California Polytechnic University SLO

Final Group Project

Team 1: Car Crash

Will Kapner, David Greco, Ryan Neely, Ethan Schultz

GSB 530 Data Analytics and Mining

Dr. Leida Chen

9 December 2024

The "Car Crash" dataset provides a wealth of information that can be leveraged to improve road safety, inform policy decisions, and enhance resource allocation strategies. In this project, we will systematically analyze this dataset to extract actionable insights and support data-driven decision-making. Our approach will follow a structured process, beginning with Business Understanding, where we will identify the opportunities the data presents and formulate relevant business questions. Next, in the Data Understanding phase, we will explore the dataset using descriptive analytics tools, examine the potential for both supervised and unsupervised analysis techniques, and identify key variables for further investigation.

In the Data Preparation phase, we will perform the necessary data wrangling and preparation tasks to ensure the data is clean, complete, and ready for analysis. This will set the stage for the Modeling phase, where we will implement appropriate analytical techniques, justify our selections, and present the results of our analysis, including identifying the most effective models.

Following this, in the Evaluation phase, we will revisit the business objectives defined at the outset and assess whether the models have successfully addressed them. Based on this evaluation, we will formulate actionable recommendations rooted in the findings. Finally, in the Deployment phase, we will communicate the insights and recommendations through a comprehensive written report designed for a managerial and non-technical audience. This report will incorporate statistical information and visualizations to effectively convey the findings and prioritize actionable strategies for improving road safety and other business goals related to motor vehicle accidents.

Business understanding:

The primary goal of analyzing the "Car Crash" dataset is to identify the factors that contribute to severe car crashes on roadways, with a focus on improving road safety and reducing fatalities. By understanding which variables influence the likelihood of severe accidents, law enforcement, insurance companies, and public safety organizations can take more targeted actions to reduce car crash-related injuries and deaths.

One key opportunity lies in traffic safety improvements. Identifying patterns and trends in accidents, such as common violations, crash types, and weather conditions, can help develop targeted traffic safety initiatives. To explore this opportunity we need to ask ourselves which of the variables within the data set are linked to the most severe car crashes. Once we understand this linkage we can provide actionable insights if there are types of accidents that are far more common than others.

Another insight we aim to provide is how emergency response teams can prioritize cases when multiple accidents are reported simultaneously. By analyzing crash types and violation categories, we hope to identify key factors that predict the severity of an accident, allowing responders to focus resources on the most critical situations. To address this, we will explore the question: How can crash type influence emergency response priorities?

Traffic enforcement and policy reform can also benefit from accident data. Law enforcement and policymakers can analyze violations, such as speeding or driving under the influence, that are frequently linked to severe accidents. This information can guide stronger enforcement strategies, such as increased fines or sobriety checkpoints in high-risk areas. To explore this we can ask: How can law enforcement and policy improve road safety by focusing on high risk driving behaviors?

We can also explore the question of “Are severe car crashes more common on public streets or highways”? This can show public services where they might need to allocate resources for overall safety on the road if accidents are skewed to one vs the other.

We also believe that weather plays a crucial role in shaping road safety measures. To explore this, we will analyze the provided weather data to determine its correlation with accident severity. By asking questions such as, "How do clear versus inclement weather conditions affect the severity of accidents?" We can help public services prepare and implement strategies, such as enhanced road maintenance or increased safety measures, to mitigate accident severity during adverse weather conditions.

Data Understanding:

The Data Understanding phase involves exploring the "Car Crash" dataset using descriptive analytics tools to extract meaningful insights and provide a foundation for further analysis. This step examines the possibility of applying supervised and unsupervised learning techniques, identifies potential variables for analysis, and aligns these efforts with the business opportunities and questions defined earlier.

The dataset includes several variables with diverse types and purposes. For instance, the ID variable, a categorical identifier, uniquely distinguishes each record but is not used for analysis. Geographical variables such as County and City categorize accidents by location, enabling regional analysis. Weekday and Month, quantitative variables, capture temporal information, while Severity, a binary variable, categorizes accidents as severe (1) or non-severe (0). Other categorical variables include ViolCat, which lists specific traffic violations, and CrashType, describing the nature of the crash. Environmental factors such as ClearWeather, Highway, and Daylight are binary indicators assessing the impact of weather, road type, and lighting conditions on accidents.

