APAC Datathon Spring 2023

Team 10

Chen Si An Amber, Ong Shao Yong, Tan Shannon

Table of Content

Non-Technical Executive Summary

I. Introduction of Problem

Philadelphia has poor road safety with a high fatality rate and major injury rate. Compared to other major cities, Philadelphia has the highest rate of traffic-related deaths per 100,000 residents (Philadelphia Vision Zero Annual Report, 2022).

Studies have found that severe road accidents are correlated with factors such as type of collision location of the road, the condition of the driver and external conditions such as weather and hour of day (Eboli L et al., 2020). To investigate this, we conducted exploratory data analysis on variables from the ‘crash_general_info.csv’ dataset.

II. Problem Statement

Given that there are many factors that can influence the maximum severity level of a crash, we decided to investigate the following questions:

1. What are the most important factors in predicting the maximum severity of a crash?
2. Given these factors, can we accurately predict the maximum severity of a crash using existing machine learning models?

III. Key Findings

In this paper, we present a prediction model for the severity of vehicle crashes in Philadelphia using data from 2010-2021.

By using a Random Forest classifier, we ranked the features by importance to evaluate their significance in the determination of severity of vehicle crashes (Figure 5). The selected top features were then fed into 4 different machine learning models to train this data. Gradient Boosting (LightGBM) and Neural Networks (TensorFlow) were found to produce significantly higher balanced accuracy scores than the Random Forest and Decision Tree models.

Among the given features, factors such as time of day, speed of vehicle and driver age were found to be particularly significant. Analyzing these factors, we then make recommendations that can be considered by authorities to mitigate the risk of high severity vehicle crashes supported by existing research to decrease severity of vehicle crashes.

Technical Exposition

I. Exploratory Data Analysis

To gain a better understanding of the patterns and relationships between variables of the dataset, we conducted preliminary exploratory data analysis on the data. This section presents our findings which will guide our subsequent data cleaning process, analysis and modeling.

Label	Maximum Severity of Crash
0	Property Damage Only
1	Fatal
2	Suspected Serious Injury
3	Suspected Minor Injury
4	Possible Injury
8	Injury – Unknown Severity
9	Unknown if Injured

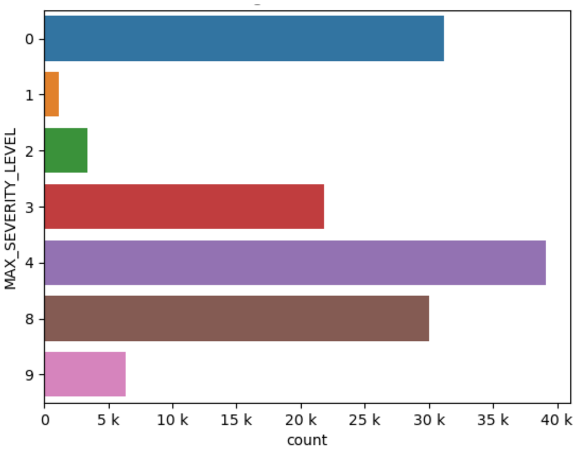


Figure 1: Count of MAX_SEVERITY_LEVEL

Given that the target variable of this research is MAX_SEVERITY_LEVEL, understanding the distribution of categories is crucial. Crashes of severity level "Possible Injury", which is represented by "4", have the highest occurrence in the dataset. On the other hand, a very small proportion of crashes are of severity levels 1 and 2, representing "Fatal" or "Serious Injury".

