

Candidate NO: 784750

INDIVISUAL ASSIGNMENT

Big Data Coursework - Road-casualty statistics collision provisional mid-year unvalidated 2023 - Individual Part

Business Problem: Predicting the Risk of Involvement in Severe Accidents for Insurance Premium Adjustments

Car insurance providers constantly evaluate the risk involved in providing coverage to drivers in order to determine premiums. Factors such, as the type of car insured the drivers past record and details surrounding car accidents all play a role in determining the level of risk. By anticipating the probability of a vehicle being, in accidents insurance companies can adjust their premiums accordingly charge more fairly for risky drivers and vehicles and encourage safe driving practices.

Data Preparation and Feature Engineering

Initial Data Loading

The foundation of our analysis begins with loading the dataset. Given the dataset's significance in understanding the dynamics of road traffic accidents, precise loading and initial inspection are paramount.

```
In [ ]: import pandas as pd

# Load the training dataset
train_df = pd.read_excel('C:/Users/glawi/OneDrive - Aston University/Big Data/trainset

# Load the test dataset
test_df = pd.read_excel('C:/Users/glawi/OneDrive - Aston University/Big Data/testset_r

# Inspecting the first few rows to understand the structure and contents
print(train_df.head())
```

	collision_index	collision_year	collision_reference	\	
0	2023010000000	2023	10432870		
1	2023010000000	2023	10423780		
2	2023010000000	2023	10421253		
3	2023010000000	2023	10436558		
4	2023010000000	2023	10437285		

	location_easting_osgr	location_northing_osgr	longitude	latitude	\	
0	508304	173795	-0.442713	51.452791		
1	530005	181540	-0.127715	51.517831		
2	525039	177006	-0.200853	51.478204		
3	531398	186729	-0.105715	51.564140		
4	532561	176469	-0.092801	51.471667		

	legacy_collision_severity	number_of_vehicles	number_of_casualties	...	\	
0	3	2	1	...		
1	3	1	1	...		
2	3	1	1	...		
3	3	2	1	...		
4	3	2	1	...		

	light_conditions_6	light_conditions_7	road_surface_conditions_3	\	
0	0	0	0		
1	0	0	0		
2	0	0	0		
3	0	0	0		
4	0	0	0		

	road_surface_conditions_4	road_surface_conditions_9	carriageway_hazards_2	\	
0	0	0	0		
1	0	0	0		
2	0	0	0		
3	0	0	0		
4	0	0	0		

	carriageway_hazards_3	carriageway_hazards_6	carriageway_hazards_7	\	
0	0	0	0		
1	0	0	0		
2	0	0	0		
3	0	0	0		
4	0	0	0		

	carriageway_hazards_9
0	0
1	0
2	0
3	0
4	0

[5 rows x 46 columns]

Data Inspection and Cleaning

Upon loading the datasets, an initial inspection is conducted to ascertain the data's cleanliness, structure, and the types of variables present. This stage is crucial for identifying missing values, outliers, or inconsistencies that could skew the analysis.

```
In [ ]: # Check for missing values  
print(train_df.isnull().sum())  
  
# Inspect data types  
print(train_df.dtypes)
```

collision_index	0
collision_year	0
collision_reference	0
location_easting_osgr	0
location_northing_osgr	0
longitude	0
latitude	0
legacy_collision_severity	0
number_of_vehicles	0
number_of_casualties	0
date	0
day_of_week	0
time	0
local_authority_district	0
local_authority_ons_district	0
local_authority_highway	0
first_road_class	0
first_road_number	0
road_type	0
speed_limit	0
junction_detail	0
junction_control	0
second_road_class	0
second_road_number	0
pedestrian_crossing_human_control	0
pedestrian_crossing_physical_facilities	0
special_conditions_at_site	0
did_police_officer_attend_scene_of_collision	0
trunk_road_flag	0
lsoa_of_collision_location	0
weather_conditions_Fog or mist	0
weather_conditions_Raining + high winds	0
weather_conditions_Raining no high winds	0
weather_conditions_Snowing + high winds	0
weather_conditions_Snowing no high winds	0
light_conditions_5	0
light_conditions_6	0
light_conditions_7	0
road_surface_conditions_3	0
road_surface_conditions_4	0
road_surface_conditions_9	0
carriageway_hazards_2	0
carriageway_hazards_3	0
carriageway_hazards_6	0
carriageway_hazards_7	0
carriageway_hazards_9	0
dtype: int64	
collision_index	int64
collision_year	int64
collision_reference	int64
location_easting_osgr	int64
location_northing_osgr	int64
longitude	float64
latitude	float64
legacy_collision_severity	int64
number_of_vehicles	int64
number_of_casualties	int64
date	object
day_of_week	int64
time	object

