Predicting the cost of an insurance claim: Stacking Ensemble

GIAMPINO ALICE[1], MELOGRANA FEDERICO[2],
MISSINEO NICHOLAS[3], SOMASCHINI BEATRICE[4]
GROUP: GOLDFISHER

[1]Università degli Studi di Milano-Bicocca
[2]Università degli Studi di Milano-Bicocca
[3]Università degli Studi di Milano-Bicocca
[4]Università degli Studi di Milano-Bicocca

Abstract. Predicting the cost of an insurance claim based on important variables is the key to have a great predictive power. Therefore, accurate predictions, often, do not derive from model with high interpretability. For this reason, we spent a lot of time thinking about new variables that can help to develop a better model. In our analysis, SAS Viya and R were fundamental tools not only for manipulating and exploring data, but also for creating new variables and evaluating our models.

Keywords: Data engineering. Predictive power. Feature extraction. Regularization.

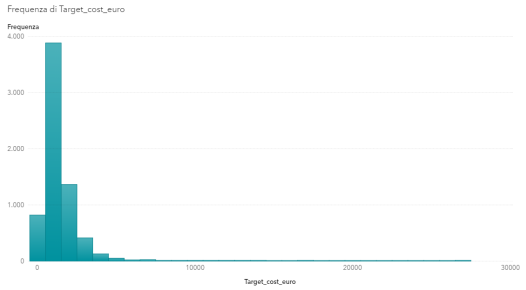

Figure 1: Histogram of target variable

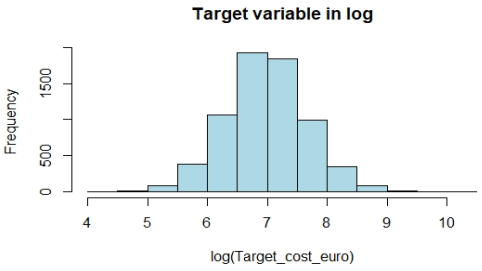

Figure 2: Histogram of the target variable in logarithm

1 PREPARE DATA

At the beginning of our analysis, we start combining train set and test set. In this way, every changing in the covariates is equal in both. Following, we look at the complete dataset and the analysis begin. For our goal, we decide to use R Studio and SAS Viya.

2 DATA EXPLORATION

At first sight,the dependent variable has an asymmetric distribution, as it can be seen in Figure 1. After a logarithm transformation, the distribution results more regular. Therefore, we decide to apply the logarithm for the target variable, Figure 2.

Looking at the entire dataset, some variables emerge to be important and this is a considerable starting point for data engineering and feature extraction. We use SAS Viya as a powerful tool for determining which ones to select. Thanks to the quickness of this platform we have been able to collect meaningful insight of our data and to discriminate between variables. This can be seen in Figure 3.

Joint analysis of the variables is useful to understand the multivariate relationship between them. Hence, a heat map is the right tool to explore in a visual way the correlation, Figure 4.

Another important issue to deal with is the presence of missing value (Figure 5) that can compromise the goodness of our analysis. For this reason, in a following section a detailed explanation is reported.

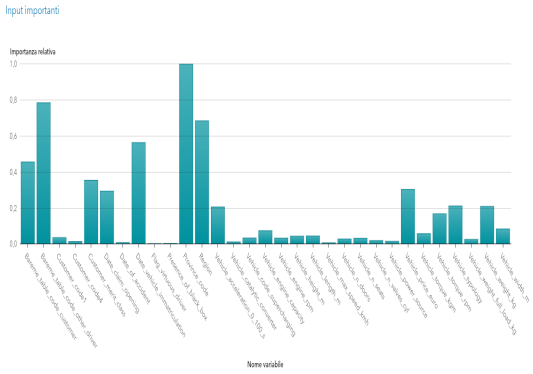

Figure 3: Important inputs

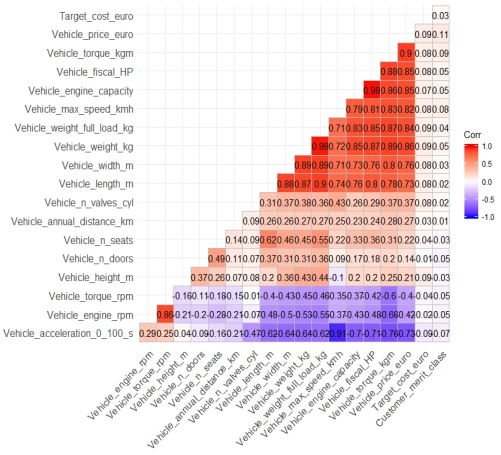

Figure 4: Correlation Plot

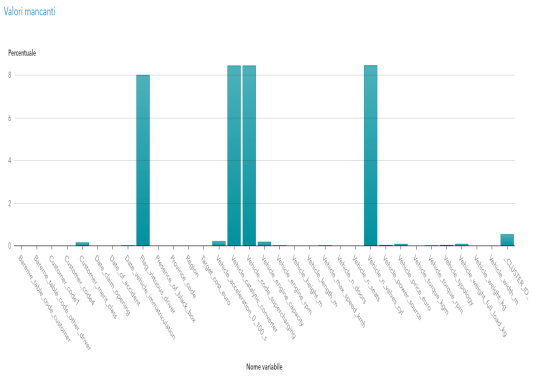

Figure 5: Percentage of missing values

3 DATA ENGINEERING

To improve the information that could have been useful to develop the model, we have searched for additional information in internet that could provide extra features. We found some information about accident rate detailed for each Province, and also in a larger aggregation for regions. Also, we found a rank for assurances policies prices in a province level of aggregation. That variable has proof to be very useful to predict the output. In addition to web searching, we have also made a huge effort trying to extract different features and information from existing covariates. We have extract Year, Month, Day, and WeekDay from both AccidentDate and ClaimDate. Those variables have later proved useless. Instead, we compute also the difference, in days, between ClaimDate and AccidentDate, that has been very useful for our purpose. Other tries has lead to nothing except wasting of time, like splitting the year in the four season, that has not been included in the model. We have also extract the year in which the car has been registered, that had resulted very important. In conclusion we have tried to extract as much information as possible from our data, even merging different sources to improve our pool of possible important predictors. We strongly believe that data engineering is a really important, and perhaps underrated, step in a Machine Learning pipeline. It is also a very time-consuming step, that, if faced up in the right way can lead to very satisfying results.