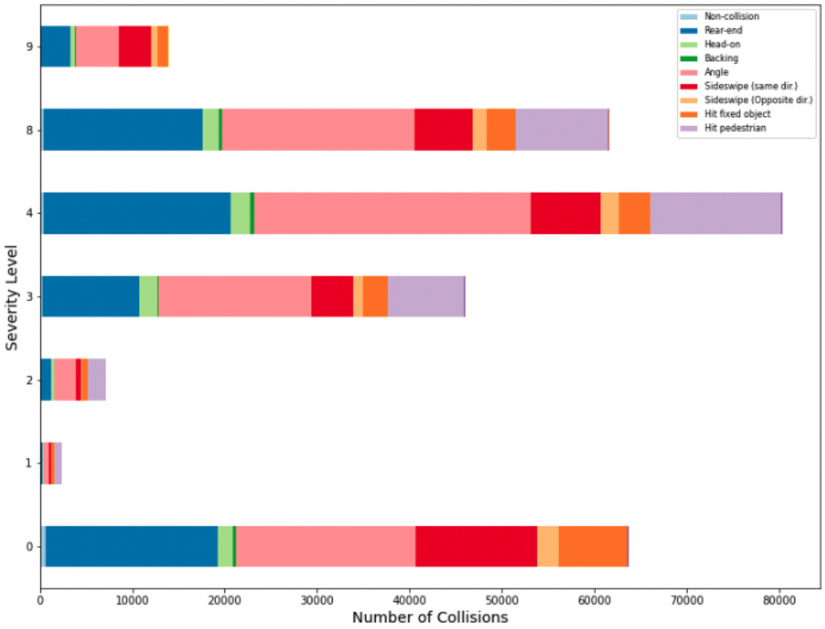


Figure 2: Stacked Barplot of Collision Type against Max Severity Level

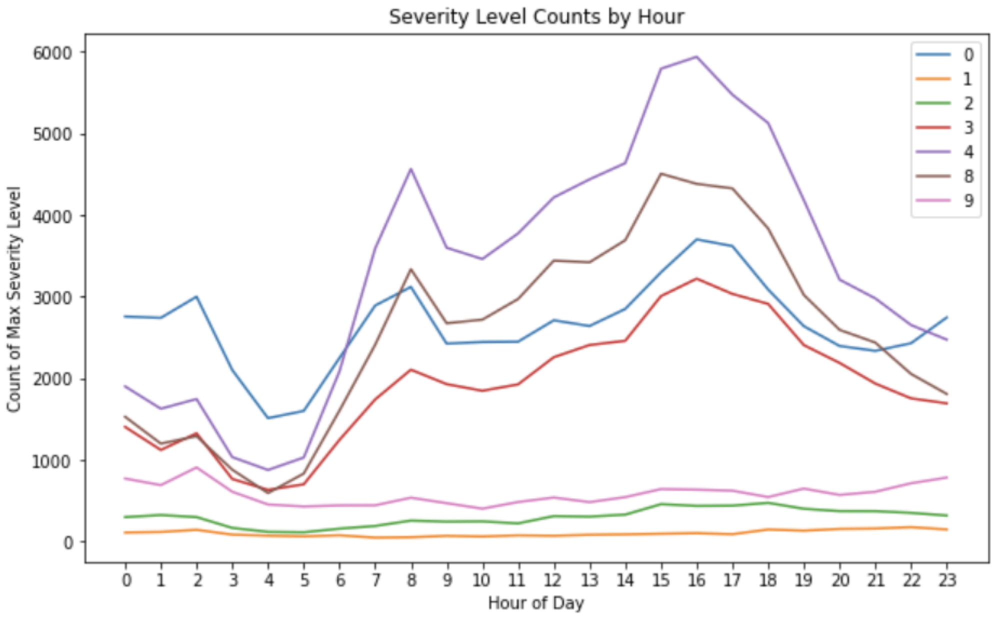


Figure 3: Line graph of Max Severity Level against Hour of Day

We visualized several features against MAX_SEVERITY_COUNT. The proportion of collision types vary across different severity levels (Figure 2). There is a common trend between severity levels of suspected minor injury and possible injury, with a higher incidence of such accidents at 0800, increasing in the afternoon to 1600 (Figure 3).

II. Data Cleaning and Wrangling

Data Merging

We used 3 datasets, namely 'crash_info_general.csv', 'crash_info_flag_variables.csv' and 'crash_info_vehicles.csv' for ease of access. For each dataset, we removed insignificant variables that were unlikely to have a significant impact on the MAX_SEVERITY_LEVEL and would not provide useful information for the model. The datasets were then merged on CRN, the primary key of the crash datasets.

For the purpose of this research, we will be focusing on vehicles that *cause* the crash. Therefore, in the rest of our analysis, we will only be looking at vehicles that are either striking or struck with values of either 1.0 (Striking) or 3.0 (Striking and Struck) respectively in VEH_ROLE as we focus on vehicles that cause the crashes, as well as only non-secondary crashes.

Data Standardisation

To improve performance of subsequent machine learning models, we ensured that the data was of a consistent numeric type, changing “Y”, “N” and “U” values to 1, 0 and -1 respectively. One hot encoding was also performed to transform categorical variables into numeric binary variables. This is important for machine learning models that can only work with categorical data when it is encoded, such as Gradient Boost.

Missing Data

Missing data was addressed using a combination of strategies. We conducted zero imputation on the null values in the dataset, for variables without a significant number of null values.

For other null values, we assigned the corresponding "Unknown" value to them (9 or 99). Thereafter, unknown values for 'MAX_SEVERITY_LEVEL', 'ILLUMINATION', 'VEHICLE_TYPE' were removed.

