Exploring Road Traffic Accident Data and Text Analytics

Table of contents

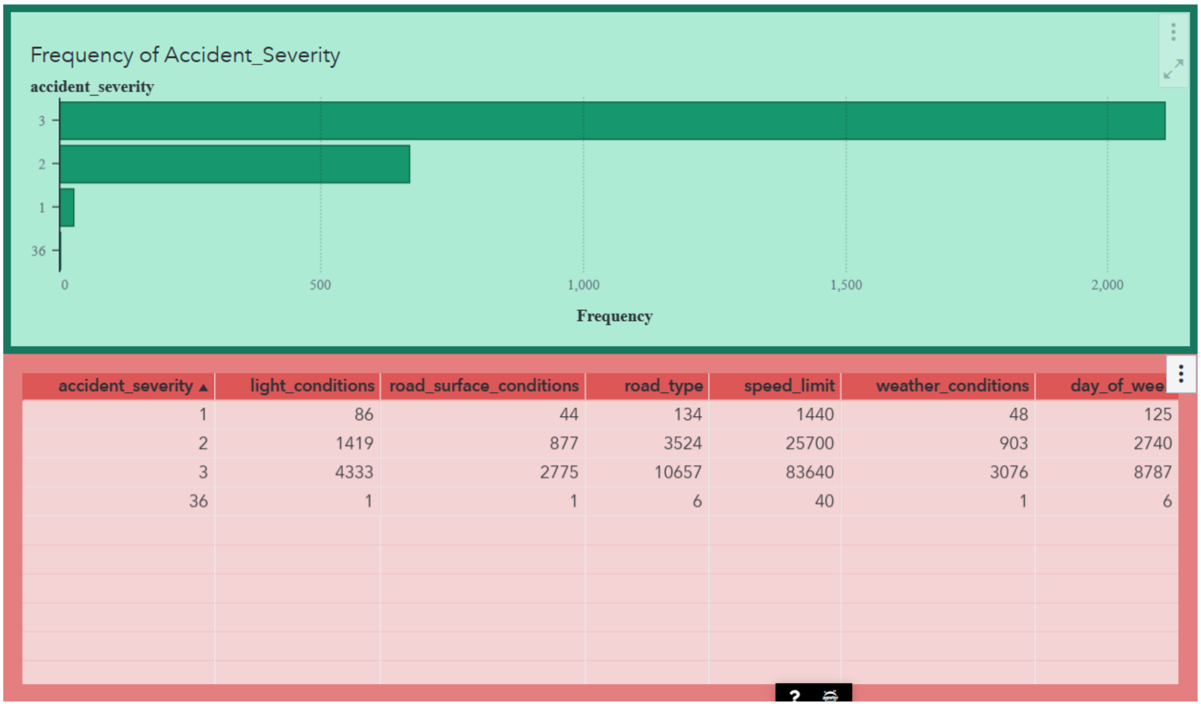
Table of figures

List of tables

1 Data Exploration

The Road Accident 2022 Surrey dataset has been successfully uploaded. Summary statistics for key variable was generated to understand data's central tendency and dispersion. The dataset consists of 2806 records and 35 variables. The target variable is accident severity and the variables used to analyze the severity are day of the week, light conditions, weather conditions, road surface conditions, road type and speed limit.

1.1 Explanatory Data Analysis



accident_severity ▲	light_conditions	road_surface_conditions	road_type	speed_limit	weather_conditions	day_of_wee
1	86	44	134	1440	48	125
2	1419	877	3524	25700	903	2740
3	4333	2775	10657	83640	3076	8787
36	1	1	6	40	1	6

Figure 1 Accident Severity Details

Figure 1 depicts histogram of target variable accident severity. With lower value 1 and upper value 2, we have got a frequency of 28. Lower value 2 and upper value 3 has got frequency of 668 and the lower value 3 and upper value 4 has got the most, 2108. There is also an outlier for accident severity 36.

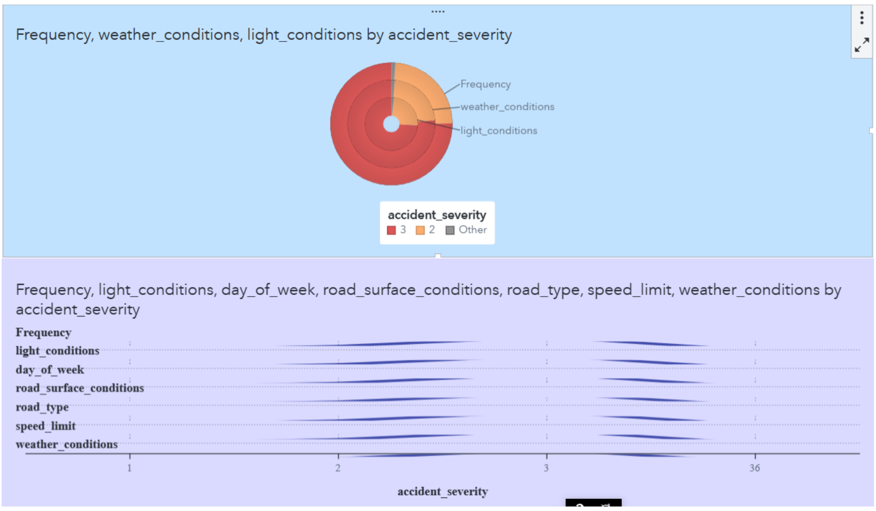


Figure 2 Pie chart to explore the set of variables

Figure 2 takes into account of how light condition and weather condition together has caused the accident. It is clearly seen that weather condition and light condition have caused major accidents with accident severity as 3 and reported. Where as for accident severity as 3, few accidents were reported.

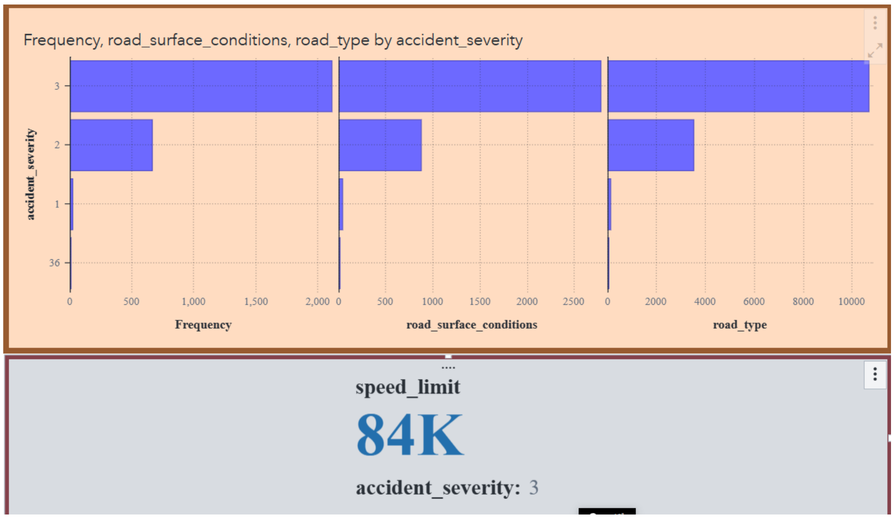
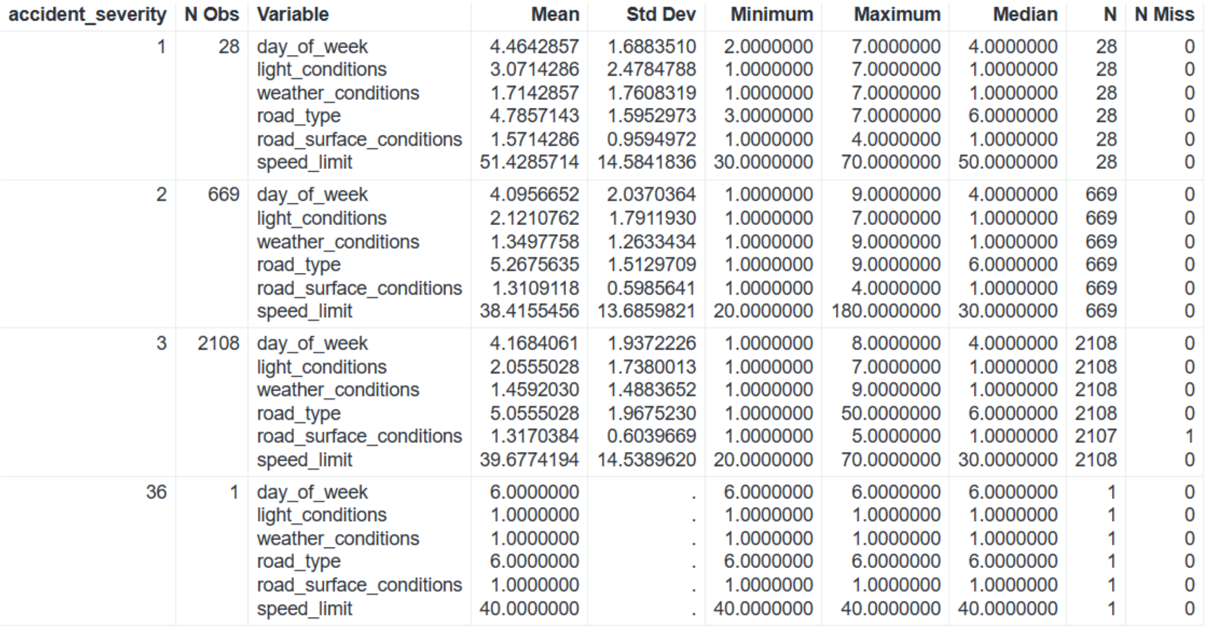


