Capstone Project – Predicting accidents with victims on Brazilian federal roads

Udacity - MACHINE LEARNING ENGINEER NANODEGREE

Carlos Eduardo de Souza
São Paulo, Brazil
https://github.com/SouzaCadu
https://www.linkedin.com/in/carlos-eduardo-de-souza/?locale=en_US

June 28, 2020

ABSTRACT

This project analyses accident data on Brazilian federal roads, collected by the Federal Highway Police, from 2017 to 2020 to develop a machine learning able to predict multi-class accident type. This dataset contains accidents recorded along the more than 70,000 km of monitored highways, every day of the year; also, are recorded the weather conditions, runway conditions, type of accident, number of people involved and victims, among other information. Will be performed data cleaning, treatment of categorical variables, exploratory analysis and selection of the information that will go through the model to predict if given these characteristics an accident result in "no victims", "injured victims" or "fatal victims". A regression model will be defined to estimate the probability of accidents occurring to victims. This information can assist in planning the distribution of checkpoints, improving road actions or triggering an alert system so that drivers drive more carefully.

Keywords: Exploratory Data Analysis · Supervised Learning

Definition

Project Overview

In Brazil, the large number of deaths on federal highways is alarming, in the period 2010 - 2017 there were 62,120 deaths, with an average of 21 per day, almost one victim per hour; in addition to 201,006 seriously injured and 578,954 lightly injured as a result of 2,392,205 accidents.

In this project, will be analyzed the conditions in which accidents on Brazilian roads took place between 2017 and 2020, to build a model capable of predicting whether in a given accident there will be fatal victims, injured or no victims. The set of information that describes this problem contains variables about the road conditions, climatic conditions, number of passengers, number

of vehicles, cause of the accident, information about the roads and observed damages, these data were obtained from the Federal Highway Police website.

There are four data files associated with this project:

- datatran2017.csv – Traffic accident data on Brazilian federal roads in 2017 grouped by occurrences; 89.563 (accidents) x 30 features
- datatran2018.csv – Traffic accident data on Brazilian federal roads in 2018 grouped by occurrences; 69.295 (accidents) x 30 features
- datatran2019.csv – Traffic accident data on Brazilian federal roads in 2019 grouped by occurrences; 67.446 (accidents) x 30 features
- datatran2020.csv – Traffic accident data on Brazilian federal roads in 2020 grouped by occurrences; 19.573 (accidents) x 30 features

Problem statement

The objective of this project is to predict based on road conditions, weather conditions, number of vehicles involved, cause of the accident, on the highways where the accidents occurred, whether the result will be an accident without victims, with injured victims of fatal victims.

If it is possible to predict this result with a certain degree of confidence, the model can be used as a support tool in the management of accident prevention resources, such as determining inspection points, distributing rescue teams, and warning systems for drivers.

This project is, therefore, a multi-class classification problem and fits within the supervised learning group. Several machine learning algorithms can be used in this task; for this project, we will focus on the linear regression model, as a baseline. From the observed results, we will seek to improve the prediction with a more complex model, the XGBoost.

Metrics

For this project, the metric used to measure the quality of the model will be Precision. The reason for choosing this metric is to minimize the type I error, which occurs when the model predicts an event as accurate while the observation is false.

This type of error is captured by Precision, given its formulation:

$$Precision = {}^{True\ Positive}/_{(True\ Positive + False\ Positive)}$$

In other words, Precision talks about how precise/accurate the model is out of those predicted positive, how many of them are positive.

For this problem, it is particularly important to correctly predict actual events, accidents than with victims, as the costs associated with false positives are quite high. One cannot calculate what the loss of human life represents or the psychological damage and traumatic stresses to which traffic victims and their families are subjected after events of this nature. On the other hand, there is also the formation of economic and financial costs that directly impact families, as well as society in general, and that can be estimated using specific calculation methodologies.

The details of this methodology can be found in the document ***Impactos sociais e econômicos dos acidentes de trânsito nas rodovias brasileiras*** (Social and economic impacts of traffic accidents on Brazilian highways), prepared by competent bodies of the Brazilian government to assist in the elaboration of public policies on the subject.

Two other metrics will be calculated as support and measures of comparability with other studies.

$$Recall = {True\ Positive}/{(True\ Positive + False\ Negative)}$$

$$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1\ (\hat{y}_i = y_i)$$

Finally, the confusion matrix with the three classes will be analyzed to help interpret the results and identify opportunities for improvement.

Analysis

Data Exploration

A brief description of these features is provided below:

Variable	Description
id	Variable with numerical values, representing the accident identifier.
data_inversa	Date of occurrence in dd / mm / yyyy format.
dia_semana	Day of the week of the occurrence.
horario	Time of occurrence in hh: mm: ss format.
uf	Federation Unit.
br	Variable with numerical values, representing the BR identifier of the accident.
km	Identification of the kilometre where the

	accident occurred, with a minimum value of 0.1 km and with a decimal point separated by a point.
município	Name of the municipality where the accident occurred.
causa_acidente	Identification of the main cause of the accident. In this data set, accidents with the main cause variable equal "No" are excluded.
tipo_acidente	Identification of the type of accident. E.g., frontal collision, lane departure, etc. In this data set, types of accidents with an order greater than or equal to two are excluded. The order of the accident shows the chronological sequence of the types present in the same occurrence.
classificação_acidente	Classification according to the severity of the accident: Without Victims, With Injured Victims, With Fatal Victims and Ignored.
fase_dia	The phase of the day at the time of the accident.
sentido_via	The direction of the road considering the collision point: Ascending and decreasing
condição_meteorologica	Meteorological condition at the time of the accident.
tipo_pista	Track type considering the number of tracks
tracado_via	Description of the route layout.
uso_solo	DescAbouofbout the characteristics of the accident site: Urban = Yes; Rural = No.
latitude	Latitude of the accident site in decimal geodetic format.
longitude	Longitude of the accident site in decimal geodetic format.
pessoas	Total people involved in the incident.
mortos	Total dead people involved in the occurrence.
feridos_leves	The total number of people with minor injuries involved in the incident.
feridos_graves	Total of people with serious injuries involved in the occurrence.
feridos_total	Total number of injured persons involved in the incident (this is the sum of the light and the serious injured).
ilesos	Total unscathed people involved in the incident.
ignorados	Total of people involved in the occurrence and the physical state was not known.
veiculos	Total vehicles involved in the occurrence.

Data for the years 2017 to 2020 were used, with a total of 245.877 records. Originally the result of the accident was recorded in the variable "classification of the accident", with three categories "With Fatal Victims", transformed into "0", with "With Injured Victims", turned into "1" and with "Without Victims", transformed in 2". The new values were stored in the "target" variable.

