Report on Road Traffic Accidents in Great Britain (2020)

Introduction

This report presents analysis and modelling of road traffic accidents in Great Britain for the year 2020. The objective is to provide insights into accident patterns and offer recommendations to government agencies for improving road safety.

Data Overview

- **Total Accidents**: 91,199
- **Total Vehicles Involved**: 167,375
- **Total Casualties**: 115,584
- **Accident Involvement**:
    - Two vehicles: 62.9%
    - One vehicle: 28.2%

Apriori Algorithm Analysis

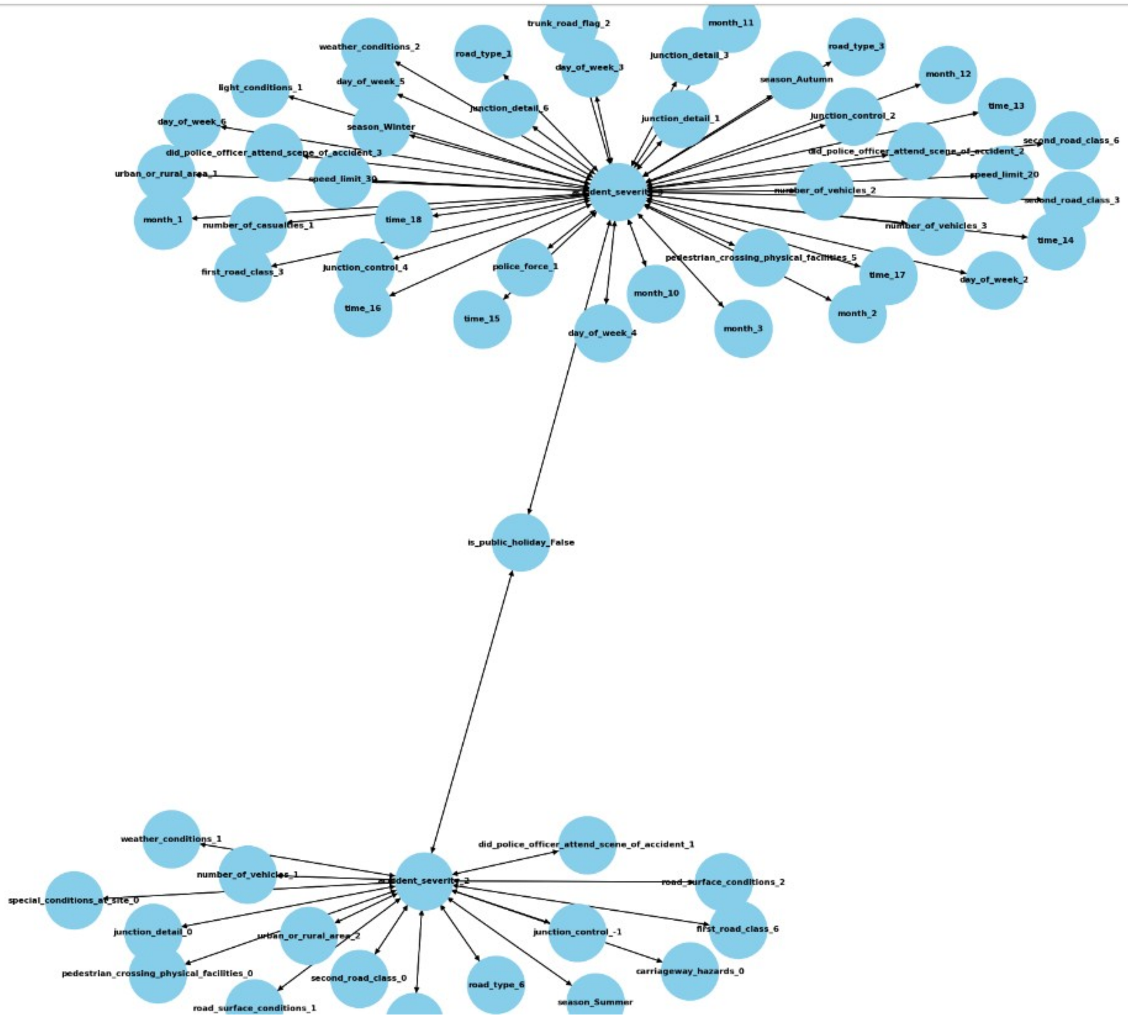
	antecedents	consequents	support	confidence	lift
0	accident_severity_2	is_public_holiday_False	0.201263	1.000000	1.000000
1	accident_severity_3	is_public_holiday_False	0.783484	1.000000	1.000000
2	accident_severity_2	pedestrian_crossing_human_control_0	0.195737	0.972542	1.027060
3	accident_severity_2	carriageway_hazards_0	0.195068	0.969218	1.005812
4	accident_severity_2	special_conditions_at_site_0	0.194388	0.965840	1.008873
...	...	...	...	...	...
115	accident_severity_3	month_3	0.056634	0.072285	1.006311
116	accident_severity_3	number_of_vehicles_3	0.053827	0.068703	1.003942
117	accident_severity_3	time_14	0.053707	0.068549	1.001050
118	accident_severity_3	road_type_1	0.051393	0.065596	1.072088
119	accident_severity_3	time_13	0.050143	0.064000	1.016678

