Final Report: Predicting Insurance Claim Cost

Linh Nguyen

12/13/2020

1 Exploratory Data Analysis

Data exploration and visualisation is the first important step to provide us with important information about the data structure and to identify any unique features to be aware of during model fitting. The train dataset, ”InsNova_train”, included 22,610 cases with no missingness and 11 different variables:

1. *id*: unique policy identifiers, ranging from 3 to 67,852;
2. *veh_value*: vehicle’s value $10,000’s, ranging from 0 to 26 with a mean of 1.87 (SD = 1.28);
3. *exposure*: the basic unit of risk, ranging from .003 to 1 with a mean of .48 (SD = .28);
4. *veh_body*: 13 categorical type of vehicles;
5. *veh_age*: age of vehicles coded as 1 (Youngest) to 4 (Oldest) with a mean of 2.68 (SD = 1.07);
6. *gender*: gender of the drivers with 12,850 females and 9,760 males;
7. *area*: 6 categorical driving area of residence;
8. *dr_age*: driver’s age category coded as 1 (Young) to 6 (Old) with a mean of 3.49 (SD = 1.43);
9. *claim_ind*: the indicator of claims coded as either 0 (no claim) or 1 (at least 1 claim);
10. *claim_count*: the discrete number of claims ranging from 0 to 3 with a mean of .07 (SD = .27);
11. *claim_cost*: the claim amount, ranging from 0 to 57,896 with a mean of 140 (SD = 1123.34).

First, we were interested in the pattern of bivariate correlations (Table 1) as well as the distributions of individual variables, especially the three variables of interest: *claim_ind*, *claim_count*, *claim_cost*. As expected, there was a high, near perfect positive correlation between *claim_ind* and *claim_count*, which is understandable because they are not independent: *claim_count* only gives a bit more information when *claim_ind* = 1. Numerous statistical significant ($p < .01$) correlations were reported; however, this was due to the large sample size that may attribute significance to even small effects. Among the other predictor variables, the only large (negative) correlation was between *veh_value* and *veh_age*. This is also understandable, because older vehicles generally decrease in market value. The main variable of interest, *claim_cost* showed the highest correlations with the other claim variables along with a small yet significant correlation with *exposure* (positively) and *dr_age* (negatively). We can then infer that on average cost tends to decrease with age and increase with risk.

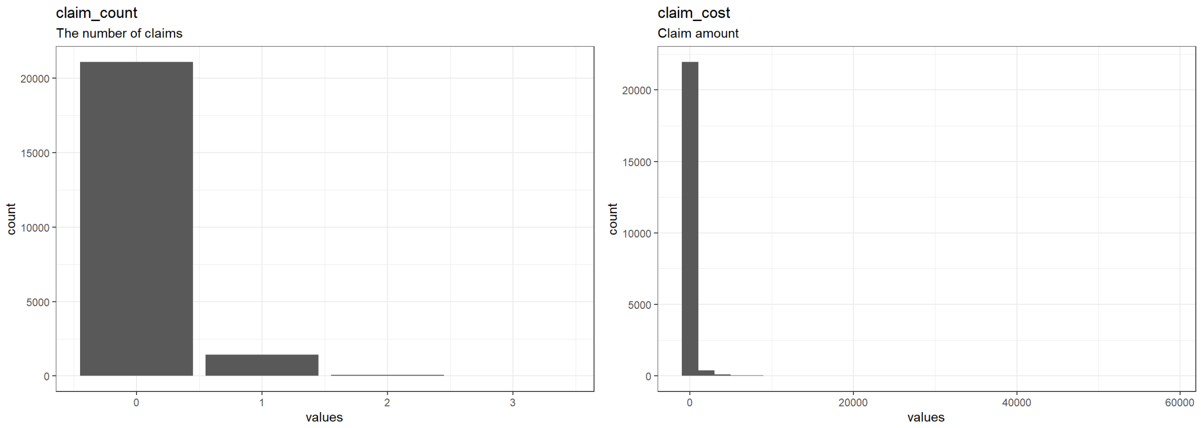
Table 1. Bivariate correlations between all numerical variables

Variable	veh_value	exposure	veh_age	dr_age	claim_ind	claim_count
veh_value						
exposure	-.00					
veh_age	**-.55	**.04				
dr_age	**-.06	**.03	**.02			
claim_ind	.01	**.14	.00	**-.04		
claim_count	.00	**.14	.00	**-.03	**.97	
claim_cost	.00	**.03	*.01	**-.03	**.46	**.46

Note. * indicates $p < .05$. ** indicates $p < .01$.