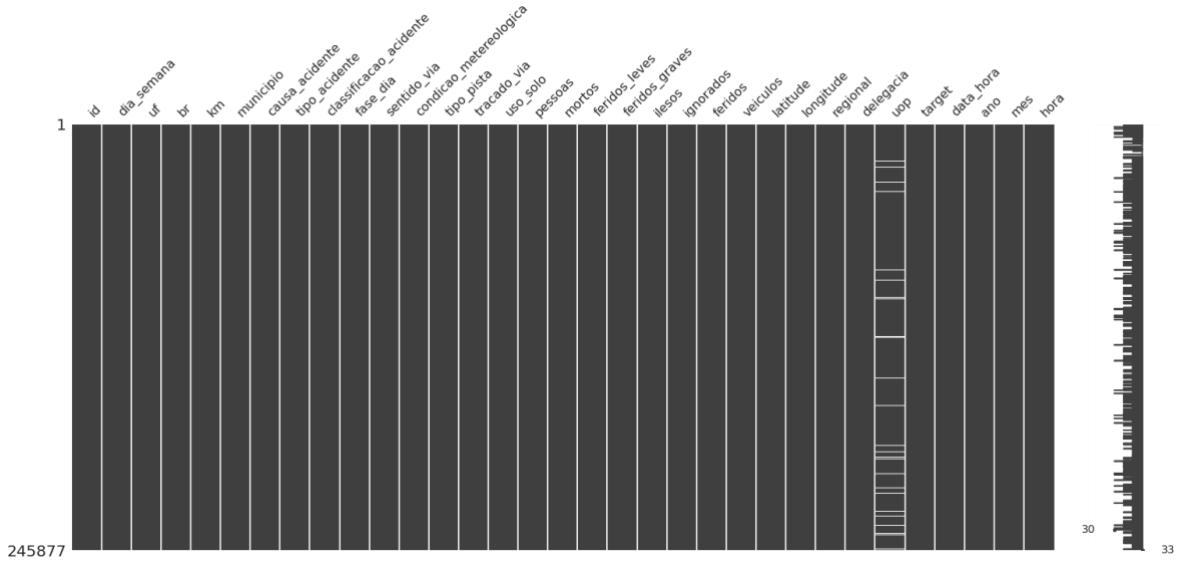


Figure 1 - Distribution of missing data before cleaning

The lines without information in the variables "uf" and "br" were removed. The columns with the variables "id", "km", "municipio", "regional", "delegacia", "uop", were also removed, "uop" due to the amount of missing data and the others for not bringing relevant information to the problem.

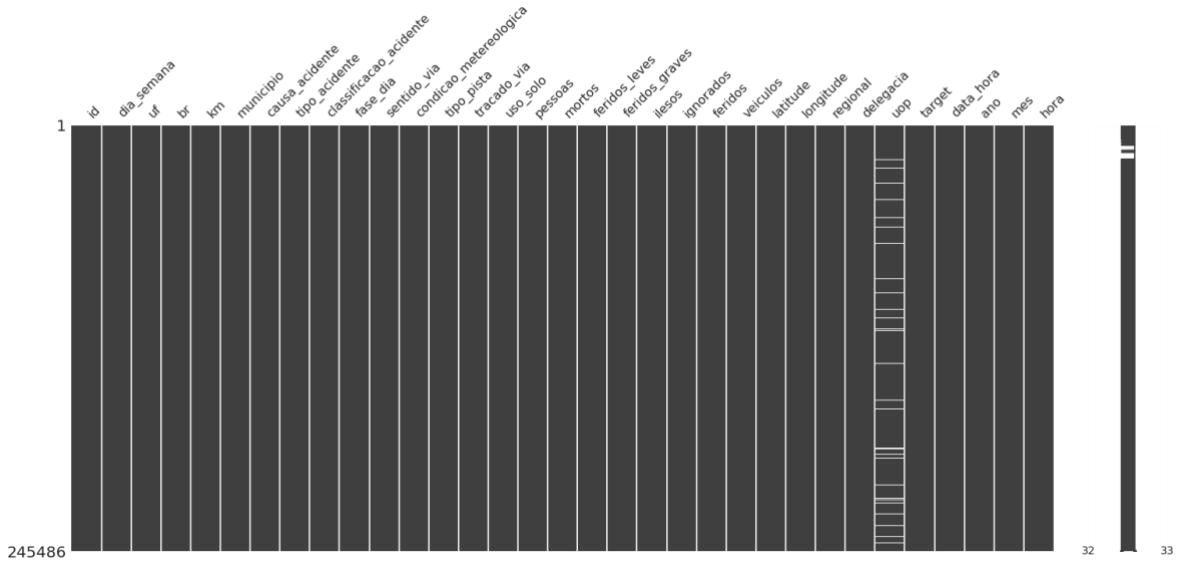


Figure 2 - Distribution of missing data after cleaning

Exploratory Visualization

After this cleaning, the domain descriptions of the variables "causa_acidente", "tipo_acidente", "condicao_metereologica" were analyzed. In the variable "tipo_acidente", some descriptions are variations of the same type of accident and, therefore, were grouped to reduce the dimensionality of this variable. In the others, the number of domains was considered acceptable, and no adjustment was made.

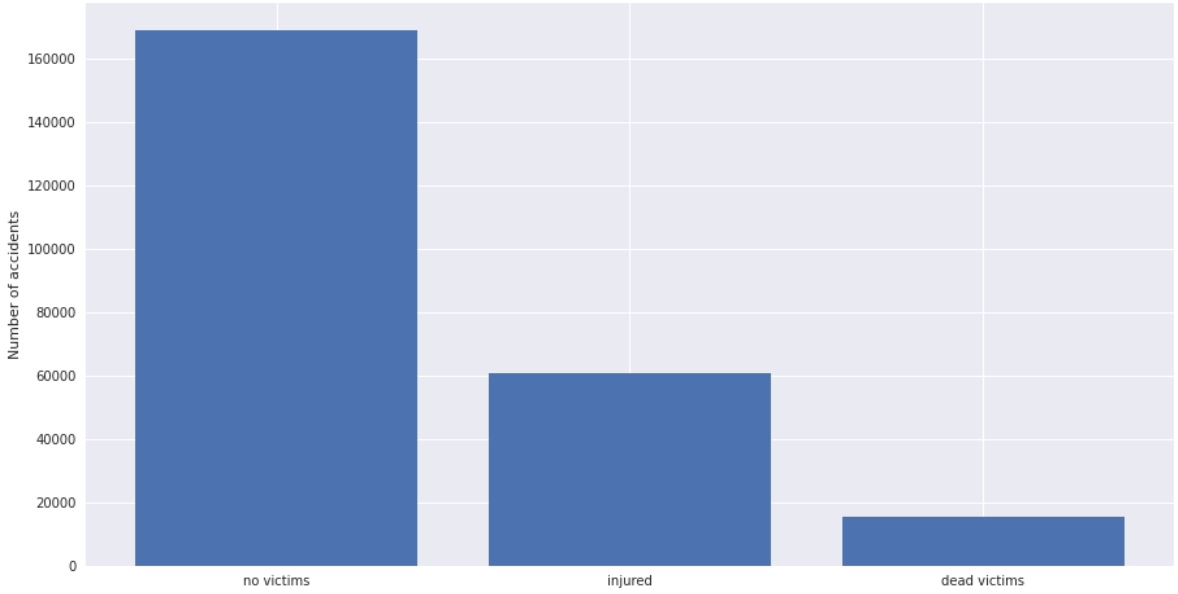


Figure 3 - Target variable distribution

The first analysis of the "target" variable showed a significant imbalance between classes. The class "without victims" has approximately 65% of the records, "with victims" almost 24% and "fatal victims" are less than 1%.