The top rules by lift and confidence seem to be:

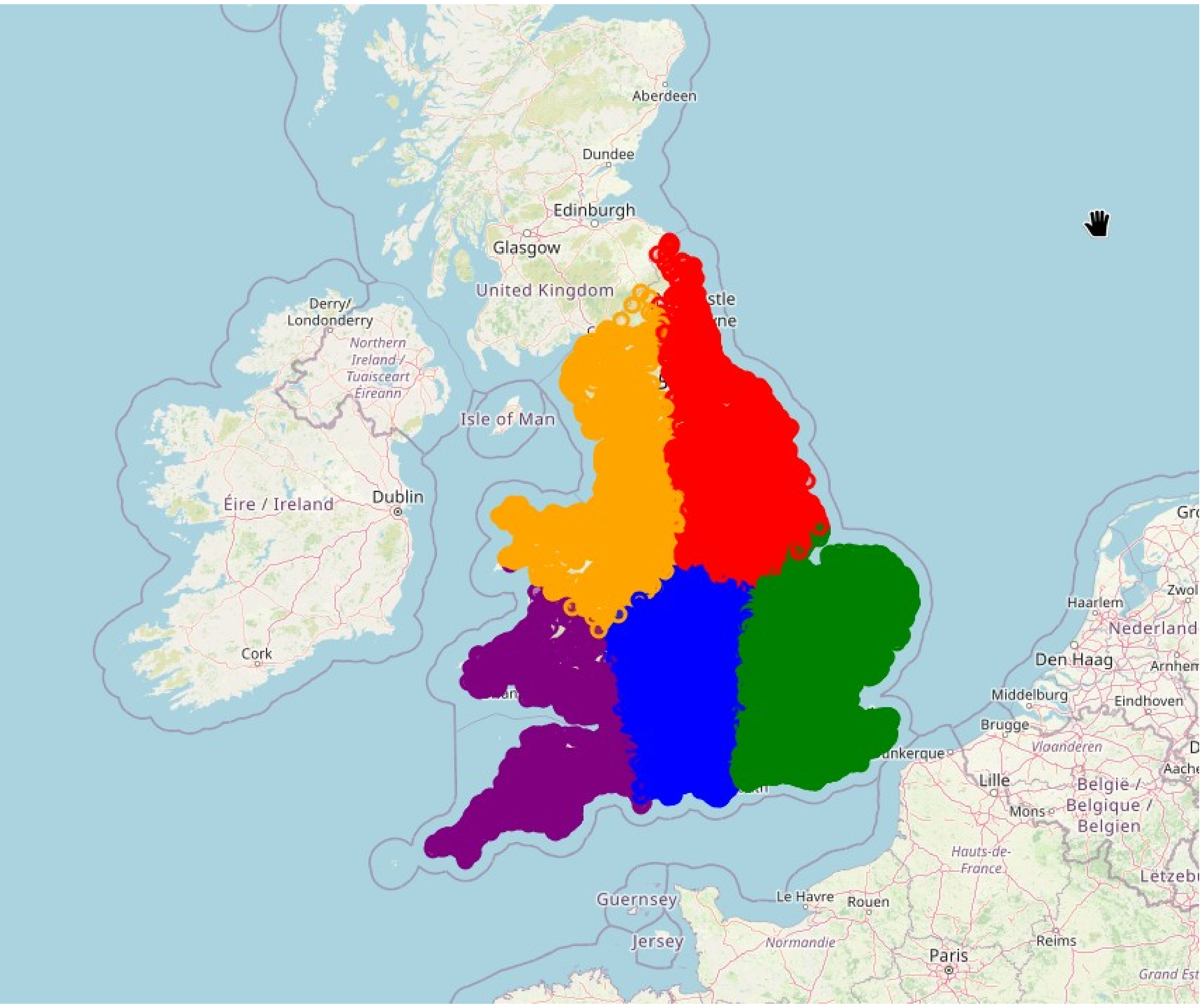
- accident_severity_2 -> is_public_holiday_False (confidence: 1.0, lift: 1.0)
- accident_severity_3 -> is_public_holiday_False (confidence: 1.0, lift: 1.0)
- accident_severity_2 -> pedestrian_crossing_human_control_0 (confidence: 0.97, lift: 1.02)