Figure 3 Bar Chart with key column

In **figure 3,** bar chart has been plotted for road surface conditions and road type with respect to accident severity to understand the frequency of each variable when associated with target variable. Also key column is set to speed limit and for accident severity 3, speed limit has got major impact.

1.2 Summary Statistics

Table 1 Summary Statistics of key variables



accident_severity	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss
1	28	day_of_week	4.4642857	1.6883510	2.0000000	7.0000000	4.0000000	28	0
		light_conditions	3.0714286	2.4784788	1.0000000	7.0000000	1.0000000	28	0
		weather_conditions	1.7142857	1.7608319	1.0000000	7.0000000	1.0000000	28	0
		road_type	4.7857143	1.5952973	3.0000000	7.0000000	6.0000000	28	0
		road_surface_conditions	1.5714286	0.9594972	1.0000000	4.0000000	1.0000000	28	0
		speed_limit	51.4285714	14.5841836	30.0000000	70.0000000	50.0000000	28	0
2	669	day_of_week	4.0956652	2.0370364	1.0000000	9.0000000	4.0000000	669	0
		light_conditions	2.1210762	1.7911930	1.0000000	7.0000000	1.0000000	669	0
		weather_conditions	1.3497758	1.2633434	1.0000000	9.0000000	1.0000000	669	0
		road_type	5.2675635	1.5129709	1.0000000	9.0000000	6.0000000	669	0
		road_surface_conditions	1.3109118	0.5985641	1.0000000	4.0000000	1.0000000	669	0
		speed_limit	38.4155456	13.6859821	20.0000000	180.0000000	30.0000000	669	0
3	2108	day_of_week	4.1684061	1.9372226	1.0000000	8.0000000	4.0000000	2108	0
		light_conditions	2.0555028	1.7380013	1.0000000	7.0000000	1.0000000	2108	0
		weather_conditions	1.4592030	1.4883652	1.0000000	9.0000000	1.0000000	2108	0
		road_type	5.0555028	1.9675230	1.0000000	50.0000000	6.0000000	2108	0
		road_surface_conditions	1.3170384	0.6039669	1.0000000	5.0000000	1.0000000	2107	1
		speed_limit	39.6774194	14.5389620	20.0000000	70.0000000	30.0000000	2108	0
36	1	day_of_week	6.0000000	.	6.0000000	6.0000000	6.0000000	1	0
		light_conditions	1.0000000	.	1.0000000	1.0000000	1.0000000	1	0
		weather_conditions	1.0000000	.	1.0000000	1.0000000	1.0000000	1	0
		road_type	6.0000000	.	6.0000000	6.0000000	6.0000000	1	0
		road_surface_conditions	1.0000000	.	1.0000000	1.0000000	1.0000000	1	0
		speed_limit	40.0000000	.	40.0000000	40.0000000	40.0000000	1	0

The classification variable was set as "accident severity" and the analysis variables were, day of week, light conditions, weather conditions, road type, road surface conditions and speed limit. Accident severity has 3 categories. For **Fatal** category, 28 observations has been recorded. Speed limit with higher average mean **(51. 4285714)** than all other factors, day of week **(4.4642857)**, light conditions **(3.0714286),** weather conditions **(1.7142857)**, road type **(4.7857143)**, and road surface conditions **(1.5714286)**. For **Serious** category, a total of 669 observations is recorded. Speed limit has again had the greater impact with mean value of **(38.4155456)** compared with day of week **(4.0956652)**, light conditions **(2.1210762),** weather conditions **(1.3497758)**, road type **(5.2675635)**, and road surface conditions **(1.3109118)**. Similarly for **Slight** category, we have got large number of observation of 2108. The speed limit has contributed more with average mean value **(39.6774194)**. 36 is an outlier for accident severity which must be removed. The road surface condition has 1 missing value.

1.3 Visualization

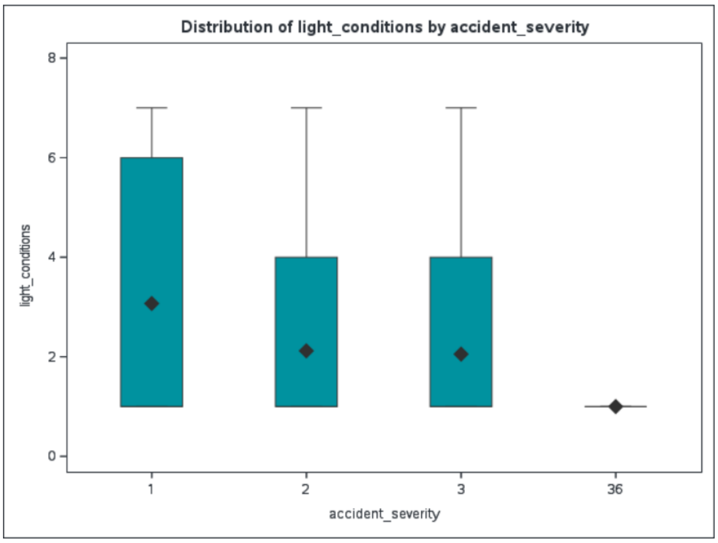


Figure 4 Outliers for accident_severity

In the above figure **(figure 4)**, there is no outlier for all the three categories of accident severity. It is to be noted that accident severity value of 36 is an outlier occurred. This outlier should be removed.

