DATA SCIENCE PROJECT

Guilherme Pires
Miguel Ferreira
João Cruz

Data Science project

Introduction

This project objective is for us to critically analyze two datasets. We are challenged to understand the data we are dealing with, process it, create models and then hypothesize what could be done to improve them. Since we are following the **CRISP-DM** process for each dataset, this report is divided into two major parts, one for each dataset, **NYC Motor Vehicle collisions to Person** and **Air Quality in China**, respectively. **Let's start!**

Dataset 1

Data profiling

For the first dataset, **NYC Motor Vehicle Collisions to Person**, regarding supervised classification methods, we are going to use as target the variable **PERSON_INJURY**, and, regarding forecasting, we will be using **NR_COLLISIONS**. We want to be especially alert to these features when analysing our data.

Data Dimensionality

As it is possible to see in [Figures 1-3 in the appendix], we have almost **6568** records with missing values for the **VECHICLE_ID** and **421** for the **PERSON_AGE**. We have **21** variables of which 15 are Symbolic, 2 are Binary and 4 are Numeric, and 45650 records. This is already presenting us with some problems for us to deal with: how to treat the high number of records with missing values (can they be simply dropped, or should we don't something else? Regarding the symbolic variables, how are we going to dummify them?). Lastly, our target variable has two values: Injured and Killed. On the dataset there are around 45 thousand records for Injured and around 250 for Killed.

Data Granularity

Before analysing the data granularity, we need to understand how our data is distributed.

The first thing we are able to notice is that there are some **incorrect values** (not noise!) on the PERSON_AGE feature, as the maximum value is 9999, and, in fact, we have 29 records with PERSON_AGE > 140 & < 0. For the remaining of the report, these incorrect values were simply removed. Nonetheless, a note must be done that these removed records were all 'Injured' in the target variable. In a bit we are going to see that this is the majority class, and that this removal doesn't impact - at all - our work (the features distributions).

This first work, of analysing the distributions, is particularly important for us to understand what values for each feature may be further divided or grouped. For each feature we compare its distribution when the records are Injured and when the records are Killed (from the target feature). Note that on the appendix is possible to see this analysis for all the features, here we are just going to present some.

Regard time hierarchy, some ideas arise:
- CRASH_DATE records could be grouped into 'Weekday', 'Weekend' and 'Holiday'. (We discovered two python libraries that could do exactly this for both USA and Chinese holidays, for both datasets).
- CRASH_TIME could also be grouped into 'Dawn', 'Morning', 'Lunch time', 'Afternoon', 'Dinner time' and 'Night'.

Both suggestions didn't arise from nothing. They are quite intuitive. We may think that, for example, we would find more crashes on Weekends, Holidays and Fridays. The same for the CRASH_TIME, probably there would be more crashes at Dinner and Night time slots. As we will figure out during this report, what one may think is **intuitive**, more often than not, isn't what we think.

Regarding granularity, there is a lot of ways for us to tackle it: for example, BODILY_INJURY could be grouped depending on the severity of the issue. SAFETY_EQUIPMENT could be grouped as well, depending on the part (or parts) of the body that a given equipment protects. For the PERSON_TYPE, Bicyclists and Motorised could also be grouped, if we view them as individuals who are using a transportation vehicle with only two heels. For all other variables (for instance CONTRIBUTING_FACTOR_1 and _2 which have many unique values), we don't really see any way of grouping the data and gathering more substantial and completer groups...

As we can see in [Figure 1], there really isn't any way for us to group the CRASH_TIME in any meaningfull way that preserves the correct distribution. If we do it, we obtain [Figure 2], which won't give us, we think, enough detail. We must also be aware that reducing unique values on Symbolic variables, just because we can, isn't necessarily what we should do. Obviously, one possible grouping could be to make two groups, one from 00h-> until 11h and another with the remaining hours. This would give us the distinction we are looking for on Figure 1 for the yellow figure. However, this wouldn't be correct as we would simply be manipulating our data to lead to the results we desire.

Figure 1: Left: Diference on distributions between Injured and Killed individuals regarding CRASH_TIME. Right: Diference on distributions between Injured and Killed individuals after grouping.

For example, regarding the feature BODILY_INJURY we observe its distribution in [Figure 2]. As we can see there's clearly a majority of records Killed (read) with injures on "Head" and "Entire Body". For Injured individuals, the distribution is much more well behavied, in sense that there isn't a majority of injured invididuals with injuries just on Head or Entire Body. This means that we could try to make a grouping between body parts like: Head, UpperBody_Front, UpperBody_Back, LowerBody and Entire Body. However, we won't proceed with this since the amount of granular detail we would loose is not a good tradeoff.