Duplicated Data

For duplicate/similar columns such as DEC_LAT vs LATITUDE, or for features with very high correlation, we chose to keep one column and remove the other.

After merging the general dataset with the 'crash_info_flag_variables.csv' and 'crash_info_vehicles.csv', rows with duplicate CRNs were removed.

Dataset Rebalancing

Given how unbalanced the data is, since the majority of the crashes that occur are less severe, adjusting the class distribution of a dataset to reduce any bias or imbalance in the data was necessary.

To address the unbalanced dataset, we changed the parameter for the machine learning models to class_weight = 'balanced', which automatically adjusts the class weights in a model to account for class imbalance in the training data. Moreover when calculating precision, recall and F1 score to evaluate our mode

III. Feature Engineering

Here, we created new features based on existing ones to improve the quality of the data for the machine learning algorithms, aiming to represent the data in a way that can be more easily understood by the models.

HOUR_OF_DAY: Values were adjusted to either 1, representing Dawn (6am to 6pm exclusive) or 0, representing Dusk (6pm to 6am exclusive).

DVR_PRES_IND: 2 (Had Been Drinking), 3 (Illegal Drug Use), 7 (Medication) Had Been Drinking were mapped to 1 in a new DRUNK/DRUGGED variable, while 4 (Sick), 5 (Fatigue), 6 (Asleep) were mapped to 1 in a new SICK/FATIGUE/ASLEEP variable.
Duplicate columns IMPAIRED_DRIVER and FATIGUE_ASLEEP were removed upon creation of the two new columns.

ILLUMINATION: Values were mapped to brightness levels in a new column BRIGHTNESS for values 1, 2, 3, 4 representing increasing levels of brightness from Daylight, Dimmed (Dusk/Dawn), Dark (streetlights/unknown roadway lighting), Dark.

BODY_TYPE: Reclassified into individual columns representing Car, Motorcycle, Bus, Van, Trucks/Heavy Duty vehicles respectively. While there were many similar columns that provided information about the type of vehicle, BODY_TYPE was the most specific and allowed us to recategorise our data effectively. Other vehicles that are irrelevant such as trains and snowmobiles are excluded.

BACKUP_PRIOR, BACKUP_NONRECURRING, BACKUP_CONGESTION: As the reason for the traffic backup is unlikely to contribute to severity level of a crash, we merged the 3 columns containing information about traffic backup into a new variable BACKUP with values 1 and 0 corresponding to 1 and 0 in any of the 3 columns. (1 representing congestion at the site of crash, 0 if no congestion).

MAX_SEVERITY_LEVEL:
In our analysis, we will only be focusing on injuries that are known to have occurred as well as injuries that have occurred with a known severity level. As such, we filtered out unknown severity levels (Injury - Unknown Severity and Unknown if Injured) and regrouped and relabelled the remaining severity levels as shown above, from a scale of 0 to 3 in increasing level of severity.

Initial value	New value
0: Property Damage Only	0
1: Fatal	3
2: Suspected Serious Injury	2
3: Suspected Minor Injury	1
4: Possible Injury	1
8: Injury – Unknown Severity	Removed
9: Unknown if Injured	Removed

Table 1: Updated ‘MAX_SEVERITY_LEVEL’ values

After regrouping and relabelling of the severity levels, these are the labels for the remaining data:
0 - Property Damage only
1 - Suspected Minor Injury/Possible Injury
2 - Suspected Major Injury
3 - Fatal