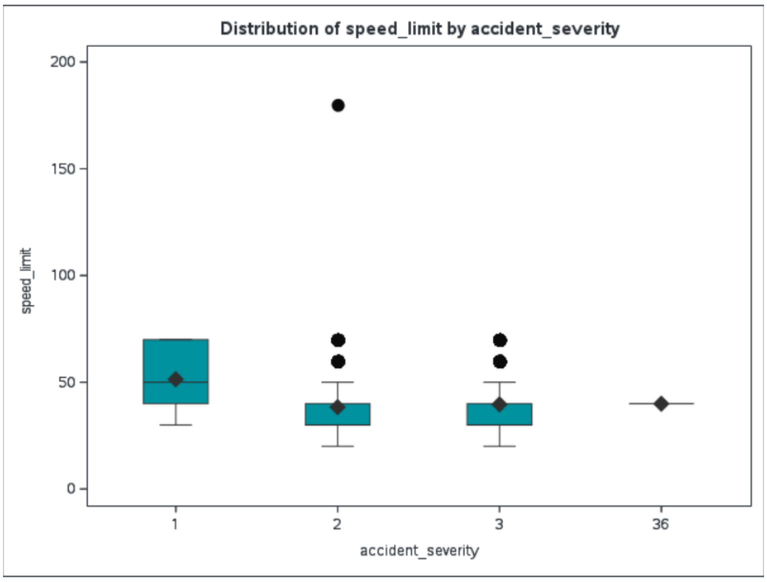


Figure 5 Outlier for accident severity, speed limit

In **figure 5**, the outliers is explained between speed limit and accident severity. Accident severity which is categorized as 1 that is **Fatal** is the only category without any outlier. Whereas category 2 and 3 has speed limit outliers. Outliers can strongly influence summary statistics such as the mean and standard deviation. There are many options to handle these outliers and that must be carried out to remove the outliers.

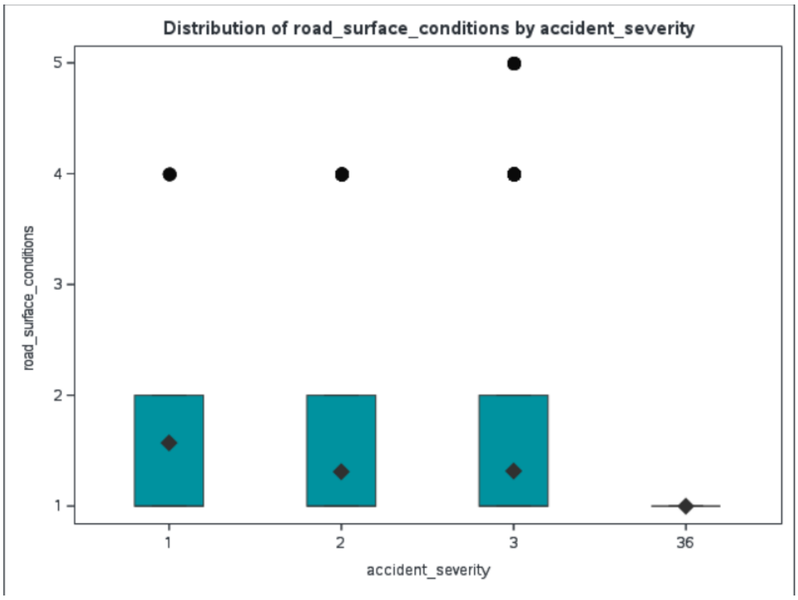


Figure 6 Outliers for accident severity and road surface condition

Outliers between road surface condition and accident severity is explained in **figure 6**. It is noted that road surface conditions have outliers in all the three categories of accident severity. This might cause greater impact while running the model. The reason for the outlier may be because of error in the dataset or missing values. The objective is to transform the data accordingly in order to build the model with good fit.

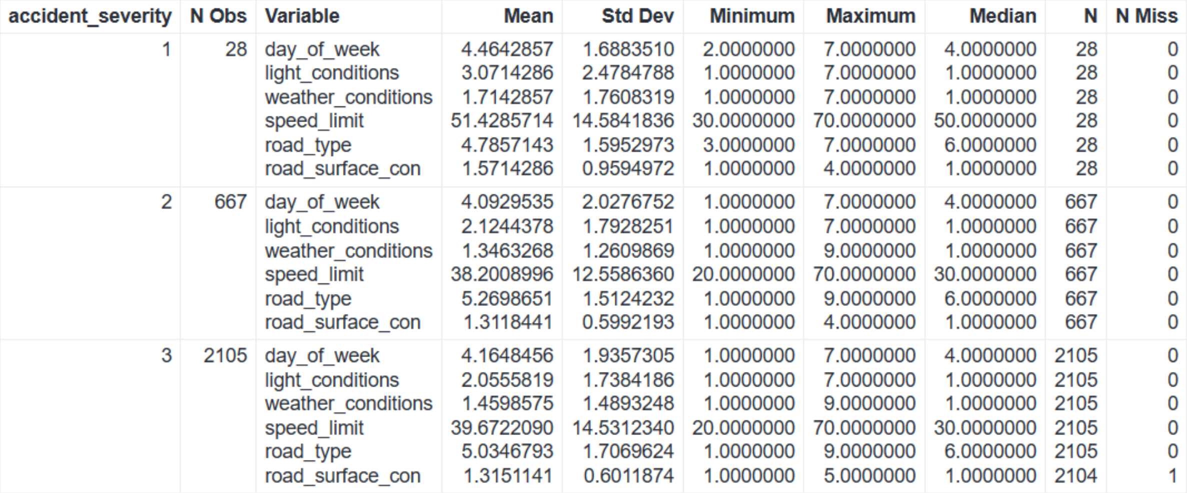
1.4 Data Cleaning

As outliers and missing values has been found out, it is essential to clean the data. List of variables that has outlier are as follows,

- Accident severity (unrecognized value of 36)
- Day of the week (unrecognized value of 8 and 9)
- Police force (unrecognized value of 47)
- Speed limit (unrecognized value of 180)
- Road type (unrecognized value of 50)

Road surface condition has missing values. There are many ways to clean the data, one such way is using the filter option in sas viya software to remove outlier. Impution node is used to fix missing values. Using more number of characters to name a variable will also impact the model, so the rename function is used to change the names of...... In order to remove the outlier, there is an option called filter under row transforms where we have to set the value not equal to so that value will be deleted from the table. The summary statistics for the dataset with renamed columns and removed outliers is executed and explained below.

Table 2 Summary statistics of table without outlier

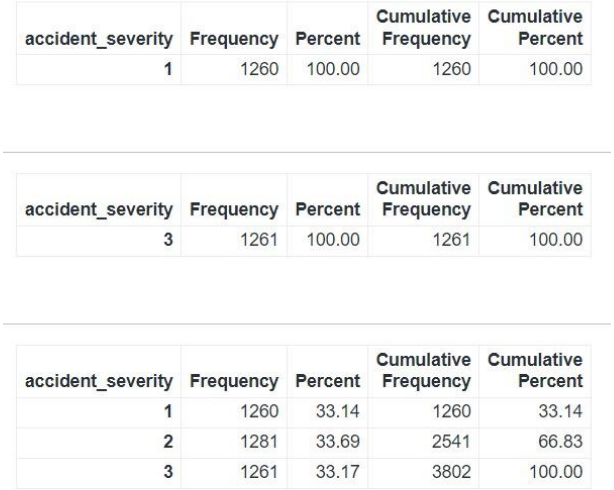


accident_severity	N Obs	Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss
1	28	day_of_week	4.4642857	1.6883510	2.0000000	7.0000000	4.0000000	28	0
		light_conditions	3.0714286	2.4784788	1.0000000	7.0000000	1.0000000	28	0
		weather_conditions	1.7142857	1.7608319	1.0000000	7.0000000	1.0000000	28	0
		speed_limit	51.4285714	14.5841836	30.0000000	70.0000000	50.0000000	28	0
		road_type	4.7857143	1.5952973	3.0000000	7.0000000	6.0000000	28	0
		road_surface_con	1.5714286	0.9594972	1.0000000	4.0000000	1.0000000	28	0
2	667	day_of_week	4.0929535	2.0276752	1.0000000	7.0000000	4.0000000	667	0
		light_conditions	2.1244378	1.7928251	1.0000000	7.0000000	1.0000000	667	0
		weather_conditions	1.3463268	1.2609869	1.0000000	9.0000000	1.0000000	667	0
		speed_limit	38.2008996	12.5586360	20.0000000	70.0000000	30.0000000	667	0
		road_type	5.2698651	1.5124232	1.0000000	9.0000000	6.0000000	667	0
		road_surface_con	1.3118441	0.5992193	1.0000000	4.0000000	1.0000000	667	0
3	2105	day_of_week	4.1648456	1.9357305	1.0000000	7.0000000	4.0000000	2105	0
		light_conditions	2.0555819	1.7384186	1.0000000	7.0000000	1.0000000	2105	0
		weather_conditions	1.4598575	1.4893248	1.0000000	9.0000000	1.0000000	2105	0
		speed_limit	39.6722090	14.5312340	20.0000000	70.0000000	30.0000000	2105	0
		road_type	5.0346793	1.7069624	1.0000000	9.0000000	6.0000000	2105	0
		road_surface_con	1.3151141	0.6011874	1.0000000	5.0000000	1.0000000	2104	1

The above table depicts the summary statistics of cleaned data. The 36 outlier has been removed from the target variable, the observation has changed for category 2 from 669 to 667 and for category 3 from 2108 to 2105. This indicates the outliers is removed from the analysis variables. The difference in the mean of all the variables listed under category 2 and category 3 has changed with respect to that of old summary statistics.