These rules suggest that:

- Most accidents, regardless of severity, occur on non-public holidays.
- Severity 2 accidents are strongly associated with no human-controlled pedestrian crossings.



Spatial Analysis and Clustering



1. Spatial Distribution of Clusters:

The map in Image 1 shows England divided into 5 distinct regions, which appear to correspond to the 5 clusters identified in your K-means clustering analysis. Each cluster is represented by a different color on the map.

2. Cluster Centroids:

From the data in Image 2, we can see the latitude and longitude coordinates for the centroids of each cluster:

- Cluster 0: (51.832978, -1.744582)

- Cluster 1: (53.747142, -1.209877)

- Cluster 2: (51.565957, -0.005725)

- Cluster 3: (51.043275, -3.798906)

- Cluster 4: (53.517308, -2.656654)

These coordinates roughly correspond to different regions in England, which aligns with the colored regions on the map.

3. Cluster Characteristics:

The cluster_characteristics data provides insights into each cluster:

- Accident Severity: All clusters show a predominant accident severity of 3, which typically represents the least severe accidents.

- Time: The average time of accidents is similar across all clusters, ranging from about 13:34 to 13:55 (in 24-hour format).

- Road Type: All clusters show a predominant road type of 6, which likely represents a specific category of road (e.g., single carriageway) that is common across England.

4. Regional Insights:

- The northern regions (red and orange on the map) correspond to clusters 1 and 4, which have slightly higher latitude values.

- The southern and eastern regions (purple, blue, and green on the map) correspond to clusters 0, 2, and 3.

- Cluster 2, with a longitude close to 0, likely represents the area around London.

5. Uniformity in Characteristics:

It's noteworthy that despite the geographical differences, the accident characteristics (severity, time, and road type) are remarkably similar across all clusters. This suggests that these factors may be consistent throughout England, regardless of the specific region.

6. Limitations:

The analysis is limited to the variables provided. Other factors like population density, traffic volume, or specific road features are not accounted for in this clustering, which could provide more nuanced insights if included.

