California Traffic Records Collision Severity Prediction

NISHANT MYSORE and MARGOT WAGNER

.Here we analyze the California Statewide Integrated Traffic Records System dataset from January 1, 2001 to mid-October 2020 and use it to predict collision severity. This prediction task was chosen as it is useful to inform which safety measures to take to reduce collision fatalities. From the dataset, we selected a variety of features to classify between collision severity ranging from property damage only to injury to fatal. We compared Logistic Regression, Decision Tree and Naive Bayes models of classification. The best performance was achieved using the Logistic Regression with the inclusion of victim data to get a balanced accuracy of 66.25%. The most important features of the ones evaluated were whether the collision was a hit and run misdemeanor, whether towing was required, and whether is was a rear-end collision.

ACM Reference Format:
Nishant Mysore and Margot Wagner. 2020. California Traffic Records Collision Severity Prediction.

1 DATASET INTRODUCTION

For this work, we focus on the State of California's Statewide Integrated Traffic Records System (SWITRS) from January 1, 2001 to mid-October 2020, which contains detailed reports of vehicle collision data in California. The specific database is an SQLite database available on Kaggle, built by Alex Gude [4]. The data is extensive, containing 9.46 million rows at 5.78 GB total.

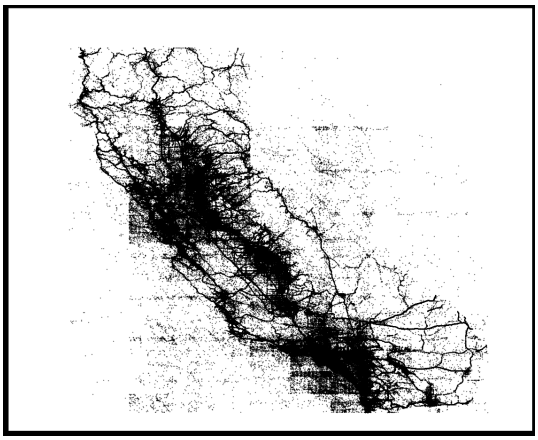

Fig. 1. Plot of geographical location of each collision

The dataset consists of four tables: "caseids", which contains the case ID and year of collision; "collisions", which contains 74 columns of collision information including location, time, severity, and environmental factors; "parties", which contains 31 columns of demographic information for all parties involved as well as vehicle information, sobriety, and other details; and "victims", which contains 11 columns of victim-specific demographic information and injury information.

Looking at the number of collisions in each category of collision severity, we see there is a lot more collisions that are of severity property damage only (PDO) or injury than fatal. It is important to keep in mind that, with regards to collision severity, the dataset is unbalanced.

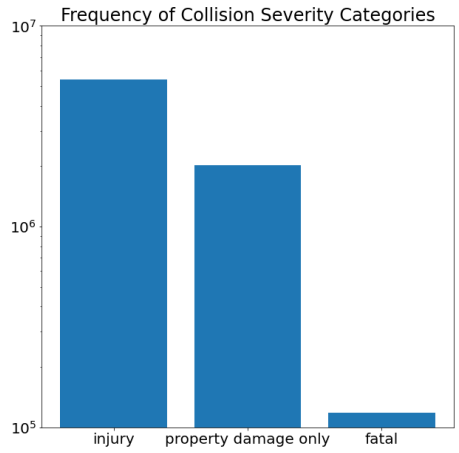

Fig. 2. Frequency of collisions of each severity type.

The primary task of exploratory analysis was to look into specific features of the data in order to motivate feature selection and engineering for the prediction task moving forward. A key component of this was comparing distributions of frequencies for certain features across the label categories. If the distribution of features changed markedly between the categories, it would likely act as a good feature for a categorical predictor.

For example, looking at the type of collision for each severity type (figure 3), fatal collision are most likely to be due to a hit object at 22.1% while both the property damage only and injury categories are most likely due to a rear end. Even still, we see a different distribution with a PDO being due to a rear end 47.3% of the time, but only 37.8% for injury. Additionally, the second most common type of collision for PDO is sideswipe while for both injuries and fatal collisions, it is broadside, suggesting broadside collisions are more dangerous than sideswipe. As the severity of the collision increases, we see an increasing percentage of hit object, head-on, pedestrian, and overturned collisions while rear end, and sideswipe collisions decrease (figure 4). Interestingly, the number of broadswipe collisions is at a maximum for injuries but then decreases again for the fatal severity case. Overall, the change in distribution of type of collision indicates that it could be a key feature in determining collision severity.

