2. Data Understanding

2.1 Understanding the Dataset

The purpose of this chapter is to gain a preliminary understanding of the Collision Dataset for the reason of selecting the label variable and potential features to predict it. In the next chapter, the selected label and feature variables will undergo pre-processing, which includes all measures of transforming and cleaning the data so that it can be read by the machine learning algorithm.

As stated before, this report uses the Collision Dataset from the city of Seattle.[1] The given dataset has 194,673 rows of observations, i.e. reported collisions, and 38 columns, which correspond to various attributes about and around the collisions. The dependent variable, also known as label, is the class that the classification algorithm should predict. In the Collision Dataset, the label is SEVERITYCODE, which can take two values:

- 1 = property damage only,
- 2 = injury collision.

Of those collisions, 136,485 belong to class 1, property damage, and 58,188 belong to class 2, injury collision. An imbalance in the labels is natural in with real-world phenomena but will bias the machine learning algorithm. The issue of imbalance will be addressed in the next chapter. The information about the label can be taken either from the metadata or from the column SEVERITYDESC.[2] Thus, SEVERITYCODE.1 being a duplicate of SEVERITYCODE can be dropped along with SEVERITYDESC and COLLISIONTYPE.

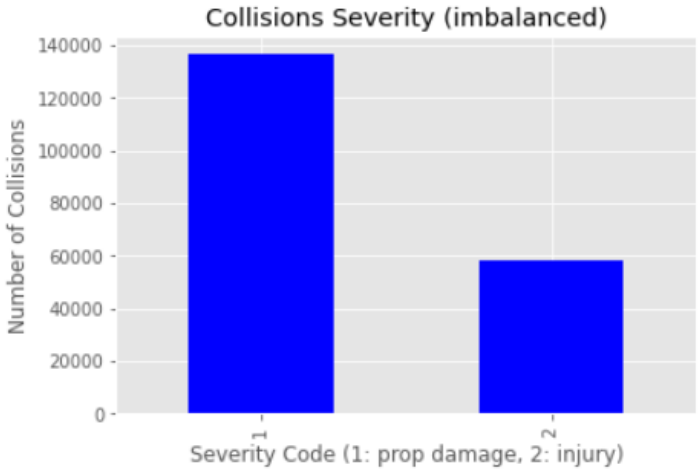

Figure 3: Collision Severity (imbalanced)

[1] The dataset was obtained here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.
An updated version of the dataset is available on the Seattle City website: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0?geometry=-123.374%2C47.452%2C-121.288%2C47.776.

[2] The metadata was obtained here: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.

The metadata also reveals that there are several columns that contain unique identification keys for each collision, but which hold no value for a machine learning algorithm.[3] They serve the authority or organization that created them for cross-referencing purposes. Among those columns are OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, and CROSSWALKKEY. As they hold no value for machine learning, they can also be dropped to narrow the feature set.

The remaining columns describe either the outcome of the collision or the conditions surrounding the accident. Among those that describe the outcome of the collision are PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, and VEHCOUNT (continuous) as well as PEDROWNOTGRNT and HITPARKEDCAR (categorical). They describe the number of various traffic participants and vehicle affected by the accident. Although continuous variables are generally a good source of information for supervised machine learning algorithm, the causality (cause-effect direction) is important. Especially, PERSONCOUNT or VEHCOUNT would make excellent target variables for prediction in a different type of supervised machine learning algorithm, namely regression. They are not truly independent of the collision or its severity but rather the result of it. Therefore, these columns are also be dropped.

2.2 Feature Selection

Now, the rest of the features only describe the conditions before or surrounding the collision. The variables given as X and Y refer to longitude and latitude respectively (and are renamed as such), and they also make a written address in the variables LOCATION and ADDRTYPE redundant. The categorical variables JUNCTIONTYPE, WEATHER, ROADCOND, and LIGHTCOND describe the environmental factors and can be considered independent variables. They will be kept as features whose variation may explain the variation in collision severity. Similarly, the variation in the INCDATE and INCDTTM may help to explain the variation in the collision severity. Likewise, the variation in human behavior as observed by the columns INATTENTIONIND, UNDERINFL, and SPEEDING may help explain the variation in collision severity. As they precede the accident, they can also be considered independent variables and kept as features.