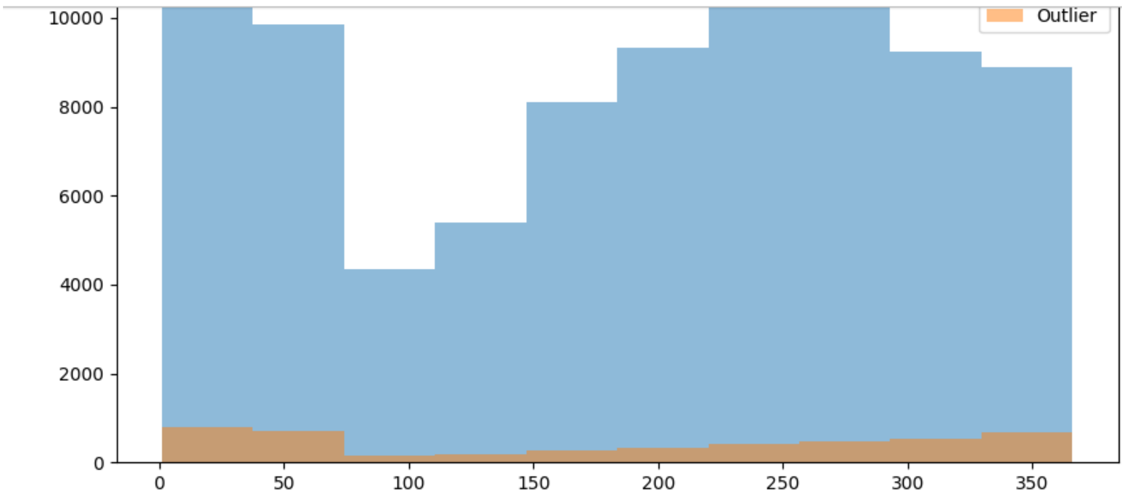
7. Further Investigation:

Given the uniformity in characteristics across clusters, it would be valuable to investigate other variables that might reveal more distinct regional differences in accident patterns. Additionally, a more granular analysis within each cluster could uncover local hotspots or unique features that are not apparent at this broader level of clustering.

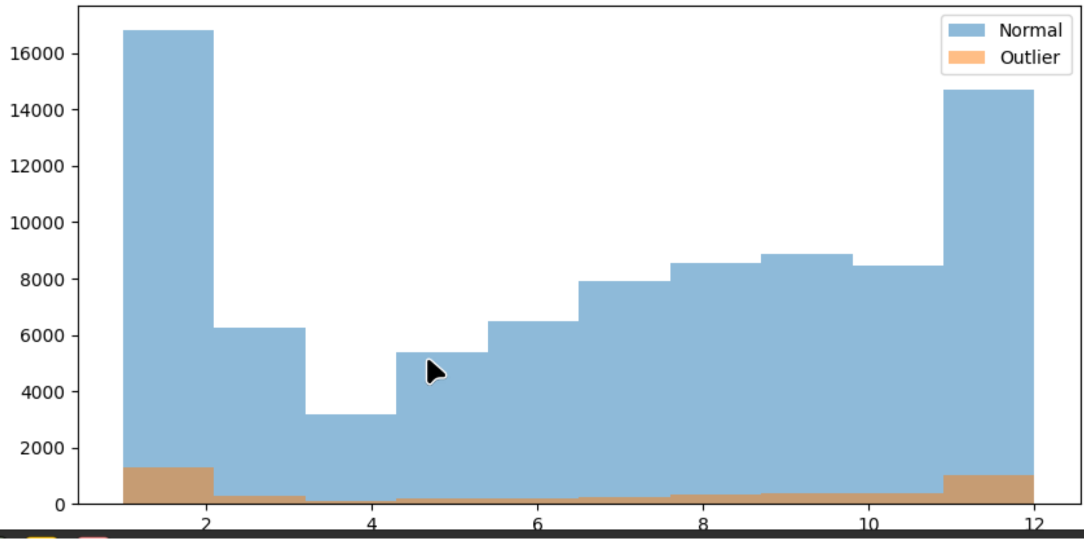
This spatial analysis provides a good starting point for understanding the geographical distribution of accidents in England, but it also highlights the need for more detailed local analyses to inform targeted road safety measures.

