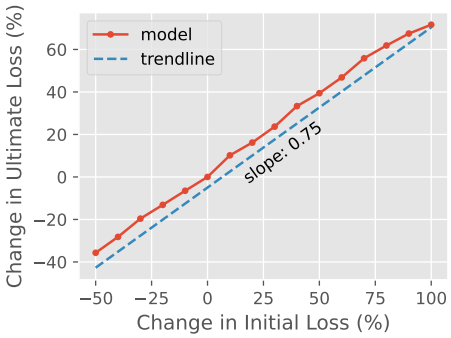

Figure 1. Effect of initial loss estimate on ultimate loss.

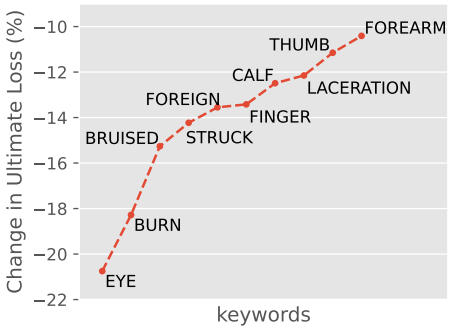

Figure 2. Top keywords associated with lower ultimate loss.

2.2. Effect of the Type of Injury

The type of injury is an important factor in predicting the ultimate loss, and is responsible for 11% of the error explained by the model. We can infer the type of injury from the claim description. If the description mentions "eye", "burn" or "bruised", the ultimate loss tends to be much lower. Whereas descriptions mentioning "head", "tissue" etc would have much higher loss. The top keywords associated with lower and higher ultimate loss are shown in Fig 2 and 3.

2.3. Effect of Age

Age also influences the expected ultimate loss. This is shown in Fig 4. A young worker at the age of 20 would have 6% lower cost than average, while one at 65 would have 6% higher cost.

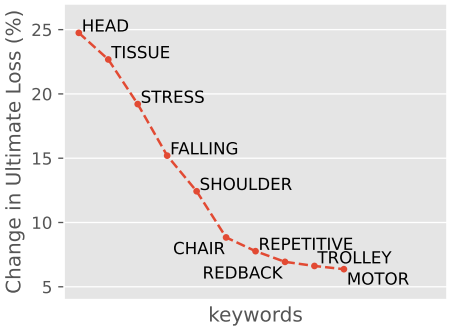

Figure 3. Top keywords associated with higher ultimate loss.

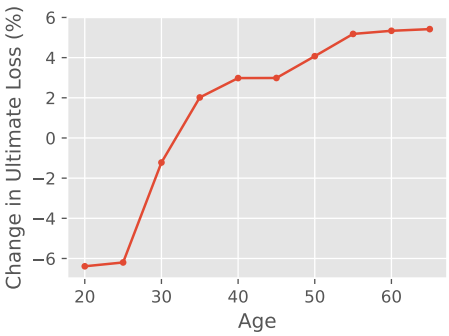

Figure 4. Effect of age on ultimate loss.

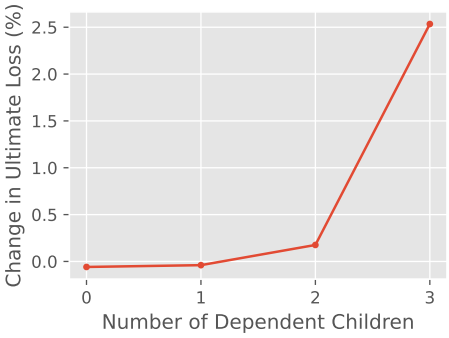

Figure 5. Effect of age on ultimate loss.

As expected, the expected loss increases as age increases, probably due to higher medical cost. There are two main increases: the expected ultimate loss rises sharply around 30 then rise moderately again at around 50.

2.4. Effect of Dependent Children

Having more dependent children is associated with higher ultimate loss. This is illustrated in Fig 5. The effect is most substantial when the worker has 3 or more children, which leads to 2.5% increase above average.

Figure 6. Effect of wage on ultimate loss.

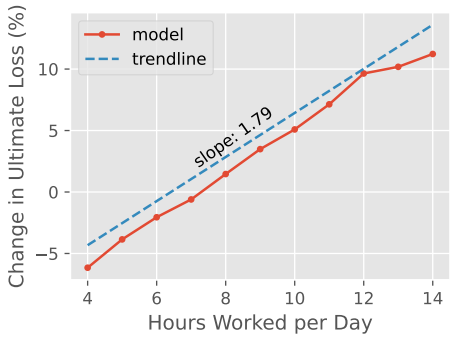

Figure 7. Effect of hours worked per day on ultimate loss.

2.5. Effect of Other Dependents

Having other dependents is associated with 2.1% higher ultimate loss, though the number of dependents does not seem to make a difference.

2.6. Effect of Wage

Higher wage is associated with higher ultimate loss, as shown in Fig 6. Given the same number of hours and days worked per week, each percent increase in wage is mostly associated with 0.18% increase in ultimate loss. However, the marginal effect diminishes for higher wage. When the wage is more than 25% higher than average, each percent increase would only lead to 0.075% increase in ultimate loss.

2.7. Effect of Hours Worked per Day

Longer working hours is associated with higher ultimate loss, as shown in Fig 7. Each additional hour worked per day is associated with 1.8% increase in ultimate loss.