1.5 Data balancing

Table 3 Data balancing



accident_severity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1260	100.00	1260	100.00

accident_severity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	1261	100.00	1261	100.00

accident_severity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1260	33.14	1260	33.14
2	1281	33.69	2541	66.83
3	1261	33.17	3802	100.00

In order to balance the target variable, accident severity, data balancing is done after cleaning the data. Hence by balancing the categories of accident severity, Fatal, Serious and Slight is equally balanced with 33% percentage.

2 Predicting Accident Severity

2.1 Explanation of the Scenario

From the dataset available, there are various variables which can used to predict the accident severity. Here for the prediction of accident severity, the surface of the road and road type is taken. How the weather conditions will play a major role in interlinking the road surface and road type. Day of the week and light condition is taken in order to categorize on which day of the week and on what light condition, the accident has taken place. Also speed might be helpful in such scenarios, so we predict the accident severity by knowing the factors like, what was the maximum speed limit on that surface of the road and on such weather conditions. To even more support this prediction, the other variables like day of the week, road type and light conditions are taken.

2.2 Predictive models and its use cases

2.2.1 Random Forest Algorithm

In random forest algorithm multiple trees are build and their predictions are combined. By choosing a random subset of features for each tree and randomly sampling data points with replacement (bootstrap sampling), it adds unpredictability to the training process. The key features of this algorithm are using bagging, it trains each tree on a distinct subset of the training data sampled with replacement. Each split in a tree takes into account a random selection of features, increasing the ensemble's diversity. In regression analysis, the final forecast is determined from averaging the individual predictions provided by each tree, but in classification assignments, the final prediction is based on the majority vote from the collective trees. Because of its adaptability, Random Forest can be applied to a wide range of issues in the classification and regression fields.

2.2.2 Neural Network

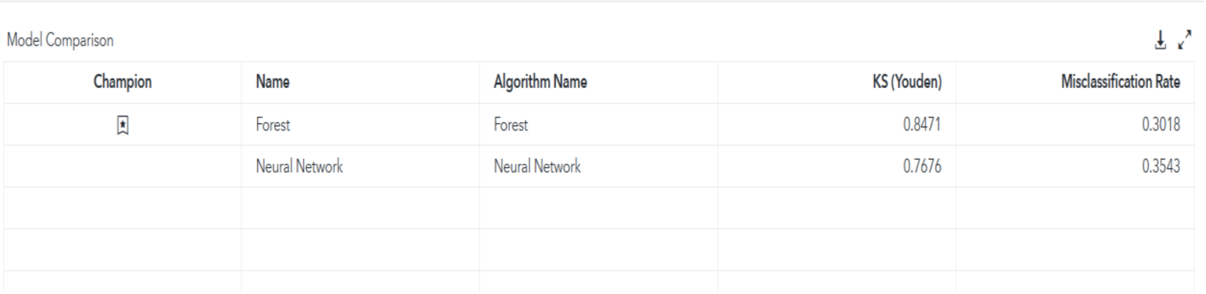
A mathematical model inspired by the structure and functioning of the human brain is the neural network. It consists of multiple layers of artificial neurons, or interconnected nodes, including an input layer, an output layer, and one or more hidden layers.One of the pillars of deep learning, a kind of machine learning, are neural networks. Neural networks are good in predictive analysis and they excel at capturing complex patterns and relationships in data.

2.3 Comparative Analysis

Random forest is an ensemble learning method which always results in more robust and accurate models. With the use of Random Forest's feature priority ranking, users can determine which features have the greatest influence on the model's predictions and random forest can capture non-linear relationships in the data and handle complex interactions between features. They also handle missing values wherein there is no need for imputation when using random forest algorithm. Even though these may be the strengths of random forest, they are weak in interpretability. Training a large number of decision trees can be computationally intensive, specifically while handling large dataset. On the other hand, neural networks can model highly non-linear relationships and capture intricate patterns in the data. They are mainly used for image recognition, natural language processing, and complex pattern recognition. Weakness of neural network may be are usually used only large datasets and training deep neural networks can be computationally intensive. Conclusion would be, each algorithm has its own benefits and the decision between Random Forest and Neural Network ultimately comes down to the particulars of the data, the issue at hand, and the trade-offs between prediction performance and interpretability.

2.4 Interpretation of the results

Table 4 Model Comparison



Model Comparison

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
★	Forest	Forest	0.8471	0.3018
	Neural Network	Neural Network	0.7676	0.3543

The main two evaluation metrics, Kolmogorov-Smirnov statistic (KS) and misclassification rate are noted to find the accuracy of each model. This metrics helps to decide which model will be the perfect fit to predict accident severity. The KS score for Forest model is **0.8471** which means the accuracy of the model is **84.71%** and misclassification rate is **0.3018** that is **30.18%.** Whereas neural network has got the accuracy of **76.76%** and misclassification rate of **35.43%**. Hence the Forest model has a higher accuracy **(84.71%)** compared to the Neural Network model **(76.76%)**. This suggests that, on average, the Forest model makes correct predictions more often that the Neural Network model. A lower misclassification rate indicates better performance, Forest model with low misclassification rate compared to Neural Network model is meant to be a better model.

2.5 Importance of different features used in predicting accident severity

While running each model, we get the scorecard having the importance score of each variable that is being analysed. This feature is helpful in finding which variable is used majorly by each model to predict the accident severity. For the Forest model, the important feature was day of the week with importance score of **139.82**, the second most important feature was light conditions with importance of **106.17**. Other features did contribute to the model's prediction but to a lesser extent. Similarly in Neural Network, the weights assigned between the input nodes and hidden nodes is considered as an important feature in predicting accident severity. Higher weighted features are expected to have a greater impact on the model's predictions. The weight of **0.6049** signifies the strength of the connection between the input feature and the prediction of accident severity level 1. A higher positive weight indicates that an increase in this feature's value will contribute more to the model predicting severity level 1. The positive sign indicates a positive correlation. If the feature's value increases, the neural network is more likely to predict accident severity level 1. Features with higher weights contribute more to the predictions made by the model. In this case, the feature associated with the weight of 0.6049 is crucial for predicting accidents with severity level 1.

2.6 Conclusion

Based on the model comparison with KS scores and misclassification rate, Forest model with **84.71%** accuracy and **30.18%** misclassification performs better than the Neural Network model, indicating that, based on these metrics, the Forest model is a better-performing model for the given task.

2.7 Recommendations in improving road safety

Results from the importance score can be used to recommend. A few recommendations to improve road safety are as follows,

- Consider focusing on specific days of the week where the accident severity is higher.
- Implement targeted safety measures or increased enforcement on these days.
- Speed limit variable also plays a crucial role. Evaluate areas where accidents are more severe at higher speeds.
- Implement speed management strategies, including speed limit adjustments, speed monitoring, and education campaigns on the dangers of speeding.
- Analyse whether certain road types or surface conditions are associated with higher severity accidents.
- Implement road maintenance, improvement, or signage enhancements accordingly.
- Implement a continuous monitoring system that tracks accident data and updates safety measures as patterns change over time.