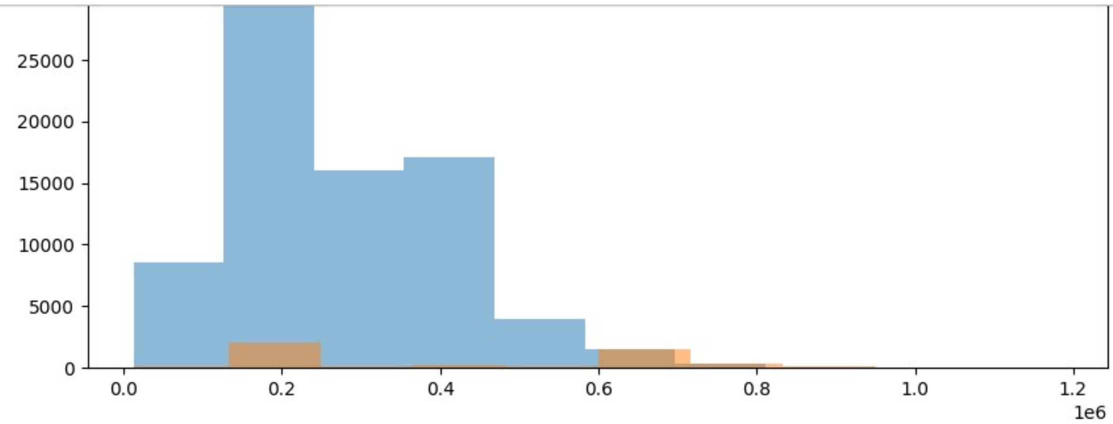
Outlier Detection and Assessment:

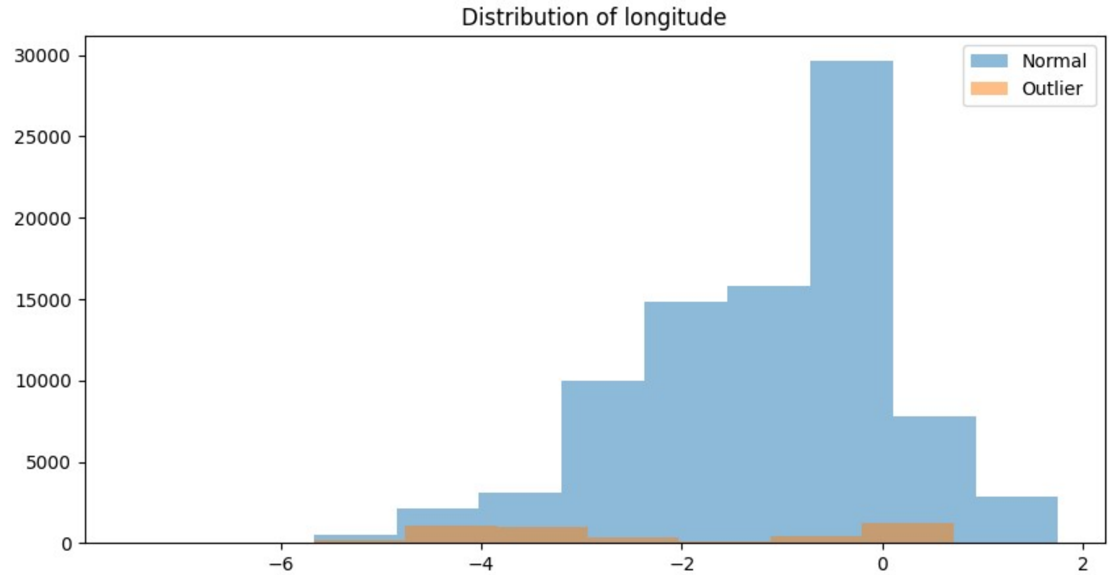
For the most part, there seems to be nothing out of the ordinary with the Outliers detected by IsolationForest algorithm. So we leave them in.