Key descriptive metrics like mean, median, and mode will help summarize quantitative variables such as Weekday and Month, while frequency distributions will analyze categorical variables like ViolCat and CrashType. Cross-tabulations, such as Severity vs. ClearWeather, will reveal relationships between variables. Visualization tools like histograms, bar charts, and pie charts will illustrate trends and distributions across variables.

In exploring analysis techniques, supervised learning methods will address specific questions, such as predicting accident severity or crash type. For example, a binary classification model using predictors like ViolCat, ClearWeather, and Daylight could estimate the likelihood of severe accidents. Algorithms such as logistic regression, decision trees, and random forests may be applied, with evaluation metrics like accuracy, precision, and recall ensuring robust model performance. Similarly, supervised learning models like multinomial logistic regression could predict crash types using features such as weather conditions, violations, and time of day.

Unsupervised learning methods will uncover hidden patterns and groupings in the data. Clustering algorithms like K-means can identify distinct accident scenarios, while DBSCAN might detect areas of high accident density. Dimensionality reduction using principal component analysis (PCA) will identify overarching patterns and correlations among variables like CrashType, ViolCat, and ClearWeather.

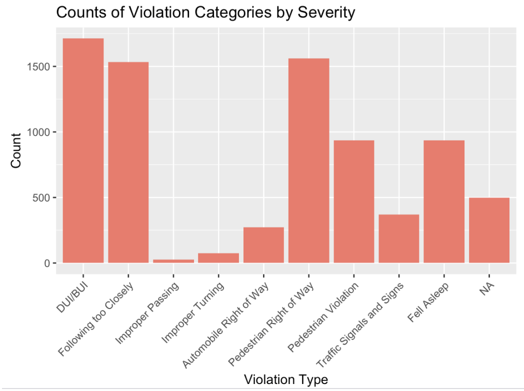
Based on the findings, variables such as Severity, ViolCat, ClearWeather, CrashType, Weekday, Month, Highway, and Daylight emerge as key candidates for further analysis. These variables align with the business questions and provide a strong basis for modeling and identifying actionable insights in subsequent phases.

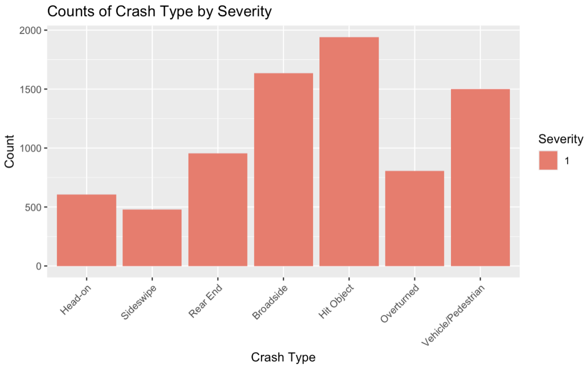
Data Preparation:

In the Data Preparation phase, we focused on preparing the "Car Crash" dataset for analysis, ensuring all variables were in a suitable format for modeling and interpretation. While the dataset was already clean, additional transformations were necessary to make it analysis-ready, particularly due to the categorical nature of most variables.

To enable effective modeling, categorical variables such as ViolCat, CrashType, County, and City were transformed into dummy variables using one-hot encoding. This step was crucial for algorithms that require numeric inputs, as it converted each category within these variables into a binary (0 or 1) format. Binary variables such as Severity, ClearWeather, Highway, and Daylight were already numeric and required no further transformation. Quantitative variables, including Weekday and Month, were left in their existing format as discrete numeric values, ready for direct use in the analysis.

By dummifying the categorical variables and maintaining the integrity of numeric variables, we ensured the dataset was structured for both supervised and unsupervised learning techniques. This preparation phase allowed us to move forward confidently into the modeling phase, knowing the data was in a working form suitable for analysis.