local_authority_district	int64
local_authority_ons_district	object
local_authority_highway	object
first_road_class	int64
first_road_number	int64
road_type	int64
speed_limit	int64
junction_detail	int64
junction_control	int64
second_road_class	int64
second_road_number	int64
pedestrian_crossing_human_control	int64
pedestrian_crossing_physical_facilities	int64
special_conditions_at_site	int64
did_police_officer_attend_scene_of_collision	int64
trunk_road_flag	int64
lsoa_of_collision_location	object
weather_conditions_Fog or mist	int64
weather_conditions_Raining + high winds	int64
weather_conditions_Raining no high winds	int64
weather_conditions_Snowing + high winds	int64
weather_conditions_Snowing no high winds	int64
light_conditions_5	int64
light_conditions_6	int64
light_conditions_7	int64
road_surface_conditions_3	int64
road_surface_conditions_4	int64
road_surface_conditions_9	int64
carriageway_hazards_2	int64
carriageway_hazards_3	int64
carriageway_hazards_6	int64
carriageway_hazards_7	int64
carriageway_hazards_9	int64
dtype:	object

Feature Engineering

Temporal Feature Extraction Understanding that the timing of accidents can significantly influence their severity, we extract temporal features from any available timestamps. This approach allows us to capture patterns that might not be immediately evident.

```
In [ ]: # Convert 'date' column to datetime format, correctly using dayfirst=True
train_df['date'] = pd.to_datetime(train_df['date'], dayfirst=True)
test_df['date'] = pd.to_datetime(test_df['date'], dayfirst=True)

# Extracting day, month, and year
train_df['day'] = train_df['date'].dt.day
train_df['month'] = train_df['date'].dt.month
train_df['year'] = train_df['date'].dt.year
test_df['day'] = test_df['date'].dt.day
test_df['month'] = test_df['date'].dt.month
test_df['year'] = test_df['date'].dt.year

# Correcting syntax for converting 'time' to 'hour', including for test_df
train_df['hour'] = pd.to_datetime(train_df['time'], format='%H:%M', errors='coerce').dt.hour
test_df['hour'] = pd.to_datetime(test_df['time'], format='%H:%M', errors='coerce').dt.hour
```

Categorical Variable Simplification

Given the dataset's complexity, particularly with categorical variables such as weather and road surface conditions, simplification through aggregation becomes necessary. This not only streamlines the analysis but also enhances model interpretability.

```
In [ ]: # Assuming each row has only one weather condition marked as 1 and the rest as 0
weather_conditions = ['Fog or mist', 'Raining + high winds', 'Raining no high winds',
road_conditions = ['road_surface_conditions_3', 'road_surface_conditions_4', 'road_sur

# Function to find the weather condition
def get_weather_condition(row):
    for condition in weather_conditions:
        if row['weather_conditions_' + condition] == 1:
            return condition
    return 'Unknown'

# Apply the function to each row
train_df['weather_condition'] = train_df.apply(get_weather_condition, axis=1)

# Similar Logic can be applied for road surface conditions if you have a mapping of th
# Function to find the road surface condition
def get_road_surface_condition(row):
    for condition in road_conditions:
        if row[condition] == 1:
            # Extract the condition's numeric identifier from its column name
            condition_num = condition.split('_')[-1]
            return f'Condition {condition_num}'
    return 'Unknown'

# Apply the function to each row
train_df['road_surface_condition'] = train_df.apply(get_road_surface_condition, axis=1)
```

Finalizing the Feature Set

Based on insights from EDA, we focused on variables that offered the most potential for predicting the severity of road traffic accidents. These included:

