**Exploring Road Traffic Accidents in Surrey: Data Mining and Sentiment Analysis Approaches**

Module : Data mining and text analytics

StudentID:6873905

Table of contents

1. Introduction
2. Data Exploration and Cleaning
3. Summary statistics
4. Explore and visualisation
5. Data cleaning strategy
6. Predicting Accident Severity
7. Text Analysis of Tweets
8. Summary and Recommendation

Introduction

Road traffic accidents are a global concern, posing significant challenges to public safety, transportation systems, and economic stability. This report examines road traffic accident data from Surrey, UK, utilizing advanced data mining and text analytics techniques within SAS Viya. The primary objective is to generate actionable insights from both structured and unstructured data, supporting data-driven decisions to improve road safety and minimize the occurrence and severity of accidents.

The dataset contains comprehensive details about personal injury road collisions in Surrey during 2021, encompassing variables such as accident severity, weather conditions, road types, speed limits, and other contributing factors. Additionally, social media data, specifically tweets related to traffic accidents in Surrey, is analyzed to uncover public sentiment and key themes regarding road safety perceptions and concerns.

# Data overview:

* **RoadAcci\_2021Surrey Dataset**
* Rows: 2480
* Columns: 35

This dataset contains structured information about road traffic accidents that occurred in Surrey, UK, during the year 2021. It includes several key attributes, categorized as follows:

1. Accident Details: Date, time, and location of the accidents which provide information about the incidents
2. Road and Environmental conditions: Variables such as road type, weather conditions, lighting conditions, and surface conditions, which are critical in understanding the external factors affecting accidents.
3. **Vehicle and Casualty Information**: Details about the number and types of vehicles involved, severity of accidents (e.g., slight, serious, or fatal), and casualty data such as age and gender.
4. **Data Quality**: The dataset appears comprehensive but requires cleaning, as it contain missing values.

* **Tweets Dataset**

The Tweets dataset consists of unstructured text data related to road traffic accidents. It provides a social perspective on road safety by capturing:

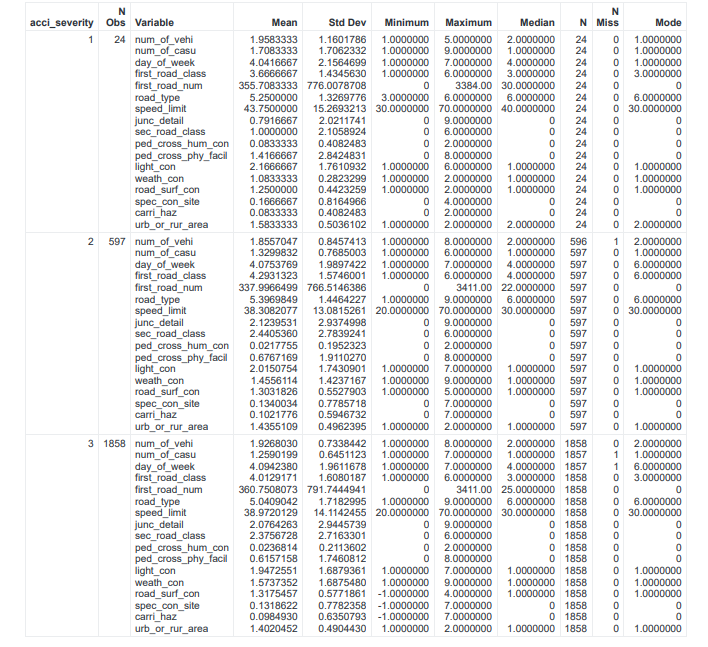
* Tweet Content: Textual descriptions of accidents or related discussions.
* Timestamp: Information about when the tweet was posted, helping to correlate public sentiment with accident occurrences.
* Potential Sentiment Indicators: Public emotions, concerns, and awareness regarding road safety, which can be extracted through sentiment analysis and text mining techniques.

Together, these datasets form a strong foundation for analyzing road safety trends, identifying risk factors, and understanding public awareness and sentiment about road traffic issues .

# Data Exploration and Cleaning

An exploratory data analysis (EDA) dataset was conducted to understand its structure, content, and quality. The dataset comprises numerous variables, including accident dates, times, locations, weather conditions, road types, and accident severities. The exploration shows that the data is well-organized but contains some quality issues, such as missing values in critical fields like number of vehicles and day of week. Summary statistics highlighted the most common accident types and severities, with the majority being classified as slight, followed by serious injuries. Visualization techniques, such as bar charts and scatter plots, identified temporal patterns, including a higher frequency of accidents during peak commuting hours and adverse weather conditions. The EDA process laid a solid foundation for identifying data quality issues and informed the development of a cleaning strategy to ensure reliable insights and accurate analysis.

# Summary statistics



The summary statistics provide an insightful breakdown of road accidents categorized by severity: slight (1), serious (2), and fatal (3). Following points will give us the overview of what is happening in the summary stat:

* Number of Vehicles and Casualties:
* Slight accidents usually involve about 2 vehicles and 2 people.
* Serious accidents involve fewer vehicles and people on average.
* Fatal accidents have similar vehicle numbers but fewer people involved compared to slight accidents.
  + Road and Area Types:
* Most accidents happen in urban areas, regardless of severity.
* Roads with speed limits of around 38–43 mph are common across all accident types.
* Weather and Light:
* Slight and serious accidents mostly occur in clear weather and during the day.
* Fatal accidents are slightly more likely to happen in bad weather or poor lighting.
* Differences Between Accident Types:
* Slight accidents show more variation in road conditions and junction types compared to serious or fatal ones.

Visualisation

The following visualizations aim to explore key factors contributing to road accidents, such as accident severity, road types and environmental conditions. These insights will help in understanding accident trends and proposing actionable recommendations.

1. Accident severity and road type distribution

A screenshot of a pie chart

Description automatically generatedA graph with blue rectangles

Description automatically generated

Fig 1 : Accident Severity Distribution (Donut Chart) Fig 2: Road Type Distribution (Bar Chart)

The visualizations highlight key insights into road accidents in terms of severity and road types. The majority of accidents (74.9%) are classified as slight in severity, with serious accidents accounting for 24.1% and fatal accidents making up only 1%. This indicates that while most incidents result in minor injuries, serious and fatal accidents still pose significant safety concerns. Additionally, the road type distribution shows that a specific road type (category 6) experiences over 1,500 accidents, far exceeding other road types, which have significantly fewer incidents. This shows that some road types might be more dangerous or have more traffic, so they should be the focus for improving safety.