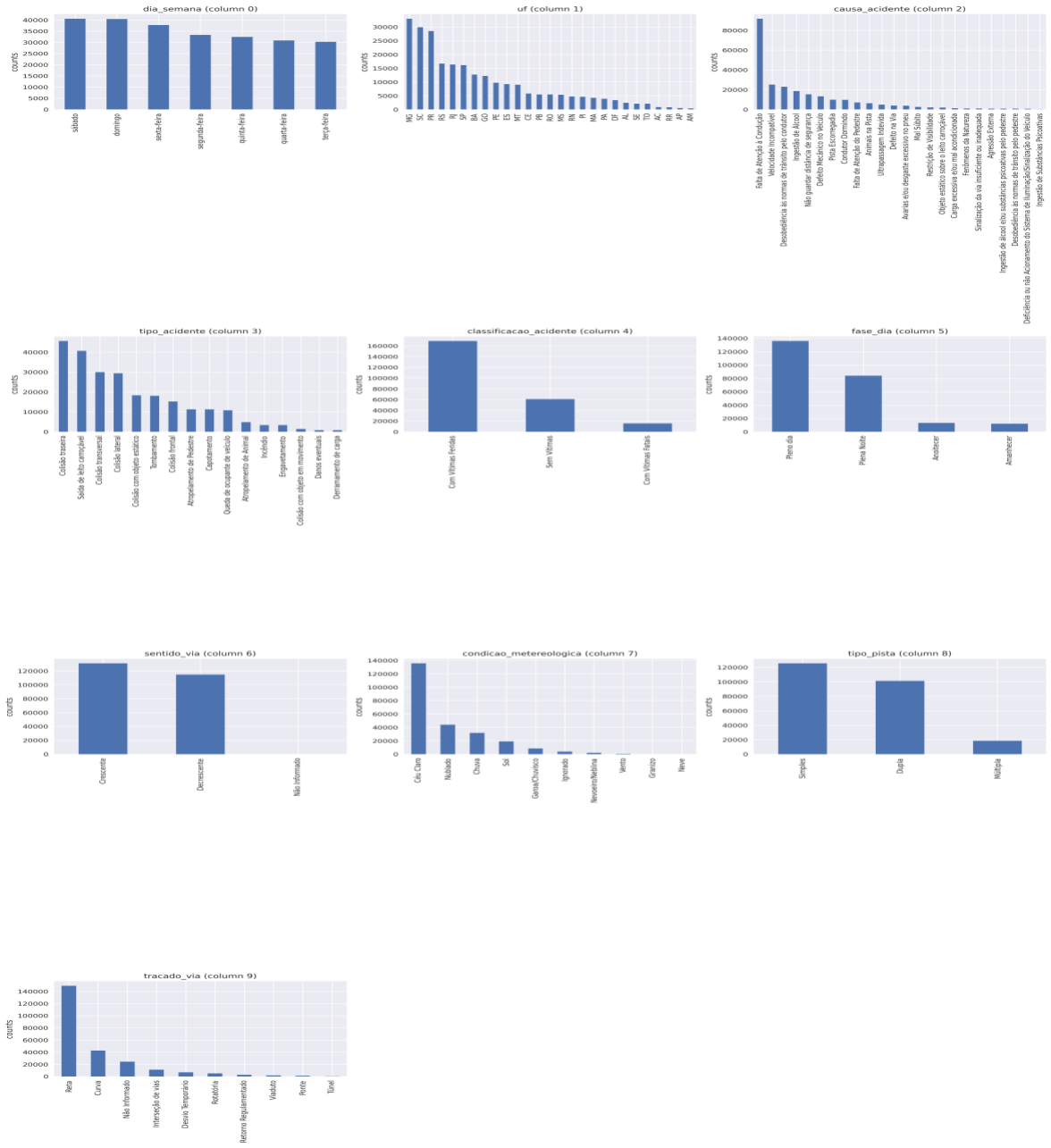


Figure 4 - Grouped graph of the first 10 variables

A graph was made with the first ten variables from the base to start the exploratory analysis. On this first set of graphics, the concentration of accidents between Friday and Sunday stands out. Three states have a high frequency of accidents. The lack of attention is identified as the main responsible for accidents. Of the five main types of crash collision appears in 4 positions, the largest of which is the rear collision and the first among all classes. It is also noteworthy that most accidents happen during the day, on straight lines on a single track.

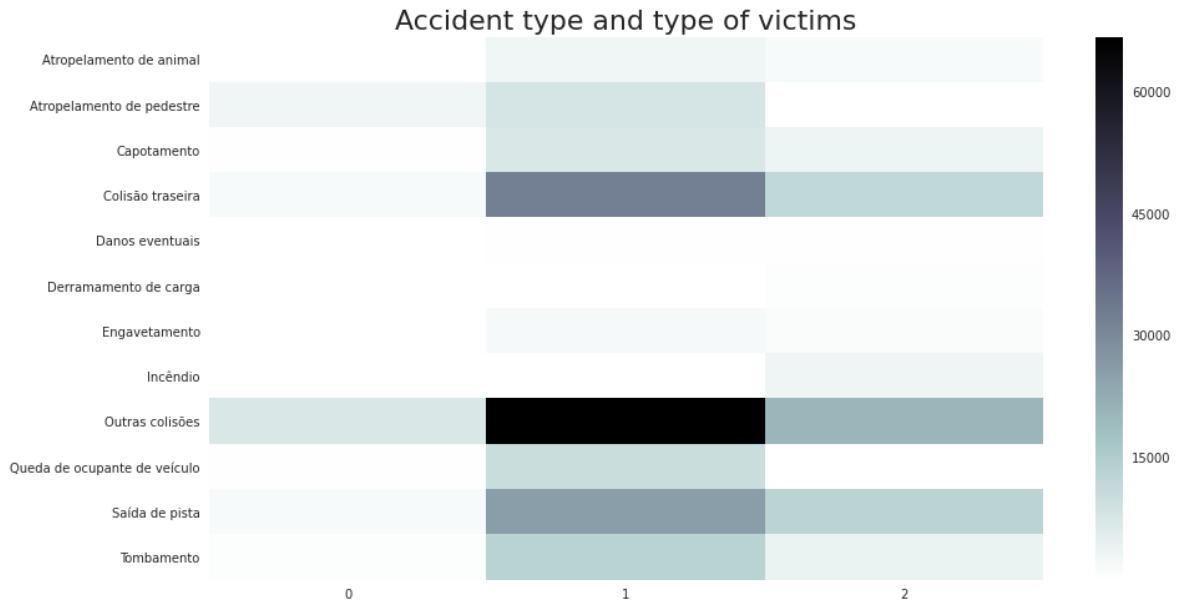


Figure 5 - Accident type and type of victims

The accident types "Collision", "Falling vehicle occupant", "Exit from the road" and "Tumbling" concentrate the most significant number of injuries. "Collisions" also appear as the most prominent type of accident with fatalities.