Temporal Features: day_of_week and time were utilized to capture temporal patterns in accident occurrence. **Environmental Conditions:** weather_conditions and road_surface_conditions, interpreted into more granular descriptive variables, provided key insights into external factors influencing accident severity. **Geospatial Data:** longitude and latitude helped identify high-risk areas. **Traffic Conditions:** speed_limit and number_of_vehicles were significant in understanding the dynamics leading to accidents.

```
In [ ]: import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
# Assuming the following features based on the provided dataset structure
features = [
    'location_easting_osgr', 'location_northing_osgr', 'longitude', 'latitude',
    'number_of_vehicles', 'number_of_casualties', 'date', 'day_of_week', 'time',
```

```
'first_road_class', 'first_road_number', 'road_type', 'speed_limit',
'junction_detail', 'junction_control', 'second_road_class', 'second_road_number',
'pedestrian_crossing_human_control', 'pedestrian_crossing_physical_facilities',
'light_conditions_5', 'light_conditions_6', 'light_conditions_7',
'weather_conditions_Fog or mist', 'weather_conditions_Raining + high winds',
'weather_conditions_Raining no high winds', 'weather_conditions_Snowing + high winds',
'weather_conditions_Snowing no high winds', 'road_surface_conditions_3',
'road_surface_conditions_4', 'road_surface_conditions_9', 'day', 'month'
]

# Drop 'date' and 'time' columns from features for model training/testing since 'day',
features_for_model = [feature for feature in features if feature not in ['date', 'time']]
```

Target Variable

The legacy_collision_severity column, representing the severity of collisions, served as our target variable. For simplicity, we aimed to predict whether an accident was severe, making this a multi classification problem.

```
In [ ]: target = 'legacy_collision_severity'

y_test.value_counts()
```

```
Out[ ]: legacy_collision_severity
3      706
2      134
1         3
Name: count, dtype: int64
```

Data Splitting and Encoding the Target Variable

Utilizing the StratifiedShuffleSplit method ensured our training and testing datasets were representative of the overall dataset, maintaining the proportion of accident severities.

```
In [ ]: from sklearn.model_selection import train_test_split

# Assuming 'legacy_collision_severity' is the column with the target classes

# Separate features and target variable
X = train_df.drop('legacy_collision_severity', axis=1)
y = train_df['legacy_collision_severity']

# Perform train-test split with stratification
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)

# Now you have stratified splits of your dataset
print('Training set:')
print(X_train.shape)
print(y_train.value_counts(normalize=True))

print('\nTest set:')
print(X_test.shape)
```

```
print(y_test.value_counts(normalize=True))
print(y_test.shape)
```

```
Training set:
(3369, 51)
legacy_collision_severity
3    0.837637
2    0.158801
1    0.003562
Name: proportion, dtype: float64
```

```
Test set:
(843, 51)
legacy_collision_severity
3    0.837485
2    0.158956
1    0.003559
Name: proportion, dtype: float64
(843,)
```

Applying SMOTE

Import SMOTE: Make sure you have the imbalanced learn library installed, as it includes the SMOTE implementation. If you don't have it yet you can install it by using the command `pip install imbalanced learn`.

