Car Insurance Modeling and Analysis

GR5293 Applied Machine Learning for Financial Modeling

Kaiqi Wang (kw2875)
Shuyuan Wang (sw3449)
Yi Nei Charlene Lau (yl4292)
Yuwei Tong (yt2713)
Tingxuan Li (tl2959)

Contents

Business Understanding

Figuring out what insurance premium should be is complicated. Theoretically, an automobile insurance company would have to calculate the individual risk of each policyholder to determine how much their rate should be — neither overcharges nor undercharges insurers. Though it is infeasible to evaluate risks for thousands of policyholders, fortunately, by means of **machine learning techniques**, we could help automobile insurance company optimize its existing vehicle insurance scoring system, making insurance pricing more reasonable and accurate, and eventually creating more revenue for the company.

Data Understanding

In this project, we use ten-year (for each calendar quarter from 2009 through 2018, intotal 1080 rows) historical claims data from Safelife's personal auto policies.

Time Related Variables
Year
Quarter
Mileage

Risk Class Variables		
Vehicle Size	Driver Age	Driver Risk
Small	Young	Low
Medium	Middle	Average
Larger	Senior	High

Number and Amount of Claims Incurred
Bodily Injury
Personal Injury
Property Damage
Collision
Comprehensive

Model 1-Regression

Investigating the relationship (Amount)

Comprehensive does not have as strong a relationship with others as the other four variables (*Bodily Injury*, *Personal Injury*, *Property Damage* and *Collision*). This is expected because the other four are caused by automobile accidents, while *Comprehensive* is caused by force majeure.

	Bodily Injury	Property Damage	Comprehensive	Collision	Personal Injury
Bodily Injury	1	0.86	0.53	0.86	0.69
Property Damage		1	0.63	0.94	0.68
Comprehensive			1	0.59	0.47
Collision				1	0.71
Personal Injury					1

Model 1-Regression

Model Comparison

Building model on **sum** of *Bodily Injury*, *Personal Injury*, *Property Damage* and *Collision* (Y) versus *Year*, *Qtr*, *Vehicle_Size*, *Driver_Age* and *Driver_Risk* (X's): **random forest** gives the best prediction on testing data: it can explain almost 97% of the variability in the testing data.

Model	Testing R-Sq	Testing Relative Error	Testing MSE root
Linear Regression	87.69	23.21	6.18
Support Vector Machine	87.46	23.96	6.23
Decision Tree	94.27	11.30	4.21
Random Forest	96.78	8.82	3.16

Model 2-Classification

Methods Description

For classification modeling, we use three criteria to divide data manually into different classes according to the value of total number and total amount of claims. Then we apply three models to each case and compare the results.

Three Classification Criteria:

- **Case 1 (Total Amount)**--based on the value of total amount of claims and divide data into 7 groups
- **Case 2 (Total Number)**--based on the value of total number of claims and divide data into 6 groups
- **Case 3 (Total Amount & Total Number)**--divide the value of total amount and total number into three groups (small, medium, and large) respectively, then combine these two factors and get 9 groups of data

Three Classification Models:

- Decision Tree
- Random Forest
- Bagging

Model 2-Classification

Results and Findings

For all cases, **random forest** seems to be the best model with smallest test error rate. And for Case 1 and Case 2, the test error ranking from low to high is random forest < bagging < decision tree. And among three cases, **Case 3** seems to be the best with relatively smaller test error.

Test Error	Case 1 (Total Amount)	Case 2 (Total Number)	Case 3 (Total Amount & Total Number)
Decision Tree	0.144	0.153	0.125
Random Forest	**0.116**	**0.116**	**0.093**
Bagging	0.139	0.144	0.444

Model 2-Classification

Unexpected Findings

In Case 3 when combining total amount and total number of claims to classify data, there are some empty groups. For example, the group LA & SN is empty. That illustrates there's correlation between total amount and total number of claims.