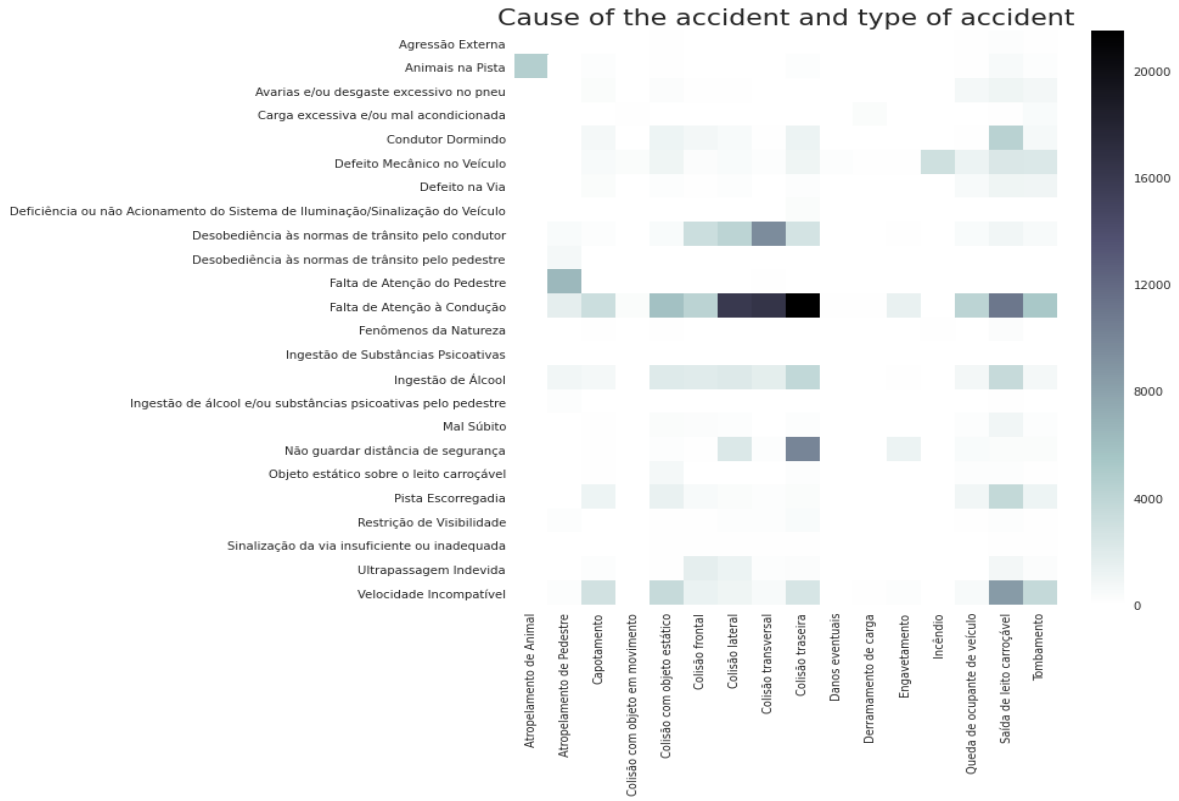


Figure 6 - Cause of the accident and type of accident

The driver's lack of attention appears to be responsible for almost all types of accidents. Also noteworthy is the lack of awareness of driving rules as the cause of most collisions.

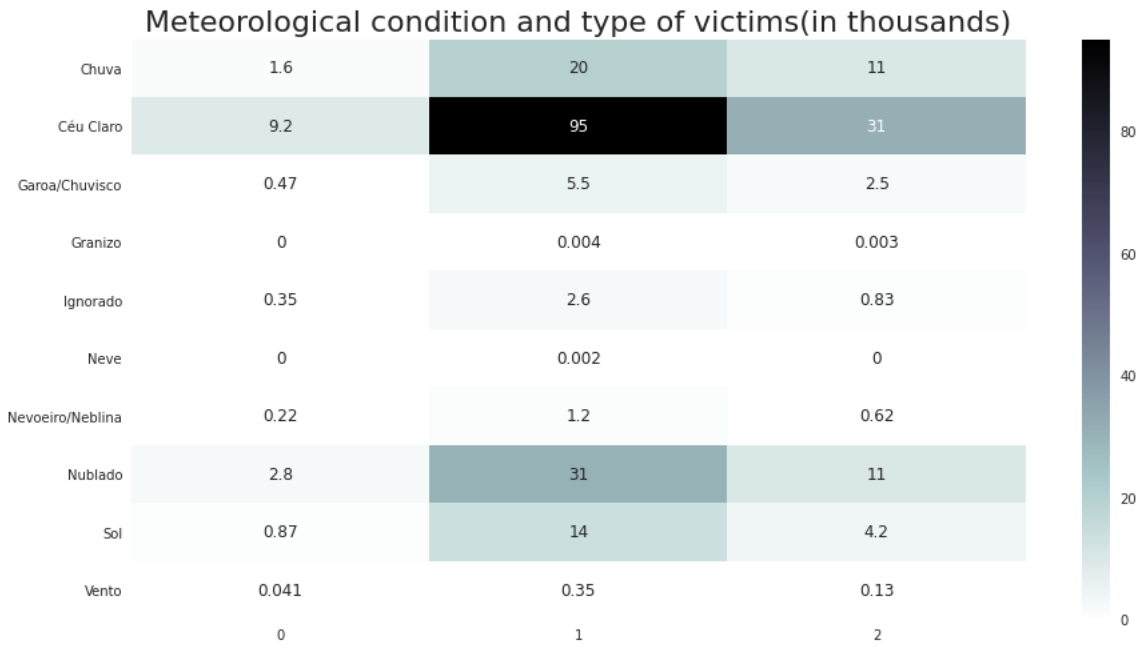


Figure 7 - Meteorological condition and type of victims (in thousands)

This graph shows that most accidents happen in clear sky conditions, which can be considered an expected result since Brazil is a tropical country.

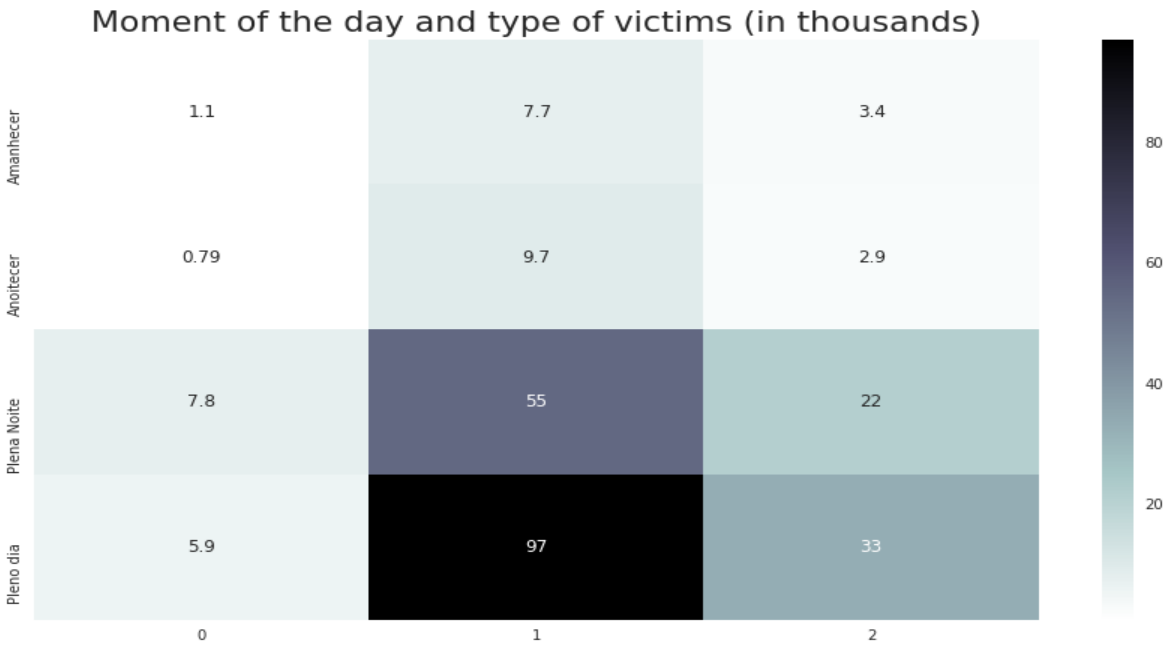


Figure 8 - Moment of the day and type of victims (in thousands)

Most accidents involving injured or fatal victims occur in broad daylight, followed by full night, with a relatively small number of cases remaining at dawn and dusk.