1. Severity based on casualities/ vehicles and speed limit

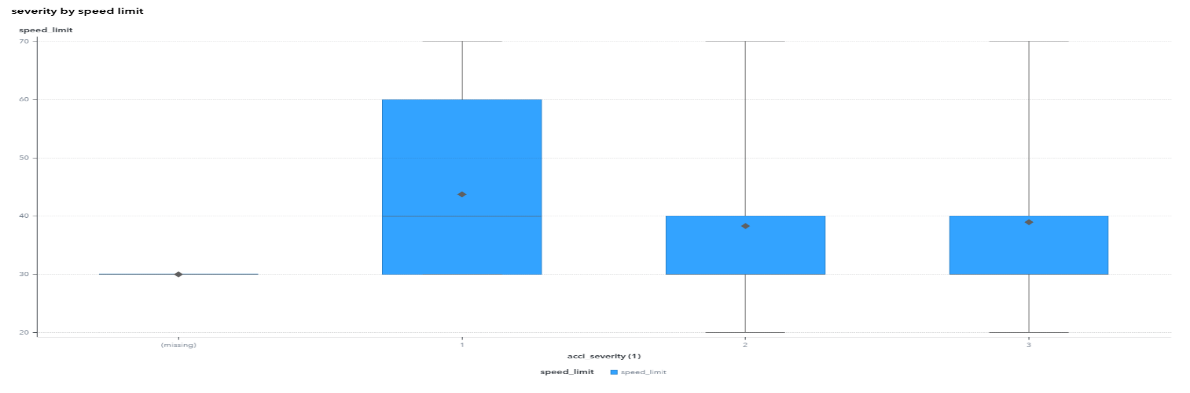


Fig 3: Accident severity based on speed limit

A graph with different colored squares

Description automatically generated   
 fig 4: accident severity based on number of casualities and vehicles

In the stacked bar plot chart, slight accidents (Severity 3) involve the most vehicles (3.6K) and casualties (2.3K), making them the most common type of accident. Serious accidents (Severity 2) involve fewer vehicles (1.1K) and casualties (794). Fatal accidents (Severity 1) are rare, with very low numbers of vehicles (47) and casualties (41).

The box plot chart shows that slight accidents happen across a wide range of speed limits, mostly around 40–50 mph. Serious and fatal accidents, however, are more common at lower speed limits, averaging closer to 30 mph. This suggests that more severe accidents often happen in areas like cities, where lower speed limits and busy traffic conditions contribute to their seriousness. These numbers highlight the need to focus on both high-traffic areas and urban zones to improve safety.

1. Analyzing weather and light conditions

A screenshot of a graph

Description automatically generated

Fig 5: Weather Condition Distribution by Urban and Rural Areas

A screenshot of a computer

Description automatically generated

Fig 7: Light Conditions Grouped by Accident Severity

The graphs show how light and weather conditions affect road accidents. Most accidents (1.8K) happen during the day, with 1.4K being minor and 435 serious, likely because there are more vehicles on the road during daylight. At night, accidents are fewer, with 420 happening on lit roads and 181 on unlit roads, but they tend to be more serious. For weather, clear conditions have the most accidents, with 49% in urban areas and 32% in rural areas, as more people drive in good weather. Rain causes around 5.5% of accidents in both urban and rural areas, while conditions like fog or snow are rare and cause less than 2% of accidents. This shows that most accidents happen in good weather and daylight, but dark and rainy conditions can lead to more serious crashes.

1. Analyzing based on time-based accident patterns and geographic accident severity distribution

A graph of different colored squares

Description automatically generated

Fig 8:Frequency of months grouped by time category

A diagram of a network

Description automatically generated with medium confidence

Fig 9: Road accidents severity across Easting and northing Coordinates

The first graph shows how accidents are distributed across the months of the year and grouped by the time of day (morning, afternoon, evening, and night). Accidents are most frequent in September, July, and June, with over 200 incidents in these months, likely due to increased road activity during summer. Afternoon and evening time slots see the highest number of accidents across all months, indicating peak traffic during these periods. Morning accidents are moderate, while night accidents are consistently the lowest throughout the year. This suggests that road safety efforts should focus on afternoon and evening times during active months to reduce accidents.

The second graph maps the geographic locations of accidents (using easting and northing coordinates) and their severity. Slight accidents dominate the map, represented by widespread purple points, indicating they occur across various locations. Serious accidents, shown in yellow, are scattered but less frequent, while fatal accidents, shown in blue, are rare and isolated to specific areas. This geographic spread highlights clusters of slight accidents in high-traffic zones and suggests that serious and fatal accidents occur in areas that may require targeted safety interventions. Together, these insights provide a detailed view of accident patterns over time and location, helping prioritize safety measures.

5 . **Analyzing geographic distribution of road accidents**

A map with many dots

Description automatically generated

This map shows where accidents happen and how severe they are. Each dot represents an accident, with the size of the dot showing the number of casualties and the color indicating the severity:

* Purple dots show slight accidents, which are the most common and spread across the region.
* Yellow dots represent serious accidents, which are less frequent but still scattered in different areas.
* Blue dots show fatal accidents, which are rare and occur in isolated spots.

The circled area focuses on a part of the region with a high number of accidents and casualties. This area was chosen because it has a lot of overlapping dots, indicating it is a hotspot for accidents. It’s likely a busy area with heavy traffic or a high population, which increases the chances of accidents.

This highlighted area is important because it shows where safety improvements are needed the most. Problems like crowded intersections, poor road conditions, or lack of clear signs might be contributing to the accidents here

DaTA CLEANING STRATEGY

Filtering rows

A diagram of a machine

Description automatically generatedA screenshot of a computer

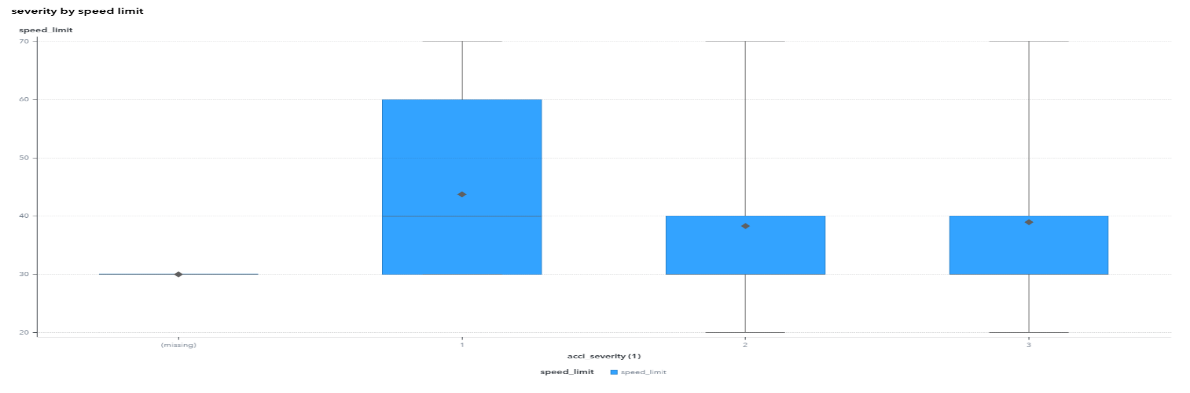
Description automatically generated

* To address the identified data quality issues, a data cleaning strategy was developed using the Filter Rows process. The goal was to ensure the dataset was clean, reliable, and ready for further analysis.
* Rows with missing or invalid values in critical columns removed to prevent errors and biases in the analysis.
* As a result, the total number of rows in the dataset was reduced from 3,523 to 3,510. This small reduction indicates that only a few records had incomplete data, ensuring that most of the dataset was preserved while improving its quality.
* The cleaned dataset, as shown, now contains only valid and complete records, making it reliable and ready for further analysis. This step ensures accuracy and consistency in subsequent modeling and insights.

Imputation :

* Variables with relatively less marginal missing values were cleared by imputation node , in which numerical variables were replaced with median . Missing values in categorical variables were replaced by mode .