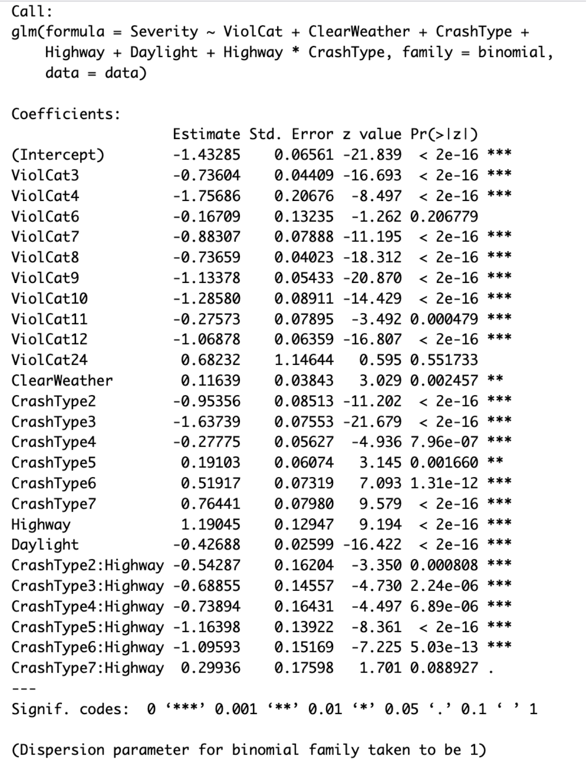


To begin analyzing our data, we made bar charts for the severity of the car crashes based on both driving violation type and car crash type. This gave us an initial idea of which variable values resulted in the most severe car crashes, and if there are extremely large or small amounts for any particular value. By looking at these graphs, we can improve our interpretations and gain general insight about the factors contributing to severe car crashes.

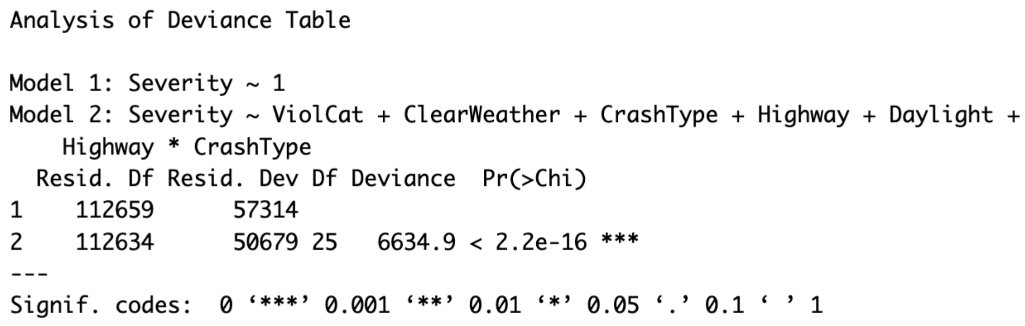
Modeling:

In the Modeling phase, our primary objective was to predict Severity, the target variable, based on key predictors such as Daylight (time of day), ViolCat (type of driving violation), and Highway (location type). To achieve this, we implemented three supervised learning models: Logistic Regression, Naive Bayes, and a Classification Tree. Each model was designed to analyze the predictors and their influence on crash severity.

We explored a Logistic Regression Model to better understand the interaction effects between Highway and each crash type (ViolCat). This model allowed us to quantify the impact of crash types on severity while accounting for whether the crash occurred on a highway. The coefficients, along with their corresponding p-values, revealed significant interactions for specific crash types, such as DUI and unsafe speeds, which were more likely to result in severe outcomes when occurring on highways.