3 Text Mining for Tweets

3.1 Exploration of dataset

The Tweets_2022 dataset is successfully loaded to the SAS Viya software. A total of 598 observations are recorded with 3 variables, The text variable is treated as target variable to execute the sentiment analysis. Source is the key column and according the pipeline is to be executed. The objective of text mining is to extract some insights about the document.

3.2 Text Preprocessing

In our text pre-processing workflow within SAS Viya, we eliminated punctuation marks like "?", "!", and "/". Additionally, we opted to exclude common stop words such as "not," "and," "the," "is," and "but" since they often lack significant meaning, streamlining the extraction of more meaningful and relevant statements. There are many ways to do text preprocessing, one such way is manually doing it in excel.

3.3 Explanatory Analysis

The Tweets_2022 dataset contains the text column and source column. After uploading the data successfully, information extraction and document categorization is carried out. Natural language processing includes two nodes, concepts and text parsing. The text variable is named as text and contains all the tweets, the concepts node in pipeline needs to run in order to make extraction simpler. There are already 9 predefined concepts available, apart from this, 3 new concepts (_Classifier_, _Miles_, _Minutes_) are included with Classifier and Regex rules.

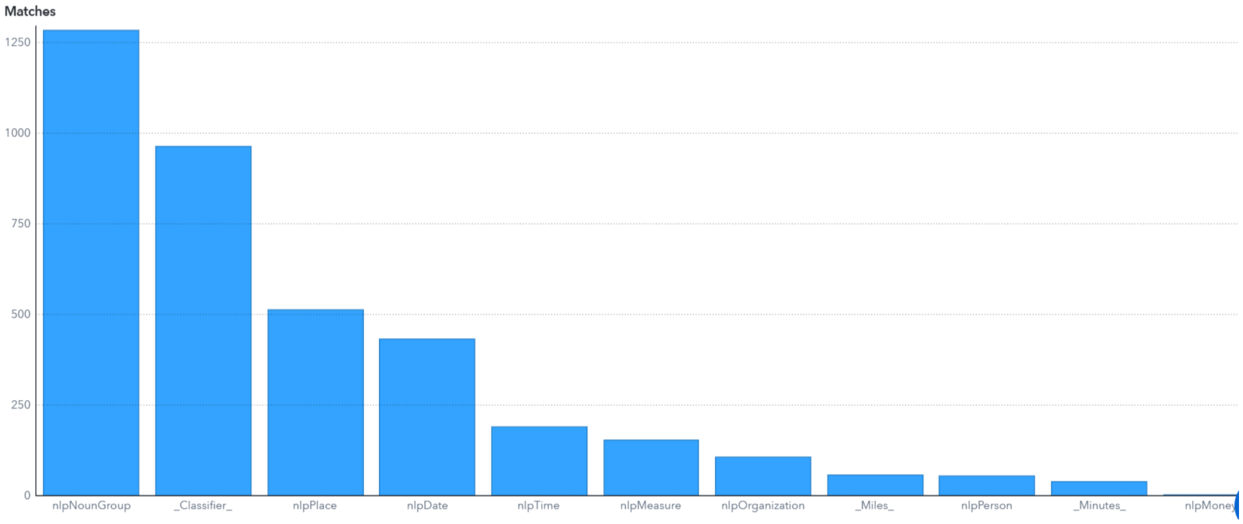


Figure 7 Concepts Node

The 3 new concepts has been successfully include with already existing 9 predefined concepts.