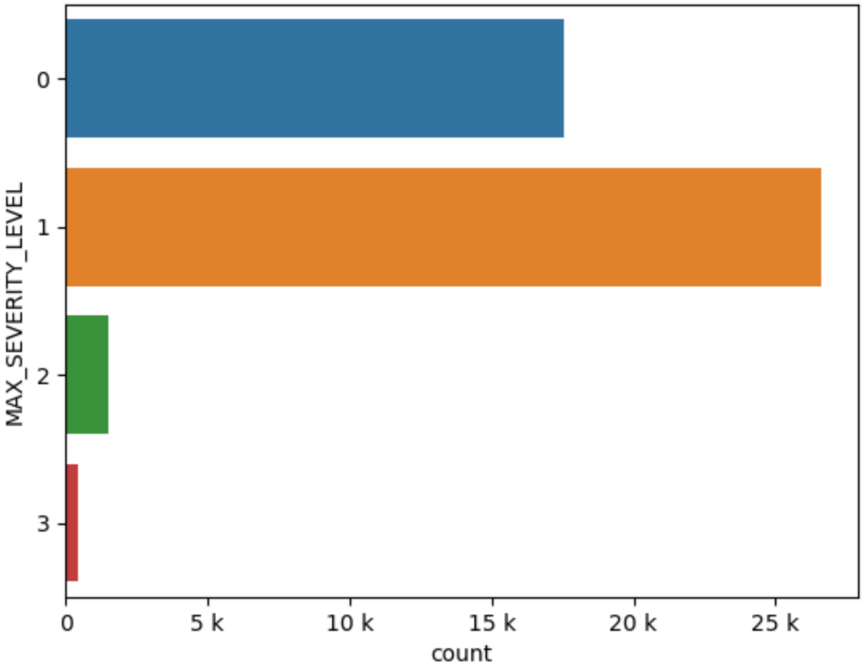


Figure 4: Count of MAX_SEVERITY_LEVEL after filtering

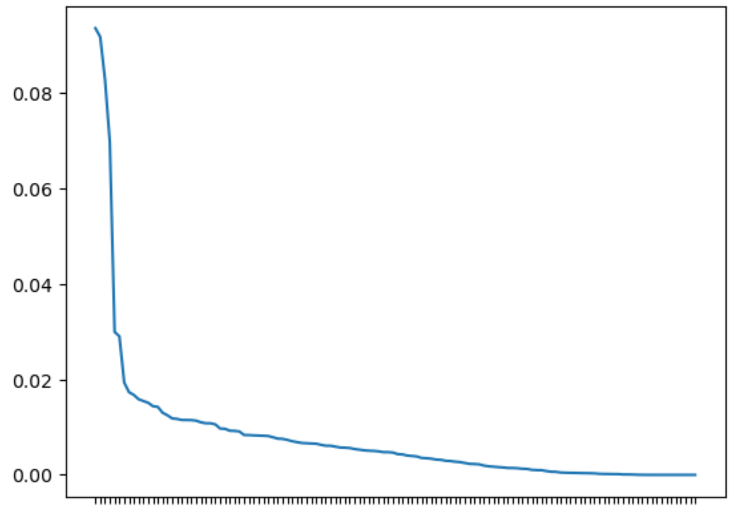
IV. Feature Selection

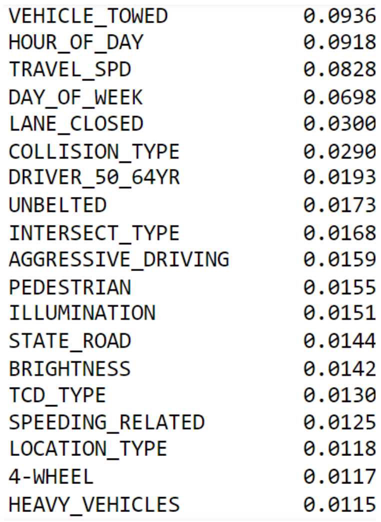
We removed all features that are directly related to the severity of the crash so as to prevent overfitting. Some of these features include ‘FATAL’, ‘POSSIBLE_INJ_COUNT’ and ‘INJURY_OR_FATAL’.

Following which, we used a random forest algorithm for feature selection. We chose to use Random Forest as our data is not continuous. Furthermore, the random forest algorithm has the ability to handle non-linear relationships between categorical variables and the target variable, which may not necessarily be captured by other feature selection methods that work better with linear models.

Due to the large size of the dataset, we decided to set test_size to 0.2 with 20% of the data being used for testing and 80% for training. This ratio allows for a relatively large test set, which can provide a reliable estimate of the model's performance while still leaving a significant amount of data available for training.

Ranking the features by importance, we decided to keep the top 50 features that contribute most to our target variable ‘MAX_SEVERITY_LEVEL’. The model's accuracy would improve more when these features are included in the model compared to when they are not.





```
VEHICLE_TOWED           0.0936
HOUR_OF_DAY             0.0918
TRAVEL_SPD              0.0828
DAY_OF_WEEK             0.0698
LANE_CLOSED             0.0300
COLLISION_TYPE          0.0290
DRIVER_50_64YR          0.0193
UNBELTED                0.0173
INTERSECT_TYPE          0.0168
AGGRESSIVE_DRIVING      0.0159
PEDESTRIAN              0.0155
ILLUMINATION            0.0151
STATE_ROAD              0.0144
BRIGHTNESS              0.0142
TCD_TYPE                0.0130
SPEEDING_RELATED        0.0125
LOCATION_TYPE           0.0118
4-WHEEL                 0.0117
HEAVY_VEHICLES          0.0115
```