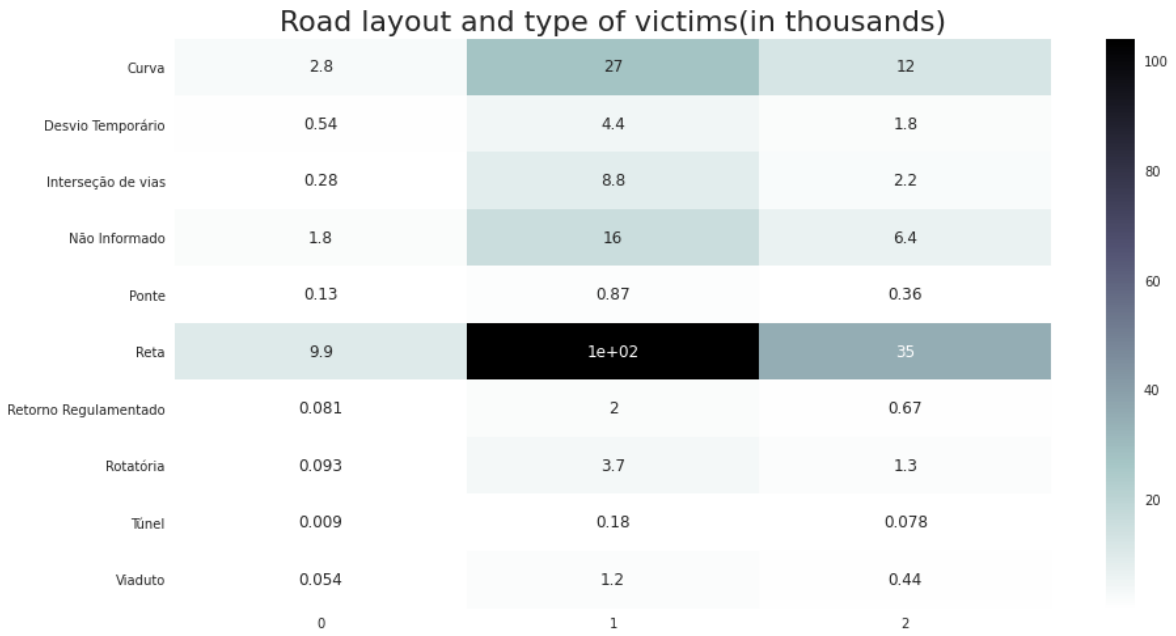


Figure 9 - Road layout and type of victims (in thousands)

The straight sections concentrate the majority of accidents in all categories.

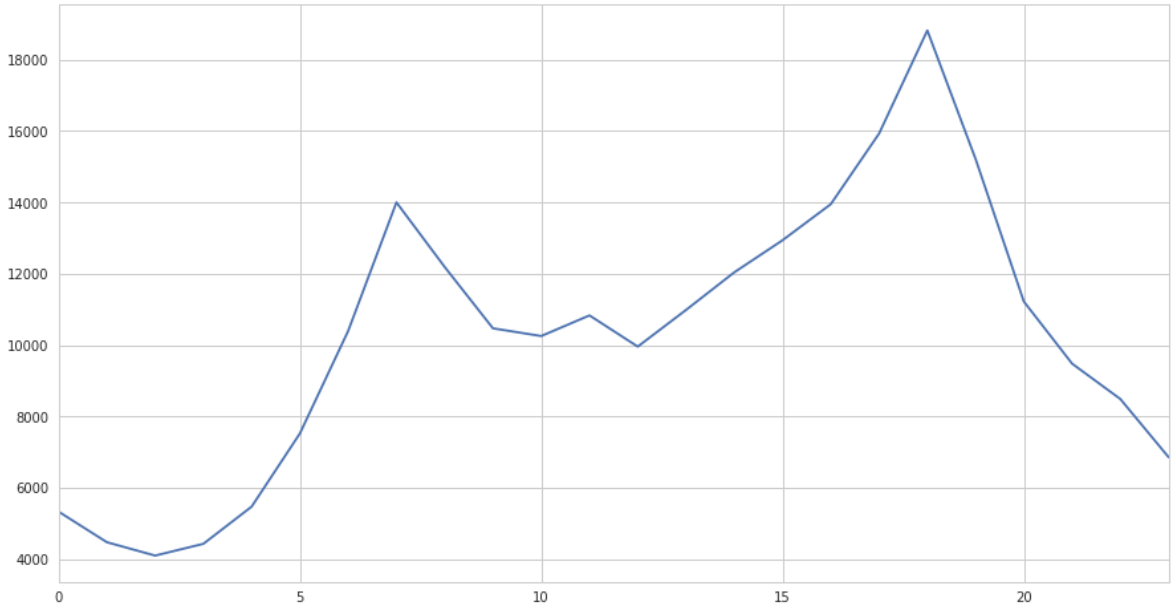


Figure 10 - Number of accidents per hour

Accidents are concentrated between 5, and 10 am, with 140 thousand cases, and between 15 and 20 at night, with just over 180 thousand cases.