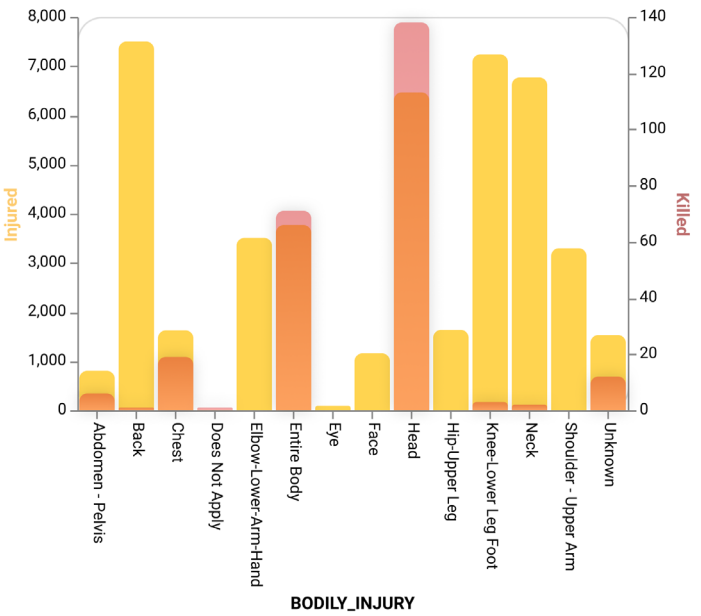

Figure 2: Diference on distributions between Injured and Killed individuals regarding BODILY_INJURY.

Similarly, for SAFETY-EQUIPMENT, [Figure 3], we are able to observe that clearly a majority of the records where there's no use of any safety equipment, the target feature is Killed. Same for who uses "Helmet (Motorcycle)". This **may** point to the fact that a majority of motorcycle accidents are much more deadly than car accidents (since for Injured, represented with Green color, a majority of the people used Lap Belt). A possible grouping , and that we will do, would be to group everything that has 'Helmet' to the group 'Helmet', and everything with 'Air Bag' to 'AirBag'.

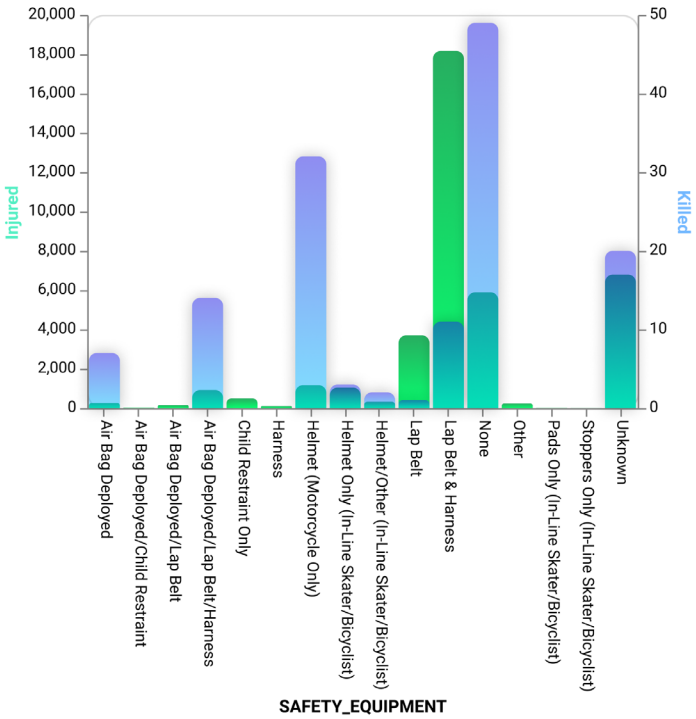

Figure 3: Diference on distributions between Injured and Killed individuals regarding SAFETY_EQUIPMENT.

Still regarding the distributions, there's a lot of curious information that should be retained! For example, if the outcome was 'Killed' then it is much more probable for the PERSON_TYPE to be 'Occupant' or 'Pedestrian', if it is 'Injured' then it is much more problem for the individual to be 'Occupant'. Similarly, for the EJECTION feature we can see that 'Ejected' and 'Not Ejected' is much more probable for 'Killed', and 'Not Ejected' is what is more probable for 'Injured'.

Where we can also retrieve valuable information is in the COMPLAINT and EMOTIONAL_STATUS features. For example, for Killed records, 60% and 20% were respectivly for 'Internal' and 'Crush Injuries'. For 'Injured', 60% is for 'Pain or Nausea', and 10% for 'None visible'. One possible grouping would be to classify these different complaints depending on their severity.

After we talked with a collegue of ours, that is on the 4th year of the medical course, it is his opinion that there isn't any clear way of grouping these values...