2.8. Effect of Days Worked per Week

As we can see in Fig 8, working more days every week is also associated with higher ultimate loss. Specifically,

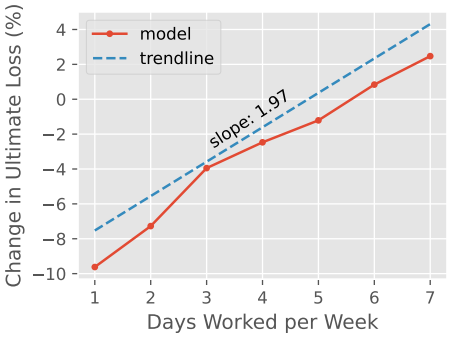

Figure 8. Effect of hours worked per day on ultimate loss.

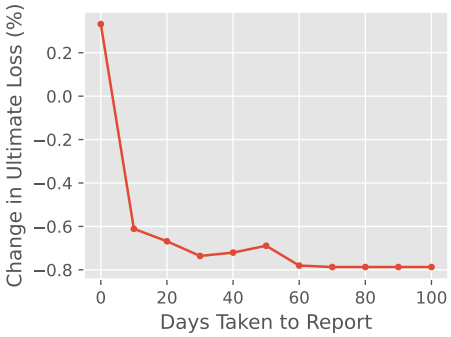

Figure 9. Effect of hours worked per day on ultimate loss.

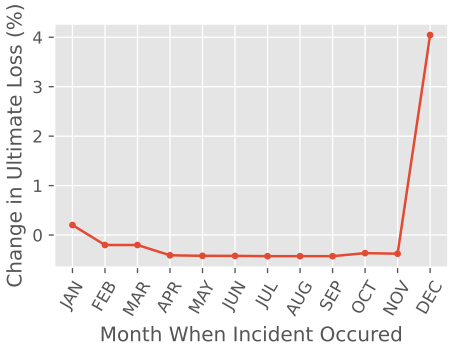

Figure 10. Effect of hours worked per day on ultimate loss.

working one more day per week is associated with 2.0% increase in ultimate loss.

2.9. Effect of Gender

Data suggests being female is associated with 1.4% higher ultimate loss than male. More specifically, ultimate loss for female is 1% higher than average while for male it is 0.4% lower.

2.10. Effect of Marital Status

Being single is associated with 0.21% decrease in ultimate loss and being married is associated with 1.42% decrease. However, claim with unknown marital status (which makes up about 10% of the record) has 0.67% higher ultimate loss.

2.11. Effect of Employment Status

Part time employment is associated with 1.7% higher expected ultimate loss, while full time employment has 1% lower loss on average.

2.12. Effect of Days Taken to Report Incident

There is a gap between the date when the injury happens and when the claim is filed. As shown in Fig 9, the fewer the days taken to report the incident, the higher the ultimate loss.

2.13. Effect of the Date and Time of Incident

Ultimate loss tend to be higher for incident occurring during winter, as seen in Fig 10. The effect is most visible in December, when the ultimate loss is about 4% higher than average.

The loss is also higher for incident happening during the night, as shown in Fig 11. However, the effect is not as substantial.

3. Discussion

Many of the findings above are not surprising. For example, having higher wage means higher compensation for the lost wage. In this case, the exploration above provides a more quantified answer and makes our intuition more precise.

Some results, however, are unexpected. While data suggests having non-children dependents is associated with higher expected loss, it is not clear why this could be the case. Dependents become a factor in the calculation of the compensation only in the unfortunate event of death. In

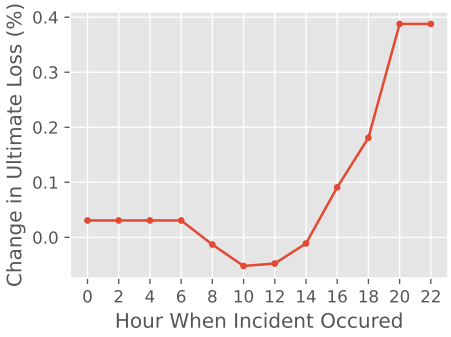

Figure 11. Effect of hours worked per day on ultimate loss.

this case, if the deceased worker has dependents, they are eligible for a death benefit. However, only 1 out of tens of thousands of cases involves death, and in that case there is no non-children dependent.

One very likely explanation is that the number of dependents is working as a proxy variable for the unmeasured working condition of the worker. Workers with dependents need to earn more to support the family and thus may choose to work in better compensated but also more dangerous jobs. This is supported by the data. Injured workers with non-children dependents tend to have 39% higher weekly wage than those without dependents, and are about 3 times more likely to mention ”hernia”, ”groin” and ”inguinal” in the description of injury. Groin hernia, including inguinal hernia, is a common but expensive medical condition associated with strenuous, physical demanding work. Other keywords with elevated frequency includes ”guns” (as in nail gun, not as a weapon), ”drums” (oil drum, large container for crude oil) and ”unloading”. Thus, it is very likely that the increase in expected loss is due to the higher risk of conditions like hernia, caused by the more demanding jobs. This suggests that we could further improve our prediction by incorporating features describing the working condition, such as a job description.

4. Conclusion

In this report, I showed how factors such as wage and type of injury influences the compensation for work related injury. The analysis sharpens our intuition by quantifying the effects of the factors. It also provides counter-intuitive new insights, namely the number of dependents may work as a proxy variable for the unmeasured working condition. This suggests the prediction can be further improved by incorporating features on working condition in the analysis.