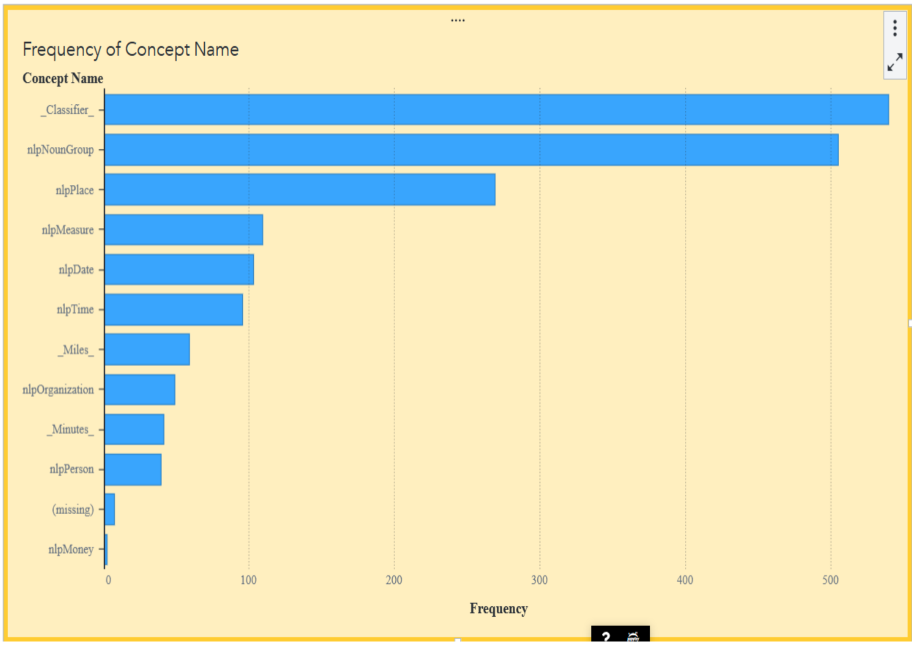


Figure 8 Histogram for word count

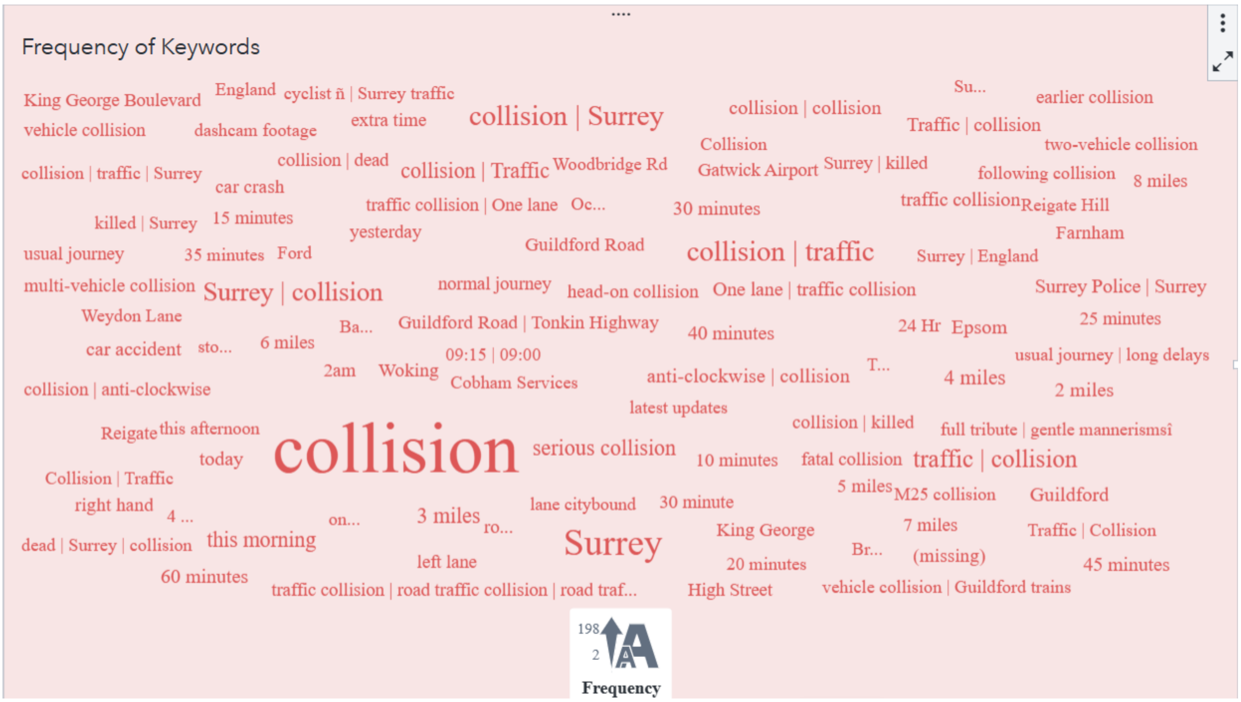


Figure 9 Word cloud visualization

Figure 8 and **figure 9** shows the calculations of word frequency and visualizing word clouds with respect to the concepts which are available. Classifier includes words like anti-clockwise, remain closed, remains closed, collision, surrey, dead, killed and traffic which means this concept rules extracts all the information regarding the mentioned key words. Frequency of 539 were matched using classifier rule. _Minutes_ and _Miles_ has the frequency of 41 and 59 respectively which uses Regex function. Major keyword is found out to be **collision**. After concepts node, text parsing is done by categorizing the kept and not kept terms in the document. _Classifier_ in total had 519 kept terms. There is a feature extraction under text parsing called topics node where the terms are clubbed together to know the sentiment whether negative, positive or neutral.

3.4 Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the sentiment expressed in a piece of text.

Types of sentiment that are available,

- **Positive Sentiment**: Expressing a favourable opinion or satisfaction.
- **Negative Sentiment**: Conveying a critical view or dissatisfaction.
- **Neutral Sentiment**: Lack of a clear positive or negative sentiment.

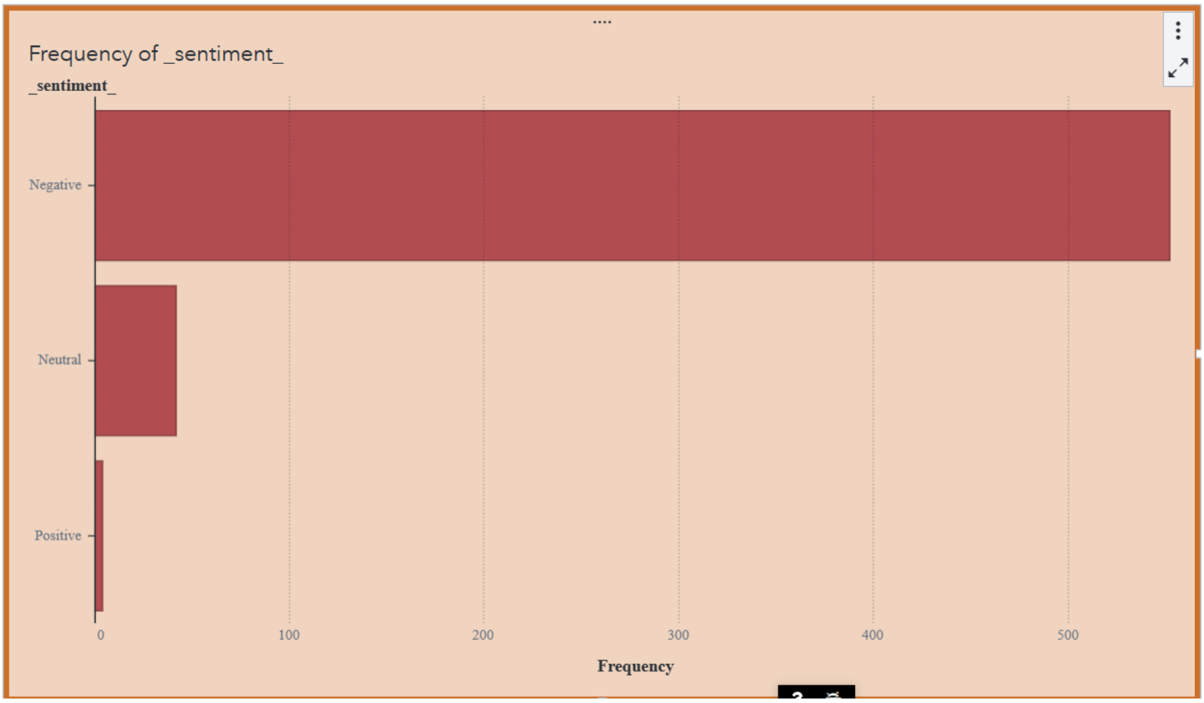


Figure 10 Sentiment Analysis Score

The above figure depicts the sentiment analysis scorecard. According to the histogram, **Negative** sentiment is recorded the most with frequency of **552**, continued by **Neutral** with frequency of **42**. The least is **Positive** with frequency of **4**.

3.5 Summarization of sentiment analysis

From the results above, it is evident that negative sentiment dominates with 552 instances. This expresses a trend of dissatisfaction tweets in the document. Positive sentiment is relatively scarce, with only 4 instances identified in the text. This suggests that positive opinions are not as commonly conveyed in the analyzed text. Neutral sentiment which indicates explicitly positive or negative sentiments has less number of instances. Hence the overall sentiment tone of the document is largely negative, so understanding tweets associated with negative sentiment provides better insights.

Task 4 Managerial Report

Summary: Comprehensive Analysis and Strategic Recommendations for Road Safety.

Overview

This summary presents key insights derived from the analysis of Surrey's 2022 road traffic data. The goal is to provide decision-makers with actionable recommendations aimed at improving road safety.

Data Exploration

We conducted a detailed review of a dataset comprising 2,806 accidents, exploring variables such as accident severity, weather conditions, road types, light conditions, speed limit, road surface conditions and day of the week. This comprehensive analysis offers a nuanced understanding of the factors influencing road incidents.

Model Prediction

Two predictive models were used and their results were compared. The Forest and Neural Network model were used, out of which Forest algorithm had better accuracy and efficiency compared to Neural Network. This implies that it could be useful for recognizing high-risk situations and creating preventative safety measures.

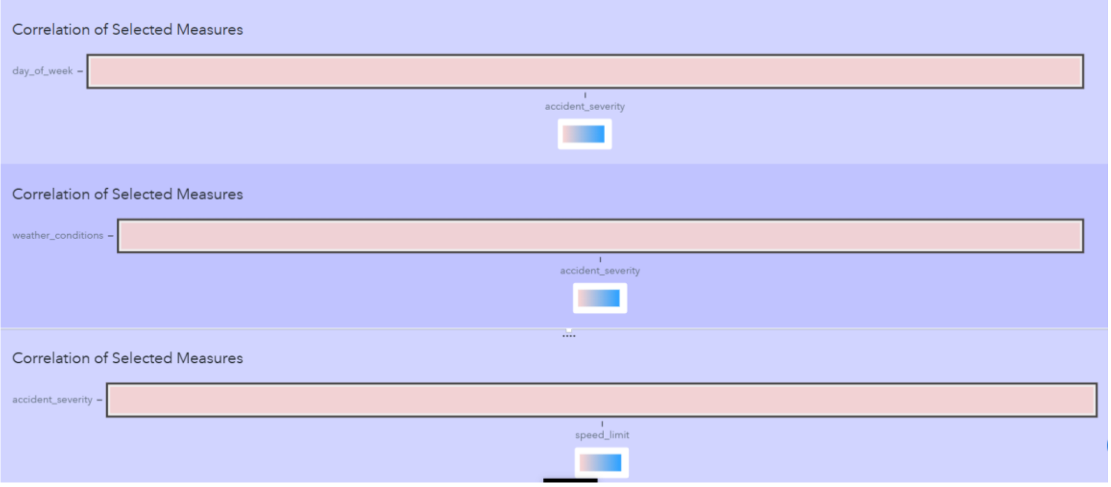
Key Findings

Impact of speed limit on Accident Severity: A direct correlation exists between speed limits and the severity of accidents. Areas with higher speed limits demonstrate increased accident severity.

Temporal Trends: Specific times, particularly weekends, witnessed a higher frequency of accidents. This highlights the need for targeted traffic management strategies during these periods.

