Car Accident Severity

Capstone Project

Week 1_Assignment

Report_Section: Data

2. Data

2.1. Data Source

Followed by the Cross-industry Standard Process for Data Mining (CRISP-DM), it is now time to understand the data and then prepare it to be fed into the modeling tools. The given dataset used in this project (provided by the SDOT Traffic Management Division, Traffic Records Group, timeframe: 2004 to present, Seattle, United States) can be downloaded here.

The shared dataset (Data-Collisions.csv) contains 194673 and 38 columns including the labeled data. The labeled data is the "Severity Code", which describes the fatality of an accident. In the shared dataset, the severity code column consists of two values: 1 for property damage and 2 for injury. The dataset includes different attributes, describing a variety of conditions e.g. location, weather, light, road, collision types, and so forth that may influence the severity of the accidents. The attributes are of the types of int64, float64, or object.

2.2. Data Balancing

In the given dataset, the targeted variable (Severity Code) has more observations in one specific class (1, corresponding to approx. 70% of the collision severity accompanied by property damage) than the other (2, corresponding to approx. 30% of the collision severity accompanied by injury). Figure 2.a shows the initial imbalanced dataset in this work.

This imbalanced dataset needs to become firstly balanced, otherwise, the models that will be developed later will be biased. Resampling is a widely adopted technique to address this issue. It consists of randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm (under sampling) or randomly duplicating observations from the minority class to reinforce its signal (over sampling).

By under sampling, a large amount of data, which can be later used for the prediction of severity will be lost. Accordingly, the oversampling technique is preferred and implemented in this work. Figure 2.b shows the resulting balanced dataset, combining the majority class with over sampled minority class.