Figure 5: Feature importance and the top 20 features

V. Model Selection

Data factors

After data cleaning and feature engineering, we were left with 46,061 data points and 128 labeled data columns. Due to the binary nature of most features, the dataset is sparse. Given that our target variable, ‘MAX_SEVERITY_LEVEL’, is a categorical variable, classic regression models were unsuitable. Furthermore, the features in our dataset are a mix of binary and continuous variables, hence classification training models were likely to be more effective. In the interest of time, we chose the following models to train our dataset:

- Decision Trees
- Random Forest
- Neural Networks
- Gradient Boosting

Hyperparameter tuning

For a more efficient and flexible way to search through a large hyperparameter space, RandomizedSearchCV was used for hyperparameter tuning. A subset of hyperparameters is randomly selected and the model's performance is evaluated using cross-validation, repeating the process until the optimal combination of hyperparameters is found. The best parameters chosen have been included in the table below, under “Results”.

VI. Findings and Discussions

Results

As shown in Figure 4, the data for 'MAX_SEVERITY_LEVEL' is imbalanced. This is expected due to the natural imbalanced nature of crashes, with most of them being low impact and thus classified as low severity. As such, the classic accuracy score cannot be used in the modeling process, and a balanced test accuracy was used instead to deal with imbalanced datasets in multiclass classification problems. This was done using the 'balanced_accuracy_score' from the sklearn library, except for the TensorFlow model where a custom class was used. For the calculation for the metrics, precision, recall and F1, the average is set to 'weighted', with the calculation of each score taking into account the frequency of each class in the data. Table 2 shows the results for each of the 4 machine learning models.

Model	Balanced Test Accuracy	Precision	Recall	F1	Hyperparameters Used
Decision Trees	0.4144	0.6695	0.5703	0.5914	'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': None, 'max_depth': 15, 'criterion': 'gini', 'class_weight': 'balanced'
Random Forest	0.3854	0.6746	0.6823	0.6729	'n_estimators': 200, 'min_samples_split': 5, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 30, 'class_weight': 'balanced'
Neural Network (TensorFlow)	0.4829	0.6503	0.6718	0.6569	'epochs': 100, 'batch_size': 32, 'validation_split': 0.1
Gradient Boosting (LightGBM)	0.4945	0.6711	0.6891	0.6800	'objective': 'multiclass', 'metric': 'multi_logloss', 'num_class': 4, 'boosting_type': 'gbdt', 'num_leaves': 31, 'learning_rate': 0.05, 'bagging_freq': 5, 'verbose': 0

Table 2: Balanced test accuracy scores, precision, recall, F1 and hyperparameters

It is evident that the balanced test accuracies for Neural Network (TensorFlow) and Gradient Boosting (LightGBM) outperformed the other 2 models.

In general, Random Forest is less sensitive to the choices of hyperparameter tuning as compared to other machine learning models like Neural Network and Gradient Boosting, which may explain its low accuracy score. Furthermore, Random Forest involves fitting a decision tree on each of the selected bootstrap samples from the training dataset. However, when using imbalanced data, there is a high probability that a bootstrap sample contains few or even none of the minority class, resulting in a tree with poor performance for predicting the minority class.

Similarly, the Decision Trees model tends to only predict the majority class data. As such, features of the minority class are often treated as noisy data and hence excluded. Subsequently, there is a higher probability of misclassification of the minority class as compared to the majority class. Furthermore, the recall and F1 scores for the Decision Trees model are lower than the other 3 classifiers. This could be due to it being a simpler model that may not work as well with complex data. Decision Trees are prone to overfitting when the tree is deep and complex and thus might perform well on training data but fail to generalize well on test data. In addition, a single Decision Tree does not consider interactions between features especially for complex relationships, which can limit the accuracy of predictions. The Decision Trees model is also sensitive to data distribution and not very robust to changes in data distribution, leading to less stable and consistent predictions.

Rather than using bagging techniques like in Random Forest, Gradient Boost adopts a boosting method. Through the iterative algorithm that trains weak models sequentially, each model can correct the mistakes made by the previous models. As such, the algorithm focuses more on the minority class by assigning higher weights to the misclassified minority class examples over time.

As for Neural Network (TensorFlow), the algorithm supports distributed training, thus enabling the training of large-scale models on large datasets. This may prove effective when working with imbalanced datasets, where the minority class is represented by a small number of examples and could be a possible explanation for its relatively higher balanced test accuracy score.

In general, while Gradient Boosting and Neural Networks presented higher balanced accuracy scores, these scores of 48%-49% are still considered relatively low. This is likely due to the fact that rebalancing during data preprocessing was not executed and data cleaning is imperfect, resulting in noise which affects the readings. Furthermore, due to time and computational constraints, the hyperparameters for each model are not fully optimized. With more time and resources, further analysis can be conducted to understand the nuances of the models' behavior and optimize the model to increase the accuracy and precision of the models.