Another feature that changes across collision severity conditions is whether or not towing was required for the accident. As the severity of the collision increases, the percentage of collisions that required towing also increases, which matches with intuition (figure

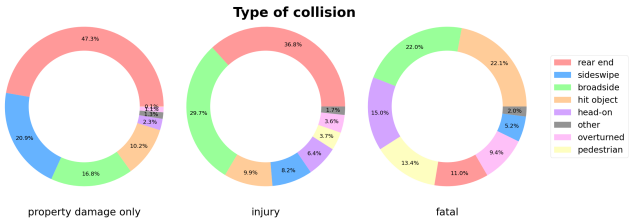

Fig. 3. Percentage of different types of collisions according to collision severity type.

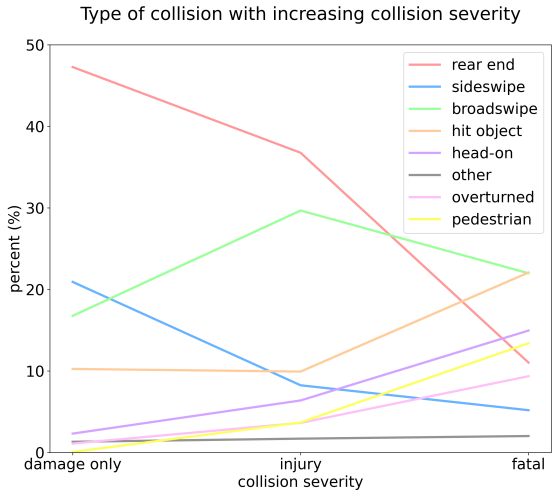

Fig. 4. Change in percentage of collision types as collision severity increases.

5 and 6). We see a dramatic increase for requiring towing from 43.5% to 91.5%. There stark difference in towing conditions paired with a continual increase in the need for towing as severity increases indicates that towing would act as a useful feature for collision severity classification.

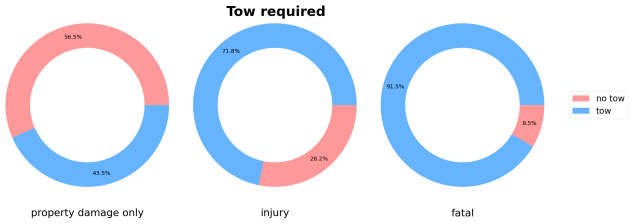

Fig. 5. Percentage of collisions that required towing according to collision severity type.

The lighting conditions at the time of collision also change markedly across severity levels. The percentage of collisions occurring in the daylight drops significantly if the collision is fatal while the percentage of collisions occurring in darkness with no street lights increases significantly (figure 7, 8). The percentage of dark with

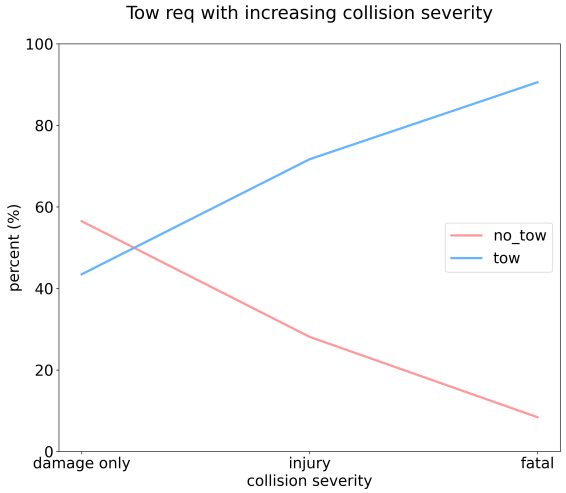

Fig. 6. Change in percentage of collisions requiring towing as collision severity increases.

street lights increases slightly as severity increases, but not as much as with no street lights, suggesting the important role street lights may play in collision fatality.