Distribution of month







Predictive Model Development:

CatBoost and LGBM algorithms were chosen for this project. Let's analyze their performance:

CatBoost:

- Accuracy: 0.52
- Weighted F1-score: 0.60

```
[35]: print(classification_report(y_valid, cat_pred))
```

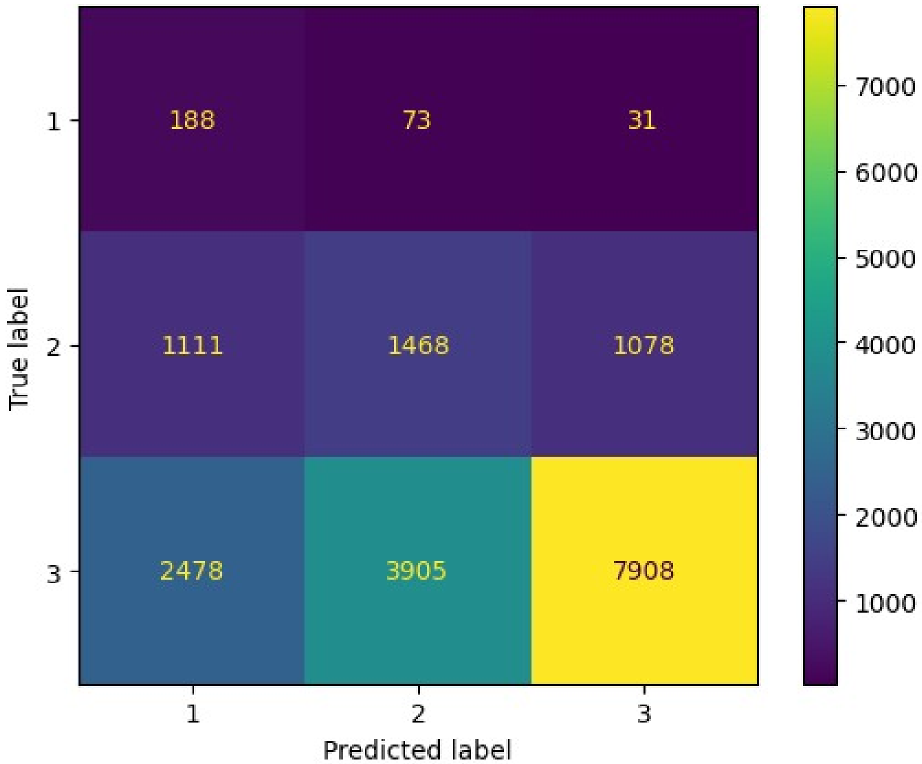
```
              precision    recall  f1-score   support

           1       0.05      0.64      0.09       292
           2       0.27      0.40      0.32      3657
           3       0.88      0.55      0.68     14291

    accuracy                           0.52     18240
   macro avg       0.40      0.53      0.36     18240
weighted avg       0.74      0.52      0.60     18240
```

```
[36]: cm = confusion_matrix(y_valid, cat_pred, labels=cat.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=cat.classes_)
disp.plot()

plt.show()
```



LightGBM:

- Accuracy: 0.66
- Weighted F1-score: 0.68

[39]:
```
print(classification_report(y_valid, lgb_pred))
```

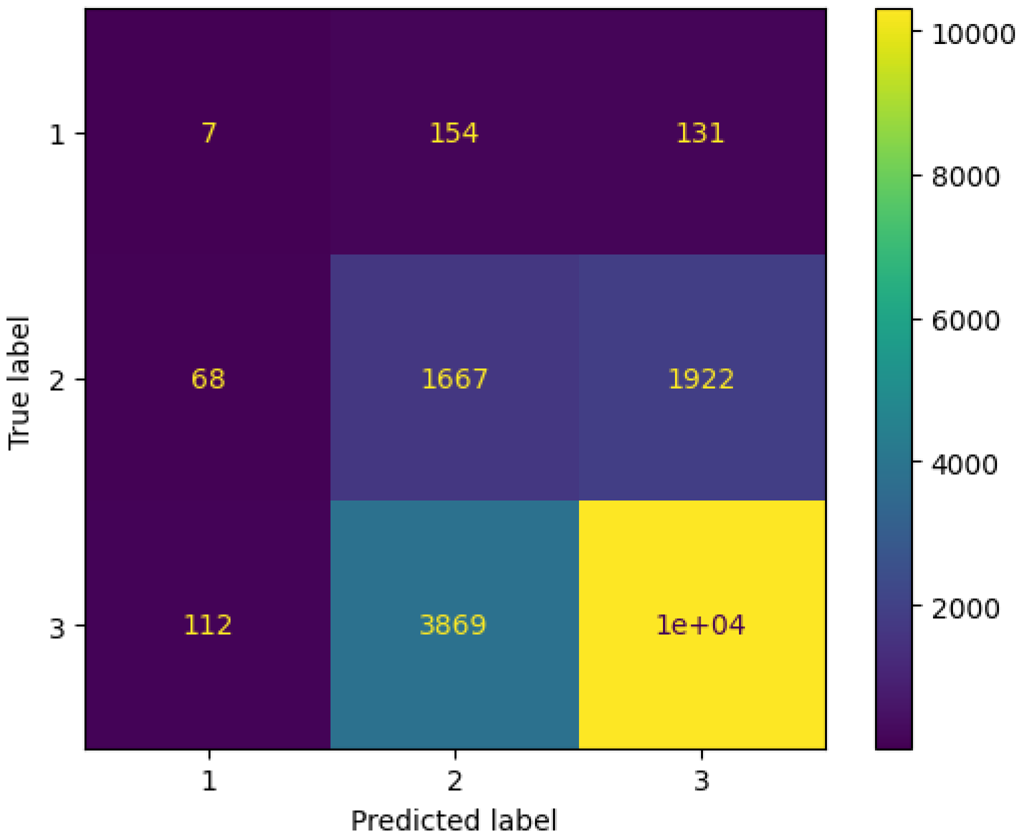
```
              precision    recall  f1-score   support

           1       0.04      0.02      0.03       292
           2       0.29      0.46      0.36      3657
           3       0.83      0.72      0.77     14291

    accuracy                           0.66     18240
   macro avg       0.39      0.40      0.39     18240
weighted avg       0.71      0.66      0.68     18240
```

[40]:
```
cm = confusion_matrix(y_valid, lgb_pred, labels=cat.classes_)
disp = ConfusionMatrixDisplay(confusion_matrix=cm,
                              display_labels=lgb.classes_)
disp.plot()

plt.show()
```



The LightGBM model seems to perform slightly better. However, both models struggle with predicting the least common class (severity 1). This suggests class imbalance issues.

Recommendations for Government Agencies

1. **Enhanced Traffic Management**:
   - Implement stricter speed limits and traffic calming measures in high-risk urban areas.
   - Increase police presence during peak hours to manage traffic flow and deter reckless driving.
2. **Infrastructure Improvements**:
   - Improve road surface conditions, especially in high-accident areas.
   - Enhance lighting at junctions and non-junction areas to reduce accidents during low visibility conditions.
3. **Public Awareness Campaigns**:
   - Target younger drivers with campaigns on safe driving practices.
   - Promote the use of pedestrian crossing facilities and educate the public on road safety.
4. **Policy Interventions**:
   - Introduce stricter regulations for older vehicles and those with larger engine capacities.
   - Implement policies to encourage the use of public transport during peak hours to reduce traffic congestion.
5. **Socioeconomic Support**:
   - Focus on improving road safety in lower-income areas through targeted interventions.
   - Provide support for road maintenance workers to ensure safer working conditions.
6. **Technology and Innovation**:
   - Invest in advanced traffic monitoring systems to identify and address accident hotspots.
   - Encourage the adoption of vehicle safety technologies, such as automatic emergency braking and lane-keeping assist.