April 21, 2024 Industrial data mining - IND 6212

Polytechnique Montréal

Yassine Azrou 2265579
David De Blas 2003038
Yannis Redjah 1748777

PREDICTION OF THE NUMBER OF ROAD ACCIDENT VICTIMS IN QUEBEC

INTRODUCTION

Both summer and winter present challenges for drivers in Quebec, requiring heightened caution on the roads. Despite an overall reduction in road accidents in 2022 compared to the previous year, there has been a 13.2% increase in fatal accidents compared to the average from 2017 to 2021.

Using datasets available on the Données Québec website, we selected a dataset provided by the Société de l'assurance automobile du Québec (SAAQ), which is openly accessible. The most recent dataset, corresponding to accident reports from 2022, serves as the basis for this analysis.

This report aims to analyze the dataset and explore the variables it contains. Following this, we will employ a statistical model to predict the number of victims based on explanatory variables. Finally, the results of the model will be evaluated to assess its predictive capabilities, and we will discuss the limitations of the approach.

DESCRIPTION DES DONNÉES

The dataset contains 108,185 accident records, each with 25 variables describing the incident, such as the year of occurrence, time of day, location details, severity, and the number of victims involved.

Table 1 : Variables Description

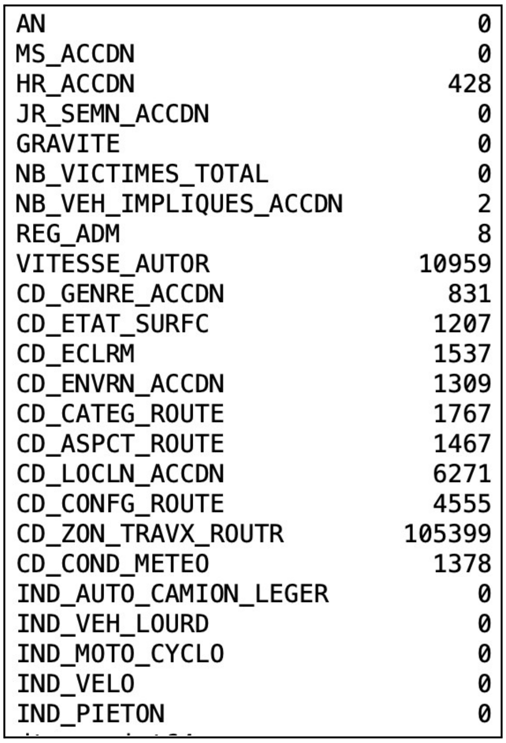
Nom de la variable	Description	Type
AN	Year of the accident	Numeric
NO_SEQ_COLL	Unique accident identifier	Categorical
MS_ACCDN	Month of the accident (coded 1 to 12)	Categorical
HR_ACCDN	Four-hour interval of accident occurrence	Categorical
JR_SEMN_ACCDN	Day of the week (weekday or weekend)	Categorical
GRAVITE	Severity of the accident (fatal, serious, or property damage)	Categorical
NB_VICTIMES_TOTAL	Total number of victims	Numeric
NB_VEH_IMPLIQUES_ACCDN	Number of vehicles involved	Numeric
REG_ADM	Administrative region of Quebec	Categorical
VITESSE_AUTOR	Speed limit at the accident location	Categorical
CD_GENRE_ACCDN	Collision type (vehicle, pedestrian, cyclist, etc.)	Categorical

CD_ETAT_SURFC	Road surface condition (dry, wet, etc.)	Categorical
CD_ECLRM	Lighting conditions during the accident	Categorical
CD_ENVRN_ACCDN	Surrounding environment type (school, residential, etc.)	Categorical
CD_CATEG_ROUTE	Road aspect at the site (straight, curved, etc.)	Categorical
CD_ASPCT_ROUTE	Longitudinal position of the accident	Categorical
CD_LOCLN_ACCDN	Position longitudinale de l'accident (intersection, entre intersections, etc.)	Categorical
CD_CONFG_ROUTE	Road configuration	Categorical
CD_ZON_TRAVX_ROUTR	Construction zone presence	Categorical
CD_COND_METEO	Weather conditions	Categorical
IND_AUTO_CAMION_LEGER	Presence of a light vehicle or truck (Yes/No)	Binary
IND_VEH_LOURD	Presence of a heavy vehicle (Yes/No)	Binary
IND_MOTO_CYCLO	Presence of a motorcycle or moped (Yes/No)	Binary
IND_VELO	Presence of a bicycle (Yes/No)	Binary
IND_PIETON	Presence of a pedestrian (Yes/No)	Binary