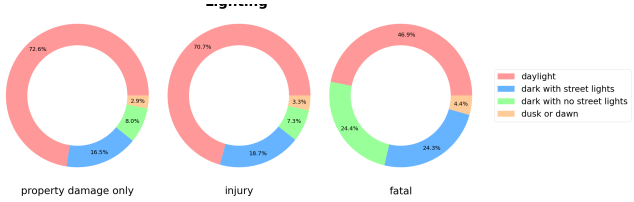

Fig. 7. Percentage of different lighting conditions according to collision severity type.

Some features which would intuitively seem to have a strong effect on determining collision severity do not appear to change significantly. The weather conditions appear mostly constant, with clear conditions remaining a strong majority across all collision severity types and all conditions remaining constant. This suggests it may not be as strong of a feature for predicting collision severity as type of collision, tow away, or lighting.

In the table in Figure 11, we have summarized the condition that makes up the majority percentage for each feature of interest across the severity conditions, excluding at-fault driver and victim age, which are described as average ages. Similar analysis above was done with each feature and related figures are included in the appendix.

Another important aspect of feature exploration is to determine if transformations are required. We can see the age data is skewed and non-negative, therefore, log-transforming it provides a better feature to work with. We can see in Figure 12 that a log transformation provides a more normal distribution than the non-transformed data

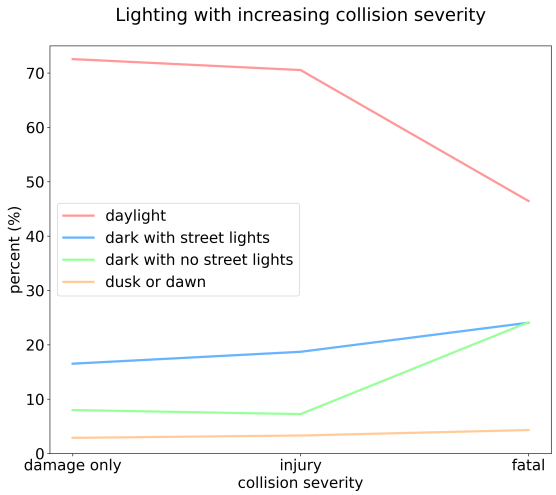

Fig. 8. Change in percentage of lighting conditions as collision severity increases.

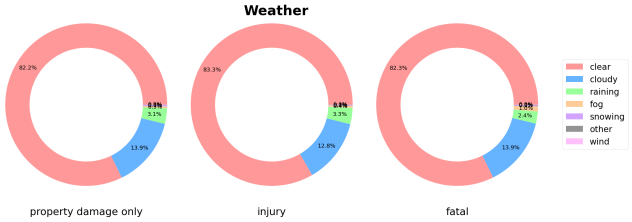

Fig. 9. Percentage of different lighting conditions according to collision severity type.

for both the at-fault driver age and the victim ages, suggesting this transformation would be important to include going forward.

2 PREDICTIVE TASK

For our predictive task, we chose to predict collision severity. From the data, this column could take on 5 different variables - fatal, pain, property damage only, severe injury, and other injury. We grouped these into 3 larger categories: property damage only, injury, and fatal, which reformed the problem into a 3-way classification problem. A list of features that we found useful from our exploratory analysis as well as literature review can be found in figure 11. Our main performance metric uses balanced accuracy, as well as observing the precision and recall. We also use a confusion matrix in order to visualize the true and false positives for each class. The baseline accuracy for our model should simply be 1/3 or 33% for randomly selecting the collision severity.

We performed other additional preprocessing of our dataset. For example, extraneous classes in some categorical variables which had no meaning were removed. Rows which held a NaN (not a number) values or mistaken entries were also removed. Additionally, some classes encoded different classes for effectively the same variable.

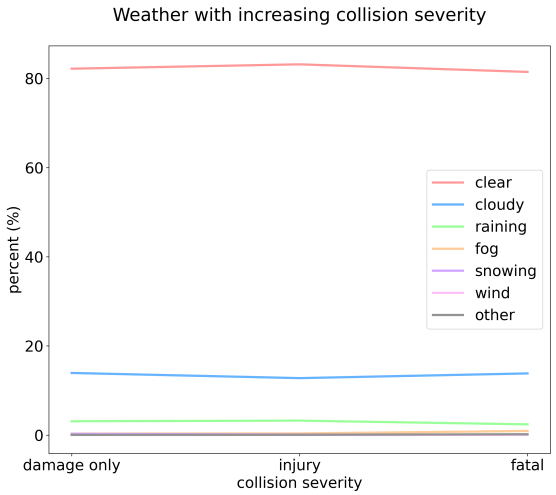

Fig. 10. Change in percentage of lighting conditions as collision severity increases.

