Road Traffic Conditions – Predictive Analysis

1. Introduction:

I was given the assignment for calculating the probability of Fatal Auto accidents in the region of Cambridgeshire, UK. There are multiple dependent and independent factors to consider. Amongst them are Road Conditions, Visibility, Time of the Day, Day of the Week, Number of Vehicles Involved, the Road Type, Speed Limit (in Kilometers), the Weather and Road Number (Identification).

It is my responsibility to create the best possible model to perform Predictive Analysis on the test case. Having said that, I will be using the Road Traffic Conditions data set (RTC Location 2018.csv). This data file is available from the Cambridgeshire County Council web site. I will be testing the data from 2018. This data set contains all of the factors that I've previously mentioned, and maybe included on my model.

1. **Data Description:**

|  |
| --- |
| I will be using the Road Traffic Conditions data set (RTC Location 2018.csv). This  Data file is available from the Cambridgeshire County Council web site. I will be testing the data from 2018. This data set contains all of the factors that I’ve previously mentioned, and maybe included on my model.  The Comma Separated Values (CSV) file will require data cleansing (preparation) in order to convert strings (defined as a Python object) into an Integer.  Once these columns are converted to an Integer, the columns can be fitted into a train-test-split regression model.   1. **Methodology** |
|  |

1. Methodology:

|  |
| --- |
| My approach was to code in Python and running the object using Jupyter Notebook on an IBM Watson platform. There were four columns that required data cleansing from a Python object into an Integer.  These columns were Severity, Road Conditions (ROAD\_CON), Visibility and Weather.  In addition, I assigned arbitrary numeric values to each of these four columns.  For Severity (Accident Score), a value of a 1 was assigned to Fatal. A value of a 2 for Serious and a 3 for Slight. |
| A similar rating scale was created for ROAD\_CON and Weather.  To calculate the accuracy, I calculated the F1 Score and Jaccard Similarity.   1. **Results**:   I decided to use regression models, Logistic Regression and Support Vector  Machine. |

The F1 Scores and Jaccard Similarity Score from both the Logistic Regression and SVM did not yield high accurate scores.

|  |  |  |
| --- | --- | --- |
|  | F1 Score | Jaccard Similarity |
| Logistic Regression | 0.6076 | 0.7236 |
| Support Vector Machine | 0.6284 | 0.7358 |

This scatter plot shows the relationship between Road Conditions and Casualties: The highest accidents occurred on Dry Conditions. This includes the highest fatal counts (20, 11).

A screenshot of a cell phone

Description automatically generated

This scatter plot shows the relationship between Weather Conditions and Casualties. The highest accidents happened on a Fine day with no high winds.

A screenshot of a cell phone

Description automatically generated

This next scatter plot describes the relationship between Visibility and Casualties. The lowest number (10) represents Daylight. Ironically, the highest casualty count also occurred in Daytime.

A screenshot of a cell phone

Description automatically generated

This last scatter plot shows the pattern of Speed Limit versus Casualties. As predicted, the highest accidents occurred when the Speed Limit was over 60 KM/hour.

A screenshot of a cell phone

Description automatically generated

This heat map shows the relationship of Road Conditions and Weather against Casualties.

A picture containing drawing

Description automatically generated

|  |
| --- |
| 1. **Discussion:**   **Observations:** |

t

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| More accidents occurred on a Dry Road Condition (defined as a 20 on the Road Dictionary. The highest value is 80 for Frosty or Icy Road. | | | | | |
| In general, there were more casualties when the Speed Limit was either 60 KM (kilometers) or 70 KM. | | | | | |
| There were more casualties when the Weather was "Fine with No High Wind". This was defined as a 10 on the Weather Dictionary. | | | | | |
| There were more accidents in Daylight (defined as a 10 on the Visibility Dictionary). | | | | |
| The highest casualty per incident (20, 11) happened on a Dry road, in broad Daylight, and on a clear (Fine with no high wind) day. | | | | | |
| Road number 0 (zero) had the greatest number of casualties at 223. This was followed by Road Number 47. The casualties were considered Fatal or Serious. | | | | | |
| There appears to be more road volume during Daylight and when the conditions are Clear (Weather) and Dry (Road). | | | | | |
| Severity 3 (defined as Slight Casualties) has the highest volume of personal accidents. | | | |
|  | | | |
| **Recommendations**: | | |
| I recommend that the Cambridgeshire County needs to capture more information on their Collision data set. | | | | | |
| What was the extent of the personal injury? | |
| What was the damage to the vehicles? |
| Was the driver or drivers under the influence of alcohol? Was the driver under the influence of drugs? | | | | | |

How fast was the vehicle driving above the speed limit?

1. **Conclusion:**

The F1 Score and Jaccard similarity index calculations suggest that this data set may not be a good model to accurately predict Road Traffic Accidents for this county.

The root causes for accidents on Dry, Clear and in broad daylight maybe a direct of the following:

More cars are on the during these ideal conditions.

There are less cars when the road is Frosty or Icy.

|  |  |
| --- | --- |
| The vehicles may have experienced a problem. The vehicle's age was not available for interpretation. Perhaps, the driver had a braking malfunction. | |
| Collisions may have been a result of reckless driving. This is common when the Speed Limit is 60 KM or above. | |
| The driver may have been distracted or texting. This information was not available on the data set. | |
| The driver's age was not disclosed on this data set. |
| If the driver is 18 or below, this may suggest that the driver does not have extensive driving experience to avoid an accident. | |
| The driver's health condition was also not available. The driver may have suffered a medical condition that led to the accident. | |
| The driver may have slept on the wheel, and as a result, may have caused the accident. | |
| If these recommended data responses are made available, another regression can be performed. | |

The accuracy score should be much higher than the ones calculated during this time.