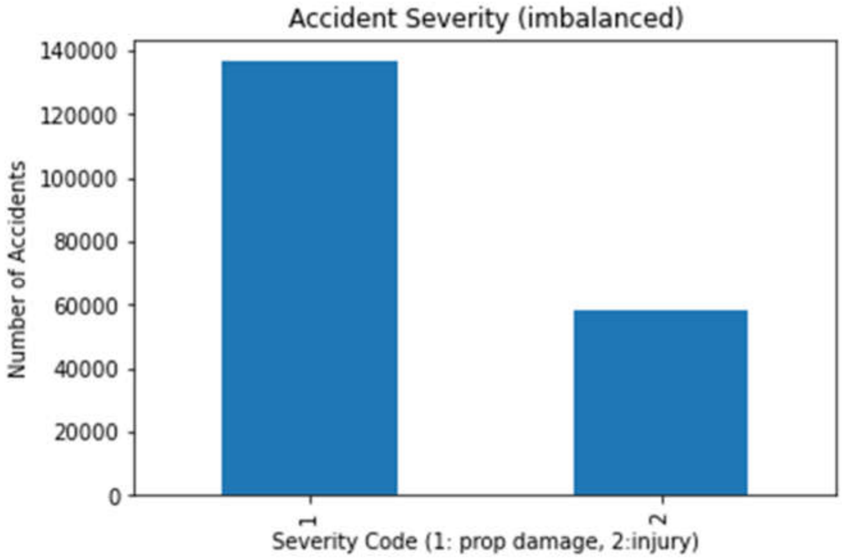

Figure 2.a, Imbalanced dataset

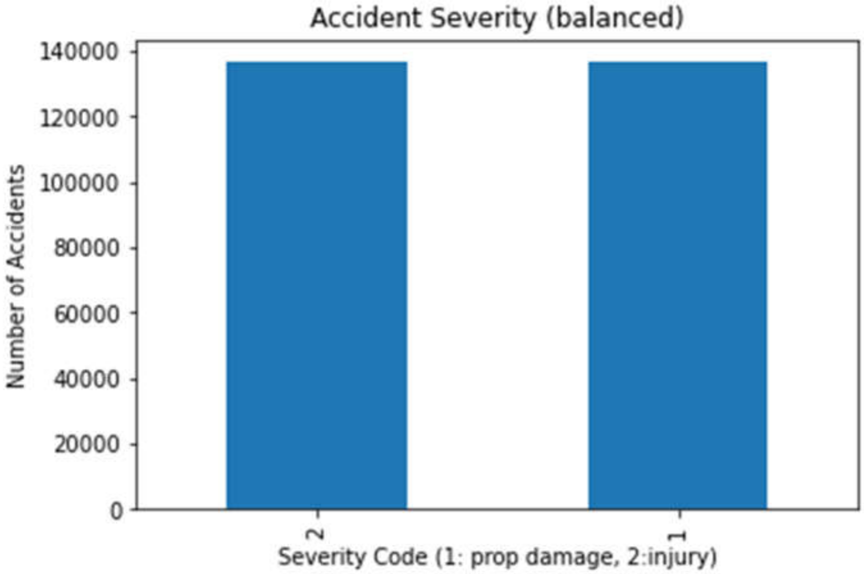

Figure 2.b, balanced dataset

2.3. Data Cleaning

There are several issues which are needed to be addressed during the data cleaning. One issue is many cells with missing values. The other issue with these missing values is that they are widely spread within 19 columns out of 38 columns in the dataset coming with a “NaN” mark. As this distribution ratio is considerably high, the replacement of the missing data with reasonable new values is a better option as far as possible.

The other issue is the presence of both numerical and categorical data in the dataset. To this effect, the replacement is done by the frequency for the categorical variables and by mean for the numerical values. The missing categorical values that are replaced by the largest frequency belongs to the columns WEATHER, SPEEDING, LIGHTCOND, ROADCOND, JUNCTIONTYPE, INATTENTIONIND, COLLISIONTYPE, and ADDRTYPE. The missing numeric values in columns X and Y are replaced by the mean of the belonging columns, respectively.

What should be also taken into account, specifically for processing the data in the next steps, is the incompatibility of categorical variables with the predictive model analysis tools. For example, to develop regression models and being able to use packages such as Sklearn, these variables are converted into indicator variables during the data cleaning after handling the missing data.

2.4. Feature Selection

Taking a closer look into the dataset reveals that many of the columns contain inter-organizational codes which are not relevant to the case of this study and are deleted. These columns include OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY, EXCEPTRSNCODE, SDOT_COLCODE, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, and CROSSWALKKEY. For example, the column SDOT_COLCODE refers to the codes given to the collision by SDOT or the columns INCKEY and COLDETKEY contain the ESRI unique identifier and so on.

Some of the columns also consist of redundant or not enough useful information. For example, there is a second SEVERITYCODE.1 in addition to the SEVERITYCODE which will be deleted. The redundancy in data is also observed for columns such as EXCEPTRSNDESC with no description. The other example is the column UNDERINF which addresses the question:” Whether or not a driver involved was under the influence of drugs or alcohol?” There is however another column named INATTENTIONIND addressing the question: “Whether collision was due to inattention?”

The level of attention in people is usually decreased upon consuming drugs or alcoholic drinks. Accordingly, UNDERINF is deleted and INATTENTIONIND is remained to count for the level of attention. The same analogy is considered for the column LOCATION, as both the X (longitude) and Y (latitudes) are given. Working with X and Y coordinates has also the advantage of a more precise description of the places where the accident occurred. For more clarity, X and Y are also renamed to LONGITUDE and LATITUDE. The same analogy is also considered for the columns STATUS, INCDATE, INCDTTM, SDOT_COLDESC, PEDROWNOTGRNT, ST_COLDESC, PEDCYLCOUNT, HITPARKEDCAR, SEVERITYDESC, ADDRTYPE, and PEDCOUNT. The 10 features selected at the end of this step are listed in Table 1.

Table 1, **List of features being selected in the feature selection**

	Feature	**Description**
1	LONGITUDE	longitude
2	LATITUDE	latitude
3	PERSONCOUNT	total number of people involved in the collision
4	VEHCOUNT	the number of vehicles involved in the collision
5	JUNCTIONTYPE	category of junction at which collision took place
6	INATTENTIONIND	whether or not collision was due to inattention
7	WEATHER	a description of the weather conditions during the time of the collision
8	ROADCOND	the condition of the road during the collision
9	LIGHTCOND	the light conditions during the collision.
10	SPEEDING	whether or not speeding was a factor in the collision