	No injury	Injury	Fatal
weather	clear	clear	clear
highway	yes	no	no
party count	2	2	2
lighting	daylight	daylight	daylight
tow req.	no	yes	yes
sobriety (at fault)	A	A	A
road surface	dry	dry	dry
hit and run	no	no	no
sex (victim)	female	female	male
seat position (victim)	3	1	1
avg age (at fault)	35	37	38
avg age (victim)	25	33	37

Fig. 11. Most common value for categorical features across different collision severity conditions.

For example, in the lighting class, "dark with no street lights" was different from "dark with street lights not functioning". These labels were condensed to improve model accuracy. Finally, both the victim age and the at fault age were both transformed logarithmically using the following transformation $x = log(x + 0.01)$ in order to lessen the distribution skew.

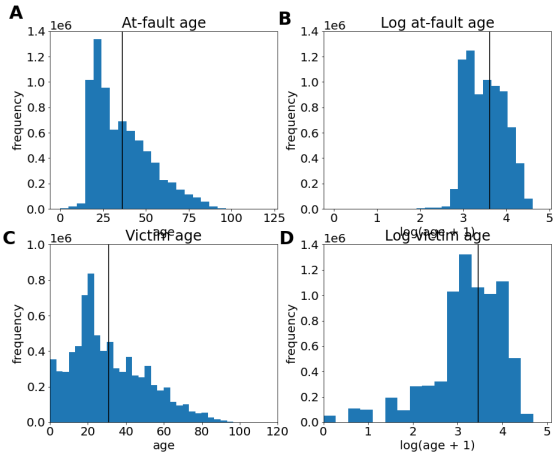

Fig. 12. (A) Distribution of at-fault driver age for all collision severity conditions and (B) the log-transformed distribution. (C) Distribution of victim ages and (B) the log-transformed distribution.

All categorical variables were encoded to one hot vectors, which served as replacement features. The exception to this were categorical variables which took on only 2 types, such as victim sex. These variables were mapped to 0 and 1 instead.

We asses the validity of our predictions by comparing to the baseline model, and comparing to hidden data from the dataset. Performance metrics were analyzed with balanced accuracy as well as observing the confusion matrix of the results, the overall precision, and the recall of the model.

3 MODEL

We trained several different models for our classification task. Since our goal is to predict a variable that can take on only 1 of 3 values, we expect a baseline model that randomly guesses to have a 1/3 or 33% balanced accuracy.

The evaluation of our model will be based on standard classification evaluation, including balanced accuracy, Precision, Recall, and a confusion matrix.

We chose three different models of classification - Logistic Regression, Decision Trees, and Naive Bayes. These were chosen as they are generally strong classification models while maintaining simplicity. Because our dataset was large, it was important to select models that were readily trainable and scalable. Logistic Regression is trained using one vs Rest, which means that the model is able to create separate classifiers for each class. The decision tree was best when fit with a depth of 10.

We created a balanced test set of about 330236 samples, of which 20% were reserved for the test set, and 20% for the validation set. The test set was never touched until the final model was determined by adjusting based on the accuracy of the validation set. Initially, we chose a set of features from only the collisions and the parties, and ignored any features from the victims. However, we then created a model that included the victim seat position, sex, and age. This generated the best results for our prediction task for every classifier.

Model Type	Balanced Accuracy	Avg Precision	Avg Recall	F1 Score
Logistic Regression (No Victim Data)	59.6%	60%	60%	62%
Decision Tree (No Victim Data)	55.3%	67%	55%	65%
Naive Bayes (No Victim Data)	52.5%	52%	53%	47%
Logistic Regression (With Victim Data)	66.25%	64%	66%	68%
Decision Tree (With Victim Data)	61.2%	69%	61%	70%
Naive Bayes (With Victim Data)	63.6%	61%	64%	63%

Table 1. Results for each model on the test dataset

Table 1 shows the results of the model on the test dataset for each model that we created. Each model was evaluated with and without having the victim data. These results were chosen after choosing the best model performance on the validation set. Our test performance is quite similar to the performance on the training and validation data, which means that we adjusted our regularization constant and decision tree depth such that we prevented too much overfitting. We had no issues due to missing data, since we removed all rows with a missing entry as part of the preprocessing.