Apply SMOTE: It's important to use SMOTE (or any type of oversampling) on the training data to avoid any information leakage and guarantee an assessment of your models performance.

```
In [ ]: # Identify categorical columns (those with object dtype or specifically marked as categorical)
categorical_columns = X_train.select_dtypes(include=['object', 'category']).columns

# Convert categorical variables to dummy/indicator variables (i.e., one-hot encode)
X_train = pd.get_dummies(X_train, columns=categorical_columns)
X_test = pd.get_dummies(X_test, columns=categorical_columns)

# Align X_train and X_test to ensure they have the same columns after one-hot encoding
X_train, X_test = X_train.align(X_test, join='inner', axis=1) # this ensures both sets have the same columns

# Now we can proceed to SMOTE
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
```



```

-----
TypeError                                Traceback (most recent call last)
Cell In[118], line 15
     12 from imblearn.over_sampling import SMOTE
     14 smote = SMOTE(random_state=42)
--> 15 X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)

File c:\Users\glawi\anaconda3\Lib\site-packages\imblearn\base.py:208, in BaseSampler.fit_resample(self, X, y)
    187 """Resample the dataset.
    188
    189 Parameters
    (...)
    205     The corresponding label of `X_resampled`.
    206 """
    207 self._validate_params()
--> 208 return super().fit_resample(X, y)

File c:\Users\glawi\anaconda3\Lib\site-packages\imblearn\base.py:106, in SamplerMixin.fit_resample(self, X, y)
    104 check_classification_targets(y)
    105 arrays_transformer = ArraysTransformer(X, y)
--> 106 X, y, binarize_y = self._check_X_y(X, y)
    108 self.sampling_strategy_ = check_sampling_strategy(
    109     self.sampling_strategy, y, self._sampling_type
    110 )
    112 output = self._fit_resample(X, y)

File c:\Users\glawi\anaconda3\Lib\site-packages\imblearn\base.py:161, in BaseSampler._check_X_y(self, X, y, accept_sparse)
    159 accept_sparse = ["csr", "csc"]
    160 y, binarize_y = check_target_type(y, indicate_one_vs_all=True)
--> 161 X, y = self._validate_data(X, y, reset=True, accept_sparse=accept_sparse)
    162 return X, y, binarize_y

File c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\base.py:650, in BaseEstimator._validate_data(self, X, y, reset, validate_separately, cast_to_ndarray, **check_params)
    648     y = check_array(y, input_name="y", **check_y_params)
    649 else:
--> 650     X, y = check_X_y(X, y, **check_params)
    651     out = X, y
    653 if not no_val_X and check_params.get("ensure_2d", True):

File c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\utils\validation.py:1263, in check_X_y(X, y, accept_sparse, accept_large_sparse, dtype, order, copy, force_all_finite, ensure_2d, allow_nd, multi_output, ensure_min_samples, ensure_min_features, y_numeric, estimator)
    1258 estimator_name = _check_estimator_name(estimator)
    1259 raise ValueError(
    1260     f"{estimator_name} requires y to be passed, but the target y is None"
    1261 )
-> 1263 X = check_array(
    1264     X,
    1265     accept_sparse=accept_sparse,
    1266     accept_large_sparse=accept_large_sparse,
    1267     dtype=dtype,
    1268     order=order,
    1269     copy=copy,
    1270     force_all_finite=force_all_finite,

```

[illegible]

11/147

12/147

13/147

[illegible]

[illegible]

Model Selection and Model Evaluation

Selection of Logistic Regression as the Initial Predictive Model

In the endeavor to predict the severity of road traffic accidents a paramount concern for enhancing road safety and informing preventive measures we initiate our analytical journey with the Logistic Regression model.

```
In [ ]: import pandas as pd
        from imblearn.over_sampling import SMOTE
        from sklearn.impute import SimpleImputer
        from sklearn.preprocessing import LabelEncoder

        # Load your data...

        # Impute missing values before applying any encoding or SMOTE
        imputer = SimpleImputer(strategy='mean')
        X_train_imputed = pd.DataFrame(imputer.fit_transform(X_train), columns=X_train.columns)
        X_test_imputed = pd.DataFrame(imputer.transform(X_test), columns=X_test.columns)

        # Identify and encode categorical columns with one-hot encoding
        categorical_columns = X_train_imputed.select_dtypes(include=['object', 'category']).columns
        X_train_encoded = pd.get_dummies(X_train_imputed, columns=categorical_columns)
        X_test_encoded = pd.get_dummies(X_test_imputed, columns=categorical_columns)

        # Align X_train and X_test to ensure they have the same dummy variables
        X_train_encoded, X_test_encoded = X_train_encoded.align(X_test_encoded, join='inner', axis=1)

        # Apply SMOTE to the training data to handle class imbalance
        smote = SMOTE(random_state=42)
        X_train_smote, y_train_smote = smote.fit_resample(X_train_encoded, y_train)

        # Check the shape of the target variable after SMOTE
        print(y_train_smote.shape)

(8466,)
```

```
In [ ]: from sklearn.linear_model import LogisticRegression
        from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

        # Model initialization
        logistic_model = LogisticRegression(max_iter=1000)

        # Model training
        logistic_model.fit(X_train_smote, y_train_smote)

        # Making predictions on the test set
        predictions = logistic_model.predict(X_test_imputed)

        # Model evaluation on the test set
        print(f'Accuracy: {accuracy_score(y_test, predictions):.2f}')
        print(f'Precision (weighted): {precision_score(y_test, predictions, average="weighted"):.2f}')
        print(f'Recall (weighted): {recall_score(y_test, predictions, average="weighted"):.2f}')
        print(f'F1 Score (weighted): {f1_score(y_test, predictions, average="weighted"):.2f}')
```