The three variables of interest are all highly skewed, with an abundance of zero values. Only 6.78% of the data had a claim ($claim\_ind = 1$) and there were high clusters around 0 for both count and cost, as shown below:

Figure 1. Distribution of claim_count and claim_cost



(a) Claim_count (b) Claim_cost

2 Model Fitting

All analyses were done in R. The main packages used for analyses and data wrangling were: tidyverse [Wickham et al., 2019], cplm [Zhang, 2014], and logistf [Heinze et al., 2020]. The analysis script is version controlled on the GitHub repository.

2.1 Multi-step approach

The submission dataset does not include the three claim variables. As a result, although the main variable of interest was *claim_cost*, we decided to follow a multi-step approach to initially predict *claim_ind* and *claim_count* to maximize prediction power, as these two variables have the highest correlation with cost.

The first stage included fitting models to predict each of the three variables separately. The remaining 8 variables are included in all models. Due to the minimal additional cost of data collection (already included in the administrative process) along with low correlations among numerical predictors (except for *veh_value* and *veh_age*), we have decided not to exclude any of these variables for model fitting. In addition, *claim_ind* was included as a predictor for both *claim_count* and *claim_cost*, and *claim_count* was also included as a predictor for *claim_cost*.

The second stage included predicting *claim_ind* and *claim_count* in the submission dataset using models fitting in stage 1. Then, the 8 predictor variables along with these 2 predicted values are used to predict *claim_cost*.

2.2 Predicting claim indicator

As previously discussed, the binary variable *claim_ind* is considered a very rare event, with approximately 6.78% of endorsement. Usually, a more even binary event would be fitted using logistic regression, using a logit link to transform the binary range (0,1) to the real number range $(-\infty, \infty)$ suitable for regression estimates. In this case, however, the rare nature of the event would bias predictions towards 0, and our model needs to use a penalty term to offset this bias. Thus, we used Firth logistic regression [Firth, 1993], which uses the square root of the determinant of the Fisher Information Matrix as the penalty term. As a result, bias is reduced from the maximum likelihood estimator and coefficients are shrunk to aid interpretation. A summary of model results is displayed in Table 2, with *exposure* and *dr_age* showing a significant effect ($p < .001$): the probability of having a claim increases with exposure and decreases with driver's age.

Table 2. Model results for claim_ind

	Coefficient	Std. Error	p-value
Intercept	-2.752	.748	<.001
veh_value	.055	.035	.133
exposure	1.944	.109	<.001
veh_age	.046	.039	.236
dr_age	-.106	.021	<.001
genderM	-.039	.063	.535

Note. For ease of presentation, the two categorical variables *veh_body* and *area* are not presented due to (1) their numerous factor levels and associated estimates and (2) the lack of any statistical significance. Nonetheless, they were included in the model.

2.3 Predicting claim count and claim cost

These variables were similar in that they both indicated very high clusters of 0 values, with the rest being positive numeric values. We thus used the Tweedie Compound Poisson Model for both variables [Zhang, 2013], which uses a log link function for parameter interpretation. The index parameter (p) indicates the shape of the dependent variable distribution. As expected from exploratory plots, $p = 1.55$ for *claim_count* and $p = 1.77$ for *claim_cost*, which indicated a compound Poisson-Gamma or Gamma distribution [Prokhorov, 2011]. Model results are presented in Table 3 and 4. Overall, the number of claims tends to increase with exposure and decrease with vehicle values. The amount of overall claim cost tends to increase with vehicle values, age, and male drivers, and decrease with exposure and driver's age. There were some differences due to the types of vehicles and area of driver's residence; however, the difference was not consistently present across all different factor levels for these two variables.