Table 1: Pre-Selected Features

	Variable	**Description**
1	SEVERITYCODE	code that corresponds to severity of the collision: • 1 = property damage • 2 = injury
2	LONGITUDE	longitude
3	LATITUDE	latitude
4	JUNCTIONTYPE	category of junction at which collision took place
5	WEATHER	description of the weather conditions during the collision

[3] Ibidem.

6	ROADCOND	condition of the road during the collision
7	LIGHTCOND	light conditions during the collision
8	INCDATE	date of the incident
9	INDTTME	date and time of the incident
10	INATTENTIONIND	whether or not collision was due to inattention
11	UNDERINFL	whether or not driver was involved under the influence of drugs or alcohol
12	SPEEDING	whether or not the speeding was a factor in the collision

3. Data Preparation

The ultimate purpose of this chapter is to ready the pre-selected variable set for machine learning. Data preparation, also called pre-processing, includes steps such as exploratory data analysis (EDA) to confirm the selection of features, dealing with missing values, and converting variables into machine-legible data types. These steps will be applied to the environmental variables, human behavior variables and date-time variables. At the end, this chapter will also address the issue of label imbalance and employ a resampling method for balancing.

3.1 Environmental Variables

LONGITUDE and LATITUDE are continuous, numerical variables that together pinpoint the location of each recorded collision. Their minimum and maximum value map a square-shaped area of Seattle in which all collisions have been recorded. LONGITUDE and LATITUDE have 5,334 missing values. Common strategies of dealing with missing values in continuous variables are by replacing them either with their median or their mean. In this case, the median would just the point to the center of the square-shaped map of Seattle. But the mean would generally point to the area were collisions happen most frequently (see Figure 4). LONGITUDE has a mean of approximately -122.33052 and LATITUDE has a mean of approximately 47.61954. Their missing values will be replaced with their mean values respectively.

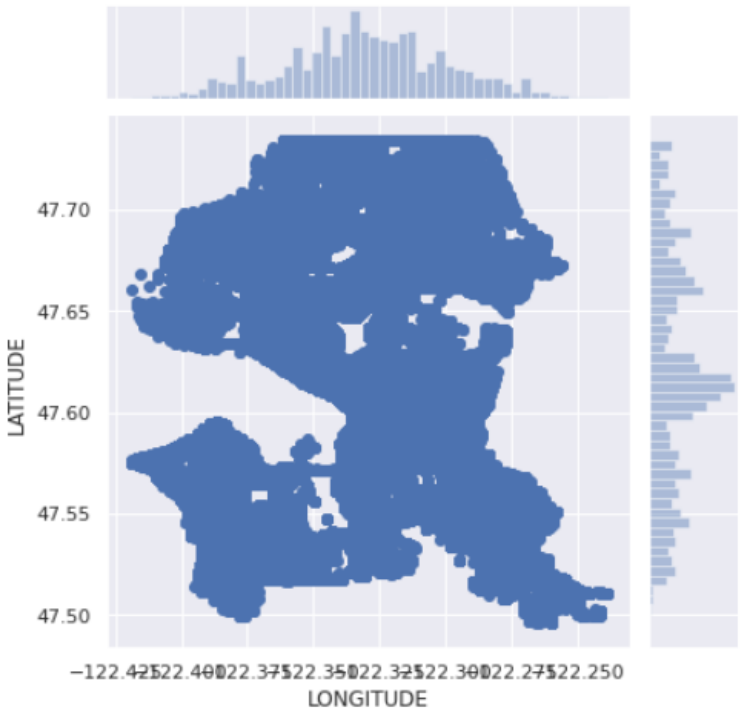

Figure 4: Location and Distribution of Collisions (by LONGITUDE and LATITUDE)

Next, JUNCTIONTYPE, WEATHER, ROADCOND, and LIGHTCOND are all categorical variables that describe the place and surrounding conditions of the accident. A common strategy of replacing missing values is to replace with the mode, the most frequent value.

For JUNCTIONTYPE, the mode is “Mid-Block (not related to intersection)” and it replaces 6,329 missing variables. “Unknown” values are dropped because their weight is insignificant. Afterwards, the different categories are converted into dummy variables for the machine learning algorithm, and the original variable JUNCTIONTYPE is dropped.