Weather and Light Conditions: Adverse weather and inadequate lighting significantly contribute to accidents, emphasizing the role of infrastructure in preventing road incidents.

Correlation of impacted variables


Correlation of Selected Measures
day_of_week
accident_severity
Correlation of Selected Measures
weather_conditions
accident_severity
Correlation of Selected Measures
accident_severity
speed_limit

The above figure shows the correlation between accident severity and the variable which were more impactful, day of the week, speed limit and weather condition. It is said to have positive correlation, hence the results are derived.

Recommendations for Action

Speed Restriction: Implement more robust speed monitoring and enforcement measures, especially in identified high-risk zones.

Adaptive Traffic Management: Improve traffic control strategies during peak accident times based on insights from our temporal analysis.

Infrastructure Upgrades: Enhance road lighting and surface conditions to create safer driving environments, particularly in adverse weather.

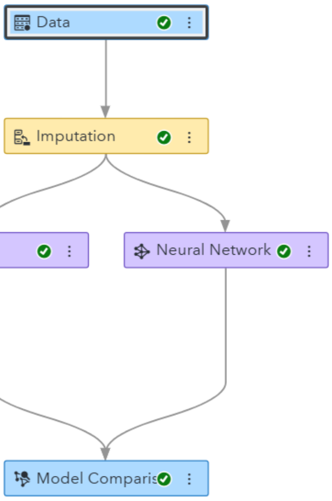
Pedestrian Safety Initiatives: Strengthen protocols and infrastructure for pedestrian crossings in busy urban areas.

Community Awareness Programs: Launch road safety campaigns targeting identified risk factors to instill a culture of safety among road users.

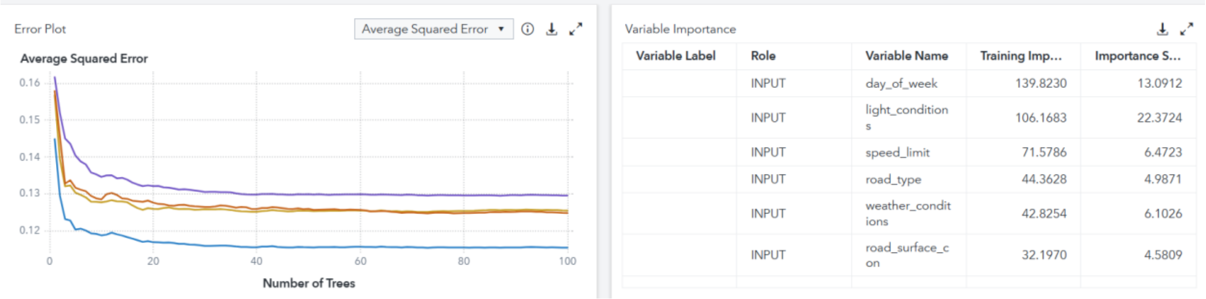
Conclusion

The proactive implementation of these data-driven recommendations is expected to significantly reduce both the frequency and severity of road accidents in Surrey. This approach underscores our commitment to ensuring the well-being of our community and enhancing the overall quality of road travel.

Appendix



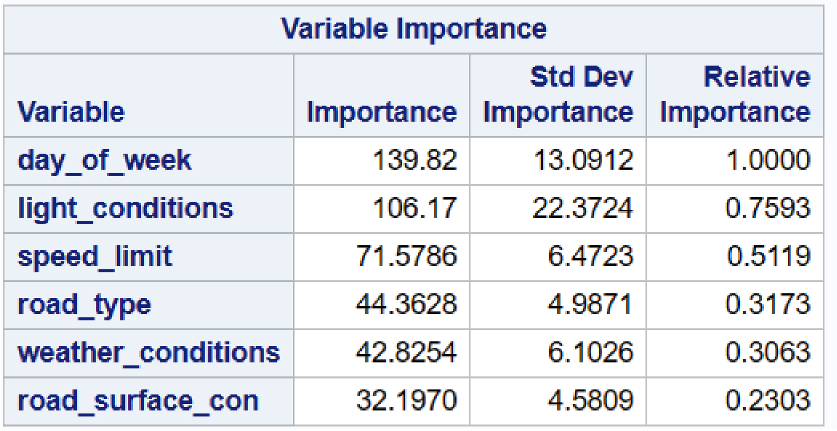
Model pipeline



Variable Importance

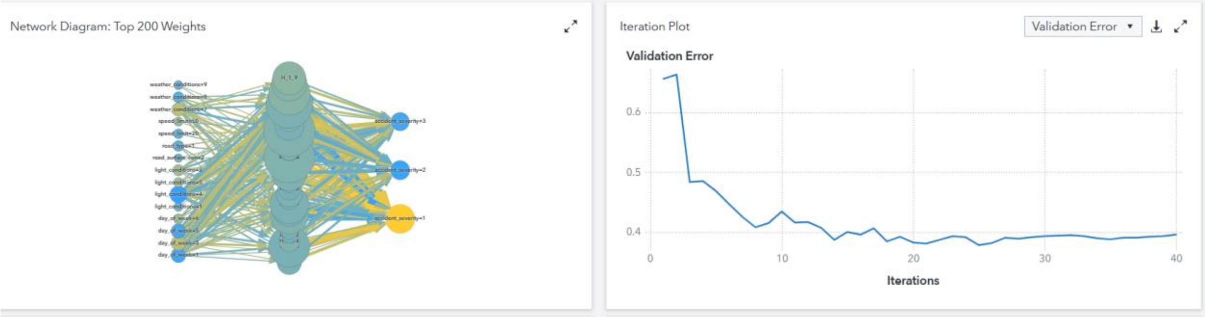
Variable Label	Role	Variable Name	Training Imp...	Importance S...
	INPUT	day_of_week	139.8230	13.0912
	INPUT	light_condition s	106.1683	22.3724
	INPUT	speed_limit	71.5786	6.4723
	INPUT	road_type	44.3628	4.9871
	INPUT	weather_condit ions	42.8254	6.1026
	INPUT	road_surface_c on	32.1970	4.5809

Forest Model graphs

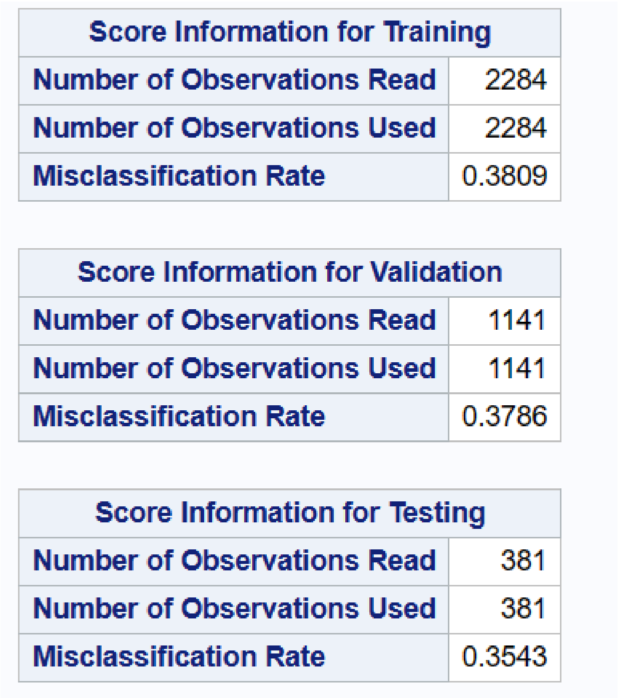


Variable Importance			
Variable	Importance	Std Dev Importance	Relative Importance
day_of_week	139.82	13.0912	1.0000
light_conditions	106.17	22.3724	0.7593
speed_limit	71.5786	6.4723	0.5119
road_type	44.3628	4.9871	0.3173
weather_conditions	42.8254	6.1026	0.3063
road_surface_con	32.1970	4.5809	0.2303

