Homework 2: Car Insurance Prediction

Simone Luzi *782201*

The dataset used in this study comprises a total of 19 variables, including 11 numeric and 8 categorical ones. No duplicates are found in the dataset; however, missing values are identified in the "Credit Score" and "Annual Mileage" columns. A detailed graphical representation of each variable was provided as part of the preliminary analysis. Examination of the correlation matrix did not reveal significant relationships between the variables; however, a more in-depth analysis uncovered relevant insights; for instance, observations related to individuals aged between 16 and 25 consistently showed an incidence of accidents equal to zero. Additionally, it was observed that with an increase in the number of observations, cases of speeding incidents rose, while instances of driving under the influence remained relatively constant.

Subsequently, missing values were addressed. An analysis of the distribution of variables with missing values and outliers was conducted to determine the most suitable method for imputing these values. The missing fields were then populated with the median values of the respective variables, as both were non-normally distributed and featured outliers. The choice of median was dictated by its lesser sensitivity to such characteristics.

Categorical variables were transformed into numerical forms using ordinal encoding for "orderable" variables and one-hot encoding for "non-orderable" variables (the latter were all binary).

Following that, the model development stage ensued. Logistic regression was preferred over other models as it adapts well to binary classification problems and is computationally efficient. Data were standardized to avoid scalability issues; the variable "Outcome" was defined as the dependent variable (y), while all other variables (excluding "ID" as it is non-informative) were considered independent input variables. The dataset was split into a training set (80%) and a test set (20%). After model training, predictions were calculated, yielding satisfactory results in terms of accuracy, precision, recall, F1-score, and MCC. Finally, an illustrative confusion matrix was created.

The process was repeated with hyperparameter optimization; however, both the initial results and those after optimization demonstrated substantial stability in the model evaluation metrics. The confusion matrix, accuracy, precision, recall, and F1-score remained almost unchanged. Despite hyperparameter optimization, the model appears to maintain similar overall performance, suggesting that the initial configuration may already be close to optimal for the given dataset. Further analysis may be required to identify potential avenues for improvement.