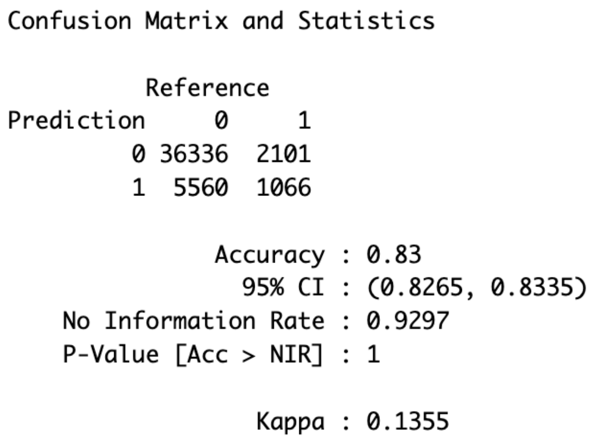
```
Call:
glm(formula = Severity ~ ViolCat + ClearWeather + CrashType +
    Highway + Daylight + Highway * CrashType, family = binomial,
    data = data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.43285    0.06561 -21.839  < 2e-16 ***
ViolCat3           -0.73604    0.04409 -16.693  < 2e-16 ***
ViolCat4           -1.75686    0.20676  -8.497  < 2e-16 ***
ViolCat6           -0.16709    0.13235  -1.262 0.206779    
ViolCat7           -0.88307    0.07888 -11.195  < 2e-16 ***
ViolCat8           -0.73659    0.04023 -18.312  < 2e-16 ***
ViolCat9           -1.13378    0.05433 -20.870  < 2e-16 ***
ViolCat10          -1.28580    0.08911 -14.429  < 2e-16 ***
ViolCat11          -0.27573    0.07895  -3.492 0.000479 ***
ViolCat12          -1.06878    0.06359 -16.807  < 2e-16 ***
ViolCat24           0.68232    1.14644   0.595 0.551733    
ClearWeather        0.11639    0.03843   3.029 0.002457 ** 
CrashType2         -0.95356    0.08513 -11.202  < 2e-16 ***
CrashType3         -1.63739    0.07553 -21.679  < 2e-16 ***
CrashType4         -0.27775    0.05627  -4.936 7.96e-07 ***
CrashType5          0.19103    0.06074   3.145 0.001660 ** 
CrashType6          0.51917    0.07319   7.093 1.31e-12 ***
CrashType7          0.76441    0.07980   9.579  < 2e-16 ***
Highway             1.19045    0.12947   9.194  < 2e-16 ***
Daylight           -0.42688    0.02599 -16.422  < 2e-16 ***
CrashType2:Highway -0.54287    0.16204  -3.350 0.000808 ***
CrashType3:Highway -0.68855    0.14557  -4.730 2.24e-06 ***
CrashType4:Highway -0.73894    0.16431  -4.497 6.89e-06 ***
CrashType5:Highway -1.16398    0.13922  -8.361  < 2e-16 ***
CrashType6:Highway -1.09593    0.15169  -7.225 5.03e-13 ***
CrashType7:Highway  0.29936    0.17598   1.701 0.088927 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```



```
Analysis of Deviance Table

Model 1: Severity ~ 1
Model 2: Severity ~ ViolCat + ClearWeather + CrashType + Highway + Daylight + 
    Highway * CrashType
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1    112659      57314                          
2    112634      50679 25   6634.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

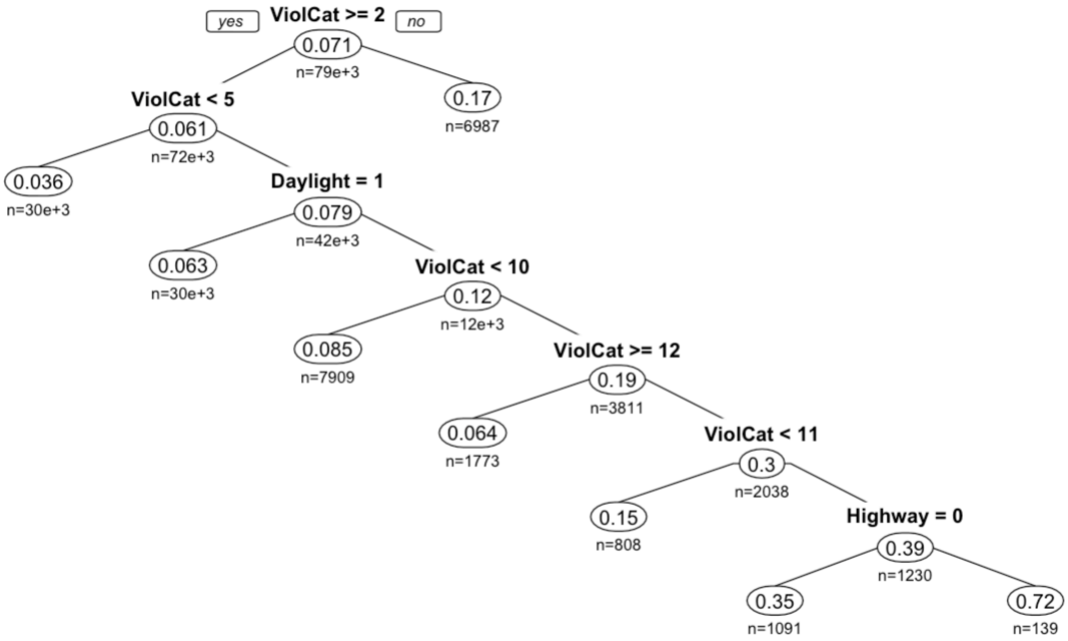
The p-values' significance levels were indicated using stars (*, **, ***), showing their statistical importance. Below is a table summarizing the estimates, standard errors, z-values, and p-values for the interaction terms. This logistic regression model provided additional depth by highlighting how specific violations and the highway environment interact to influence crash severity, offering actionable insights for targeted interventions.

