1. Introduction

It has long been recognized across that weather and subsequent road conditions play a major factor in road accidents (e.g., Andrey, 2001; Edwards, 1998, 1999; Malin et al., 2019; Pisano et al., 2008). Pisano et al., (2008) reported that 25% of all US public road accidents are weather-related, translating into over 1.5 million crashes annually resulting in hundreds of thousands of injuries and thousands of deaths. The type of adverse road condition and current weather influences the driver's perception of the overall risk, with weather conditions that impact visibility (e.g., rain and snow) causing an increase of driver caution over just wet road conditions (Pisano et al., 2008). Malin et al., (2019) cite snowy and icy road conditions as having the highest accident risk compared to a clean road surface along with a corresponding increase in the risk for fatal accidents during slushy conditions. As such, when road conditions deviate from bare road conditions and good weather with high visibility, the prevailing wisdom is the risk for an accident increases which includes a potential for injuries or fatalities.

One potential mitigation strategy is dissemination of information that conditions are hazardous with a high potential for an accident to both drivers and emergency responders. This would warn drivers to take extra caution and allow for extra time when on the road or consider postponing their trip if possible. For emergency responders, this would provide information when to expect a potentially higher number of automobile incidents and accidents, allowing them to prepare to have resources available to respond to increased incidents based off weather forecasts and other information sources. The objective of this work is to predict severity of an automobile accident based on weather and road conditions. The goal of this work is *not* to determine which specific intersections or locations are the most dangerous nor likelihood of an accident, but what overall conditions would lead to different severity of accidents with the type of location factored into the analysis.

2. Data

2.1 Overview of the Data Used

The data used in this work is the accident severity data provided within the Capstone project. The data describes characteristics of accidents from January 1, 2004 to May 20, 2020 in Seattle, WA, USA including a severity index and numerous columns describing ambient

conditions surrounding the incidents. This section describes the initial processing and culling of the features of the dataset to allow for complete data analysis in later sections.

2.2 Feature Selection

Given the objective of modeling potential accident severity based on the road and overall ambient conditions; a number of the initial 37 features were filtered as redundant or non-explanatory within the context of this analysis. Features with less than 25% data coverage were investigated more closely to understand why this was the case (Figure 1, Table 1). Of these four, only EXCEPTRSNDESC was determined to be not useful since no information about it was in the metadata. The other three columns only contained data if the condition were true; the missing data were filled with false values to build a continuous feature set for that column header. Of the other columns, three had data coverage of under 95% (Table 1). EXCEPTRSNCODE was removed as no information about this column was in the metadata; SDOTCOLUMN doubles the OBJECTID column as an incident identifier, and INTKEY is a code for the intersection that does not provide more detail to the overall ambient conditions contributing to the incident that is not available in other features.

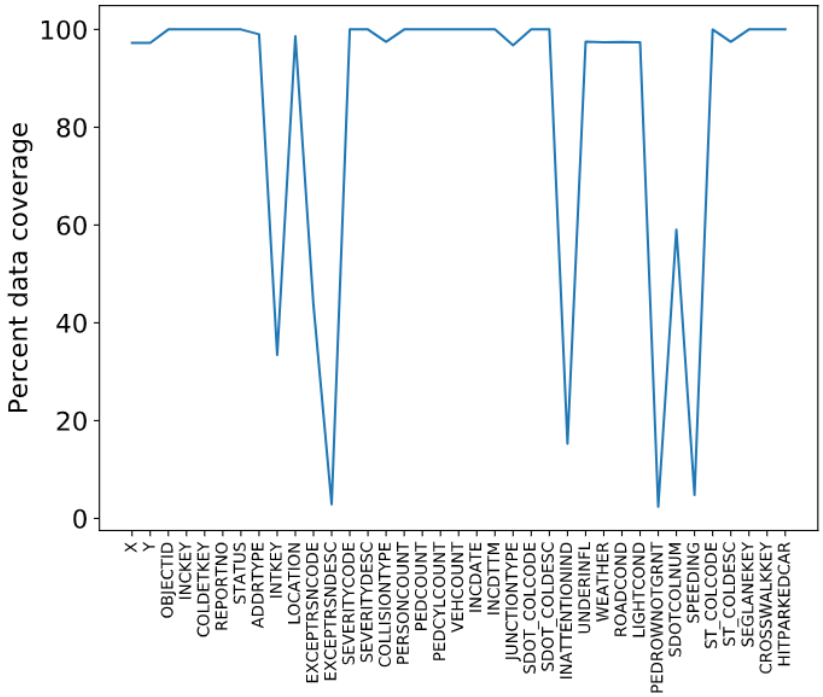

Figure 1. Percent of data coverage for each column header within the initial dataset. Percentage was determined by dividing the count of each column by the total length of the dataset.