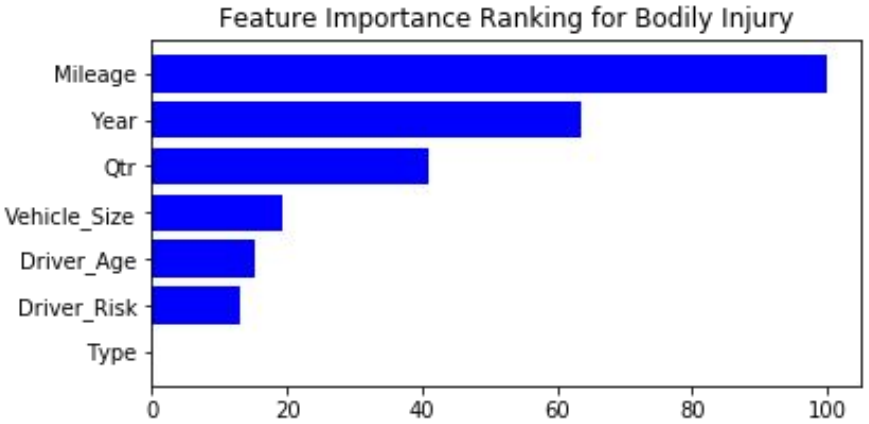
Group Name	Number of Data
SA & SN	664
MA & SN	0
LA & SN	0
SA & MN	321
MA & MN	60
LA & MN	0
SA & LN	0
MA & LN	26
LA & LN	9

Group Name Notation: SA/MA/LA means small/medium/large total amount, and SN/MN/LN means small/medium/large total number.

Feature Importance

Methodology

Feature Importance Ranking:
All features were ranked according to their relative percentage importance. Each target variable was analyzed separately to compare the rankings for each target variable. Rankings exhibited similar behavior with similar relative percentage importances.

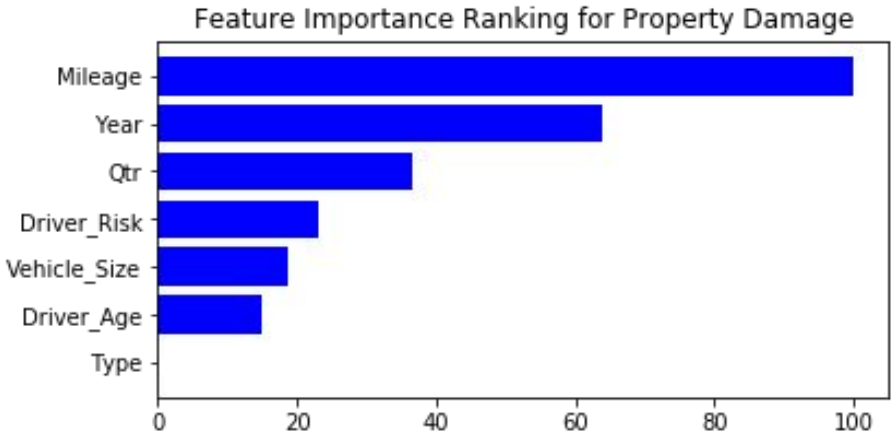


Feature ranking for Bodily Injury:
1. feature 5 (0.000000)
2. feature 4 (12.962134)
3. feature 3 (15.237442)
4. feature 2 (19.486313)
5. feature 1 (41.044426)
6. feature 0 (63.706108)
7. feature 6 (100.000000)

Feature Importance

Unexpected Findings

Feature importance rankings indicate that mileage is the most important factor. Driver Risk and Driver Age seems to have relatively lower importance, so it might be suggested that the characteristics of the vehicle outweigh those of the driver when it comes to vehicle insurance claims.



Feature ranking for Property Damage:
1. feature 5 (0.000000)
2. feature 3 (15.040907)
3. feature 2 (18.734408)
4. feature 4 (23.010867)
5. feature 1 (36.626818)
6. feature 0 (63.996687)
7. feature 6 (100.000000)

Improvement of Risk Level

Motivation

When we calculated the average value of each indicator according to the risk level classification, we found that the drivers of the average risk level caused the various losses to be comparable to the drivers of the high risk level. Even the personal injury item significantly surpassed the high-level driver. Therefore, we further decompose, compare and analyze drivers with average and high-level risks

Driver Risk	Low	Average	High
Mileage	8145.032	11530.357	8370.15
Bodily Injury	2067121	4339599	4733172
Property Damage	2125727	3716432	4070842
Comprehensive	1966201	2102473	2057019
Collision	3165114	6499239	6264212
Personal Injury	855859.8	3268671.1	2049467.1

