**Accident Severity Prediction Data**

1. **Data Acquisition, Selection, and Cleaning**
   1. **Data Sources**

Dataset for this model are accidents report recorded in Seattle City between January 1st, 2004, and May 20th, 2020. This dataset is available as example dataset in Coursera Applied Data Science Capstone Course, and can be downloaded in [https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass /DP0701EN/version-2/Data-Collisions.csv.](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) This dataset contain many features that are described in its metadata. The metadata for this dataset can be downloaded in [https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf)

[DP0701EN/version-2/Metadata.pdf](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf). Not all features in the dataset will be utilized in the model. The next sections will describe the features selection, and data cleaning.

* 1. **Features Selection**

The machine learning model will be used to predict the severity of the accident. Therefore, the dataset attribute “SEVERITYCODE”, or a code that corresponds to the severity of the collision will be maintained as training data. After checking the features in the dataset, features are selected for accident severity prediction model building. Features selected are features that can be related with road condition, and driver condition during the accident. These selected features are summarized in the table below:

Table 1. Summary of selected features with their description

|  |  |
| --- | --- |
| Feature Name | Description |
| “ADDRTYPE” | Collision address type. |
| “PERSONCOUNT” | The total number of people involved in the collision |
| “VEHCOUNT” | The number of vehicles involved in the collision. |
| “INATTENTIONIND” | Whether or not collision was due to inattention. |
| “UNDERINFL” | Whether or not a driver involved was under the influence of drugs or alcohol. |
| “WEATHER” | A description of the weather conditions during the time of the collision. |
| “ROADCOND” | The condition of the road during the collision. |
| “LIGHTCOND” | The light conditions during the collision. |
| “SPEEDING” | Whether or not speeding was a factor in the collision. |

There are several features that have been drop because of several reasons which are: unknown features, features with very few data, redundant features, unused identification codes, description that cannot be quantified with number, date and location features that could not be used for prediction, and features that not related to road condition, and driver condition during the accident. Dropped features, their descriptions, and the reason why those features are dropped is summarize in the table below:

Table 2. Summary of dropped features with their description and reason for dropping

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Reason for dropping |
| “X” | Location’s latitude | Location feature |
| “Y” | Location’s longitude | Location feature |
| “OBJECTID” | Unique identifier | Unused identification codes |
| “INCKEY” | A unique key for the incident | Unused identification codes |
| “COLDETKEY” | Secondary key for the incident | Unused identification codes |
| “REPORTNO” | Report number | Unused identification codes |
| “STATUS” | Data status | Unknown feature |
| “INTKEY” | Key that corresponds to the intersection associated with a collision | Unused identification codes |
| “LOCATION” | Description of the general location of the collision | Description that cannot be quantified |
| “EXCEPTRSNCODE” | Unknown | Unknown feature |
| “EXCEPTRSNDESC” | Unknown | Unknown feature |
| “SEVERITYCODE.1” | A code that corresponds to the severity of the collision | Redundant feature with “SEVERITYCODE” feature |
| “SEVERITYDESC” | A detailed description of the severity of the collision | Description that cannot be quantified |
| “COLLISIONTYPE” | Collision type | Not related to road condition |
| “PEDCOUNT” | The number of pedestrians involved in the collision. | Redundant feature with “PERSONCOUNT” feature |
| “PEDCYLCOUNT” | The number of bicycles involved in the collision | Redundant feature with “VEHCOUNT” feature |
| “INCDATE” | The date of the incident. | Date and time feature |
| “INCDTTM” | The date and time of the incident. | Date and time feature |
| “JUNCTIONTYPE” | Category of junction at which collision took place | Redundant feature with “ADDRTYPE” feature |
| “SDOT\_COLCODE” | A code given to the collision by SDOT. | Unused identification codes |
| “SDOT\_COLDESC” | A description of the collision corresponding to the collision code. | Unused identification codes |
| “PEDROWNOTGRNT” | Whether or not the pedestrian right of way was not granted. | Features with very few data < 5% occurrence |
| “SDOT\_COLNUM” | A number given to the collision by SDOT. | Unused identification codes |
| “ST\_COLCODE” | A code provided by the state that describes the collision. | Unused identification codes |
| “ST\_COLDESC” | A description that corresponds to the state’s coding designation. | Unused identification codes |
| “SEGLANEKEY” | A key for the lane segment in which the collision occurred. | Unused identification codes |
| “CROSSWALKKEY” | A key for the crosswalk at which the collision occurred. | Unused identification codes |
| “HITPARKEDCAR” | Whether or not the collision involved hitting a parked car. | Redundant feature with “COLLISIONTYPE” feature that has type “Parked Car” |

* 1. **Data Cleaning**

It is clear that after just checking few rows in the dataset, the dataset needs some cleaning to be done. The first step is to change or remove features with NaN or empty value. All features that have NaN or empty values have to be identified, and will be dealt accordingly. In the feature “ADDRTYPE”, there are 1,926 rows with empty values. These empty rows will be dropped because there is no way to check this data. There are no missing values in feature “PERSONCOUNT”, and “VEHCOUNT”. The features “INATTENTIONIND”, “UNDERINFL”, and “SPEEDING” are features with yes or no value. The missing data can be interpreted as “no” value, and string “Y” can be interpreted as “yes” value. Yes or no value will be converted to integer 1 for “yes”, and 0 for “no”, so that it could be used in the predictive model. There are a few exceptions in the “UNDERINFL” feature since multiple formats are used to represent yes and no values. In the “UNDERINFL” feature, the value ‘Y’, and ‘1’ will be converted to integer 1, then, the missing values, ‘N’, and ‘1’ will be converted to integer 0. The features 'WEATHER', ‘ROADCOND’, and ‘LIGHTCOND’ have missing values of 5,081; 5,012; and 5,170, consecutively. The missing values in these features will be change to “Unknown” since values “Unknown” are already in the data. After this data cleaning process the data will be ready to explore, and transform as needed by the machine learning algorithm.