We also implemented a Naive Bayes Model to classify crash severity, leveraging its simplicity and effectiveness for categorical data. This probabilistic model evaluated the likelihood of severity based on predictors such as ViolCat, Daylight, and Highway. The resulting confusion matrix highlighted the model's performance, showing a balanced classification across severe and mild crashes. With an overall accuracy of 83%, the Naive Bayes model proved effective in capturing the relationships between predictors and crash severity, although it occasionally struggled with classifying rare severe crashes due to their lower prevalence in the dataset.

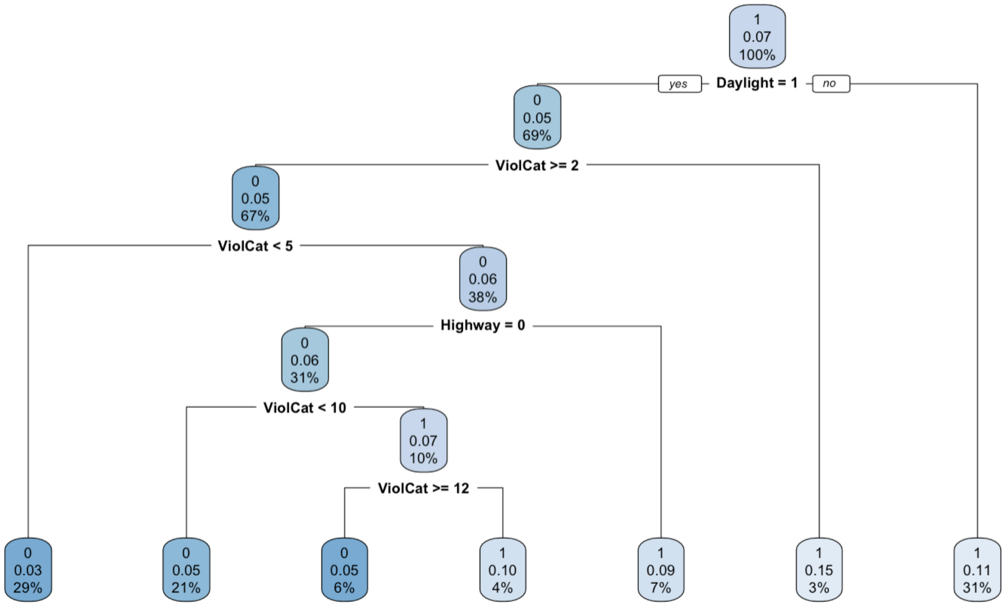


```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 36336  2101
         1  5560  1066

               Accuracy : 0.83
                 95% CI : (0.8265, 0.8335)
    No Information Rate : 0.9297
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1355
```