Table 3. Model results for claim_count

	Coefficient	Std. Error	p-value
Intercept	-27.401	.175	<.001
veh_value	-.004	.002	.050
exposure	.088	.007	<.001
veh_age	-.001	.002	.417
dr_age	.000	.001	.690
genderM	.003	.004	.344
claim_ind	27.351	.170	<.001
COUPE	.111	.046	.017
PANVN	.146	.046	.002
RDSTR	.704	.068	<.001
SEDAN	.110	.044	.012
TRUCK	.092	.045	.040
UTE	.116	.044	.008
areaB	-.026	.005	<.001
areaC	-.035	.004	<.001
areaD	-.037	.007	<.001
areaE	-.042	.007	<.001

Note. For ease of presentation, two categorical variables *veh_body* and *area* only included factors for which there was a significant effect. The reference level for *veh_body* was ”BUS” and *area* was ”A”.

Table 4. Model results for claim_cost

	Coefficient	Std. Error	p-value
Intercept	-28.355	.381	<.001
veh_value	.102	.916	<.001
exposure	-1.000	.046	<.001
veh_age	.210	.016	<.001
dr_age	-.077	.009	<.001
genderM	.126	.025	<.001
claim_ind	34.316	.174	<.001
claim_count	.632	.045	<.001
COUPE	1.244	.348	<.001
HBACK	1.015	.333	.002
HDTOP	1.368	.337	<.001
MIBUS	1.604	.353	<.001
PANVN	.811	.351	.021
SEDAN	.976	.333	.003
TRUCK	1.101	.340	.001
areaF	.190	.056	<.001

Note. For ease of presentation, two categorical variables *veh_body* and *area* only included factors for which there was a significant effect. The reference level for *veh_body* was ”BUS” and *area* was ”A”.

2.4 Evaluation

The train dataset was divided into 2 parts: 80% for model training and 20% for model testing and evaluation. We mainly examined the Root Mean Square Error (RMSE): $\sqrt{\frac{\sum_{i}^{N}=1(Predicted_i - Actual_i)^2}{N}}$, the normalized NRMSE: $\frac{RMSE}{max(Actual)-min(Actual)}$, and the Gini coefficient. A function to compute the normalized Gini values was created according to Kaggle instructions [Batzner, 2017].

In addition, because predicted values for *claim_ind* and *claim_count* were used to predict the final *claim_cost* values for submission, these two variables were deleted in the test subset in order to accurately evaluate the entire prediction process, because evaluating our *claim_cost* model using actual values for *claim_ind* and *claim_count* would provide a much better fit than using those predicted values, as they were susceptible to compounded errors from multiple models and stages.

Model fit statistics in the training subset and testing subset for *claim_ind* and *claim_count* are presented in Table 5. Overall, we can observe that fit did not decrease significantly for the test subset compared to the train subset, indicating that there was no serious overfitting. Of note is the seemingly very good fit for predicting *claim_count*; however, this is because we included actual values of *claim_ind* in the model, which shared a near perfect correlation (r = .97). Thus, this high fit is slightly artificial and will decrease significantly in testing the entire process.

Table 5. Model fit statistics for claim_ind and claim_count

Table 1: Claim indicator

	R2	RMSE	NRMSE	Gini
Train (80%)	.022	.507	.507	.334
Test (20%)	.018	.510	.510	.300

Table 2: Claim count

	R2	RMSE	NRMSE	Gini
Train (80%)	.945	.065	.022	.997
Test (20%)	.949	.061	.030	.997

Note. R2 = R-squared, RMSE = Root mean squared error, NRMSE = Normalized root mean squared error, Gini = Gini coefficient.

