THERE ARE NO [SEVERE] ACCIDENTS - *MASTER OOGWAY*

TJ HART AND ADAM WARD

ABSTRACT. In this study, we explore whether it is possible to classify the severity of a car accident in Utah generally and Utah County specifically. To answer this question, a public dataset of accident information is analyzed. However, owing to the skewed nature of the data, classification becomes a challenge. This study attempts to overcome the data skewness problem by using a grid search over multiple classification models and clever data downsampling techniques. In addition, ethical concerns about the data in question and the models trained are addressed to ensure correct understanding and use of the work presented.

1. Research question and overview of the data

The Emergency Medical Services (EMS) national website states: "EMS clinicians respond to nearly 1.5 million motor vehicle crashes on the nation's roadways every year ..." [EMS]. In light of the frequency of these events, the goal of this data project is to classify the severity of motor vehicle accidents to inform police and emergency personnel in accident response efforts.

The data found in this project originates from the Kaggle dataset "US Accidents (2016 - 2023)"[MSPR19, MSP+19], which contains data for 7.7 million vehicle accidents in the contiguous US states from 2016-2023. The dataset includes measurements such as accident location, time, temperature, and the presence of traffic lights or other notable road features in the vicinity. Also included is the severity of the accident, measured as "a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay)" [MSPR19]. We seek to answer the question: *Given that an accident has occurred, can we predict its severity?* As stated above, the motivation behind this question is to enable a more data-driven approach to the allocation of police and other emergency resources in the event of an accident on the roadway. With the initial data provided on an accident, a prediction could be made about its severity and the amount of resources to be allocated to manage the accident and surrounding traffic.

Prior research performed with this accident dataset similarly sought to classify accident severity using several combinations of features, models, and data engineering techniques. In our research, we chose to consider a

Date: December 13, 2024.

relatively small subset of the full dataset, namely, accident data for Utah County, since we live in Utah County, and are personally impacted by traffic safety in this region, as are the professors and students at Brigham Young University. However, due to concerns with the small amount of data in this subset, we also ran our analysis on accident data for the entire state of Utah. Since our model is only being trained on a subset of the full dataset, a natural question we address in Section 5 is whether our model generalizes to all counties in Utah, as well as to other US states.

As part of our analysis, we also obtained geographic datasets containing road lines for Utah County [UCB23] and all US county boundaries [Bur16]. These aid in our data visualizations found in Sections 3 and 4.

2. Data Cleaning / Feature Engineering

The full accident dataset boasts a total of 46 columns or features. We started our data cleaning by considering which features should be dropped both to reduce feature redundancy and decrease model complexity.

We dropped the "Country" feature because it is homogeneous for the entire dataset. The "State", "County", "City", "Timezone", "Street", "Zipcode", and "Airport_Code" columns were dropped because these features are specific for each data subset, and we wanted to build a model that could generalize to other data subsets. The "Source" feature contains an indicator based upon which of three sources the data point came from. In our early tests, it consistently was the best predictor of accident severity, and we feared the original sources were disproportionately skewed e.g. one source contained data from only high severity accidents. Thus, it was dropped as well. The columns "Description", "Weather_Timestamp", "Wind_Direction", "Nautical_Twilight", and "Astronomical_Twilight" were dropped because they contained redundant information.

Now, we chose to keep and one-hot encode some categorical data, specifically the "Month", "Day", "Civil_Twilight", and "Sunrise_Sunset" features. These served as proxies to see whether certain months, days of the week, or times of day would help predict severity.

Also, the "Weather_Condition" column had around 40 unique values. We engineered seven simpler, more informative classifications to minimize feature complexity, namely: **snow/hail, rain, fog, dust/smoke, storm, cloudy, fair**. This feature engineering allowed for more information in the model because some types of weather fall into two categories (e.g. 'Wintry Mix' falls into both 'snow/hail' and 'rain'). These features act as proxies to describe vague weather conditions.

Finally, the columns were formatted/cleaned as follows:

- **Start_Time**: Transformed into Year, Month, Day, and Hour columns.
- **Civil_Twilight/Sunrise_Sunset:** If null, filled with "Unspecified".
- **End_Lat and End_Lng**: If null, filled with Start_Lat and Start_Lng.