Importance of features

Following which, we decided to look into the following features that are identified as more important (during feature selection) for modeling, and their relationship with “MAX_SEVERITY_LEVEL”.

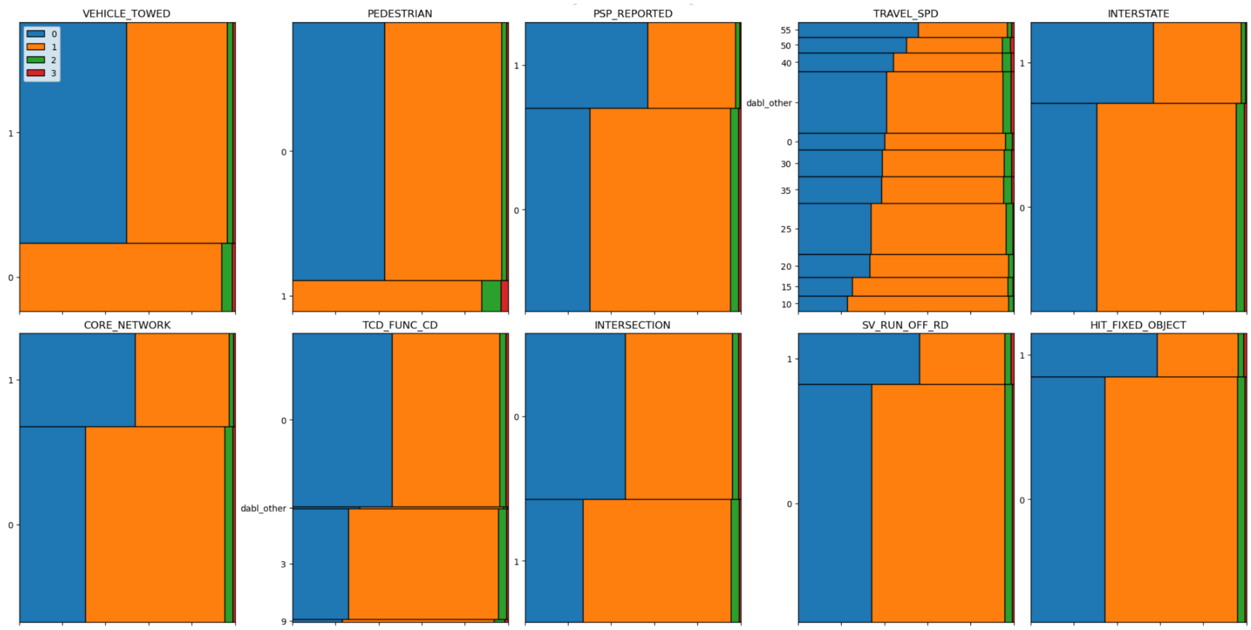


Figure 6: Plots of categorical features vs Target Variable (‘MAX_SEVERITY_LEVEL’)

We plotted some variables with interesting trends against maximum severity level.

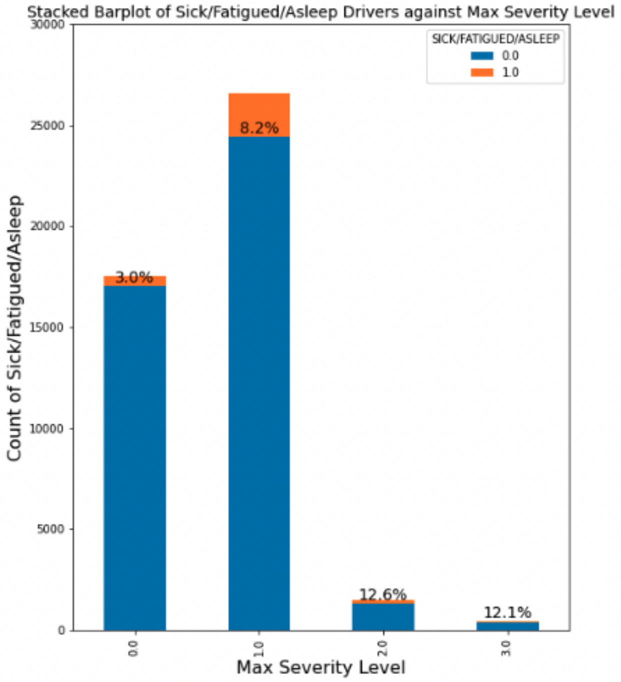


Figure 7: Stacked barplot showing number of Sick/Fatigued/Asleep drivers against Maximum Severity Level

Crashes of higher severity levels are shown here to have a higher percentage of drivers who are sick, fatigued or asleep. The alertness of drivers has been shown to be negatively correlated with the severity level of a car crash. Earlier prevention can potentially reduce the incidence of related injuries by up to about 20% or 30%. (Abdulbari Bener et al., 2017).

