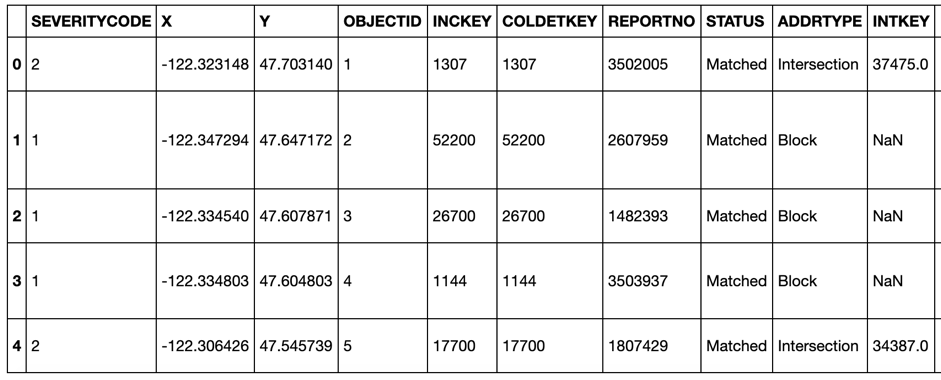
**Predicting Traffic Accident Severity in Seattle**

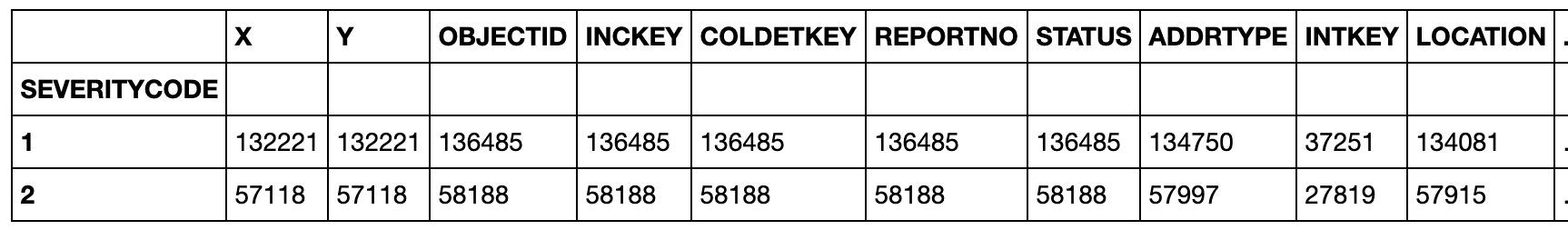
1. **Introduction**  
     
   Our client, Business X, has retained our services to help them predict traffic accident severity in Seattle, where they currently operate. Business X recently began revamping their employee safety program and discovered that their employees were involved in an alarming amount of severe traffic accidents. Business X would like to mitigate this problem and has identified three potential countermeasures:
   1. Allow employees to work from home on rainy days to avoid hazardous road conditions.
   2. Deploy an application onto their employees’ cell phones to prevent distracted driving.
   3. Shift their operating hours by two hours from 9am-5pm to 7am-3pm so that their employees can commute before the morning and evening rush hours.

Business X will make their decision on which countermeasure to choose based on which of the three has the highest probability to reduce the severity of traffic accidents. Thus, we will build a machine learning model to predict the severity of traffic accidents in and around Seattle and use the results to advise Business X as to which countermeasure is likely to be the most successful at reducing the frequency of severe traffic accidents for their employees.

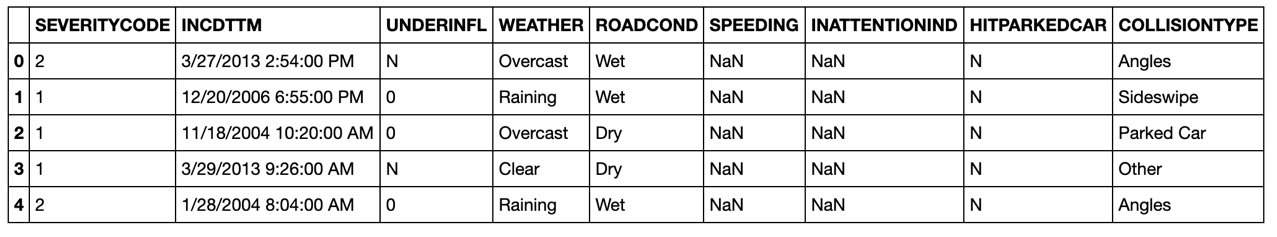
1. **Data**  
     
   The data we will use to build and train our machine learning model comes from the Seattle Department of Transportation and contains information related to traffic collisions and accidents in and around the Seattle area. Critically, this dataset includes a severity classifier (SEVERITYCODE) for each traffic accident (1 for minor vs. 2 for severe) as well as attributes related to Business X’s three potential countermeasures, namely WEATHER for weather conditions at the time of the accident, the time the accident occurred (INCDTTM), and whether or not the accident was due to an inattentive driver (INATTENTIONIND). In all, the dataset contains 37 attributes, some of which will be used to refine the dataset while others will not be relevant or necessary for our analysis.