Addressing outliers :



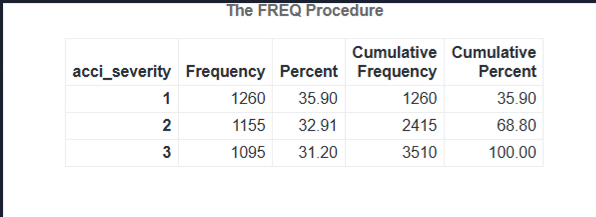
* The box plot above illustrates the distribution of speed limit values for different levels of accident severity. Each box represents the range of speed limits within which most accidents occurred for each severity level (1, 2, and 3).
* A few outliers are observed in these categories as well, but they are less extreme.
* Additionally, understanding these extreme cases provides insights into accident patterns under atypical conditions, which may inform preventive measures.

Imbalanced target variable

### 

* Addressing the balanced data involves ensuring that the target variable, such as accident\_severity, has an equal or proportional representation of all its categories.
* This is important because imbalanced data can cause predictive models to focus heavily on the majority class while ignoring the minority classes.
* By balancing the data, we create a fair representation of all categories, allowing the model to learn patterns effectively across all severity levels. This results in improved accuracy and fairness in predictions, especially for underrepresented categories, and ensures that all aspects of the data are considered equally during analysis and decision-making processes.

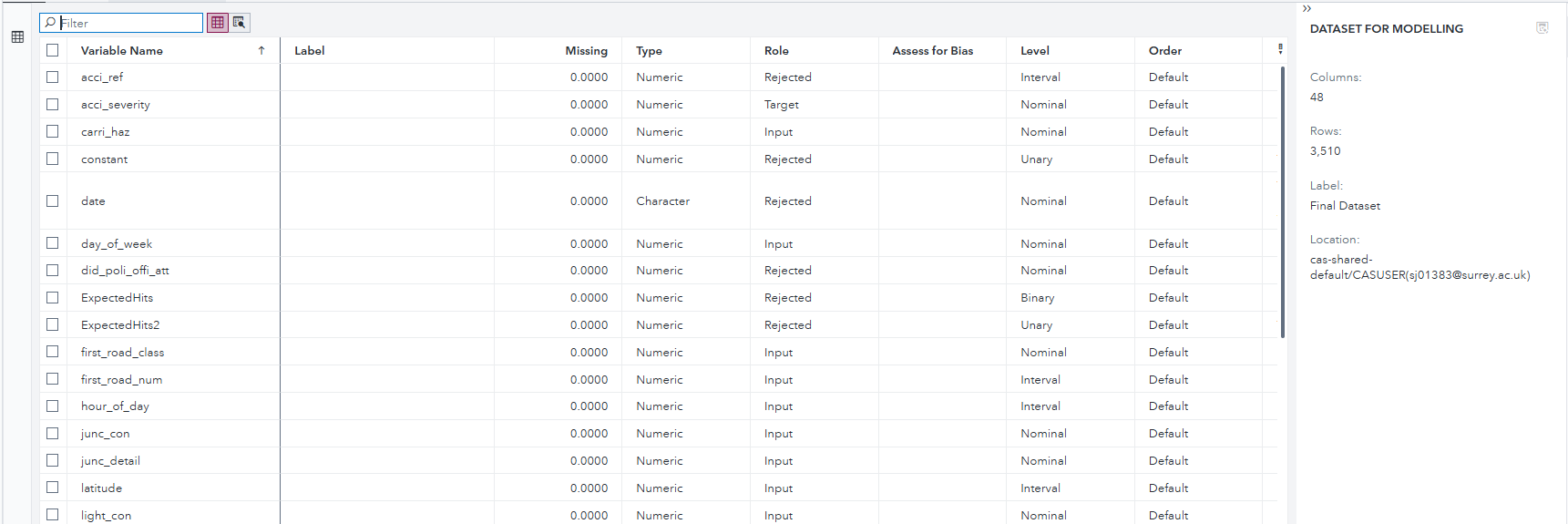
# Balancing the dataset



* The table above shows the frequency distribution of the acci\_severity variable after balancing the dataset. The goal of balancing is to ensure that each category in the target variable has a relatively equal representation, which prevents bias in the analysis or modeling process.
* In this case, the acci\_severity categories (1, 2, and 3) have frequencies of 1,260 (35.90%), 1,155 (32.91%), and 1,095 (31.20%), respectively. This ensures that all severity levels are fairly represented, allowing the model to learn equally from each category and make accurate predictions.
* By performing this step the data now is ready for further analysis for making the predictions using machine learning techniques .

# Predicting Accident severity

Data utilisation

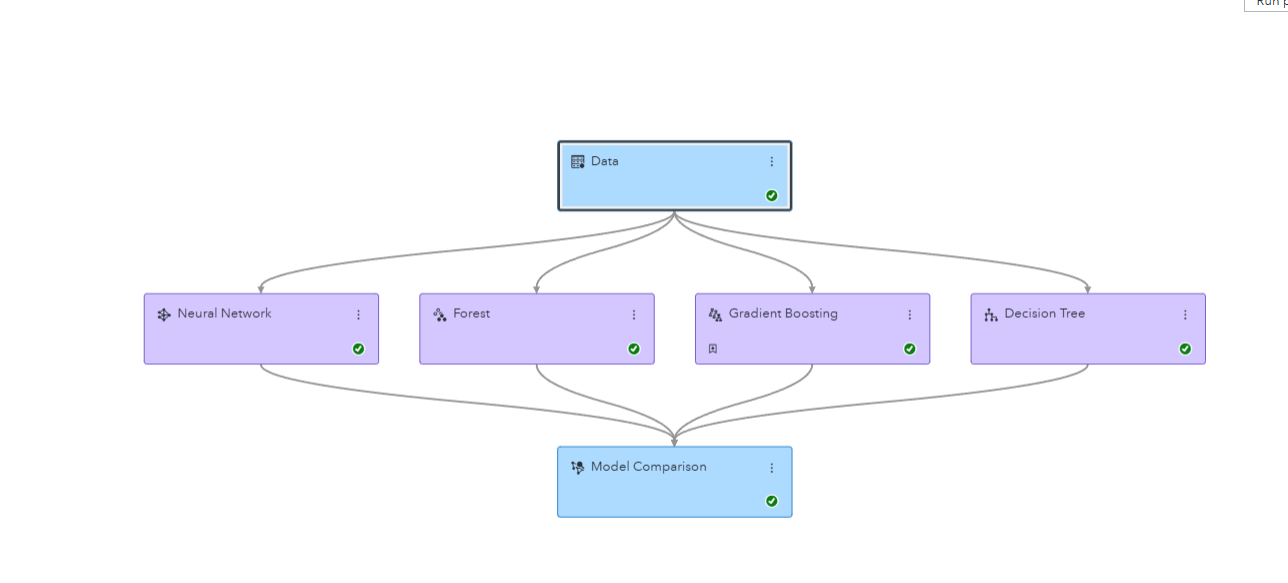


* Once the dataset is balanced , the columns and the rows gets boosted due addition of more data in order to cope up with imbalance . Now , we have the clean and balanced 48 columns and 3510 rows on which the machine learning models will be implemented . The new variables such as Expected hits , sampling weights are the example of the added variables .

Developing the Scenario

To begin the analysis, a real world scenario was made up . The focus was on understanding factors contributing to accident severity, which can help improve road safety and inform preventive measures. To achieve this, the most relevant variables were carefully selected based on their potential influence on accident outcomes. Under the observations , we got to know that accidents commonly occur during afternoon or evening hours in urban areas in wet road conditions in high risk locations as mentioned in geo-maps . This analysis gives importance to target specifically in urban hotspots and careful examination of weather and time -specific conditions to reduce accident occurrences .