- **Weather Related Numerical Values (e.g. Precipitation(in)):** If null, filled with median value of the column for the given month.
- **Weather_Condition:** If null, filled with the mode condition of the column for the given month.

3. Data Visualization and Basic Analysis

To begin, we used the GeoPandas Python package to plot the location of accidents on a map of Utah County (see Figure 1). We noted that visually, the majority of higher severity accidents happen along the freeway, I-15.

We also immediately noticed there are significantly more blue dots (Severity 2 accidents) than any of the others. We found that the Utah dataset alone has about **five times** more Severity 2 accidents than any other severity label, leading to a disproportionate weighting of the model on that label. To remedy this, we made plans to experiment with downsampling the data to obtain uniform label proportions in our training dataset (see Subsection 4.2).

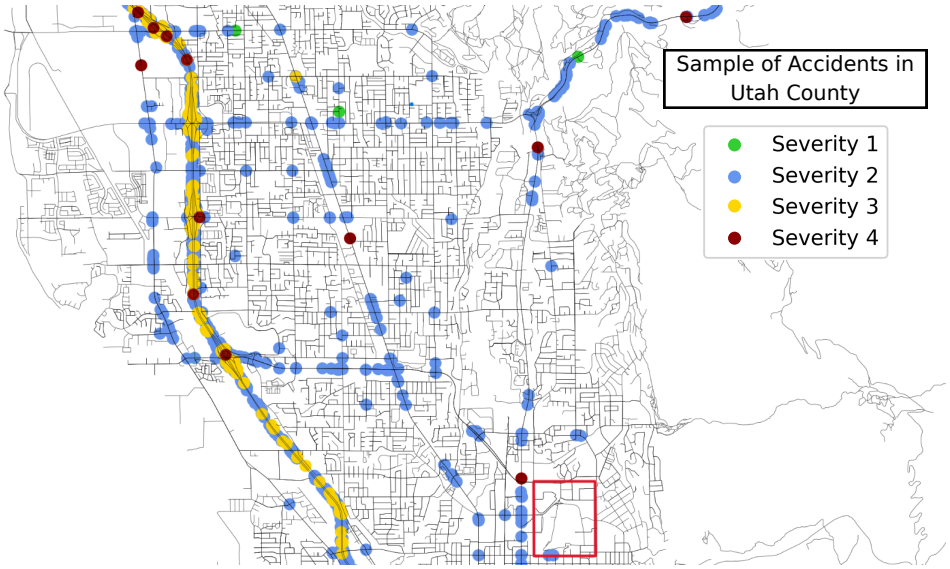

Figure 1. Subset of US Accidents Dataset containing BYU campus (red box). Note the comparatively high frequency of Severity 2 accidents.

4. Learning Algorithms and In-depth Analysis

4.1. **Model Selection.** To predict severity of car accidents, we searched over a variety of model types (learned in Math 402) and many model parameters for each type. Given the time and computer resources available to us, we acknowledge our model search and their respective parameter grid

searches are incomplete. However, we believe our limited search provided insights into our research question and a useful foundation for future research, as described below.

4.1.1. *Grid Search.* We began our model search by first restricting our list of models to the following common tree methods: Decision Tree (DT) Classifier, Random Forest (RF) Classifier, Gradient Boosted (GB) Classifier, and XGBoost (XGB) Classifier. We chose these tree methods because of our time and computational resource constraints, and the fact they perform generally well under a smaller grid search. For each classifier, we did a grid search over a medium-sized set of parameters (see code for full list). In total, we performed 25,460 model fits.

4.1.2. *Scoring.* After a small test run of our models, we quickly found that all model types could easily achieve an accuracy of about 80%. However, this was because our datasets are heavily skewed and the models quickly learned to predict the two most common labels (severity 2 and 3) more often than not. To avoid this issue, we used the macro F1-score to evaluate the best models. The macro F1-score finds the F1-score (which accounts for precision and recall) for each label and computes the unweighted average. This metric gave us a more realistic understanding of prediction capability even though the macro F1-score values were not as high as the accuracy score. We achieved a score of .63 (RF) on the Utah data subset and a score of .61 (GB and RF) on the Utah County data subset. Thus, our model is moderately effective at predicting accident severity, as measured by the macro F1-score. The implications of this are addressed in Section 5.