4 LITERATURE

As previously mentioned, this was an existing dataset hosted on Kaggle by Alex Gude. Gude did basic exploratory analysis including mapping the collision locations and collisions over time, specifically looking at which days of the week collisions occur and analyzing the effect of daylight savings on car crashes [4]. Additional analysis notebooks have been posted on Kaggle for this dataset by other users. One of which aims to predict the number of killed victims from the data using collision data, weather, road surface and road condition as features and predicting using random forest, logistic regression, gradient boosting, and multilayer perceptron [6]. Feature importance was not reported. Another analysis did statistical analysis of motorcycle crashes specifically, looking for the impact of day of the week, motorcycle type, weather, and road conditions on motorcycle crashes [9]. Stewart continued by predicting whether a motorcycle crash was fatal or not using a multitude of features and trained a random forest classifier to achieve 70% model accuracy [8]. The most important features were whether towing was required, whether alcohol was involved, the party count, and lighting, all of which we used as features in our model, and the models used in both these works provide some inspiration for our model selection. The Transportation Injury Mapping System was another tool developed at UC Berkeley to generate summary plots of fatalities and injuries using the same Statewide Integrated Traffic Records System (SWITRS) dataset [7]. The data is also used to project estimated number of fatalities and serious injuries in the future using linear regression from previous years.

Similar datasets have also been used to answer the question of collision severity using a variety of different models with varying degrees of success, including both statistical models and machine learning models. Wu et al. us mixed logit models to analyze driver injury severity in single-vehicle and multi-vehicle crashes on rural two-lane highways in New Mexico from 2010 to 2011 and found dark lighting and dusty weather conditions good predictors for injury severity in multi-vehicle crashes and alcohol involved in both single- and multi-vehicle crashes [11]. In another study, Chen et al used Decision Table and Naive Bayes methods to predict driver injury severity in rear-end crashes and found that poor traffic environment, poor lighting, poor roadway condition, increased vehicle damage, and increased number of vehicles involved all increased the severity of collisions [3].

Comparing statistical models with machine learning models, it has been shown on similar datasets that machine learning models perform better while also being generally easier to implement.. Li et al. compared SVM and ordered probit (OP) models to predict injury severity of individual crashes and found SVM performed better, even for five injury-severity levels [5]. Additionally, they found that two-level prediction resulted in a higher accuracy than five levels. Ahmadi et al. also compared multinomial logit, mixed multinomial logic, and SVM models to predict severity of rear-end crashes and again found SVM outperformed the logit models [1].

The state-of-the-art models for predicting collision severity and related tasks appear to be machine learning classification models. SVM outperforms statistical models, but other classification models have shown to display better performance even still. Wahab and Jiang compared three different machine learning motorcycle collision severity classification models, random-forest models showed the highest accuracy [10]. Features of particular importance for severity prediction were location and time of crash, collision type, road surface type, and shoulder condition. Another study compared Bayesian Network, Artificial Neural Networks-Multi-Layer Perceptron (ANN-MLP), Artificial Neural Networks-Radial Basis Function (ANN-RBF), Support Vector Machine (SVM)-Polynomial and Support Vector Machine (SVM)-Sigmoid models for predicting crash severity and found ANN-RBF models to display the best performance [2].

Overall, from literature it is clear that typical machine learning classification models are the state-of-the-art for the prediction of collision severity, which is clear for the relatively high accuracy achieved with our models. Additionally, the features used in our model were highly motivated by those used in similar types of analyses. This helped us narrow down the dataset from approximately 200 features to the 12 that we used. A main difference between our work and that in literature is the diversity of data. The dataset we used contains data from all types of vehicles and all types of collisions. In order to achieve higher accuracy, it may be beneficial to narrow down the scope of the predictive task in future work to a more specific type of crash, such as sideswipe or motorcycle.

5 RESULTS

Our best model in terms of accuracy was the logistic regression multi class classifier. The confusion matrix for this model can be seen below.