Data Preparation and Formatting

Before beginning the data analysis, it was essential to prepare and format our database to facilitate processing. First, we removed all information contained in the columns **AN** and NO_**SEQ**_COLL, as they added no value to our predictive model. Additionally, we observed that the **CD_ZON_TRAVX_ROUTR** variable had a significant amount of missing data. While construction zones impact traffic flow, they can also influence other variables in our dataset, such as **VITESSE_AUTOR**. Considering these complexities and for the sake of simplicity, we decided to exclude this variable from the analysis. Furthermore, some variables, such as **VITESSE_AUTOR**, also contained missing data for certain accidents. To maintain precision, we opted to delete all rows with missing data. This step reduced the sample size by approximately 18%. While this reduction is substantial, the dataset remains large enough to build a robust model.

Figure 1: Nombre de données manquantes



```
AN                              0
MS_ACCDN                        0
HR_ACCDN                      428
JR_SEMN_ACCDN                   0
GRAVITE                         0
NB_VICTIMES_TOTAL               0
NB_VEH_IMPLIQUES_ACCDN          2
REG_ADM                         8
VITESSE_AUTOR               10959
CD_GENRE_ACCDN                831
CD_ETAT_SURFC                1207
CD_ECLRM                     1537
CD_ENVRN_ACCDN               1309
CD_CATEG_ROUTE               1767
CD_ASPCT_ROUTE               1467
CD_LOCLN_ACCDN               6271
CD_CONFG_ROUTE               4555
CD_ZON_TRAVX_ROUTR         105399
CD_COND_METEO                1378
IND_AUTO_CAMION_LEGER           0
IND_VEH_LOURD                   0
IND_MOTO_CYCLO                  0
IND_VELO                        0
IND_PIETON                      0
```

To ensure consistency across the dataset, it was crucial that all variables followed the same format. Therefore, we transformed all categorical variables into numerical ones so that our model could interpret them correctly. To achieve this, we grouped all categorical variables and applied one-hot encoding to each. One-hot encoding generates a binary column for each unique value of a given variable. For instance, the **CD_GENRE_ACCDN** variable, which categorizes the type of accident, illustrates the necessity of this transformation. For example, a collision with a stationary object does not have the same human or material impact as a multi-vehicle collision. *One-hot encoding* allowed us to treat these categories independently and equitably in our analysis. For binary variables, we transformed them into Boolean values so the model could process them effectively (instead of representing the states as strings like ‘yes’ or ‘no,’ we now represent them as 1 or 0).

After standardizing the format of our data, we divided it into a training set and a testing set. We used a traditional split, allocating 80% of the data to training and the remaining 20% to testing. The training set was used for learning, identifying trends and correlations inherent in our data. The testing set, on the other hand, was used to evaluate the model's performance on unseen data. This step is crucial because it helps us determine whether the model is adequate and ensures that its performance is not merely the result of overfitting the training data. In summary, we aim to reliably identify the model’s predictive value and, if necessary, make adjustments to improve its accuracy.

PREDICTIVE MODEL: RANDOM FORESTS

We selected the Random Forest model to predict the number of victims in road accidents in Quebec. The choice of this model was based on the nature and type of data in our dataset, which includes both numerical and categorical variables. Random Forests are an excellent model for handling such data structures. Moreover, this method is particularly effective for analyzing large datasets, both in terms of the number of variables and the volume of observations, which aligns with the configuration of our dataset.

The advantages of Random Forests extend beyond their ability to manage diverse types of data. They are also valued for their fast-training speed and their efficiency in performing parallel computations, which is especially advantageous when conducting an exhaustive search for hyperparameters—a process that can be computationally expensive in terms of time and resources. Additionally, Random Forests are robust against irrelevant variables, the challenges posed by multicollinearity, and overfitting, even with large datasets. These characteristics justify our choice of this model to address the problem at hand.

To refine our model, improve its performance metrics, and reduce the likelihood of overfitting, we employed a technique called **Grid Search**. This method systematically explores combinations of a wide range of hyperparameters to identify the configuration that offers the highest precision. However, relying solely on a single division of training and validation data during this process could lead to misleading conclusions. Therefore, we combined Grid Search with cross-validation.

Cross-validation allows the dataset to be partitioned into multiple subsets. In our case, we divided the data into three subsets (3-fold cross-validation). The model was trained on every possible combination of these subsets and tested on the remaining subset.