A number of different identifiers of location and incident type is provided within the dataset, of these, only one provides a unique identifier (INCKEY) to match specific incident. Otherwise, other incident identifier coders were filtered out. Features describing the general location were kept (ADDRTYPE), the rest were filtered as well (Table 2). Other redundant columns for the type of collision, either an integer code or text description were removed to reduce confusion of which system is being used (Table 2). The SDOT_COLCODE was used as the description of the collision as it was relatively complete and had clear descriptions of each incident type. This left 18 features including the SEVERITYCODE being used as the dependent variable on which to base the model predictions. All others passing the initial filtering are potential explanatory variables. Any further filtering will occur within the initial data exploration described in Section 3.

Table 1. Listing of columns with less than 25% and 95% data coverage and their total coverage.

Description	**Column Header**
Less than 25% data coverage (Percentage of data available)	EXCEPTRSNDESC (2.9%), INATTENTIONIND (15.3%), PEDROWNOTGRNT (2.4%), SPEEDING (4.8%)
Less than 95% data coverage (Percentage of data available)	EXCEPTRSNCODE (44%), SDOTCOLUMN (59%), and INTKEY (33%)

Table 2. Listing of features in the entire dataset and if they were removed for analysis or retained for the final analysis. More details on why different columns were removed are contained within the text.

Description	**Column Header**
Removed for redundant location and object identifiers	X, Y, OBJECTID, COLDETKEY, INTKEY, LOCATION, SEGLANEKEY, INCDTTM, REPORTNO, JUNCTIONTYPE, SDOTCOLUMN, CROSSWALKKEY
Removed for redundant collision severity and type columns	ST_COLCODE, ST_COLDESC, SDOT_COLDESC
Removed for lack of metadata	STATUS, EXCEPTRSNDESC, EXCEPTRSNCODE
Retained for analysis	INCKEY, ADDRTYPE, SEVERITYCODE, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT_COLCODE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, HITPARKEDCAR, INATTENTIONIND, SPEEDING, PEDROWNOTGRNT, INCDAT

2.3 Data Processing and Cleaning

The dataset required a few instances of data processing and cleaning during the initial data filtering and feature selection phase (Section 2.2). The columns not retained (Table 2), were dropped from the pandas dataframe for ease of use and reduced computational complexity. As shown in Figure 1; not every column retained was continuous due to lack of complete coverage. The three columns with minimal data indicating only a positive value were filled with an ‘N’ to indicate a negative to generate a continuous feature column. After these values were filled, every column being used had at least 97% of the labels fully filled with data across all the features. To make sure all labels had data in every feature, labels that contained any missing data across the features were removed. This left 187,504 labels with full data coverage across all features, representing 96.31% of the original labels to be used for the analysis to answer the objective of the work. A deeper exploration of the dataset will be accomplished in Section 3 as part of the model development.

2.4 Final Feature Usage and Dependent Variable Description

The SEVERITYCODE (SC) will be used as the main dependent variable on which the analyses and model(s) will be built. There are only two SCs that are available within the dataset, '1' or '2' indicating a non-injury collision (1) or a collision with injuries (2). Due to this limitation, the overall severity of the collisions will be limited to a binary state, either involving an injury or not regardless of seriousness of the injury. There are a number of variables that interact with each other that could lead to a collision and all will be related back to the SC. The initial data exploration will determine the level of correlation between the SC and the features being used to guide model building. Likelihood of accident will not be possible to determine as there is no data on the total amount of miles driver per incident in which to compare. The final output will be a determination that if there is a collision occurs given a set of road and weather conditions what the potential that the collision will involve an injury or not and with what external factors (speeding, intoxication, etc.,) may contribute to the collision.

3. Methodology and Exploratory Data Analysis

4. Results

5. Discussion

6. Conclusion

7. References

Andrey, D. J. (2001). *Weather Information and Road Safety*. 39.

Edwards, J. B. (1998). The Relationship Between Road Accident Severity and Recorded Weather. *Journal of Safety Research*, *29*(4), 249–262. https://doi.org/10.1016/S0022-4375(98)00051-6

Edwards, J. B. (1999). The temporal distribution of road accidents in adverse weather. *Meteorological Applications*, *6*(1), 59–68. https://doi.org/10.1017/S1350482799001139

Malin, F., Norros, I., & Innamaa, S. (2019). Accident risk of road and weather conditions on different road types. *Accident Analysis & Prevention*, *122*, 181–188. https://doi.org/10.1016/j.aap.2018.10.014

Pisano, P. A., Goodwin, L. C., & Rossetti, M. A. (2008). Highway crashes in adverse road weather conditions. *Proceedings of the American Meteorological Society 88th Annual Meeting. Retrieved September 23, 2009 from Http://Ams.Confex.Com/Ams/88Annual/Techprogram/Session_20929.Htm*.