For WEATHER, the mode “Clear” will replace all 5,081 missing values. Since there is no indication what “Other” weather is, it is merged with the category “Unknown”. “Unknown” is renamed as “Unknown Weather” to distinguish it from the variable “Unknown Roadcond” to be created in the next step. Additionally, as “Partly Cloudy” is insignificant to it is merged with “Overcast”. WEATHER is also dropped in place for its dummies.

“Dry” is the mode of ROADCOND and will replace all the 5,012 missing values. Again, “Other” is merged with “Unknown”. Moreover, “Standing Water” is added to “Wet” and “Snow/Slush” is added to “Ice” as the former in those pairs are very similar to the latter. “Unknown” is renamed “Unknown Roadcond”. It should not be dropped as it is the third largest category.

Lastly, LIGHTCOND is encoded as a single dummy, where 0 stands for “Daylight” and 1 for all other conditions. 5,170 missing values are also replaced with “Daylight,” the most frequent category.

3.2 Human Behavior Variables

The variables INATTENTIONIND, UNDERINFL, and SPEEDING describe human behaviors that are assumed to increase the chances of an accident.

For INATTENTIONIND, 29,805 observations are given as "Y", which means the missing 164,868 will assumed to be "N" and filled in accordingly. They are encoded as 1 and 0 respectively. It is very intuitive as this suggests that most accidents happen even when people are not paying attention to traffic. Since attention span may be highly depended on the time of the day, the variable INCDTTM is dropped.

The column UNDERINFL has four values ("Y", "N", 1, and 0). Since no further indication is given in the metadata, it is assumed that "Y" equates to 1 and "N" equates to 0. Here a highly counter-intuitive picture emerges, most accidents happen when people are not under the influence of drugs or alcohol (see Figure 5).

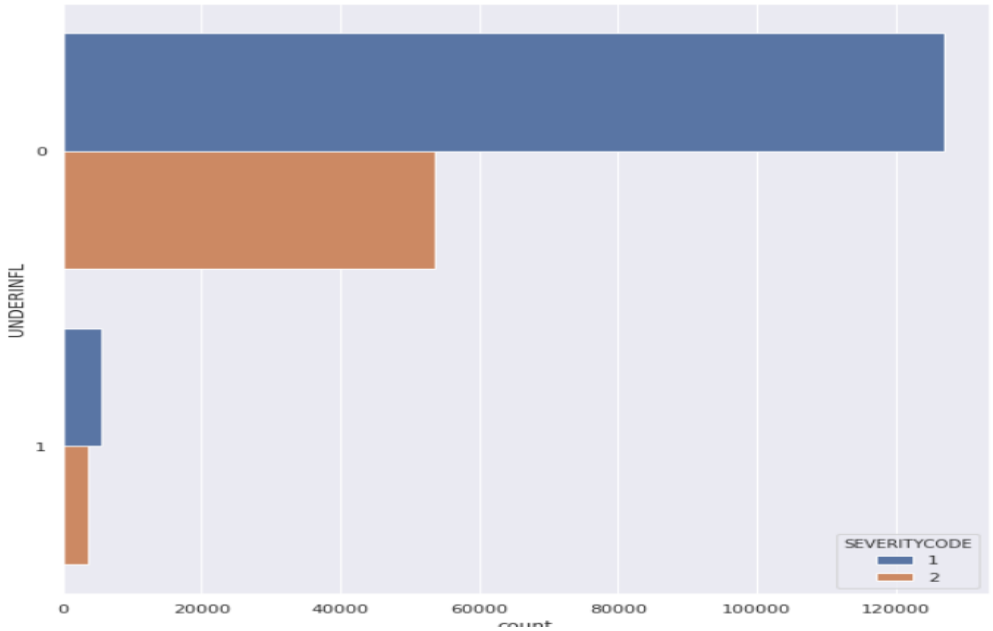

Figure 5: Influence of Drugs or Alcohol on the Frequency of Collisions

For SPEEDING, 9,333 observations are given as "Y", which means the missing 185,340 will assumed to be "N". They are also encoded as 1 and 0. Counter-intuitively, most collisions occurred without speeding.

3.3 Date-Time Variables

The assumption about the date of the collision is closely related to those about the human behavior variables, i.e. during certain days of the week, people are less attentive, for instance, due to exhaustion from work. A cursory analysis confirms that most collisions happen on Fridays while the fewest happen on Sunday (see Figure 6). The variable INCDATE is converted into dummy variable, where 1 stands for weekend and 0 weekdays.