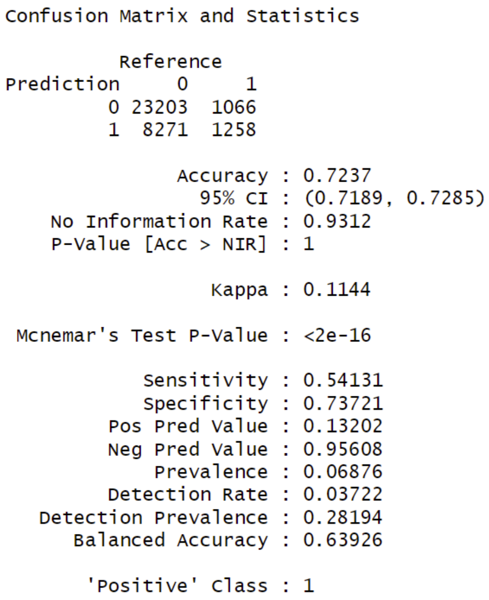
The Classification Tree (Below) provided the most informative insights into the predictors, allowing us to determine the relative importance of variables. Initially, the tree highlighted ViolCat as the most important predictor for determining crash severity, followed by Daylight and Highway. This made intuitive sense, as driving violations such as DUI and speeding are strongly associated with severe outcomes, while visibility conditions and highway environments contribute to crash dynamics.



However, the initial tree faced an issue: it misclassified severe crashes at a high rate. This was likely due to the imbalance in the dataset, with severe crashes being far less common than mild ones. As a result, the model defaulted to predicting the more frequent class—mild crashes. To address this, we adjusted the misclassification weights, assigning higher importance to false negatives (misclassifying severe crashes as mild) than to false positives. This adjustment improved the model's ability to correctly classify severe crashes. In this updated model, Daylight emerged as the most important variable, indicating that visibility conditions had a strong influence on severity, followed by ViolCat.

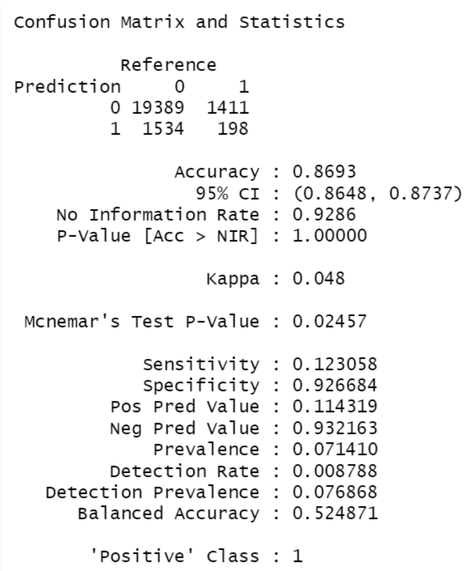


While this weighted model performed better with the class imbalance, we identified another issue: the Classification Tree treated ViolCat as a numeric variable rather than a purely categorical one. To correct this, we plan to transform ViolCat into dummy variables, ensuring the model accurately captures the categorical nature of this variable. This adjustment aims to further refine the model and improve its predictive performance.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 23203  1066
         1  8271  1258

               Accuracy : 0.7237
                 95% CI : (0.7189, 0.7285)
    No Information Rate : 0.9312
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1144

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.54131
            Specificity : 0.73721
         Pos Pred Value : 0.13202
         Neg Pred Value : 0.95608
             Prevalence : 0.06876
         Detection Rate : 0.03722
   Detection Prevalence : 0.28194
      Balanced Accuracy : 0.63926

       'Positive' Class : 1
```

The confusion matrix above is the result of our classification matrix after ensuring the ViolCat variable was treated as a categorical rather than numeric variable. We also opted to do a random forest approach which is better at handling unbalanced data sets. Originally, the data set being unbalanced was an issue with the results for our confusion matrix as nearly all of the observations were being classified as category 0. Using the random forest model we were able to weight the categories differently, giving us the results that are shown above. As shown in the output above, the accuracy of the random forest model was 0.72, the sensitivity was 0.54, the specificity was 0.73, and the precision was 0.13. Overall, this model was better than the other classification tree models at dealing with the unbalanced classes.



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19389  1411
         1  1534   198

               Accuracy : 0.8693
                 95% CI : (0.8648, 0.8737)
    No Information Rate : 0.9286
    P-Value [Acc > NIR] : 1.00000

                  Kappa : 0.048

 Mcnemar's Test P-Value : 0.02457

            Sensitivity : 0.123058
            Specificity : 0.926684
         Pos Pred Value : 0.114319
         Neg Pred Value : 0.932163
             Prevalence : 0.071410
         Detection Rate : 0.008788
   Detection Prevalence : 0.076868
      Balanced Accuracy : 0.524871

       'Positive' Class : 1
```

Lastly, we looked into classifying the severity of car accidents using a KNN classification model. To account for the class imbalance we used a weighted neighbors approach that used a K of 4 and Euclidean distance as the distance metric. Weighted KNN models are typically better for imbalanced datasets as it places a larger importance on closer neighbors. This makes the model have a higher likelihood of identifying rarer events. The results from this model can be observed in the confusion matrix above which shows an accuracy of 0.87, sensitivity of 0.12, specificity of 0.93, and precision of 0.11. Once again, the class imbalance remains an issue for the model when it comes to capturing the positive class of 1 for the severity variable.