4.2. **Downsample.** To help mitigate the effects of the label skew on the F1-score, we experimented with a downsampled version of the training dataset. We approached this in a systematic yet crude manner by selecting a label and then downsampling all other labels to have at most the same number of data points as the downsample label by randomly selecting the data points to keep (see code). We see in Figure 2 that the best scores were achieved when downsampled to labels 2 and 3, meaning all or most of the data was kept. Unfortunately, this shows that our downsampling method was unfruitful. In the future, we would instead try oversampling under-represented labels (severity 1 and 4), leading to equal proportions of each label without excluding any data points. Doing this would hopefully increase the F1-score on under-represented labels and thus increase the macro F1-score.

5. Ethical Implications and Conclusions

There are a few key ethical considerations for the use of our model. First, our model strictly predicts severity given an accident has occurred, which does not equate to predicting when an accident of a certain severity will occur. Those using the model must not assume that the model can be "dual-used" to do so.

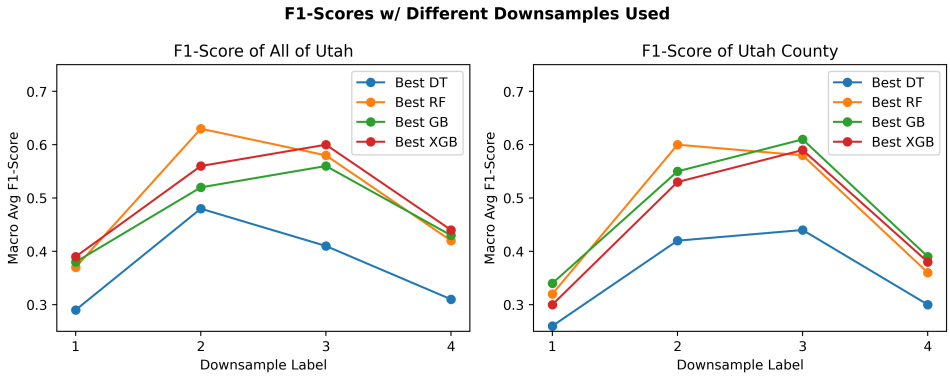

FIGURE 2. Macro F1-scores from the best models found through parameter grid search on each downsample label.

Second, given that we trained on skewed data, our model naturally includes a skew toward predicting the most represented labels. This fact is extremely important because in many cases our model could classify an accident as less severe than it truly is. If this model is used to help allocate emergency resources, this could lead to faulty prioritization of accidents. Thus, first responders should not rely on the model's prediction, though moderately accurate, more than first-hand accounts of the accident.

In addition, we ran tests on data subsets of other US states/Utah counties and found our model performs poorly on out-of-distribution data (other states average macro F1-score: 0.37, other Utah counties average macro F1-score: 0.33). Thus, our model does not generalize well and should not be used in US states/Utah counties that it was not trained on because performance is close to random there.

Also, regarding data privacy, we believe that our analysis and models do not endanger personal privacy because this data is already public and de-identified. In addition, vehicle accidents must be publicly reported by the police, so this information is generally available.

Finally, we have considered the possibility that sending more emergency resources to a misclassified "high severity" accident will cause more traffic delay, subsequently leading to a higher severity classification than it would have otherwise been. If this new data is then incorporated into subsequent training data, over time predictions of high-severity accidents may increase, thus producing a self-fulfilling feedback loop. However, this risk could be mitigated through delayed incorporation of newly recorded data in batches, e.g. add data from a given year into the model after K years, for some K.

In conclusion, despite our model's moderate performance, we believe due to the magnitude of car accidents, our model should not be used in general. However, we believe with more time and effort, we could use these findings to build a more generalized and accurate model to predict accident severity.

References

[Bur16] United States Census Bureau. http://www2.census.gov/geo/tiger/GENZ2016/shp/cb_2016_us_county_5m.zip, 2016. Accessed: 31 October 2024.

[EMS] Emergency medical services. https://www.ems.gov/issues/ems-highway-safety-and-post-crash-care/. Accessed: 25 November 2024.

[MSP+19] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. *CoRR*, abs/1909.09638, 2019.

[MSPR19] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *CoRR*, abs/1906.05409, 2019.

[UCB23] Geography Division US Census Bureau. https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2023&layergroup=Roads, 2023. Accessed: 31 October 2024.