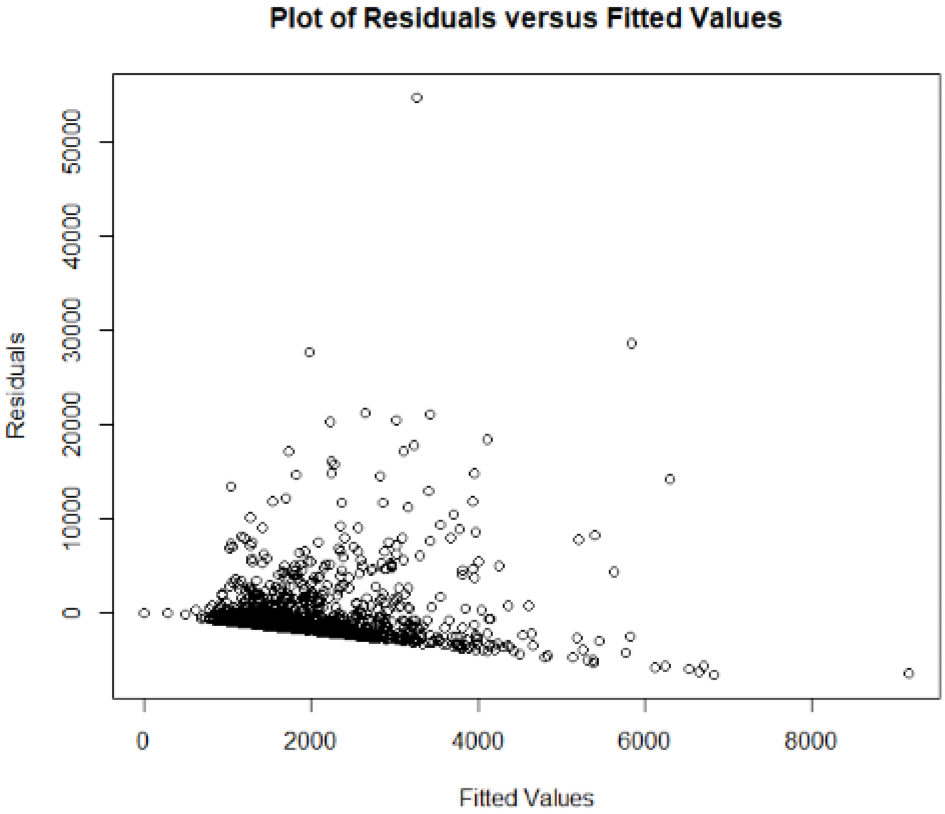
Model fit statistics in the training subset, testing subset, and testing subset without *claim_ind* or *claim_count* for predicting *claim_cost* are presented in Table 6. Similar to previous models, there was not a serious decrease in fit between the training and testing subsets. However, as is evident in evaluating the entire test process, there was a serious decrease in fit. This was expected, as noted above, because the entire testing process would involve two stages and include predicted values, not actual values, of *claim_ind* and *claim_count* in the final prediction.

Table 6. Model fit statistics for claim_cost

	R2	RMSE	NRMSE	Gini
Train (80%)	.259	958.811	.017	.971
Test (20%)	.213	1027.249	027	.965
Test whole process	<.001	1166.208	.031	.154

The plot of the predicted claim costs versus residuals is displayed in Figure 2. It is evident that there were many cases with under-prediction, in which the predicted values were much lower than the actual values. Nonetheless, there wasn’t a strong pattern seen in the residuals, and errors were generally evenly spread out, except for the very low and very high fitted values.

Figure 2. Plot of residuals versus fitted values for claim cost



3 Discussions

3.1 Other potential variables

The data are rather limited in the amount of substantive information that may help prediction. For instance, several categorical variables were coded and thus real values were obscured: age of drivers and vehicles and area. In particular, the crude age grouping may obscure any curvilinear effect that may exist in the population. Other demographic information, such as income, income-to-vehicle-value ratios, parental statuses may be helpful; and other information regarding driving behaviors, such as driving frequency/millage, traffic violation records, or any professional driving history (ride-sharing applications, taxi, or chauffeur services) would also be tremendously helpful as they may be predictive of driving risks.