**Model Building**



* Once the roles are assigned to each variable, in which Accident severity was set as the target variable , the dataset is ready for building up the pipelines . In our model we incorporated 4 supervised learning models ( Neural network , Forest , Gradient boosting , Decision trees) from which we found out the champion model to be the Gradient boosting .
* These models were selected because they are well-suited for handling structured datasets and can effectively capture complex relationships between variables.

Neural networks:

Purpose :

* Neural networks are selected because they can detect complex, nonlinear relationships between variables. They are well suited for datasets with multiple interactions, such as accident-related factors like speed limits, weather conditions, and casualties.
* Neural network Architecture

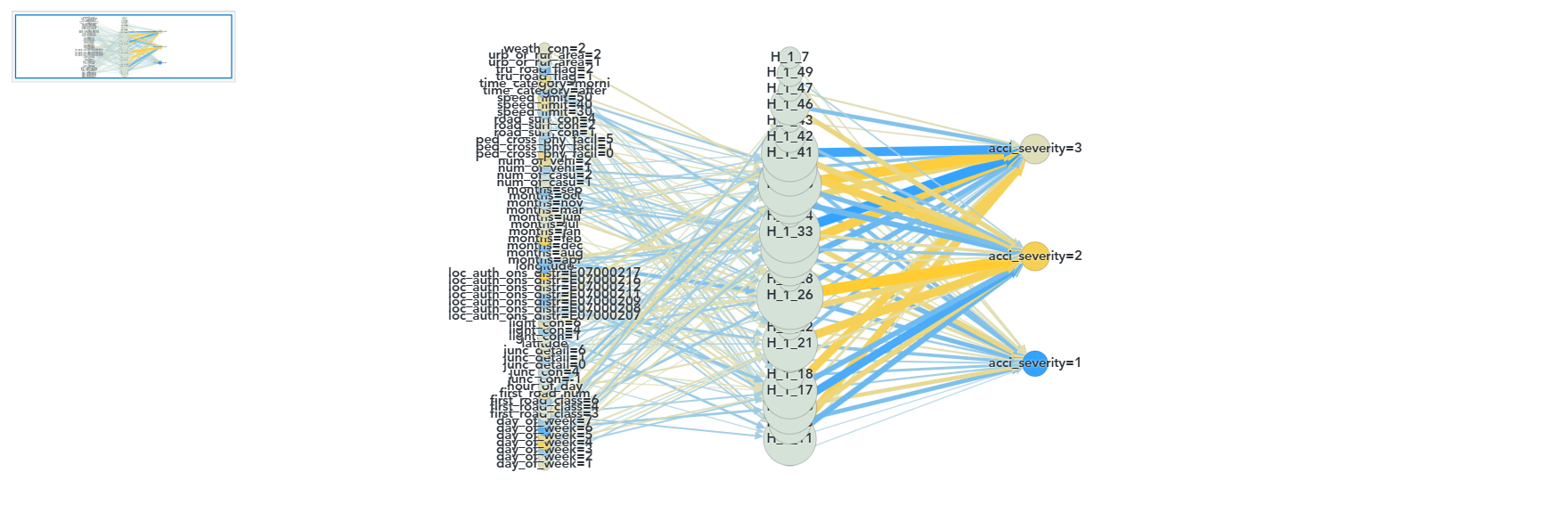


fig: neural network architecture

* The neural network used in this project consist of 3 layers , the input layer, hidden layer and output layer.
* The input layer takes in all the related variables from data sets which act as initial information .
* The hidden layer process all the input layer variables and captures all the patterns and relationships, each associated variable is assigned with weights that adjusts during training to improve accuracy .
* Finally , the output layer provides predictions according to severity.

**Evaluation metrics of Neural netwroks :**

Key strengths and benefits :

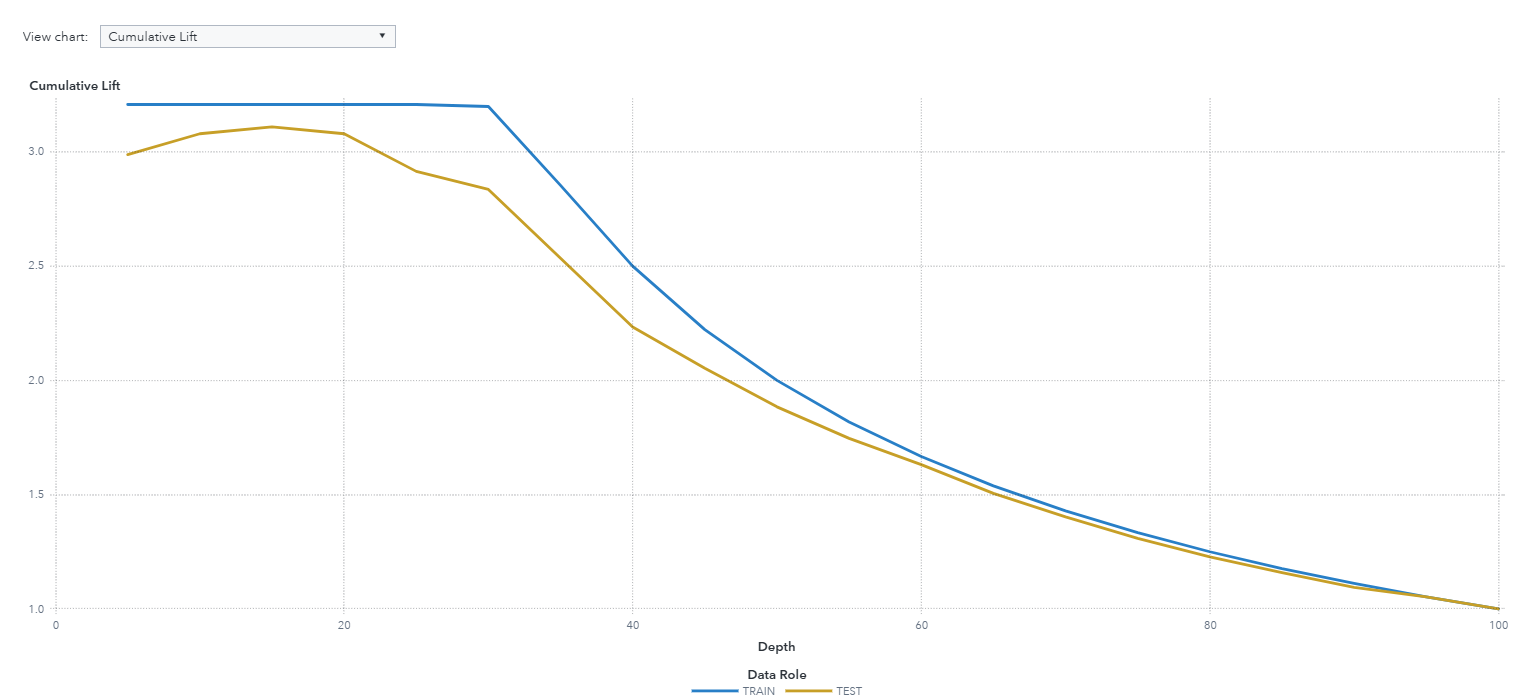


Fig : cumulative lift response

* At the top of cumulative response chart 10%-20% of the data, the cumulative lift exceeds 3, indicating that the model's predictions are three times more effective than random guessing in identifying critical cases. Thus , confirming that model outperforms from random guessing .