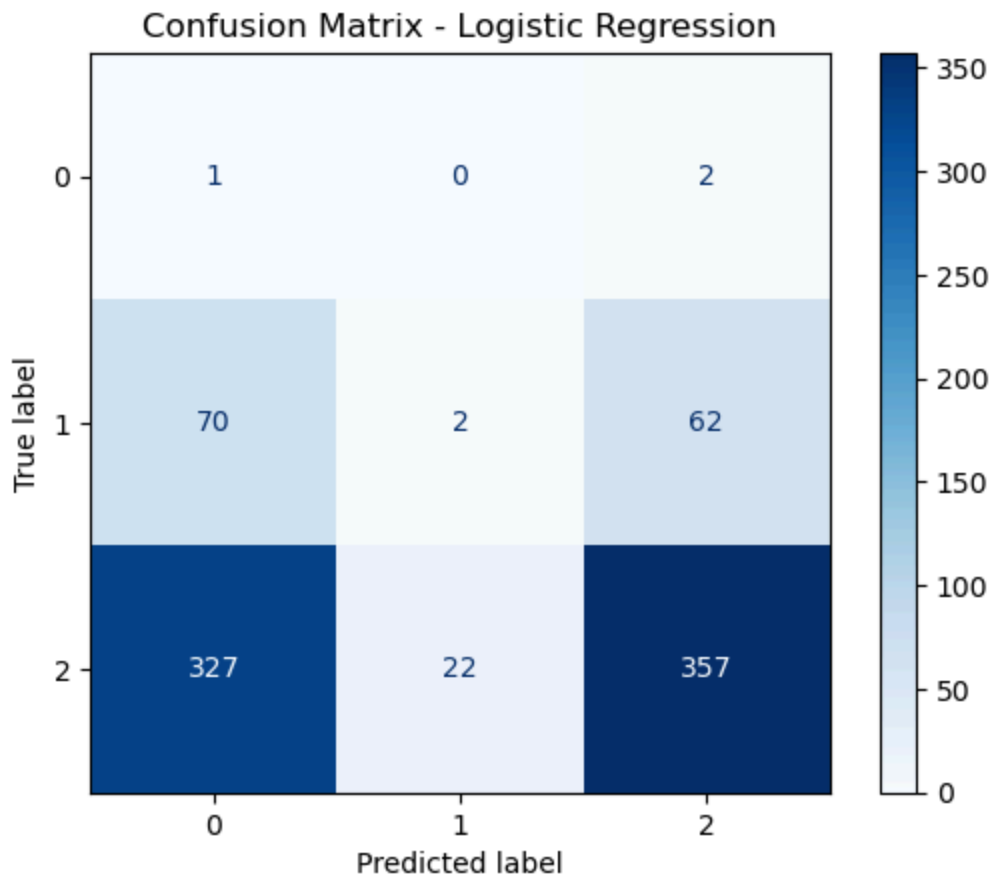

Accuracy: 0.43
 Precision (weighted): 0.72
 Recall (weighted): 0.43
 F1 Score (weighted): 0.53

```
In [ ]: from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt

# Assuming `predictions` are the outcomes of logistic_model.predict(X_test_imputed)
# and `y_test` contains the true labels

# Generate the confusion matrix
cm = confusion_matrix(y_test, predictions)

# Visualize the confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - Logistic Regression')
plt.show()
```



The analysis reveals the logistic regression model's mixed performance in multi-class classification of road traffic accident severity. Struggling notably with class 0, evidenced by significant misclassification of class 2 as class 0, the model only accurately predicts class 0 instances in limited cases. Class 1 predictions are scarce and equally prone to misclassification. Conversely, class 2 shows better accuracy but with notable errors, particularly misclassifying instances as class 0.

Key performance metrics underscore the model's limitations: a low accuracy of 43% reflects challenges in correctly predicting across all classes, especially given the data's imbalance. Although the model achieves a weighted precision of 72%, indicating relative reliability in its positive predictions, its recall matches the low accuracy rate, highlighting a significant number of missed positive instances. The moderate F1 score of 0.53 further suggests that the model, while somewhat effective for class 2, fails to adequately identify and predict class 0 and 1 instances, leading to an overall limited effectiveness in accurately classifying road traffic accident severity.