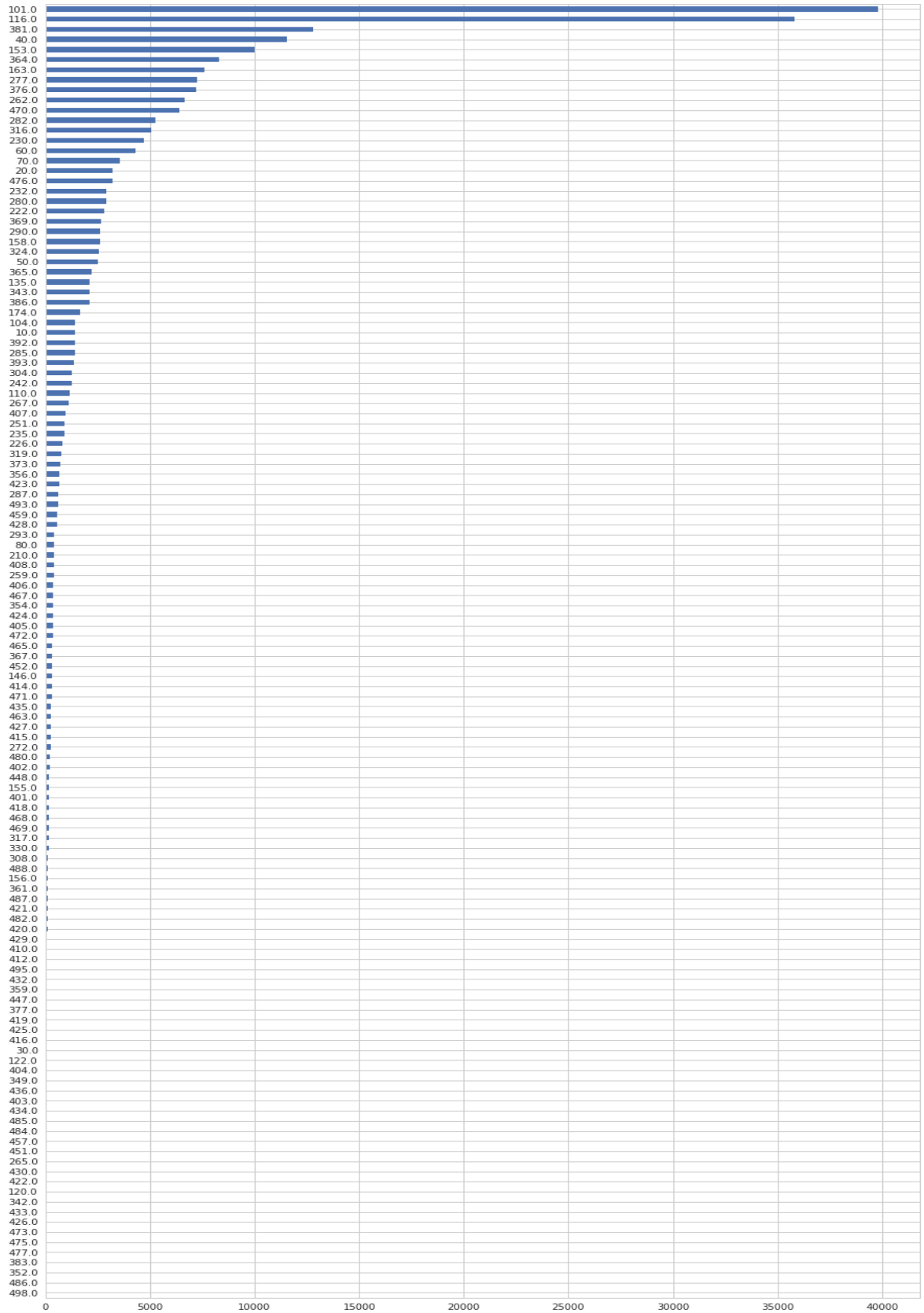


Figure 11 - Number of accidents recorded on each highway

Of the 127 highways monitored by the Federal Highway Police, two highways, BR-101 and BR-116 concentrate approximately 30% of the total accidents. This behaviour can be attributed to the fact that these two highways connect the main capitals of the country.

BR-101 is a Brazilian longitudinal highway that begins in the municipality of Touros, in the state of Rio Grande do Norte, and ends in São José do Norte, in Rio Grande do Sul. Alongside BR-116, it is one of the main road axles in the country with 4 650 km long (https://pt.wikipedia.org/wiki/BR-101).

In turn, the BR-116 passes through ten states, connecting important cities such as Porto Alegre, Curitiba, São Paulo, Rio de Janeiro and Fortaleza. The road is duplicated in the metropolitan areas, in addition to being wholly duplicated between Curitiba and Rio de Janeiro, after the completion of the Serra do Cafezal stretch, on the Régis Bittencourt Highway (https://pt.wikipedia.org/wiki/BR-116).

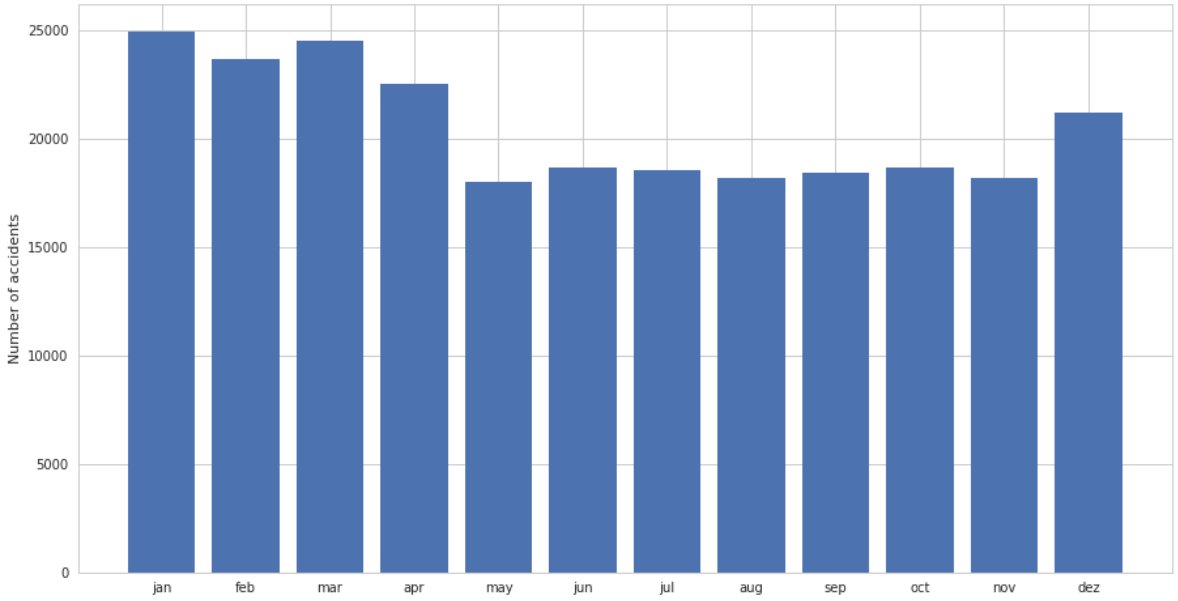


Figure 12 - Number of accidents each month

Regarding the number of accidents, they start to increase in December and grow over January, February and March, this period coincides with the Brazilian summer, vacation period and carnival. Both holidays and the carnival holiday tend to show great movement on the roads.

Feature engineering

After analyzing the base, I identified some opportunities to add information that can increase the performance of the model.

The first is to explain the dates that correspond to national holidays, where there is a more excellent circulation of people. The second is to make the months of January, July and December explicit as holiday months. January and December for the summer holidays and July for being a

month of school holidays. The third opportunity was to concentrate the information of the ten highways with more accidents on dummies-type variables. However, this represents a loss of data, using the same strategy with all 127 roads would generate an extremely sparse vector that would impair the model's performance due to the high dimensionality of that information only.

Finally, the land use information was encoded in a single variable, with the domain "Yes" for urban stretch and "No" for the rural stretch, and this variable was transformed into a dummy variable considering urban stretch as "1 "and rural as" 0 ".

Algorithms and Techniques

In this project, two models were used to classify accidents according to the target variable. Both models used the native AWS SageMaker implementation. The initial model was the LinearRegression configured for the prediction of the three categories, optimizing the Precision Score.

The second model used is the XGBoost configured for the multi-class classification problem. In the first training phase, in addition to this configuration, reference values will be used for the other hyperparameters. Finally, hyperparameters will be optimized for the problem through the adjustment process already implemented for AWS SageMaker.

Benchmark

Model evaluation is the process of objectively measuring how well machine learning models perform the specific tasks they were designed to do.

We use the Precision Score to benchmark the performance of the models. A model with the highest Precision is considered as the best performer. We train a LogisticRegression model with default parameters and use it to predict the classes for the train and test datasets. The obtained Precision score is our base score used for comparison.

As mentioned in the proposal for this project, three other pieces of information, Recall Score, Accuracy and Matrix of Confusion, will be calculated for comparison with other studies on road accidents considered as references for this project.

Methodology

Data Preprocessing

Following the analysis of the data, two main stages of preparation were performed before modelling.

When importing the clean and analyzed data, variables with the number of victims were left out, to prevent information from leaking into the model, the information with the name of the roads where the accidents occurred, the day and time of the accidents.

Both the information with the name of the highway, as well as the day and time will be captured by other variables that were already in the base, such as "fase_dia", "holiday", "vacation" and the variables corresponding to the ten roads with more accidents.

As the base has several categorical variables, the first main stage of pre-processing is to transform them into dummy type variables.

In the sequence, the data referring to the year 2020 were separated, for the validation of the model.

Finally, the SMOTE method was used to balance the data through class oversampling; formerly the "target" variable showed a significant imbalance between levels. The category "without victims" has approximately 65% of the records, "with victims" almost 24% and "fatal victims" are less than 1%. After this process, all classes started to be equally represented at the base.