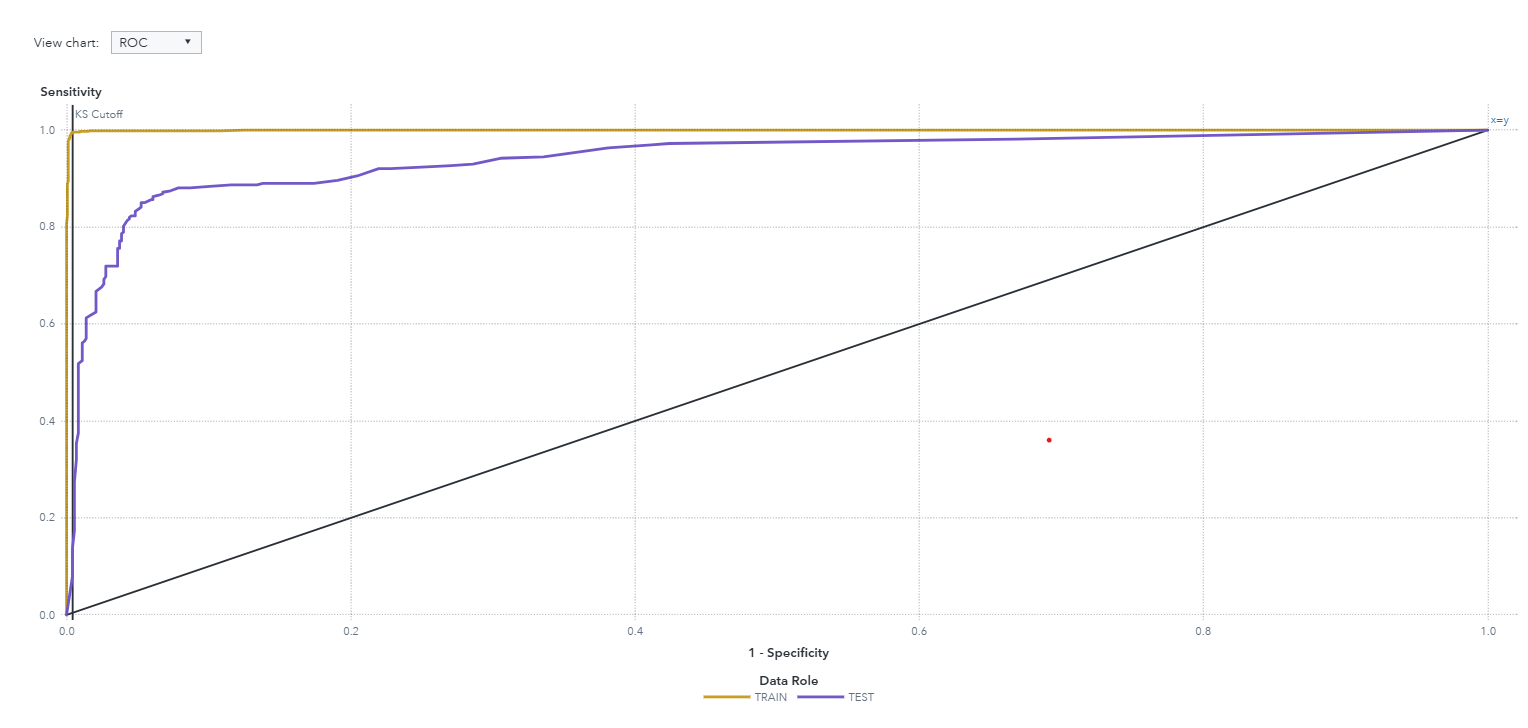
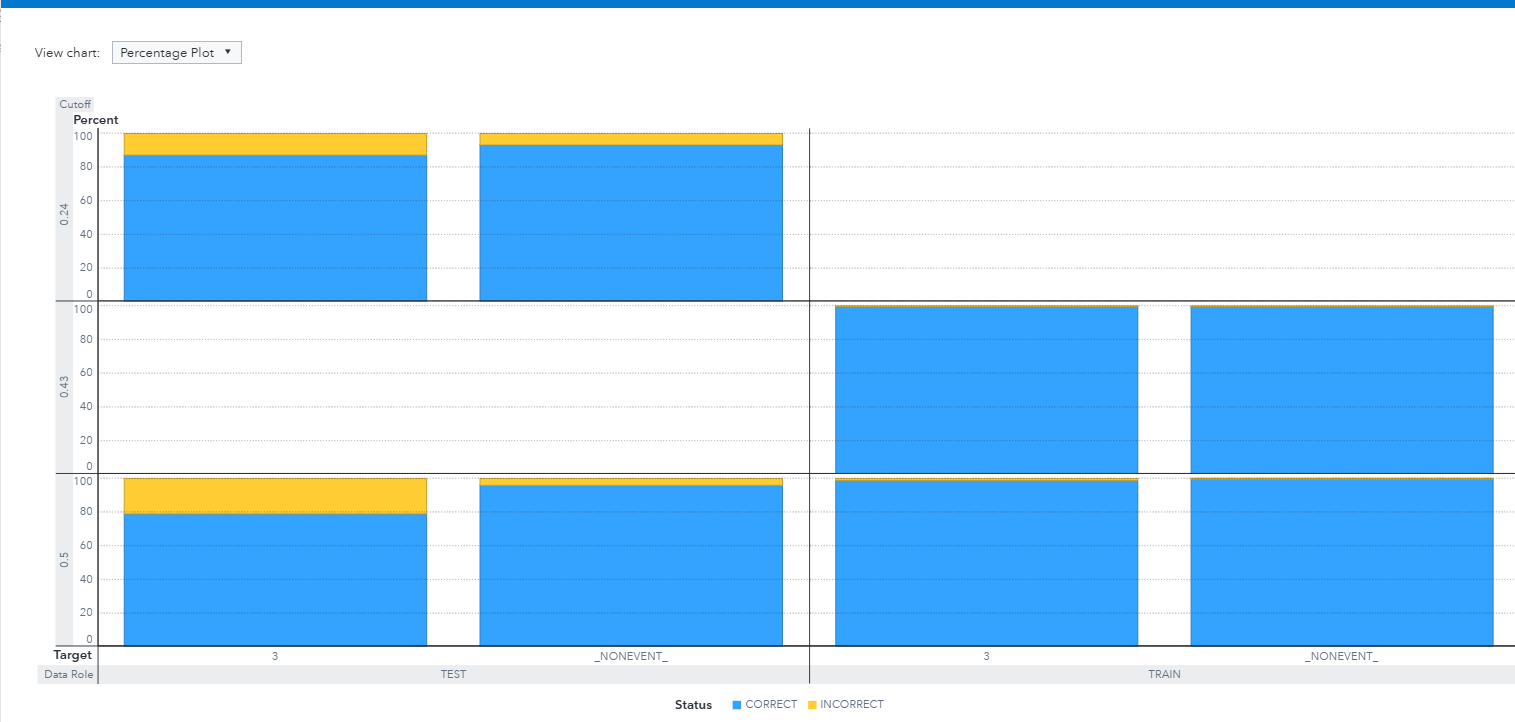