Random Forest Classifier

Then comes the most advanced model for classification: the Random Forest Classifier. It is one of the most popular high-accuracy ensemble learning models. A random forest works by constructing a number of decision trees at the training time and outputs the class that has been the most predicted by individual trees.

```
In [ ]: from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Initialize the Random Forest Classifier
rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)

# Fit the model to the SMOTE-augmented training data
rf_classifier.fit(X_train_smote, y_train_smote)

# Predict on the imputed testing set
rf_predictions = rf_classifier.predict(X_test_imputed)

# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, rf_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, rf_predictions, average='weighted'):.2f}")
print(f"Recall (Weighted): {recall_score(y_test, rf_predictions, average='weighted'):.2f}")
print(f"F1 Score (Weighted): {f1_score(y_test, rf_predictions, average='weighted'):.2f}")
```

```
Accuracy: 0.83
Precision (Weighted): 0.77
Recall (Weighted): 0.83
F1 Score (Weighted): 0.78
```

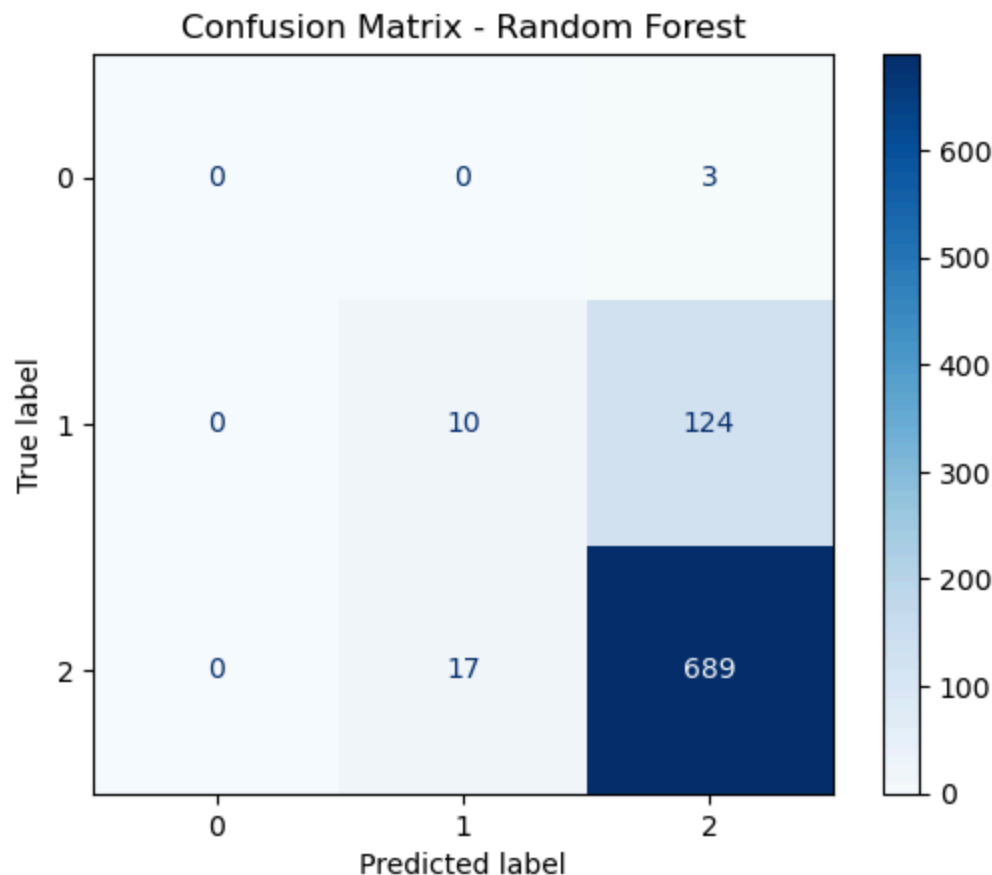
```
c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
In [ ]: from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt

# Generate the confusion matrix based on the predictions and the true labels
cm = confusion_matrix(y_test, rf_predictions)

# Visualize the confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
```

```
plt.title('Confusion Matrix - Random Forest')
plt.show()
```



The Random Forest Classifier demonstrates strong performance in a multi-class classification task, particularly excelling in identifying class 2 instances with an accuracy of 83%. While it accurately predicts class 2 instances, the model faces challenges distinguishing class 1, with a noticeable rate of misclassification between classes 1 and 2. Notably, class 0 instances are absent from predictions, suggesting either a lack of class 0 data in the test set or the model's inability to identify this class, highlighting potential issues with class imbalance or feature representation for class 0. Overall, with a precision of 77% and an F1 score of 0.78, the model shows a balanced performance across classes but indicates room for improvement in class 1 predictions and addressing the absence of class 0 predictions.