We recommend the Philadelphia traffic police to increase education of road safety and the importance of alert driving, through campaigns and initiatives. A possible example to emulate would be the Drowsy Driving Prevention Week organized by the National Sleep Foundation, which seeks to remind drivers of the dangers of driving when fatigued (*National Sleep Foundation,* 2023). This would likely reduce the number of unwell and tired drivers on the road and lower the severity of crashes.

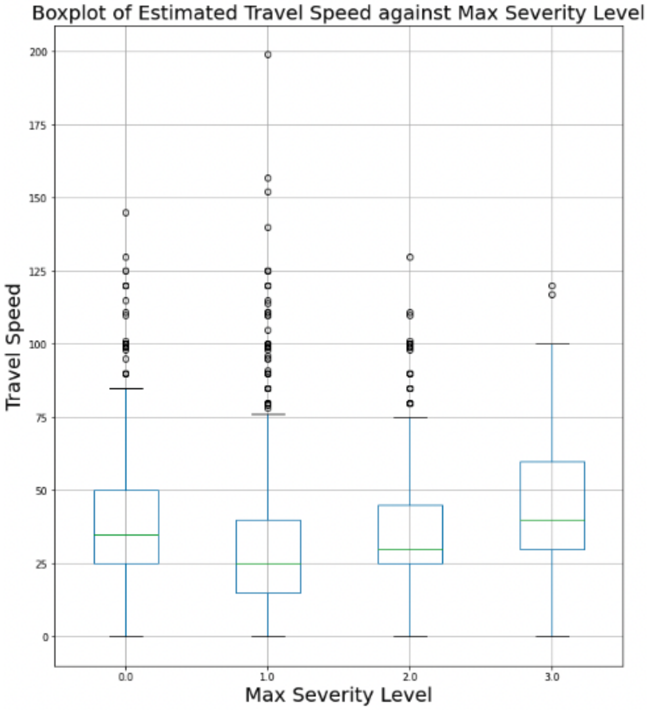


Figure 8: Boxplot of Estimated Travel Speed against Maximum Severity Level

In Figure 8, it can also be seen that travel speed is somewhat correlated with severity level of a crash. The increase in severity levels of crashes from 1 to 3 can be associated with higher median travel speed in vehicles.

Therefore, implementing more stringent speed limits could prove to be a highly effective approach towards mitigating the severity of crashes. In Toronto, a study found that on streets where speed limits were lowered from 40 kph to 30 kph, there was a 67% decline in the number of fatal and serious injuries sustained in car crashes (Fridman L. et al., 2020). As such, it is recommended that traffic police take proactive measures to identify areas with a high incidence of speeding offenses and consider reducing speed limits in such locations, provided that it is a feasible option.

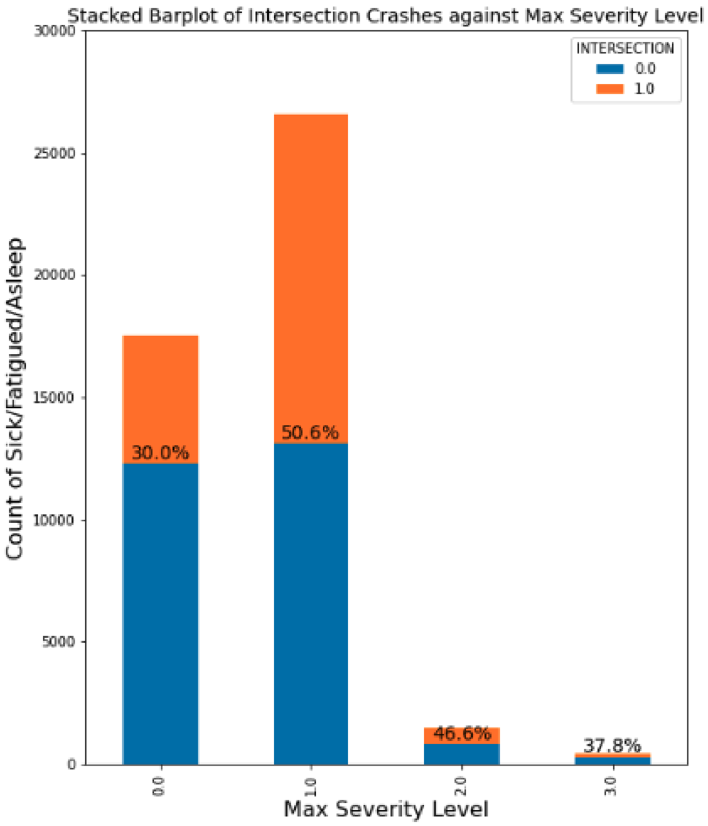


Figure 9: Stacked Barplot of Crashes at Intersections against Maximum Severity Level