Evaluation:

The main goal of analyzing this car crash data is to improve the safety of our road system by identifying factors that may lead to an increased probability of severe car crashes. To do this, it's important to understand that a number of factors contribute to car crashes, and each may have their own individual weights. Therefore, we need to determine which factors are the most important before we can think of ways to mitigate severe injuries caused by automobile collisions. From a business standpoint, the goal of every organization is to reduce the number of car crash deaths, but their motives may be different. A California Highway Patrol office may want to determine if there are specific times or road conditions where they should increase their surveillance to ensure drivers are following speed postings and DUI laws. However, an insurance company may want to know whether certain drivers are more at risk for critical injury crashes to be able to adjust rates beforehand. This data can benefit a multitude of people and groups, but our overarching goal is to improve the safety of our California roadways.

Tackling the question about how traffic safety can be improved will be the most important one we deal with because it directly relates to how we can improve the safety of people on the road. The two target groups for this question are education groups seeking to improve driving habits and law enforcement individuals trying to improve the safety of roadways for Californians including themselves. From our logistic regression results, we can see that the traffic violations such as drunk driving and driving at unsafe speeds are strong predictors of severe injuries during car crashes. To avoid these types of accidents, it would make sense for driver's education programs to emphasize avoiding these types of behavior and for patrol cars to be especially aware of these types of illegal behaviors. Similarly, there is also a strong relationship between highway and severity which signals that many of these severe car crashes occur on highways. This tells us that drivers should be extra safe on these types of roads as one mistake could have extreme consequences for themselves and any other passengers in the car.

Another factor to consider is the response time of emergency service individuals to crash scenes. In terms of crash type, certain types of crashes like vehicle-pedestrian are more likely to involve a severe injury which may give extra information to emergency service professionals and first responders. This can allow them to prioritize certain types of crashes to arrive at scenes first where there might be a critically injured person. These fast response times are critical for victims and may be the difference between life and death. Using models like the random forest classification tree can allow emergency service call details to be inputted into these models and be classified as likely to be severe or non-severe which will help prioritize cases in times of multiple calls.

Another factor to consider are the times that certain crashes occur. For example, on many holidays like the 4th of July, New Year's Eve, and Halloween, there tend to be more instances of car accidents, likely because more people are engaging in high risk behavior. Due to this cyclic behavior, this would be a strong indicator to increase patrols of police officers and emergency response workers to plan for and hopefully mitigate the increase of crashes. It may be in the best interest of other individuals who are not engaging in high risk behavior like driving under the influence of alcohol/drugs and speeding to do their best to avoid driving in these types of conditions. In our classification tree models, we can see that nighttime driving is significantly correlated with more severe car crashes.

We can also factor in the weather for our car crash decisions. In our logistic regression model, the clear weather variable has a positive coefficient and has a statistically significant p-value. We could take this to mean that education outlets are doing a good job teaching people about the dangers of driving in inclement weather. Therefore, there might be a need to switch the focus in driver's education type classes to more pressing issues like distracted driving and drunk driving. The weather coefficient was surprising, but we can still use this to extract information and make inferences to adjust our plans to reduce unsafe driving.

Conclusion:

Our models about car crashes can help us predict and prevent future crashes. It is impractical to try and eliminate crashes altogether, so we decided it was best to mitigate the severe car crash occurrences and try to limit crashes to parking lot bumps and fender-benders. By targeting predictor variables that have the biggest impact on causing crashes with severe injuries, we can offer ideas for solutions to reduce the amount of life-changing accidents. This will involve a combination of adjustments from police forces, emergency response teams, car companies, and driving education programs. By identifying factors that contribute to devastating crashes, we can take the first step to preventing them from happening in the first place.