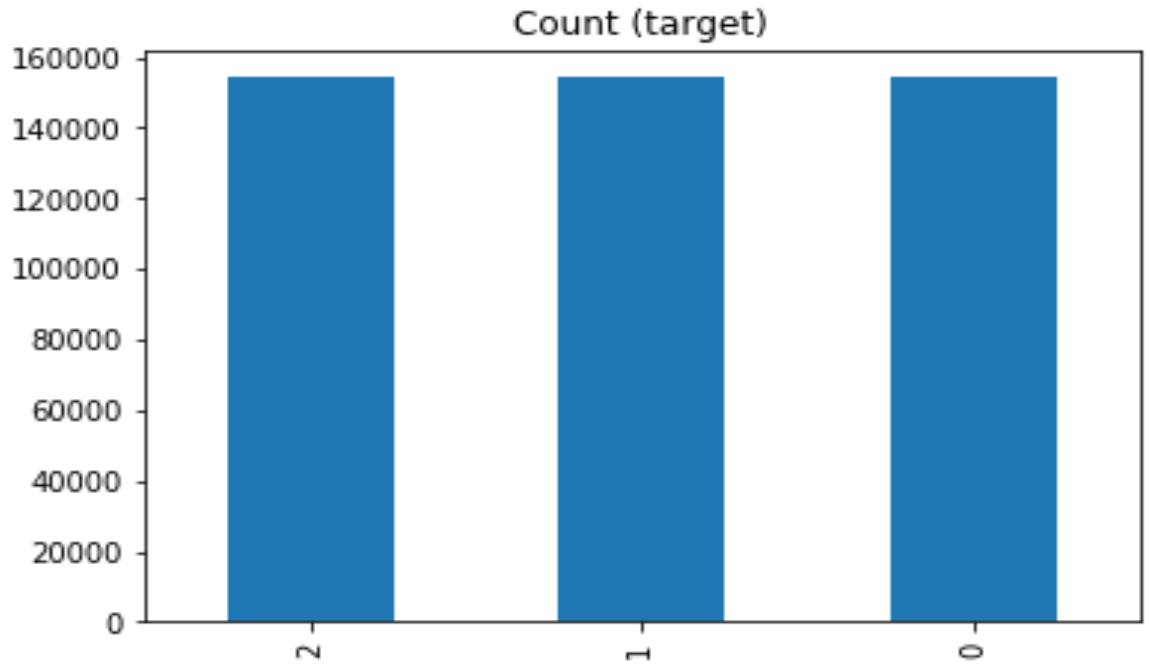


Figure 13 - Target variable after resampling process with SMOTE

Implementation

The first model trained was the Linear Regressor configured to predict three classes, while the other hyperparameters were maintained with automatic settings.

With a relatively small processing effort, we obtain a Precision Score of 0.66 as a baseline. From there, we will try to achieve higher values using more complex models and adjustment of hyperparameters.

Refinement

In this step, XGBoost will be used to improve the model's predictive capacity. In the first version, the model will be configured for multi-class mode and the standard error metric of the "merror" model.

First set of hyperparameters applied to the problem:

max_depth=6,
eta=0.4,
gamma=4,
min_child_weight=6,
subsample=0.8,
objective='multi:softmax',
eval_metric= 'merror',
early_stopping_rounds=10,
num_class=3,
num_round=200,
seed=42

With the initial configuration proposed, XGBoost was able to increase the Precision Score by 0.15, reaching the 0.77 marks, which can be considered reasonable. From now on, the effort will be to further enhance the metric by adjusting hyperparameters.

In this stage, the hyperparameters will be divided into two groups, and the first will remain fixed and equal to the one used in the first test so that we have comparability between the models, the second set will be tested at intervals to obtain the best configuration, which minimizes the evaluation metric "merror ".

Below the first set, fixed values equal to those of the first model.

```
max_depth=6,
eta=0.4,
gamma=4,
min_child_weight=6,
subsample=0.8,
objective='multi:softmax',
early_stopping_rounds=10,
num_class=3,
num_round=200,
seed=42
```

Then the hyperparameters and the respective intervals tested for optimization.

```
hyperparameter_ranges = {'max_depth': IntegerParameter(3, 15),
                         'eta': ContinuousParameter(0.05, 0.5),
                         'min_child_weight': IntegerParameter(2, 12),
                         'subsample': ContinuousParameter(0.5, 0.9),
                         'gamma': ContinuousParameter(0, 10),})
```

The hyperparameter adjustment resulted in a marginal increase in the Precision Score from 0.77 to 0.79. Although in the reference metric, the gains were not expressive, the prediction of the cases without victims increased from 0.84 to 0.89, and of fatal victims, it went from 0.70 to 0.73 in the confusion matrix. Therefore, we can consider that the investment in time and processing to adjust the model was worth it.

The last test to certify the quality of the model will be to use the data from 2020. For this test, the data were pre-processed only to be under the input expected by the model; however, the data will not be balanced. Model behaviour as if it were in a production environment.

The result observed in the Precision Score is 0.71, a relatively small drop compared to the value obtained in the development. Therefore, the quality of the prediction of the classes "without victims" and "with fatal victims", measured in the confusion matrix was 0.29 and 0.49 respectively.

Results

Model Evaluation and Validation

The final model, XGBoost with the adjusted hyperparameters, showed essential gains over the previous versions.

There was again in the evaluation metrics and compared to the reference studies, Chong, Abraham & Paprzycki, 2005 and Oudemans, 2016, this model performed better.

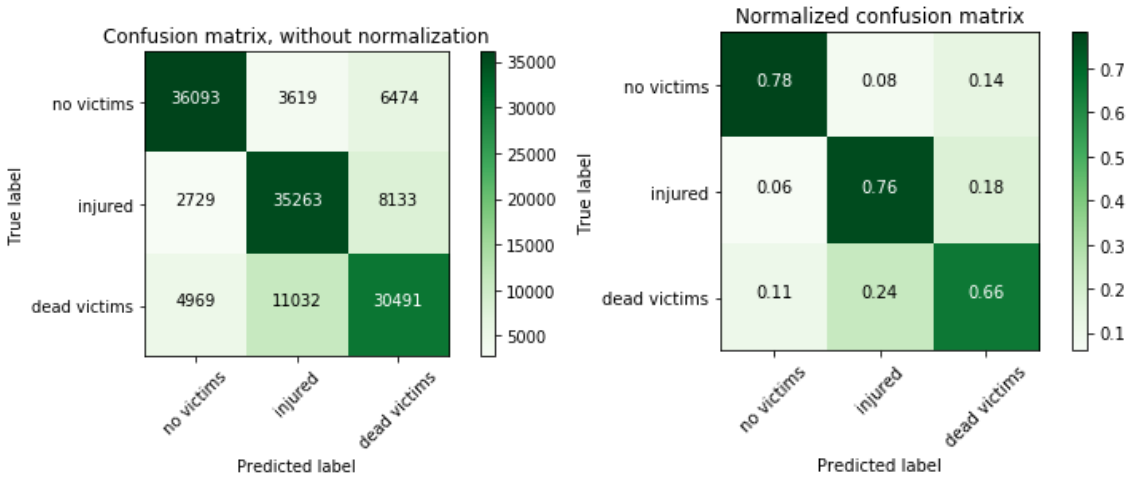
In Chong, Abraham & Paprzycki, 2005, the evaluation metric was Accuracy, and the authors obtained the result of approximately 60%, while in this project Accuracy was 0.78.