Importance feature score



Neural network graph

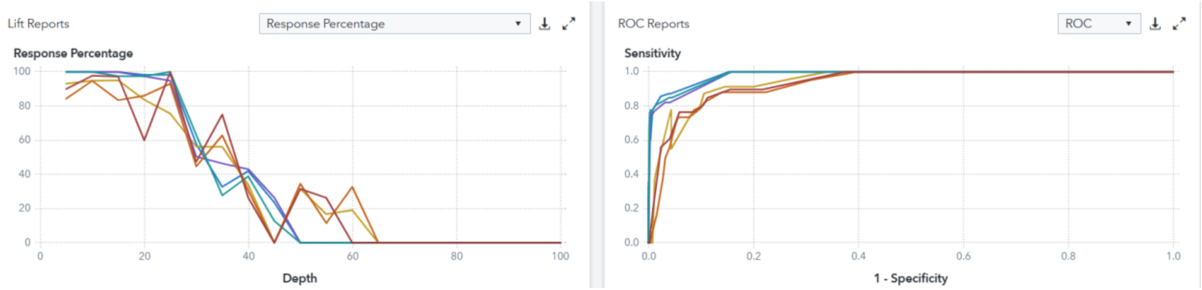


Score Information for Training	
Number of Observations Read	2284
Number of Observations Used	2284
Misclassification Rate	0.3809

Score Information for Validation	
Number of Observations Read	1141
Number of Observations Used	1141
Misclassification Rate	0.3786

Score Information for Testing	
Number of Observations Read	381
Number of Observations Used	381
Misclassification Rate	0.3543

Scores after running the model



ROC Curve for model comparison

Codes for data balancing

```
%let NumSamples1 = 70;

%let NumSamples2 = 3;

%let NumSamples3 = 1;

/* Sort the dataset by accident_severity */

proc sort data=WORK.IMPORT;

 by accident_severity;

run;

/* Use PROC SURVEYSELECT to make the dataset balanced */

proc surveyselect data=WORK.IMPORT NOPRINT out=BalancedData1

 method=urs

 seed=12345

 samprate=(1 0 0); /* Adjust the samprate to balance the strata */

 strata accident_severity;

run;

/* Create a variable with a constant value for each observation */

data BalancedData_with_constant1;

 set BalancedData1;

 constant = 1;

run;

/* Use PROC SURVEYSELECT to perform bootstrap sampling based on accident_severity */

proc surveyselect data=BalancedData_with_constant1 NOPRINT seed=1

 method=urs

 samprate=1
 OUTHITS
 reps=&NumSamples1(repname=row)
 out=BootSamp1;
 strata accident_severity;
run;
/* Overall, how often was each observation selected? */
proc freq data=BootSamp1;
 tables accident_severity;
run;
/* Use PROC SURVEYSELECT to make the dataset balanced */
proc surveyselect data=WORK.IMPORT NOPRINT out=BalancedData2
 method=urs
 seed=12345
 samprate=(0 1 0); /* Adjust the samprate to balance the strata */
 strata accident_severity;
run;
/* Create a variable with a constant value for each observation */
data BalancedData_with_constant2;
 set BalancedData2;
 constant = 1;
run;
/* Use PROC SURVEYSELECT to perform bootstrap sampling based on accident_severity
*/
proc surveyselect data=BalancedData_with_constant2 NOPRINT seed=1
 method=urs
 samprate=1
 OUTHITS
 reps=&NumSamples2(repname=row)
 out=BootSamp2;
 strata accident_severity;
run;
/* Use PROC SURVEYSELECT to make the dataset balanced */
proc surveyselect data=WORK.IMPORT NOPRINT out=BalancedData3
 method=urs
 seed=12345
 samprate=(0 0 0.9); /* Adjust the samprate to balance the strata */
 strata accident_severity;
run;
/* Create a variable with a constant value for each observation */
data BalancedData_with_constant3;
 set BalancedData3;
 constant = 1;
run;
/* Use PROC SURVEYSELECT to perform bootstrap sampling based on accident_severity
*/
proc surveyselect data=BalancedData_with_constant3 NOPRINT seed=1
 method=urs
 samprate=1
 OUTHITS
 reps=&NumSamples3(repname=row)
 out=BootSamp3;
 strata accident_severity;
run;
proc freq data=BootSamp3;
 tables accident_severity;
run;
data BootSamp;
 set BootSamp1 BootSamp2 BootSamp3;
run;
proc freq data=BootSamp;
 tables accident_severity;
run;
```

Code for sentiment score

```
/****************************************************************
* SAS Visual Text Analytics
* Sentiment Score Code
*
* Modify the following macro variables to match your needs.
***************************************************************/

/* specifies CAS library information for the CAS table that you would like to score. You must modify the value to provide the name of the library that contains the table to be scored. */
%let input_caslib_name = "CASUSER";
```

```
/* specifies the CAS table you would like to score. You must modify the value to provide the
name of the input table, such as "MyTable". Do not include an extension. */
%let input_table_name = "TWEETS_2022";


/* specifies the column in the CAS table that contains a unique document identifier. You must
modify the value to provide the name of the document identifer column in the table. */
%let key_column = "Source";


/* specifies the column in the CAS table that contains the text data to score. You must modify
the value to provide the name of the text column in the table. */
%let document_column = "text";


/* specifies the CAS library to write the score output tables. You must modify the value to
provide the name of the library that will contain the output tables that the score code
produces. */
%let output_caslib_name = "CASUSER";


/* specifies the sentiment output CAS table to produce */
%let output_sentiment_table_name = "Road_Accident_out_sentiment";


/* specifies the matches output CAS table to produce */
%let output_matches_table_name = "Road_Accident_out_sent_matches";


/* specifies the features output CAS table to produce */
%let output_features_table_name = "Road_Accident_out_sent_features";
```

```
/* specifies the language of the associated SAS Visual Text Analytics project. This should be
set automatically to the language you selected when you created your project. */
%let language = "ENGLISH";

/* specifies the hostname for the CAS server. This should be set automatically to the host for
the associated SAS Visual Text Analytics project. */
%let cas_server_hostname = "pdcesx23045.exnet.sas.com";

/* specifies the port for the CAS server. This should be set automatically to the host for the
associated SAS Visual Text Analytics project. */
%let cas_server_port = 5570;

/* creates a session */
cas sascas1 host=&cas_server_hostname port=&cas_server_port;
libname sascas1 cas sessref=sascas1 datalimit=all;

/* calls the scoring action */
proc cas;
   session sascas1;
   loadactionset "sentimentAnalysis";

   action applySent;
      param
         table={caslib=&input_caslib_name, name=&input_table_name}
```

```
        docId=&key_column
        text=&document_column
        language=&language
        casOut={caslib=&output_caslib_name, name=&output_sentiment_table_name, replace=TRUE}
        matchOut={caslib=&output_caslib_name, name=&output_matches_table_name, replace=TRUE}
        featureOut={caslib=&output_caslib_name, name=&output_features_table_name, replace=TRUE}
      ;
  run;
quit;
```

References

1. Ziegler, A. and König, I.R., 2014. Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), pp.55-63.
2. Verikas, A., Gelzinis, A. and Bacauskiene, M., 2011. Mining data with random forests: A survey and results of new tests. *Pattern recognition*, *44*(2), pp.330-349.
3. Wang, L. and Sui, T.Z., 2007, September. Application of data mining technology based on neural network in the engineering. In *2007 International Conference on Wireless Communications, Networking and Mobile Computing* (pp. 5544-5547). IEEE.
4. Ni, X., 2008. Research of data mining based on neural networks. *World Academy of Science, Engineering and Technology*, *39*(1), pp.381-384.
5. Cogburn, D. and Hine, M., 2017. Introduction to text mining in big data analytics Minitrack.
6. Nisbet, R., Elder, J. and Miner, G.D., 2009. *Handbook of statistical analysis and data mining applications*. Academic press.