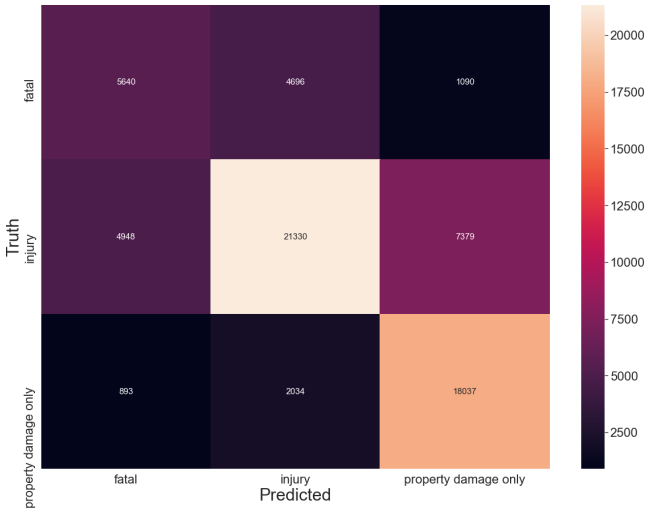

Fig. 13. Confusion matrix for logistic regression classifier

Here, we observe that the model was very accurate in classifying the injury and property damage only categories. This could be due to the imbalance in the data, since there was about double the amount of data for these two categories than the fatal category. One thing that we observe is that our false negatives for predicting fatality are fairly low. Thus, we rarely make a mistake when we predict property damage only and the real answer is fatality. However, we do make significant mistakes in predicting injury when the true value is fatality. The model needs to become more robust to these False negatives so that if it is actually used when a citizen phones into the police station, the police are notified of the correct severity, and take the appropriate action.

The most important features in the network were evaluated based on the weights with the highest absolute values. The top 3 of these were whether there was a hit and run misdemeanor, whether there was a tow away, and whether it was a rear-end collision. This is consistent with what we would expect, because these features would indicate very well the severity of the crash. A tow away and a rear-end collision might be significant predictors of injury, and a hit-and-run misdemeanor would very well predict that there was property damage only. Overall, it is clear the importance of maintaining detailed accounts of collisions in order to prepare for and avoid future collisions as much as possible.

REFERENCES

[1] Alidad Ahmadi, Arash Jahangiri, Vincent Berardi, and Sahar Ghanipoor Machiani. 2017. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. *Journal of Transportation Safety Security* 12, 4 (July 2017), 522–546. https://doi.org/10.1080/19439962.2018.1505793

[2] Amir Mohammadian Amiri, Navid Nadimi, and David Ragland. 2018. Predicting Crash Severity Based on Its Related Collision Type Using Five Data Mining Techniques. *Transportation Research Board 97th Annual Meeting* (Jan. 2018).

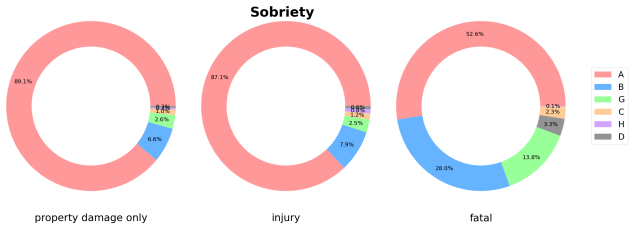

Fig. 16. Percentage of sobriety types according to collision severity type.

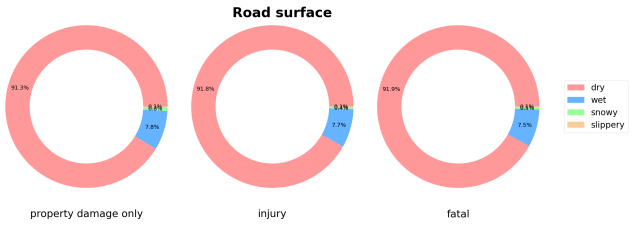

Fig. 17. Percentage of different road conditions according to collision severity type.

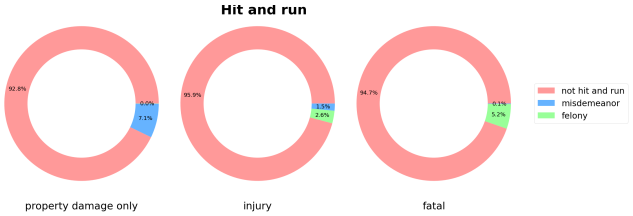

Fig. 18. Percentage of different hit and run conditions according to collision severity type.