4 MISSING IMPUTATION & OUTLIER DETECTION

After an exploration procedure, we found out that the dataset contains missing values. For the missing imputation we considered the completed dataset composed of the train set and the test set, this because it is really important not to have missing values in the test set, otherwise it becomes impossible to have a prediction.
For the most of the variables including the vehicle information we have decided to impute missing using a prediction based on a linear model, as we found out that vehicle variables are really correlated among each other, in particular we used a linear model to replace missing for the variables:

- Vehicle_max_speed_kmh (for which we still have a NA due to the fact that the regressor of the prevision model: *Vehicle_engine_capacity,Vehicle_fiscal_HP* both present a missing value at the specific observation)

- Vehicle_acceleration_0_100_s using as regressors: *Vehicle_max_speed_kmh*
- Vehicle_engine_capacity using as regressors: *Vehicle_fiscal_HP*
- Vehicle_weight_full_load_kg using as regressors: *Vehicle_weight_kg* and *Vehicle_power_source*
- Vehicle_price_euro using as regressors: *Vehicle_fiscal_HP* and *Vehicle_acceleration_0_100_s*

For the following variables we decided to use the mode of the variable, more specific:

- Customer_merit_class
- Date_vehicle_immatriculation, for which we select the most common year: 2007 and putting: 2007-01-01 (the month and day variables will be then eliminate)
- Vehicle_power_source
- Vehicle_typology

Other missing values have been replaced with their mean value:

- Vehicle_fiscal_HP
- Vehicle_engine_rpm
- Vehicle_torque_rpm

For the feature : *Vehicle_code_supercharging* and *Vehicle_catalytic_converter* we imagine that the meaning of the missing value was that the cars appertain to another class, so we have replace the missing with another class: *ALTRO*.

With the purpose of outlier detection, we have decided to consider the outlier only in a multivariate way as we consider it a more proper analysis instead of a univariate one.
We have followed a supervised approach using a linear model. The response variable for the regression is the target variable: *Target_cost_euro* and we put all the other variables in the model. We performed both an *Outlier test* and a *Influence plot* from which we have managed to identify two anomalous observations. For this reason, we have eliminate them.

5 FEATURE SELECTION

In the field of Feature Selection, after an initial exploration previously stated, we have decided to use a variable importance method, through a boosting regression. Thanks to that technique, we have been able to eliminate a lot of noisy variables that would have confound the real model. The selected variables, both numeric and categorical, have been insert in the model, because we had an evidence that were strongly associated with the dependent variable. We have also corroborate this result with other models of variable importance, as another variable importance method developed on a Random Forest. The result, the most important variables, were the same in the two setting, so we have a measure that our selection technique is not far away from the truth.

The variable used for the models are the following: *Region, Bareme_table_code_other_driver, Vehicle_price_euro, Bareme_table_code_customer, diffday, annimmatricola, RankrischioProv, Vehicle_weight_kg, Vehicle_height_m, Vehicle_catalytic_converter, Vehicle_acceleration_0_100_s, Vehicle_max_speed_kmh, Vehicle_weight_full_load_kg,Vehicle_torque_kgm, Vehicle_code_supercharging, Vehicle_width_m, Vehicle_engine_rpm,Vehicle_engine_capacity, IndiceIndProv, monthaccident*

6 MODELS

Our best model is a stacked ensemble between a Random Forest Algorithm and a Boosting Algorithm with a percentage of 60% for the first one and 40 % for the second one. We tough about these two models because we had, after selecting the variables, many interaction between them so these two allowed us to verify the significance among them.
The tuning parameters that we choose for the Random Forest is the number of trees, equal to 1200, the max depth of the trees, equal to 4, and the min size of the node, equal to 1.
The tuning parameters for the Boosting, instead, is the max depth of the trees, equal to 3, and the learning rate, equal to 0.001. For this model we used the One-hot-encoding in the step of pre-processing: thanks to him we drop out the categorical variables transforming them a dummy variables. Furthermore, we scaled the numeric variables that needed it. To find the best solution we used the Cross-validation.

Other models that we tried before finding this combination are a support vector machine and a linear model. The Support Vector Machine had *radial* as ker-

nel and a list of possible cost function equal to 0.01, 0.1,1,5,10,100 to find the best solution. The linear model, instead, gave us a good performance (even if it is only a simply model) and gave us the first path to find the best variables for our model.

7 CONCLUSION

The results of our analysis have suggested that the most important variables that help to the formations of the cost of an insurance claim, are *Bareme_table_code_other_driver* which consists of the description of the accident, *Region*, *AnnodiImmatricolazione* and other variables relied to the vehicles characteristics.

As we focus on powerful predictive models we loose part of the interpretability of the model, but, thanks to our exploration phase and in-depth study of the variables and their relationship with the response, we are able to have a better understanding of the problem.