UNIQA Data Science Challenge 2022

Amit Sahoo	Mohamed Hassan	Christoph Spiess
India	Egypt	Austria
+49 1573 9250 491	+49 1772 599 981	+43 699 11315279
amit.sahoo@rwth-aachen.de	mohamed.hassan@tum.de	christoph.spiess@tuwien.at

Overview

In this challenge the focus was on insurance pricing. To price insurance policies adequately it is important to calculate the risk that is in insurance terms, the chance that something harmful or unexpected could happen. The data set contained 138 variables and one target binary variable to predict, which is the claim, "CLAIMS_REAL". Due to data anonymity, we could only assume that there is correlation between these variables and the target variable. After generating multiple models utilizing the whole data set, we tried forward selection and backward elimination, relying solely on the GINI coefficient for assessment, but in the end and due to computation bottlenecks we decided to rely on the full model. Given that the data set had many policies lasting three months, including exposure as feature could result in a biased model. To handle this problem, we decided to fit a GAM from the Poisson family and include exposure as an offset. In other terms we would fix the exposure´s coefficient to 1 which we would later utilize by using the logarithmic properties to arrive to our final predictions which can be established by running the attached Jupyter notebook. This PDF aims to provide guidance while walking through our code. The final predictions can be found here as a csv-file:

https://www.dropbox.com/sh/unfv3ypvo8zsfgd/AABe2__WwUzf5HnI7pKSvBTJa?dl=0

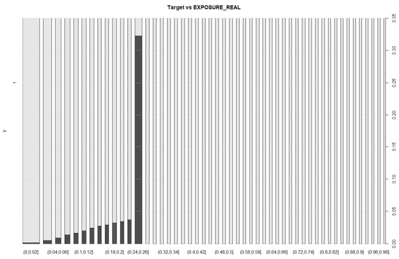
Model preparation

Variable assignments

To prepare the model we first assigned the columns to several bins. The endings of the column names implied to which bin type the column should belong. Therefore, we divided the columns into categorical *(_NOM)*, integer *(_INT)* and numerical ones *(_REAL)*. However, we made one adjustment. We noticed that PERSON_20_REAL consisted of only a few integer values! (in the train dataset as well as in the test dataset) so we decided to assign it to the integer bin.

Issues with including exposure as a standard feature

Since there were no detailed information provided about the variables, we decided not to exclude any variables, aiming for not to miss any aspects. However, we decided not to include exposure as a covariate due to strong dependency with loss occurrence. Looking at the plot below we, see that there is a peak in claims frequency around exposure of 0.25 (3 months). We do not expect that an exposure value of 0.25 should be predictive.



Formula

We tried some models removing variables in advance to reduce complexity. However, doing so had the effect of decreasing the GINI value. Furthermore, due to lack of information about the variables, we decided it is good to have all the variables in the model, as we were aiming for not to miss any aspects. Following this approach, we also did not remove the values of -1 and "000" to see if the fact that a variable is missing had an effect on the number of claims. We also included the current model as a variable as suggested by UNIQA.

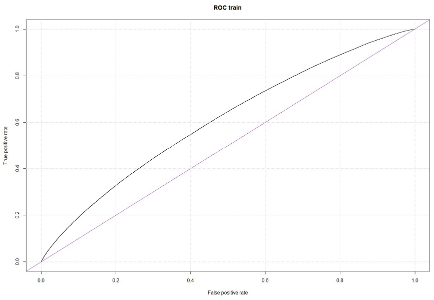
Model

Since our target variable is binary a logistic model would be suitable, however, to have the possibility to scale our predictions intuitively to annual basis and since claims are usually expected to follow a Poisson distribution, we decided to use a Poisson model. We went for a GAM, having the ability to extract not only linear dependence but also applying smoothing effects on certain variables. The division of the variables into separate bins already provided the idea to apply the smoothing effect on the variables ending in "_REAL" excluding PERSON_20_REAL which we treated as an integer variable. We generated multiple GLMs / GAMs. The initial challenge was to still make use of exposure. We decided to model rates instead of counts, hence we predicted claims divided by exposure in our Poisson model which is equivalent to including it as an offset in our GAM (we used the bam function to create a GAM, this is recommended for big data sets). This is due to the fact that e.g., the expected number of claims should be 0.4 for a given bordereau with 0.1 expected claims in 0.25 years of exposure.

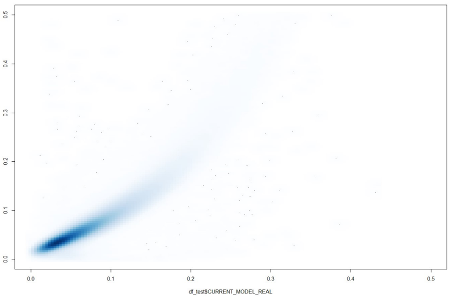
Our predictions would be interpreted as the expected number of claims per policy which we afterwards transform into probabilities. As the link function is log(.), applying the exponential function delivered the predicted annual claims. Using the Poisson distribution, we transformed the predicted claims into probabilities of one or more annual claims.

$$P(x) = \frac{\lambda^x}{x!}e^{-\lambda} \Longrightarrow P(x \geq 1\} = 1 - e^{-\lambda}$$

The GINI coefficient we obtained is around 0.219, which is slightly better than around 0.215, the one of the current model. At the right-hand side you can see the ROC curve.