The hyperparameters studied include:

- **max_depth**: Maximum depth of the trees.
- **min_samples_split**: Minimum number of samples required to split a node.
- **min_samples_leaf**: Minimum number of samples required in a leaf node.

Through this process, we aimed to identify a combination that minimizes overfitting while maximizing the model's performance.

ANALYSIS OF RESULTS

The evaluation of our Random Forest model on the training and testing datasets produced encouraging results. Below is a summary of the observed performance:

- **Mean Squared Error (MSE) on the training set:** 0.0543
- **Mean Squared Error (MSE) on the testing set:** 0.0589
- **R^2 Score on the training set:** 0.8439
- **R^2 Score on the testing set:** 0.8275

These metrics indicate that the model makes accurate predictions, as the MSE values are relatively low.

This observation is further supported by the R^2 values, showing that approximately 80% of the variability in the response variable can be explained by the explanatory variables.

Focusing on the performance indicators for the testing set, we observe the following results:

- **Mean Squared Error (MSE):** 0.0589
- **Root Mean Squared Error (RMSE):** 0.2428
- **Mean Absolute Error (MAE):** 0.0803
- **R^2 Score:** 0.8275

The RMSE provides insight into the standard deviation of residuals, or the precision of predictions for the number of victims per accident. The RMSE of approximately 0.25, given that predicted values range from 0 to 3, suggests that the model's predictions deviate on average by about a quarter of the possible value range. The R^2 score of approximately 83% means that the model's predictions closely align with actual outcomes. In other words, 83% of the observed variability in the real results can be explained by the model. These results are satisfactory regarding the overall accuracy of the model used to predict the number of victims on Quebec's roads.

We now compare these results to those of a baseline model, which systematically predicts the average number of victims observed in the training data. The baseline model yields the following results:

- **Predicted mean value: 0.2898**
- **Mean Squared Error (MSE): 0.3418**
- **Root Mean Squared Error (RMSE): 0.5847**
- **R^2 Score: 0.00034**

The analysis of these metrics reveals that our *Random Forest model* significantly outperforms the baseline model. While the baseline model barely surpasses random predictions (with an R^2 score close to 0), our predictive model shows much lower MSE and RMSE values, suggesting far more accurate and reliable predictions. Additionally, the baseline model produces an R^2 score close to 0, even slightly negative. This highlights its poor performance compared to the Random Forest model, which provides robust and precise predictions.

CRITIQUE OF RESULTS AND LIMITATIONS

The modeling results are promising, with an R^2 score approaching 80%, indicating that the model successfully explains a significant portion of the variability in the number of victims based on the defined factors. A low RMSE further suggests that the model's predictions are generally accurate, with minimal deviations from actual values. However, despite these excellent results, there remains room for improvement. The hyperparameters selected during the grid search represent only a fraction of the many possibilities. Expanding the hyperparameter search could potentially enhance the model's precision.

Furthermore, while Random Forest was chosen for its versatility and robustness, it is important to acknowledge some of its drawbacks. Random Forest models can be complex and difficult to interpret. This complexity directly impacts computation time when obtaining results. Additionally, preprocessing is necessary to handle missing data effectively before working with a Random Forest model. Random Forests are also highly sensitive to parameters such as the number of trees and tree depth. Striking the right balance in these parameters could be critical to the model's performance.

It is also worth noting that alternative methods, such as neural networks or Bayesian models, might have yielded better results than those currently achieved. While these approaches may require longer training times or different preprocessing steps, they could potentially lead to performance improvements. For example, one key factor for obtaining satisfactory results with these two methods is reducing multicollinearity between variables, as these methods are more sensitive to internal correlations that can disrupt the learning process and the validity of predictions. In contrast, Random Forests are much more robust to this issue, primarily due to their random variable selection mechanism during the creation of individual trees.

Additionally, incorporating more explanatory variables, such as driver age, could have enriched our predictive model, as age is a factor often correlated with the risk of road accidents. Statistics show that young drivers are generally more involved in accidents than other age groups. By integrating age as a variable, we could have potentially refined our predictions and gained a better understanding of the underlying dynamics of road accidents.

CONCLUSION

The modeling of the number of victims in road accidents in Quebec using Random Forests yielded promising results. The model demonstrates satisfactory accuracy, with an R^2 score close to 80% and a relatively low RMSE. Ultimately, while Random Forests are an excellent choice for addressing this problem, other approaches might have been more appropriate in certain cases. A more extensive search of hyperparameters and the incorporation of additional data could further improve the model's precision and deepen our understanding of the factors influencing the number of victims in road accidents in Quebec.