The dataset initially includes information on 194673 accidents, 132221 of which are labeled as minor (1) and 57118 as severe (2). The dataset will therefore need to be balanced in preparation for building the machine learning model to avoid bias.



The relevant attributes for our analysis and machine learning model include those that allow us to eliminate confounding variables on accident severity, such as those caused while driving under the influence (UNDERINFL) or speeding (SPEEDING), in addition to those directly related to Business X’s problem, as described above.



1. **Methodology**

The dataset was first processed by dropping columns that were not relevant to the question of predicting which potential countermeasure would have the greatest impact in reducing the frequency of severe traffic accidents. This was accomplished mainly by examining the metadata for relevant columns and eliminating the rest.

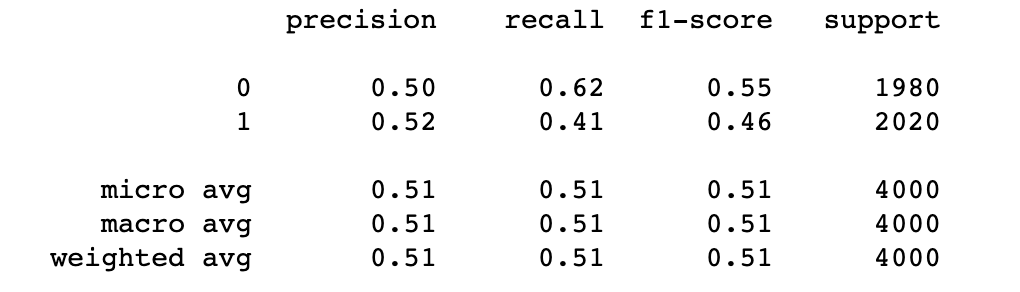
The dataset was further refined by performing exploratory analysis on the remaining attributes. For example, the attribute COLLISIONTYPE revealed >10,000 accidents included pedestrians and/or cycles. These accidents were dropped as the vast majority of Business X’s employees commute by car. Other “human factors”, independent of distracted driving, were also dropped to reduce the influence of confounding variables. This included accidents involving hitting parked cars, speeding, and driving under the influence. Such rows were intentionally removed by dropping the categorical indexes and then the columns altogether.

Next, the ROADCOND and WEATHER attributes were further examined to determine which one would provide the most accurate representation of the driving conditions when the accident occurred. Through additional exploratory analysis, it was concluded that reducing the WEATHER attribute to rows that were either clear or raining would provide the most accuracy. This is largely due to the fact that overcast days in the weather attribute could be associated with dry or wet roads in the ROADCOND attribute. Thus, eliminating this ambiguity would make the results of the analysis more readily interpretable.

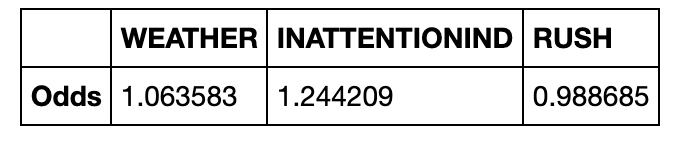
The final steps in preparing the dataset for a machine learning model included removing weekend days since Business X is not open on the weekend, filtering the remaining dataset into two time periods based on the time the accidents occurred (‘Early’ for times that would represent shifting operating hours to accommodate an earlier commute and ‘Current’ for current operating hours and commuting times), and balancing the dataset within the target variable SEVERITYCODE. The latter was accomplished by taking equally-sized random samples of the dataset based on the two SEVERITYCODE values (1 for minor, 2 for severe) and then merging the resulting samples back into a single dataset for modeling.

Logistic regression, a supervised machine learning model, was chosen to classify/predict the severity of an accident. This modeling technique was chosen primarily for two reasons: 1) the target variable, accident severity, is categorical and 2) logistic regression allows one to estimate the influence of the independent variables (WEATHER, RUSH, and INATTENTIONIND) on the dependent variable (SEVERITYCOUNT) by calculating the odds ratios for the independent variables. Thus, a logistic regression model provides Business X with the means to not only predict the severity of an accident but also determine which countermeasure is most likely to be successful in reducing the frequency of severe accidents for their employees.

1. **Results**The resulting logistic regression model has poor predictive validity on estimating the severity of an accident as demonstrated by the log-loss value of 0.69 and the following classification report:



Nevertheless, the logistic regression model was sufficient to determine and compare the relative influence of the three independent variables on the severity of an accident:



1. **Discussion**

These results indicate that, although the overall model performs poorly at predicting the severity of an accident, distracted driving has the greatest impact on increasing the odds of a severe accident, at least among the three independent variables of weather, distracted driving, and time of day. This is evidenced by the odds ratios that indicate a ~24.4% increased probability of an accident being severe rather than minor when a driver is distracted compared to an ~6.4% increase when it is raining, and a ~1.1% decrease when an accident occurs during standard rush hour versus the earlier time period.

We therefore advise that Business X deploy the app that helps to prevent distracted driving by their employees rather than shifting their operating hours or implementing a policy to allow employs to work from home on rainy days.

1. **Conclusion**

Business X may not be satisfied with the predictive validity of the logistic regression model but will find immense value from results, especially given the odds ratios that provide clear direction in terms of selecting the countermeasure with the highest probability of reducing severe traffic accidents for their employees.