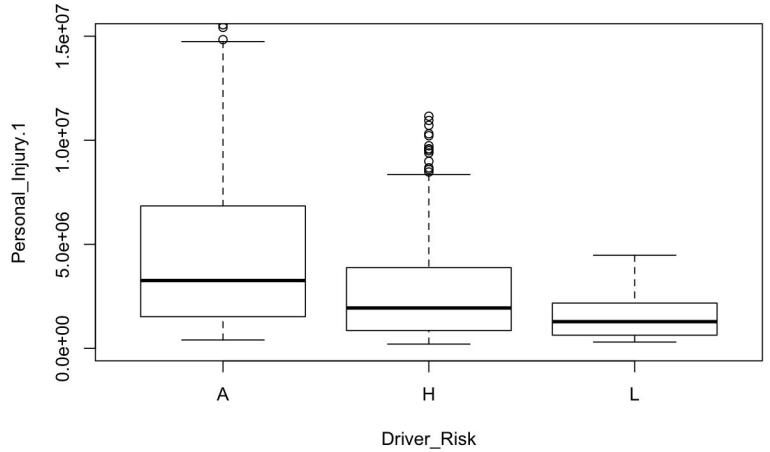
Improvement of Risk Level

Decomposed by Age

One of the most unexpected things we found is that the Middle age Average risk level's driver's personal injury amount is nearly twice about the Middle age High risk level's driver's personal injury amount.

We can clearly see the difference between the two through the box plot below.

Driver Age	Risk: Average	Driver Age	Risk: High
Young	1056080	Young	730554
Middle	5738775	Middle	2796480
Senior	3011158	Senior	2621367



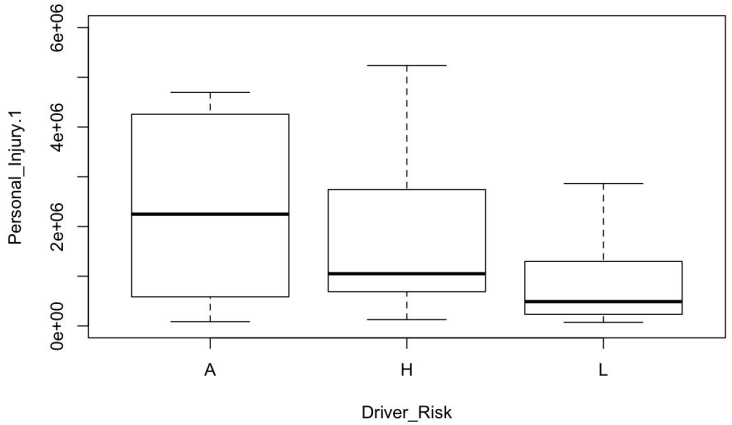
Improvement of Risk Level

Decomposed by Vehicle size

This time we found that for the Medium size vehicle, Average risk level drivers caused much more personal injury than the High risk level drivers.

Similarly, by box-plot we can see the distribution more clearly.

Vehicle Size	Risk: Average	Vehicle Size	Risk: High
Small	2041792	**Young**	701096.4
Medium	4354060	**Medium**	1960902.4
Large	3410161	**Large**	3486402.4



Improvement of Risk Level

Model Improvement

Through the decomposition analysis of the risk level, we can find that the original division of risk in the data is inaccurate. Similarly, in the ranking of feature importance, we can also see that the contribution of the risk level to the overall loss is significantly underestimated. It is also due to inaccurate grouping of risk levels.

Model Section

Improvement 1: Regroup risk level Average and High

According to the personal injury we relabel the Average and High risk level, which can help us improve the prediction accuracy. (In addition, for middle age medium size driver we default them to High risk.)

Improvement 2: Track driver's historical risks and losses

We can track and record the historical information of the same driver, and adjust the premium and the risk level accordingly according to the historical loss of the driver.

Conclusion

Based on our modeling results, we consider machine learning techniques quite useful and accurate in predicting number and amount of claims incurred, especially **random forest**. From the following analyses, we also notice some unexpected findings and dive into detailed analysis, i.e. the original classification of risk is inaccurate, the contribution of the risk level to the overall loss is underestimated.

We believe when more information is provided so that we could **regroup the risk levels**, as well as **add more of the characteristics of the vehicle as new independent variables**. In such way, when we re-apply **random forest**, we are able to obtain a better prediction results compared to what we have done at the beginning.