In Oudemans, 2016, the evaluation metric was also Accuracy. Still, the author presented the Precision Score, so for this case, we will make a direct comparison; in the project, we obtained 0.79 while Oudemans obtained 0.14.

Metrics for simple LinearLearner.

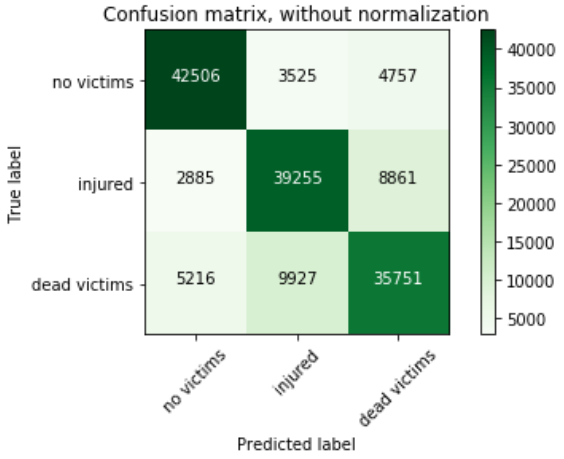
prediction (col)	0.0	1.0	2.0
actual (row)			
0.0	36093	3619	6474
1.0	2729	35263	8133
2.0	4969	11032	30491

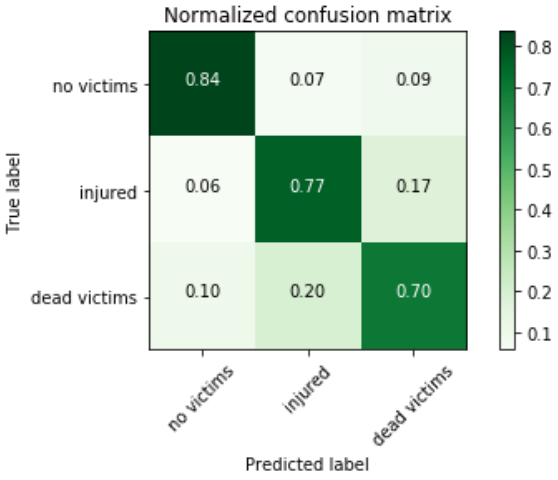
Recall: 0.647
Precision: 0.622
Accuracy: 0.625



Metrics for untuned XGBoost

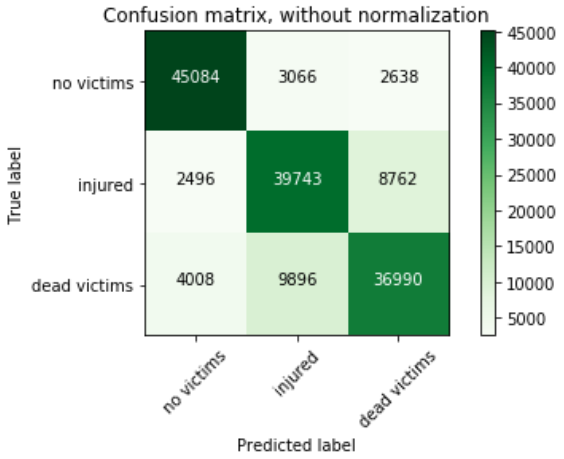
Recall: 0.770
Precision: 0.770
Accuracy: 0.770

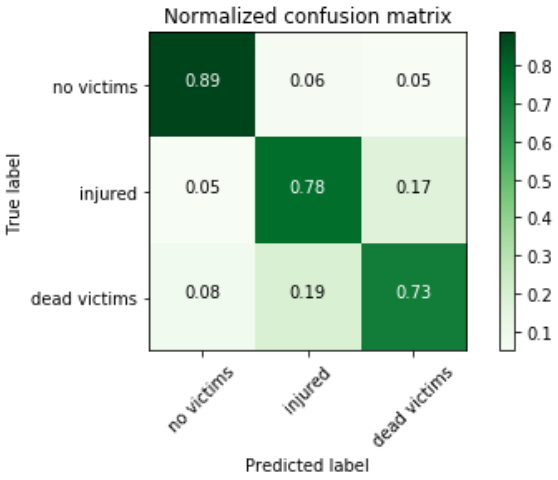




Metrics for tuned XGBoost

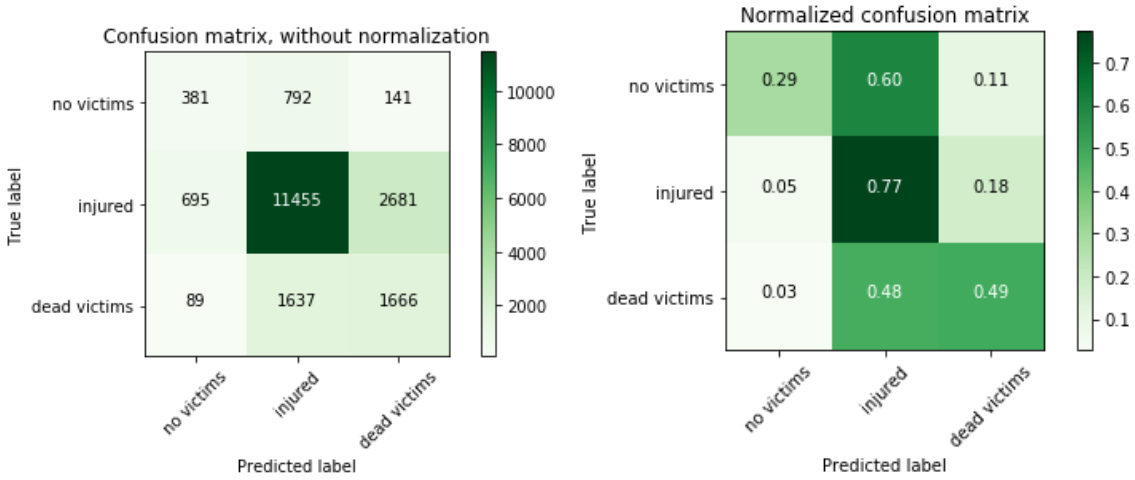
Recall: 0.798
Precision: 0.797
Accuracy: 0.798





Metrics for tuned XGBoost applied in 2020 dataset

Recall: 0.691
Precision: 0.713
Accuracy: 0.691



Conclusion

Reflection

In this project, we use accident data on Brazilian federal highways to create a model capable of predicting whether a given accident will be "without victims", "with injured victims" or "fatal victims".

In this case, the metric used to measure the quality of the model was the Precision. The reason for choosing this metric is to minimize the type I error, which occurs when the model predicts an event as accurate while the observation is false.

For this problem, it is particularly important to correctly predict real events, accidents than with victims, as the costs associated with false positives are quite high. One cannot calculate what the loss of human life represents or the psychological damage and traumatic stresses to which traffic victims and their families are subjected after events of this nature. On the other hand, there is also the formation of economic and financial costs that directly impact families, as well as society in general, and that can be estimated using specific calculation methodologies.

In the end, the result had a result of 0.71 simulating the behaviour of the model in production. This value can be considered a good result, qualifying the model as part of a system for managing resources to mitigate accidents.
The techniques applied in the development of this project were able to overcome the results achieved in other similar studies, which even used more complex models.

Improvements can be achieved with the more intensive use of feature engineering, other balancing techniques and models of neural networks capable of identifying more complex relationships between the variables analyzed.