For the EMOTIONAL_STATUS, for Killed, almost 90% of the records are for 'Apparent Death' and 'Unconscious'. For the Injured records, 92% are for 'Conscious'. However, we can't find any reasonable grouping to be made.

We took almost 2 pages and an half on data granularity and its distribution, because not only this allows us to **understand** much better the possible problems and results that will appear along the report, but also understanding what are the most important values in each feature allows us to do a much more informed dummification of these **nominal** symbolic variables.

In sum, we have that:
- We will use time hierarchy only for CRASH_TIME and not for CRASH_DATE since we don't **think** it would give us enough detail - or at least the amount of detail we want. **ALTERAR EM CIMA, DISSE QUE NAO IAMOS FAZER**

- We perform granularity for the SAFETY_EQUIPMENT since we will loosing almost no detail and gain genarality, and, it **may** facilitate our predictive models work.

- Regarding the data distribution, type, domain and range, the images needed to perform the respective analysis are all present in the appendix **ADD THE NUMBERS**. There's not much to say: for the symbolic variables, the distribution is, for all variables, very similar to a lognormal function. For the PERSON_AGE the distribution is similar to a gaussian one. In terms of ranges and domain, there's also nothing of much importance to describe.

- Regarding **outliers**, we won't be detecting the outliers now, since we think it will be more frutious to do it after the dummification step.

Data Sparsity

Regarding the numeric variables, there's no relevant information that we can retrieve from their sparsity analysis. Remember that the numeric variables are PERSON_AGE and all other fea-

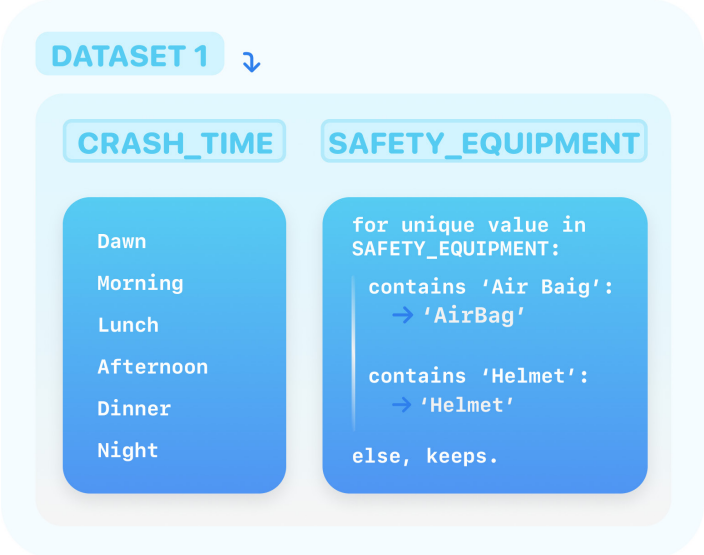

Figure 4: Granularity transformations for the CRASH_TIME and SAFETY_EQUIPMENT features.

tures that are ID's. In fact we don't gain any valuable type of information with the ID's features. In some cases, ID's may withold some type of structure, in the case of this dataset we don't observe it, remarking this features as useless.

For the symbolic features, its sparsity analysis give us the information that we have presented on the previous section (Data Granularity and Dimensinality). Being pragmatic, the sparsity graphs are quite dense, making it more difficul to do any type of analysis. Even more since this dataset is so unbalanced (more on that later).

What we can, though, observe in the sparsity graphs [**IN THE APPENDIX**], is that our data is quite sparse for both Injured and Killed values of our target feature for almost all relevant (excluding ID's) pairwise combination of features. Even though the data is sparse, we have linear 'spots' /clusters of data points for both outcomes. This already tells us that we will have positive correlations between some of the features. Whan we can also conclude by this is that we have, even though we are dealing with a lot of features and records, the data isn't uniform, and still has 'structure', so not every point is the same statiscally speaking. We are not under the curse of multidimensionality.**AREN'T WE?**

Regarding correlations, there's two curious approaches: analyse the correlation matrix for the entire dataset (both Injured and Killed records) and analyse two different correlation matrices, one for Killed and another for Injured records.

Starting by analysing the correlation matrix for the entire dataset, we identify little correlation between most of the features. Only identifying it for (PED_LOCATION x PED_ACTION), (PED_ROLE x POSITION_IN_VEHICLE), (PERSON_TYPE x EJECTION), and a few more similar combination of features. Namely, we observe correlation between the following set of features: [PED_ROLE, PED_ACTION, CONTRIBUTING_FACTOR_1, EJECTION, CONTRIBUTING_FACTOR_2, PERSON_TYPE, PED_LOCATION, SAFETY_EQUIPMENT]. This already points out that when we do feature selection, probably some of the features can be dropped since they can be described by others with little loss of information. More importantly two notes must be taken:

1) There's little to no correlation between the target feature and all other features.
2) The majority of the high correlations found, and described early, have quite intuitive domain knowledge reasons to appear. For example, for the (PED_ROLE x POSITION_IN_VEHICLE), there is the PED_ROLE 'Driver' and also the POSITION_IN_VEHICLE 'Driver'. Similar reasoning can be made for some of the other correlations.