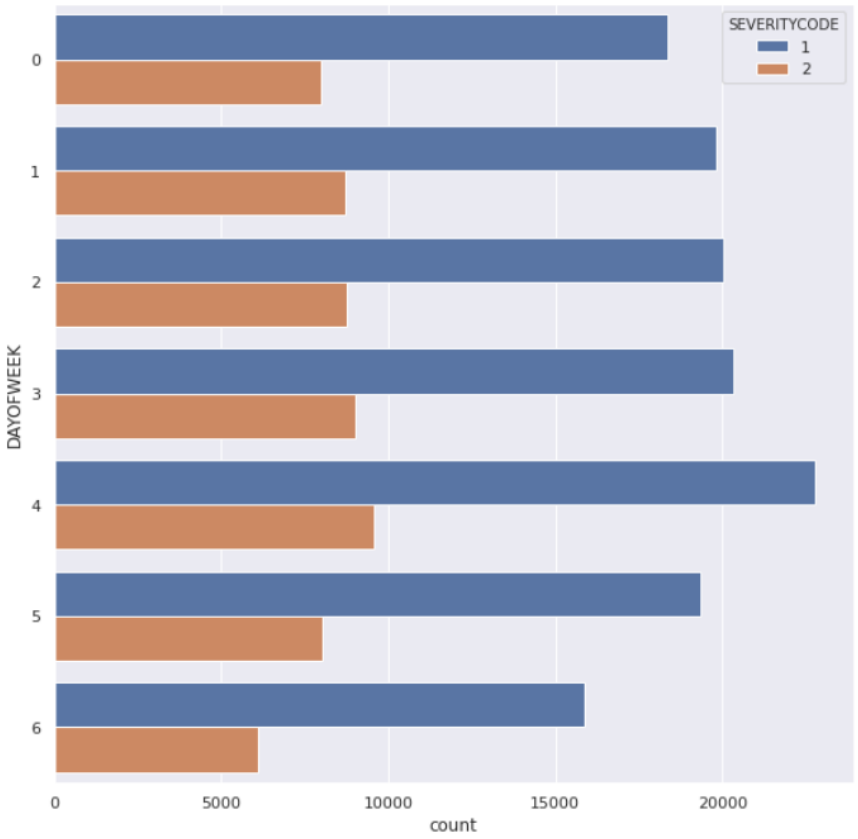

Figure 6: Collision Frequency by Day of the Week

(0 = Monday, 1 = Tuesday, 2 = Wednesday, 3 = Thursday, 4 = Friday, 5 = Saturday, 6 = Sunday)

3.4 Balancing the Dataset

In this dataset, 58,188 injury collision (class 2) and 136,485 property-damage-only collision (class 1) are recorded. Many real-world classification problems have an imbalanced class distribution such as fraud detection, churn prediction, or extreme event prediction. Slight imbalances between classes are considered to have ratio around 4:6, while severe imbalances are considered to have a ratio of 1:100 or more. Thus, the Collision Dataset with an approximate ratio of 58:137 has a severe imbalance. The abundant class, in this case property damage collision (class), is called majority class, whereas the one with fewer samples is called minority class. By definition, a minority class is difficult to predict because of its few examples. This means it is more challenging for a machine learning model to learn the characteristics of examples from this class and to distinguish it from the majority class. In fact, most classification algorithms are designed and demonstrated on problems that assume an equal distribution of classes.

The two common strategies of rectifying imbalanced classes in machine learning are random over-sampling (ROS) and random under-sampling (RUS). ROS is the process of supplementing the dataset with multiple, randomly chosen copies of cases from the minority class, until the number of samples match the majority class. RUS randomly deletes samples from the majority class until the number of samples matches the minority class. Both methods come with advantages and disadvantages. While ROS may inflate or exaggerate underlying patterns in the minority class, RUS may potentially

discard important samples of majority class und distort its underlying patterns. A rule of thumb is to use ROS when the given dataset is small and RUS when the given dataset is large. As the Collision Dataset is sufficiently large, this report employs the method random under-sampling to rectify the imbalance and thereby reducing class 1 collisions to 58,188 samples.