Comparison against the current model



As we cannot compare our model to actual claims from the test dataset, we decided to plot our predictions against the current model. We can see that especially for small amounts of expected annual claims the predictions have a high correlation. However, our model would in general tend to predict more claims than predicted by the current model.

Variable Importance

To further analyze the model, we look at the most important variables. We first examine the effect of the most important non-numerical variables in combination with their histograms within the train dataset, afterwards we will discuss the numerical ones. The claim probability increases if and only if the expected claims increase, so talking about one or the other when talking about trends does not affect the meaning.

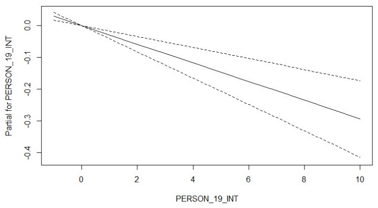
The most important non-numerical variables according to the p-values are (to be noticed by the additional ending of e.g. POLICY_DATA_3_NOM2, R delivers the p-value of the second bin of POLICY_DATA_3_NOM):

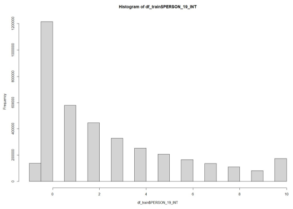
PERSON_19_INT	1.105198e-06
PERSON_18_INT	7.108405e-06
CAR_23_INT	3.031080e-05
POLICY_DATA_3_NOM2	1.148068e-04
CAR_13_NOM31	2.658480e-04

Let us examine the effect of those variables on the predictions:

PERSON_19_INT

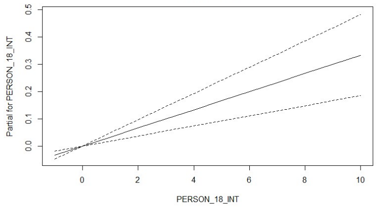
We can see that by increasing value of PERSON_19_INT, fewer claims are predicted. The histogram does not really give a hint about what the variable represents. The confidence level decreases for higher values of PERSON_19_INT due to the decrease of observations.

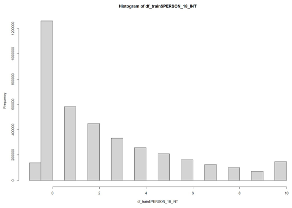




PERSON_18_INT

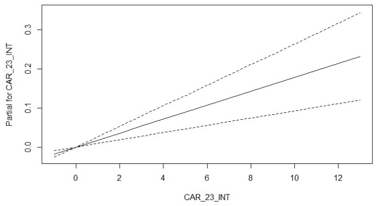
Looking at the histogram, one would expect PERSON_18_INT to maybe represent something similar to PERSON_19_INT, however, this is unlikely, since we notice an increase of predicted claims by increasing PERSON_18_INT.

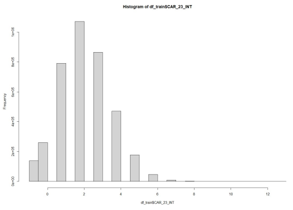




CAR_23_INT

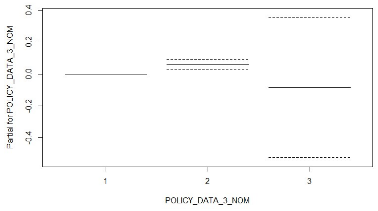
By increasing CAR_23_INT the number of expected claims rises, with a much broader confidence interval at higher values. Maybe this variable could represent the number of drivers of one car. A high frequency of 2 would be likely and the value of 0 would mean 0 expected claims, however, having 0 drivers probably an insurance is not needed.

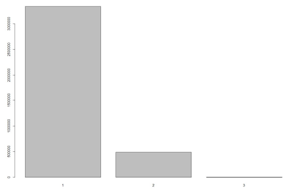




POLICY_DATA_3_NOM2

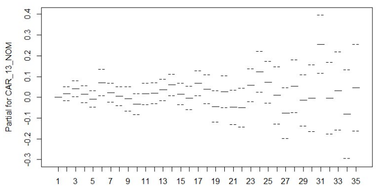
This categorical variable consists of 3 different values. The confidence interval for values of 1 is very small as there are many observations obtaining this value. The predictions differ slightly for different observations, being expected to be the lowest for values of 3, however the confidence level is low due to few observations. The bin for observations with values of 2 attained the best p-value.