Figure X: ROC Curve Showing Model Performance for Train and Test Data

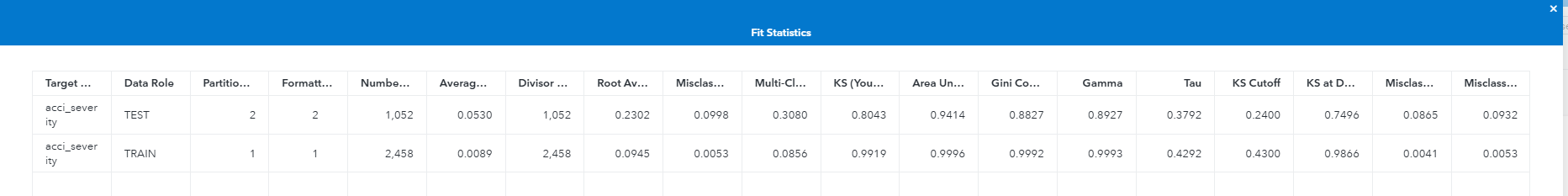
* The results show that the model significantly outperforms random guessing, as indicated by the curves being well above the diagonal baseline, which represents random performance.
* The AUC for test dataset is approximately 0.96 which is higher that baseline of 0.5 from random guessing

**RESULTS :**

Event classification :



* In the test data sets ,the majority of predictions are correct , with accuracy exceeding 90% for the severity level
* the small portion which are represented in yellow bars shows the incorrect predictions which is around 9.32% .
* similary , in training dataset , most of prediction accuracy are high , incorrect prediction contributing to just 0.53% .



* The fit statistic show that the test data achieves AUC of 0.9414, which tells us that it has strong predictive ability.
* KS statistic of 0.8 further confirms that model ability to differentiate between classes .
* Overall the results displays that model is robust ,reliable and capable of accurately picking and predicting in real world scenarios .

Gradient boosting (champion model):

Purpose :

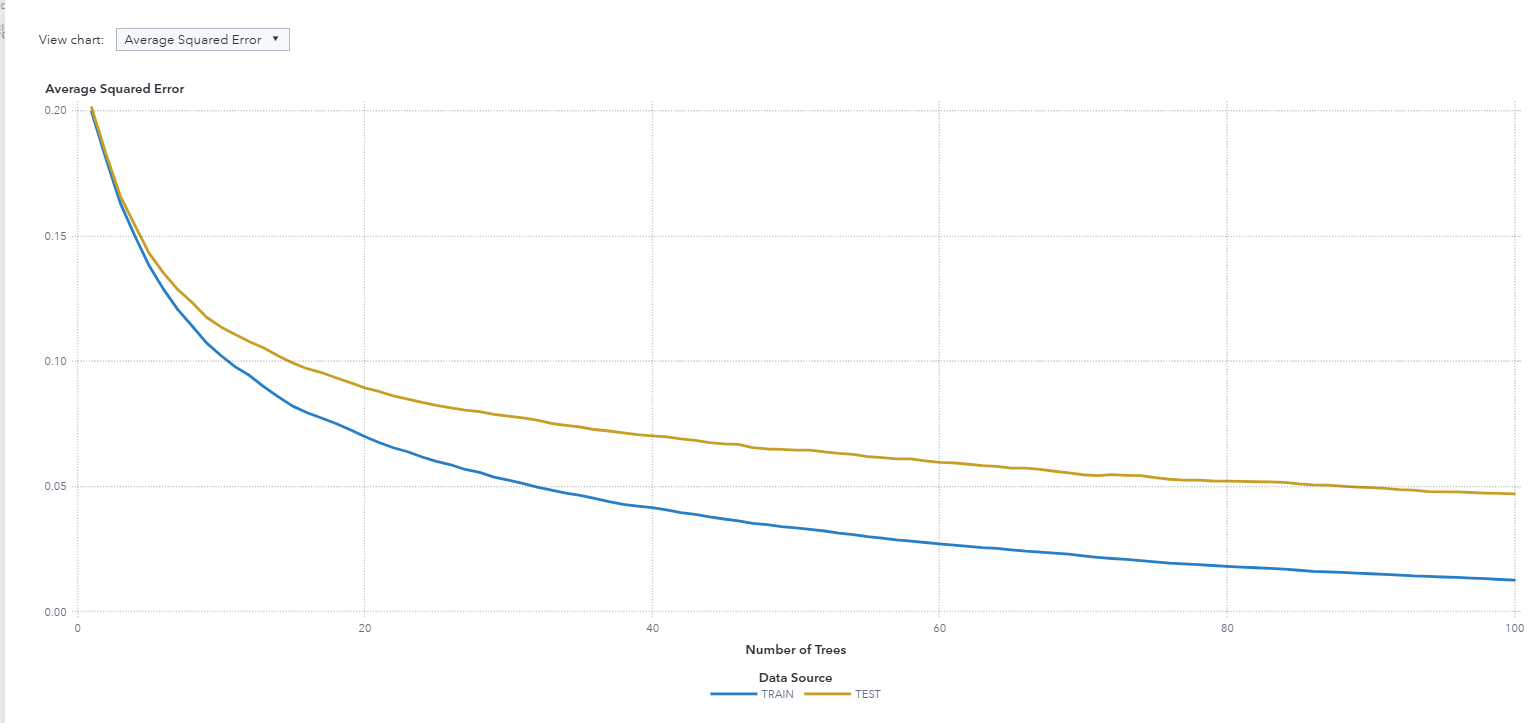
* Gradient Boosting for our dataset is highly effective at handling structured data and capturing complex patterns between variables. Our dataset involves multiple factors, such as speed\_limit, weather\_conditions, and accident\_severity, which interact in non-linear ways.

Its iterative learning pupose ensures that it minimizes errors by focusing difficult cases which results in improved prediction quality .

* Gradient Boosting is particularly useful for this type of analysis because it builds a series of small decision trees, where each tree focuses on correcting the errors of the previous ones. This step-by-step improvement allows it to create a strong model that can accurately predict accident severity.
* Gradient Boosting is useful for our dataset as it is robust to noisy data and provides high accuracy. Additionally, it can identify which features are most important for predicting accident severity, offering insights into which factors contribute most to accidents. This makes it a valuable tool not only for creating reliable predictions but also for understanding the key drivers behind accident severity, which can help in developing safety measures and policies.

evaluation of model :