3.2 Limitations

The current models assume a linear association between the age variables and all dependent variables. That is, we treated *dr_age* and *veh_age* as numeric, rather than categorical, variables. This restriction helped simplify interpretation and reduce the number of parameters; however, it may obscure any important curvilinear trend for age. For instance, risk may be highest for those in the youngest and oldest age group and lowest for those in between – such relationship would be negated in a linear analysis. Analyses that treated the age variables as factors did not yield different fit statistics and not included in this report. However, this is likely due to the crude categorization, with only a few groups for age instead of a continuous numeric variable.

In addition, the final model relied on a two-stage process and employed some predicted values in model fitting instead of observed values. As was evident in evaluating fit statistics previously, this practice significantly reduced model performance, because errors have been compounded from previous models to affect predicted values that were fed into subsequent models. Although we recognized this issue, the current model was selected because of the strong association between claim indicator/count and cost. Thus, we wanted to optimize all information provided and include these variables in the final model, albeit with their prediction errors.

Perhaps the most interesting discovery and also the biggest limitation during our process was the *id* variable. During exploration, we found that policy ID, an allegedly nonsensical variable to document different cases, was a significant predictor of cost when included in the final Tweedie model, with an estimate of $.0001, p < .001$. This was of course also a matter of the large sample size, for which a very small effect may be judged significant. However, our Gini score was significantly inflated when including this variable in the model (Table 7).

Table 7. Model fit statistics for claim_cost with and without ID

	R2	RMSE	NRMSE	Gini
Test with ID	<.001	1166.208	.031	.154
Test without ID	<.001	1166.206	.031	.239

Note. R2 = R-squared, RMSE = Root mean squared error, NRMSE = Normalized root mean squared error, Gini = Gini coefficient. Fit statistics are presented for testing the entire prediction process, which included predicted values of *claim_ind* and *claim_count* in the final model.

This sudden inflation in the Gini index − the main criteria in Kaggle leaderboard − raised many questions about the ID variable. In our opinion, there are two potential explanations for this increase in prediction: (1) Policy ID is not a purely random variable and (2) Policy ID is truly a random, nonsensical variable that coincidentally inflated model prediction due to some oddities of the model fitting process. The first reason would indicate a lack of randomization in curating the dataset, causing the ID variable to be incidentally related to some other substantively meaningful values. For instance, higher IDs could be assigned to cases later in time. However, we believe the second reason to be much more probable, and much more interesting.

Because our final model was underpredicting real cost, with a much restricted range of 0 to 39.88 compared to the given range of 0 to 57,896 in the train dataset, the small but positive coefficient for ID in the model could be successfully but artificially inflating our prediction to approach real values more closely. Nonetheless, the existence of a seemingly nonsensical variable that boosted prediction by a non-negligible amount really casts doubt into inference made for any other variables in the model. This highlights a serious problem of model fitting without substantive domain knowledge. In maximizing prediction according to some predetermined indices (Gini index, in this case), our model suffers from a loss of interpretation and meaningful inference. If we were to accept a nonsensical variable to artificially inflate prediction scores, we also have to apply much more caution in interpreting any of the other variable. In particular, after a long process of model fitting, model selection, and maximizing prediction, it is likely that our individual estimates are no longer suitable for accurate inference.

References

[Batzner, 2017] Batzner, K. (2017). Gini coefficient - an intuitive explanation. https://www.kaggle.com/batzner/gini-coefficient-an-intuitive-explanation. Accessed: 2020-11-25.

[Firth, 1993] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, pages 27–38.

[Heinze et al., 2020] Heinze, G., Ploner, M., and Jiricka, L. (2020). *logistf: Firth's Bias-Reduced Logistic Regression*. R package version 1.24.

[Prokhorov, 2011] Prokhorov, A. (2011). Gamma-distribution. http://encyclopediaofmath.org/index.php?title=Gamma-distribution. Accessed: 2020-12-01.

[Wickham et al., 2019] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

[Zhang, 2014] Zhang, W. (2014). cplm: Tweedie compound poisson linear models. r package version 0.7-2.

[Zhang, 2013] Zhang, Y. (2013). Likelihood-based and bayesian methods for tweedie compound poisson linear mixed models. *Statistics and Computing*, 23(6):743–757.