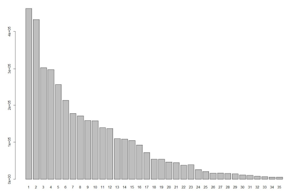




CAR_13_NOM

We can see a clear increase of the confidence interval as the number of observations decreases by an increase in CAR_13_NOM. The highest expected number of claims is attained for values of 31 by some margin. This value forms the bin with the highest p-value.



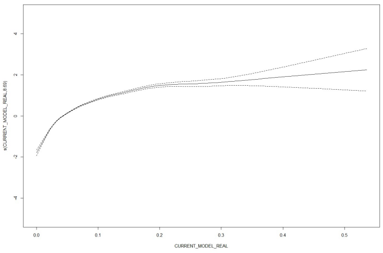


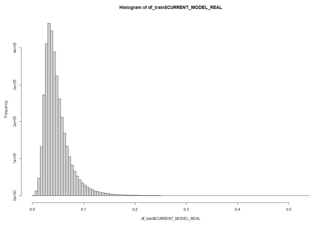
The most important numerical variables according to the chi square value are:

CURRENT_MODEL_REAL	3739.634212
INS_HISTORY_20_REAL	19.487488
CAR_4_REAL	18.077913
INS_HISTORY_24_REAL	7.182759
INS_HISTORY_26_REAL	5.623634

CURRENT_MODEL_REAL

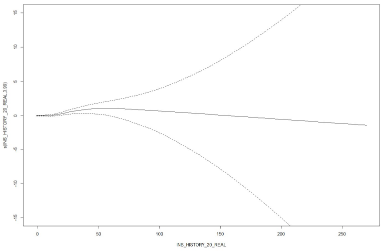
We notice a monotonical increase of our predictions by increase of the current model, just as expected. The current model is obviously the most important predictor. What we can also notice is the Poisson like behavior in the histogram, which supports our model decision.

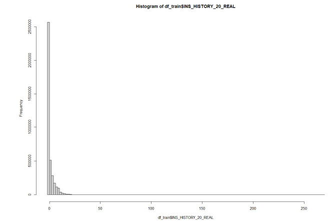




INS_HISTORY_20_REAL

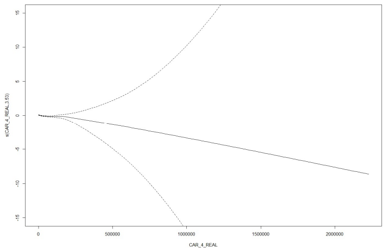
The claim probability slightly increases for values around 50. Afterwards we see a negative trend. This is an example of the advantage of using a GAM in comparison to a GLM, being able to predict positive and negative trends within the values of the same variable.

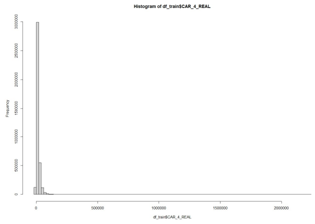




CAR_4_REAL

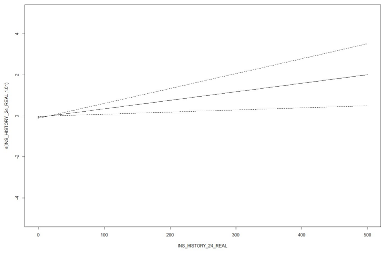
Increasing values of CAR_4_REAL lead to a decrease in claims probability, however, the confidence level gets lower quite quickly, due to the lack of high numbers of observations with high values. This variable could eventually represent car costs. There will be few high values and the security measures are likely to improve by higher price level.

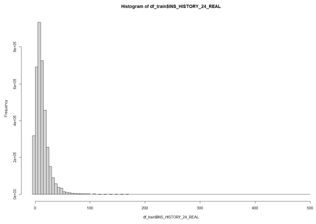




INS_HISTORY_24_REAL

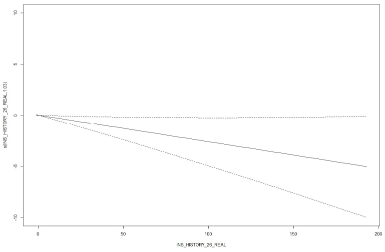
By looking at the histogram we immediately thought of the current model and Poisson distributions. This variable could represent past claims. This would as well explain the increase in claims probability by an increase in INS_HISTORY_24_REAL. However, there are extremely high values in the range of a few hundred, which we would not expect.

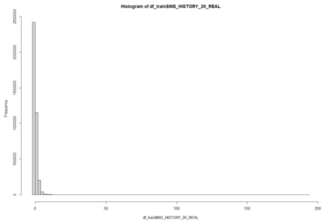




INS_HISTORY_26_REAL

Here we obtain another example. why it is important to know what the variables represent when modeling claims to set prices, outside of any competition. We could also assume this variable to represent past claims, which would contradict a decrease in claims probability.


s(INS_HISTORY_26_REAL,1.03)
INS_HISTORY_26_REAL


Histogram of df_train$INS_HISTORY_26_REAL
Frequency
df_train$INS_HISTORY_26_REAL