The second part of this analysis that is important is to understand that when we separate the dataset in two - for Killed and Injured -, the correlation matrices are quite the opposite. There's a lot more high valued correlations for the Killed (minority class) values. This is due to the fact that the data is not only much less sparse, we have a low number of records but also because, intrinsicly, there are indeed intuitive patterns for the features of these records. For example, that it is much more likely for one to be Killed if driving a Motorcycle, and that this probability will be even greater if there's no use of Helmet (this is one of the many correlations we can find). This corroborates our findings in the previous section.

Data preparation: Missing Values Imputation

As it is possible to see in [], we have a quite substancial number of records with missing values. From all features with missing values, four of them stand out: PED_LOCATION, CONTRIBUTING_FACTOR_2 and 1 and PED_ACTION.

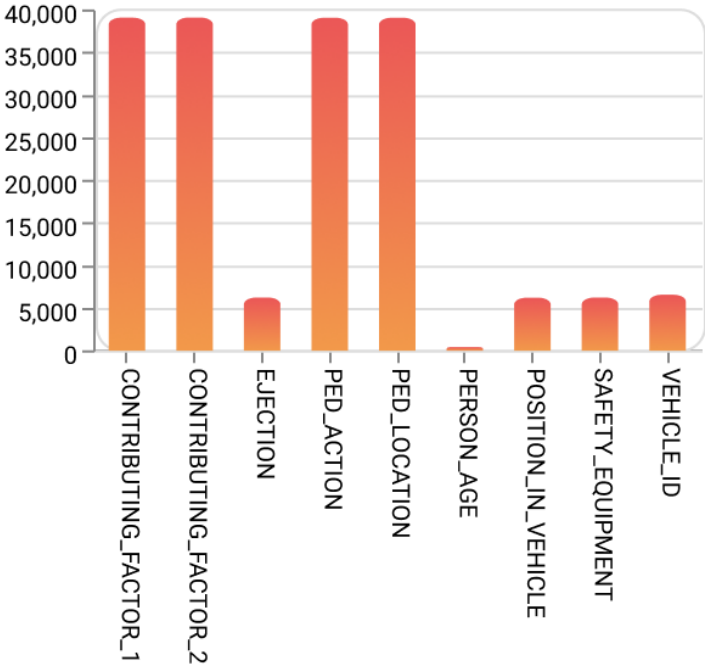

Figure 5: Number of missing values per variable.

After analysing and understanding our data, we were able to understand the reason behind the missing values. For the PED_LOCATION, there are missing values only for records where PERSON_TYPE is **not** 'Pedestrian'. And, well, this makes sense. Since you aren't a pedestrian, you can't have a pedestrian location. Every record that the PERSON_TYPE isn't 'Pedestrian', if the PED_LOCATION is null, we simply put 'NotApplicable'.

For the CONTRIBUTING_FACTOR_1 and CONTRIBUTING_FACTOR_2, every empty record is present when PERSON_TYPE is **not** 'Pedestrian'. In fact, only a few records (+/- 130) that aren't Pedestrian have CONTRIBUTING_FACTOR_1 and 2 filled. For those that haven't, we put 'NotApplicable',. Lastly, for PED_ACTION, if the PERSON_TYPE is Pedestrian, then we fill missing values with 'Unknown', otherwise, if it isn't Pedestrian, we fill with 'NotApplicable'.

For all other features(where there are few missing values), if the PERSON_TYPE is Pedestrian, excluding the features we've talked before, when there is a missing value we put 'NotApplicable'. The exception is for POSITION_IN_VEHICLE, where we fill missing values also with 'NotApplicable'.

For all other PERSON_TYPEs, if SAFETY_EQUIPMENT, EJECTION or POSITION_IN_VEHICLE is missing we put 'Unknown'. If PED_LOCATION is missing, we fill with 'NotApplicable'.

Lastly, regardless of the PERSON_TYPE, if the VEHICLE_ID is missing, we fill it with '-1'.

We end up also dropping all values where PERSON_AGE is missing and where PERSON_SEX is 'U'. When AGE is missing, the target class variable is always 'Injured', as such we aren't losing information in our minority class nor changing our distributions (we are talking about 421 records).

nyc

nyc