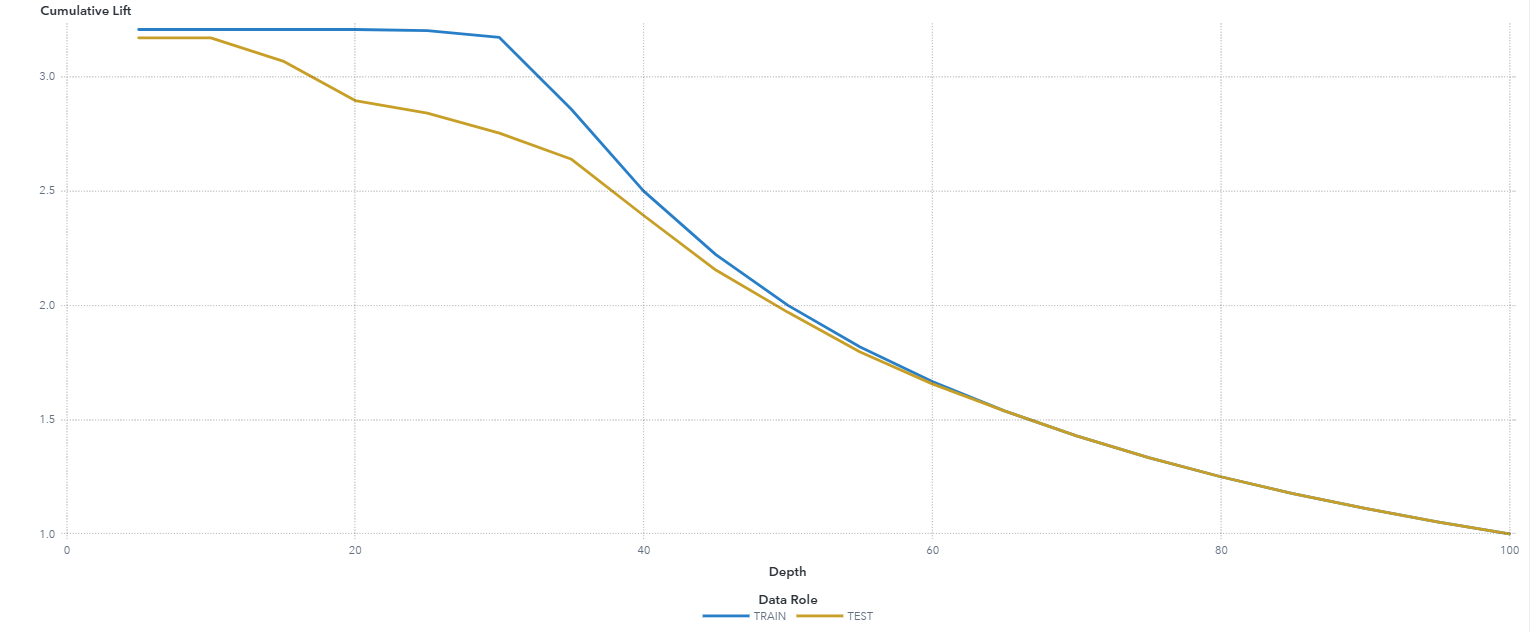
Key strengths and benefits :



* This plot shows how the average squared error changes as the number of trees in the gradient boosting model increases.
* For this model, the minimum error for the TEST partition is 0.047 and occurs for 100 trees, so the test error is still decreasing at the last tree.

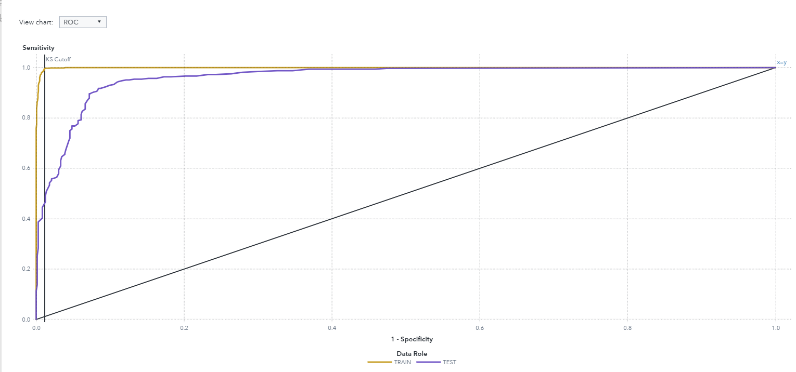
**results:**

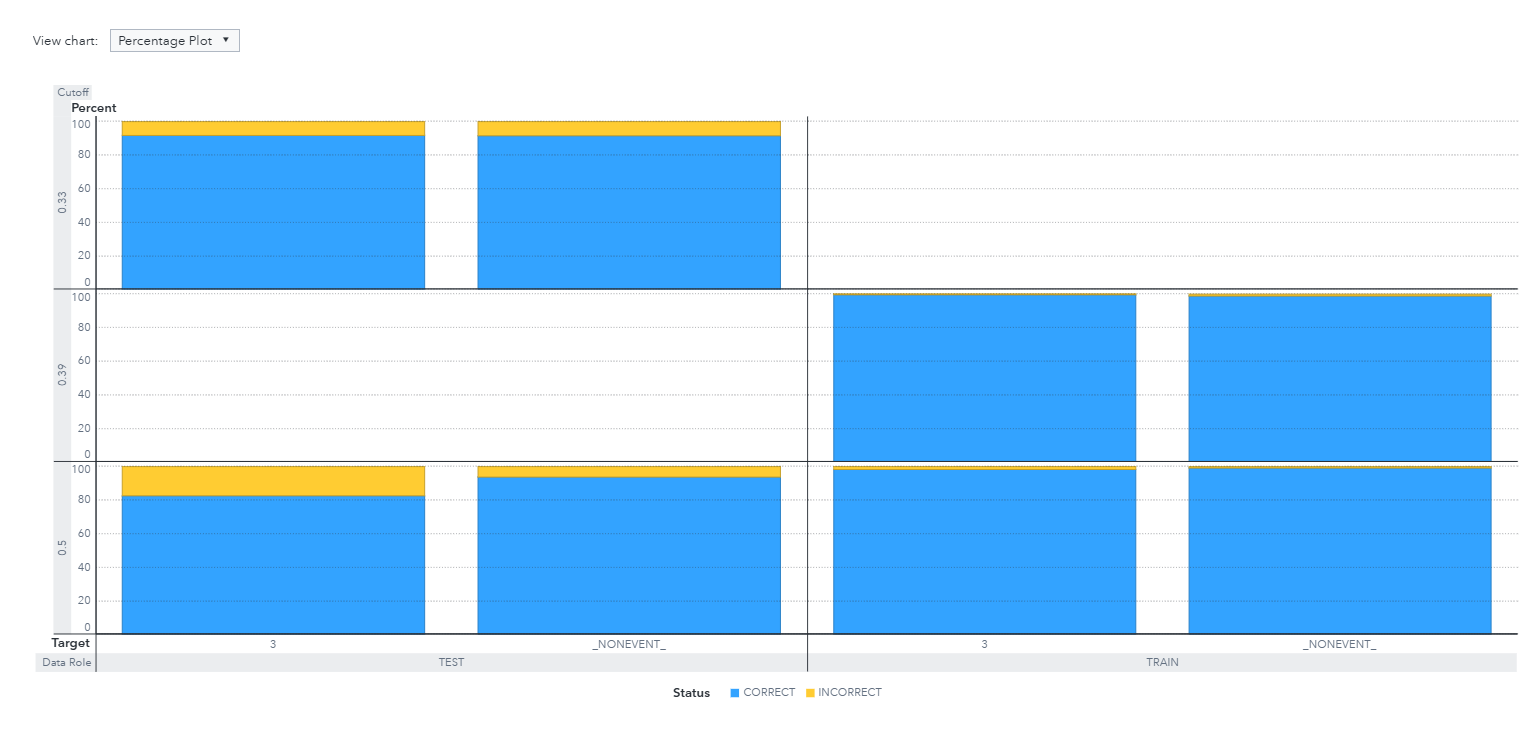
lift reports:



* The chart shows the Cumulative Lift for both the training (blue line) and testing (yellow line) datasets. The lift value measures how much better the model's predictions are compared to random guessing. At the top , the lift for the training dataset reaches over 3.0, indicating that the model is three times better than random predictions at identifying the target variable.

roc reports :





* The ROC Curve shows the model's ability to predict accident severity, with both training and testing curves well above the diagonal, indicating high accuracy. The Percentage Plot highlights over 90% correct predictions for both datasets, confirming the model's strong performance and reliability in classifying severity levels.