For crashes that took place at an intersection, there is a higher percentage of crashes with severity of at least minor injury as compared to crashes that resulted in just property damage, showing that more severe crashes are likely to occur at intersections. This is supported by research that unsignalized intersections or traffic violations at intersections can increase severity of crashes (Adanu E.K., 2021).

As such, a possible recommendation is to increase police presence at high-risk intersections. Identifying high-risk intersections with a history of severe crashes and increasing police presence in those areas can help deter risky driving behaviors and reduce the occurrence of severe crashes. Traffic police should prioritize patrolling and monitoring high-risk intersections to enforce traffic laws and ensure compliance. Future work can include identifying these high-risk intersections by utilizing spatial data or additional datasets.

VII. Conclusion: Insights & Recommendations

To recapitulate, we predicted the maximum severity level of crashes using four different machine learning models to explore the impact of different factors in the severity of crashes in Philadelphia. Using more data from external datasets can possibly improve these results.

We can conclude that more complex machine learning models are more suitable for such classification, likely because they are more robust to noise and outliers in the data, and can also capture a wider range of patterns and relationships. However, we note that these complex models are still disadvantageous in terms of computational resources used.

Since we have identified some features that influence the severity of the crash, we recommend the Philadelphia traffic police to work on reducing the presence of these factors, specifically to increase education of road safety, impose stricter speed limits and increase police presence at high-risk intersections.

VIII. Limitations & Future Work

1. Assumptions used in handling of missing data
   During data cleaning, some null values were deleted and others were assigned to a numerical value representing 'Unknown'. It is thus assumed that these values are unknown as we did not conduct data imputation.
   Moreover, keeping 'Unknown' values in the dataset could have had negative implications on the test accuracy because this could introduce bias or noise in subsequent analyses. Future work can include attempting data imputation, to reduce potential biases.

2. Possible overfitting
   During feature selection, if the number of decision trees in the random forest algorithm used is too high, this could result in selection of features that are important only in the training data and not in new, unseen data. Future work can include the use of ensemble methods such as bagging and boosting to reduce overfitting.

3. Balancing of dataset
   Due to time constraints, we did not use SMOTE (Synthetic Minority Over-sampling Technique) to rebalance this large dataset, which could have helped improve prediction of severity levels, especially given its effectiveness in addressing class imbalance when the minority class is significantly underrepresented and needs to be oversampled.

References

Adanu, E.K., Li, X., Liu, J., Jones, S. (2021) "An Analysis of the Effects of Crash Factors and Precrash Actions on Side Impact Crashes at Unsignalized Intersections", Journal of Advanced Transportation, vol. 2021, Article ID 6648523, 17 pages, 2021. https://doi.org/10.1155/2021/6648523

Bener, A., Yildirim, E., Özkan, T., Lajunen, T. (2017) Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study, Journal of Traffic and Transportation Engineering (English Edition), Volume 4, Issue 5. 2017. Pages 496-502, ISSN 2095-7564, https://doi.org/10.1016/j.jtte.2017.07.005.

Drivers are Falling Asleep Behind the Wheel. (2023). Retrieved from: https://www.nsc.org/road/safety-topics/fatigued-driver

Eboli, L., Forciniti, C., Mazzulla, G. (2020) Factors influencing accident severity: an analysis by road accident type, *Transportation Research Procedia*, Vol 47, 2020, Pages 449-456, ISSN 2352-1465, https://doi.org/10.1016/j.trpro.2020.03.120.

Fridman, L., Ling, R., Rothman, L. et al. Effect of reducing the posted speed limit to 30 km per hour on pedestrian motor vehicle collisions in Toronto, Canada - a quasi experimental, pre-post study. BMC Public Health 20, 56 (2020). https://doi.org/10.1186/s12889-019-8139-5

Keesing, A. (2020). Balanced accuracy score in TensorFlow [Answer]. Stack Overflow. https://stackoverflow.com/a/59943572

Philadelphia Vision Zero Annual Report. (2022). Retrieved from: https://visionzerophl.com/.