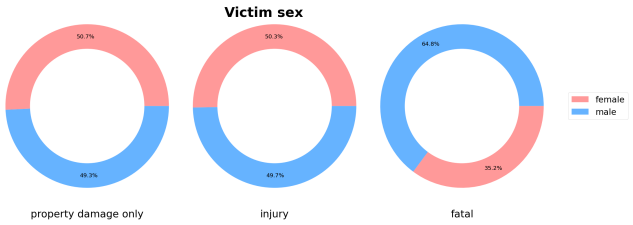

Fig. 19. Percentage of each victim sex according to collision severity type.

[3] Cong Chen, Guohui Zhang, Jinfu Yang, John C Milton, and Adélamar Dely Alcántara. 2016. An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier. *Accident analysis and prevention* 90 (May 2016), 95–107. https://doi.org/10.1016/j.aap.2016.02.002

[4] Alexander Gude and California Highway Patrol. 2020. California Traffic Collision Data from SWITRS. *Kaggle* (2020). https://doi.org/10.34740/KAGGLE/DSV/1671261

[5] Zhibin Li, Pan Liu, Wei Wang, and Chengcheng Xu. 2012. Using support vector machine models for crash injury severity analysis. *Accident analysis and prevention* 45 (March 2012), 478–86. https://doi.org/10.1016/j.aap.2011.08.016

[6] Guillame S. 2020. *Road victims prediction - WIP.* Retrieved December 5, 2020 from https://www.kaggle.com/guillaumes/road-victims-prediction-wip

[7] Safe Transportation Research Education Center (SafeTREC). 2020. *Transportation injury mapping system.* Retrieved December 5, 2020 from https://tims.berkeley.edu/help/Safety_PM.php

[8] S. Stewart. 2020. *Predicting fatalities: 70% recall accuracy.* Retrieved December 5, 2020 from https://www.kaggle.com/sstewart0/predicting-fatalities-70-recall-accuracy

[9] S. Stewart. 2020. *Statistic analysis.* Retrieved December 5, 2020 from https://www.kaggle.com/sstewart0/statistical-analysis

[10] Lukuman Wahab and Haobin Jiang. 2019. Severity prediction of motorcycle crashes with machine learning methods. *International Journal of Crashworthiness* 25, 5 (May 2019), 1–8. https://doi.org/10.1080/13588265.2019.1616885

[11] Qiong Wu, Feng Chen, Guohui Zhang, Xiaoyue Cathy Liu, Hua Wang, and Susan M. Bogus. 2014. Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. *Accident analysis and prevention* 72 (Nov. 2014), 105–115. https://doi.org/10.1016/j.aap.2014.06.014

A APPENDIX

A.1 Feature Exploratory Analysis

Here we include pie charts for the remaining features used in prediction that were not included in the body of the main text.

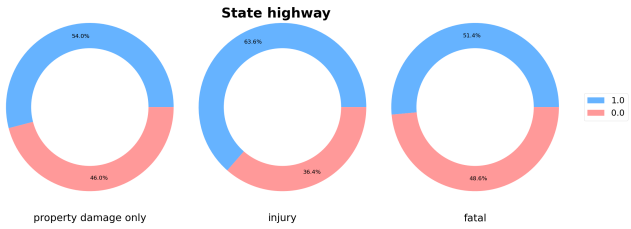

Fig. 14. Percentage of collisions occurring on state highways according to collision severity type.

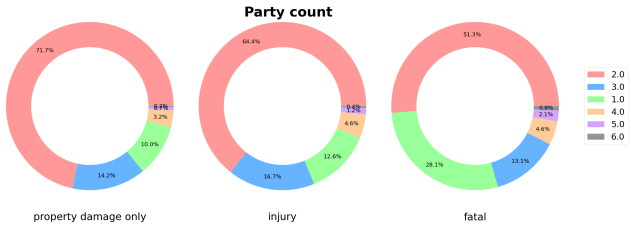

Fig. 15. Percentage of collisions different party counts according to collision severity type.