Fit statistics :

* The fit statistic show that the test data achieves AUC of 0.9621 confirming excellent model accuracy and KS of 0.83

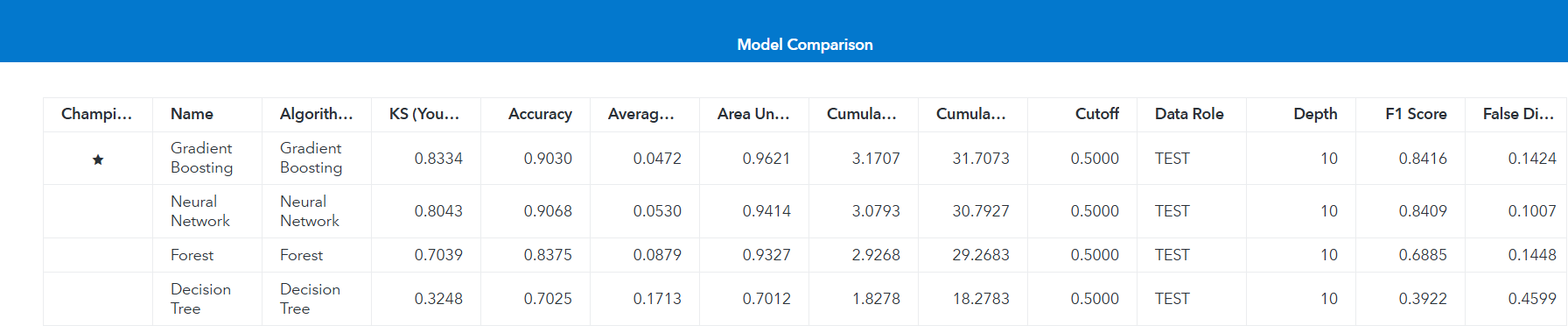
showing models ability to distinguish between classes .

**VARIABLE IMPORTANCE :**



* The feature importance analysis shows that Months and location are most critical variable in predicting accident severity showing us geographical factor and seasonal patterns .
* Followed by day of week , junction details and speed limits showing us the role of traffic patterns and road designs
* Secondary influencing factors like weather conditions and time category play supportive roles .

Model comparision :



The comparison between Gradient Boosting and Neural Network models shows their respective strengths and weaknesses.

Gradient Boosting stands out as the top-performing model, with a KS Statistic of 0.8334, an AUC of 0.9621, and a Cumulative Lift of 3.1707, demonstrating its strong predictive capabilities and consistent performance.

Neural Networks, on the other hand, achieve the highest accuracy at 0.9068 and a respectable AUC of 0.9414, showcasing their ability to capture complex patterns in the data.

Gradient Boosting has the advantage of being more interpretable and computationally efficient, while Neural Networks require higher computational resources and are more challenging to interpret. Both models effectively predict accident severity, but Gradient Boosting offers a more balanced and practical solution for this dataset.

**Recommendations for enhancing road safety based on analysis :**

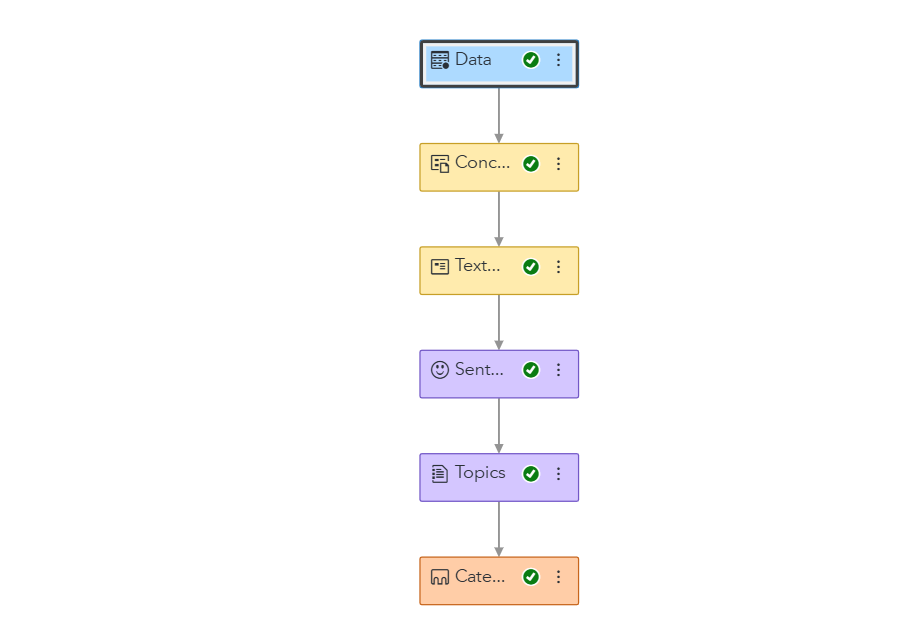
1. Top of Form
2. **Safety Measures in Accident-Prone Areas:** As we see from Geo-map visualizations it displays regions with high accident density. Authorities must concern about upgrading infrastructure, checking traffic monitoring, and ensuring traffic signs are placed at all dangerous turns.
3. **Speed Limit Adjustment:** Data analysis shows that a strong link between speed and accident severity. Administering stricter speed limits and placing speed cameras in high-risk areas can impact reducing severe accidents.
4. **Improve Lighting Conditions:** Facilitating poor lighting is a major contributor to accidents. Placing better street lighting, especially in rural and under maintained areas, can improve visibility and reduce the chances of collisions.
5. **WeatherReadiness:** Study shows that bad weather conditions are associated with higher accident rates. Authorities should raise driver awareness about weather-related risks and make sure roads are properly checked to prevent dangers like skidding.
6. **Timed measures :** Temporal analysis of accidents highlights peak times for incidents. Increased police patrols and targeted public awareness campaigns during these periods can help reduce accidents.
7. **Promote Good Driving Behaviors:** Long-term reduction in accidents can be achieved through public education campaigns that encourage responsible driving, the use of seatbelts, and avoiding distractions while driving.

# **Text analysis of the tweets**

data profiling :

* The dataset provided contains the tweets of accidents occurred in surrey state . The dataset consists of 598 text enteries . Contained in single column named as text .
* Each enteries shows a short tweets , mostly asserting to road accidents .All the datasets have non null values in the row enteries . The data type detected as object since the the data is textual. Also , the text includes , hashtags and mentions which related to specific events .

After exploring the data , we assign the roles for the variables involved , in which variable name text is given the role of “text” . When the role has been assigned , the pipeline will be created which is displayed as below :



Concept Node and Custom concepts :

One of the main and important features of text analysis is the custom node , it is powerful feature used in text analysis to categorize themes and ideas by grouping them . Its main feature is to extract patterns from the dataset .

For this kind of analysis I created 11 custom concepts tailored to the datasets unique characteristics .These custom concepts were designed based on the objective of the analysis . These variables are grouped on the basis of **Time , location , road , weather accident severity and sentiments** .

For example , in the accident severity , it is further classified as “serious , fatal , minor ,injury” . In negative sentiments it is classified as “stuck , delayed , blocked” . Similary for each events the related words are classified .