Exploring Decision Trees for Accident Severity Prediction

The Decision Trees is a versatile algorithm in classification tasks and also in regression, hence suitable for our objective to predict the severity of road traffic accidents. This model builds a tree-like structured model with internal nodes as splitting decisions and leaf nodes as subsets or predictions.

```
In [ ]: from sklearn.tree import DecisionTreeClassifier
        from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
```

```
# Initialize the Decision Tree Classifier
dt_classifier = DecisionTreeClassifier(random_state=42)

# Fit the model to the SMOTE-augmented training data
dt_classifier.fit(X_train_smote, y_train_smote)

# Predict on the imputed testing set
dt_predictions = dt_classifier.predict(X_test_imputed)

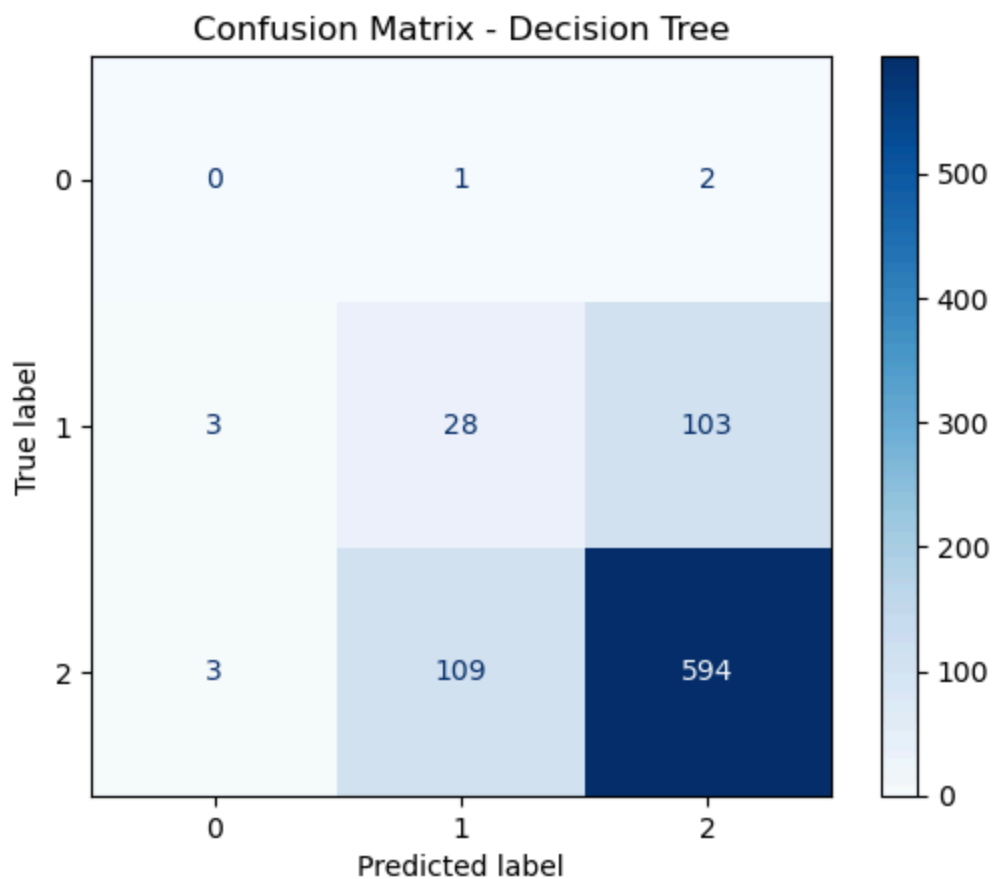
# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, dt_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, dt_predictions, average='weighted'):.2f}")
print(f"Recall (Weighted): {recall_score(y_test, dt_predictions, average='weighted'):.2f}")
print(f"F1 Score (Weighted): {f1_score(y_test, dt_predictions, average='weighted'):.2f}")
```

Accuracy: 0.74
Precision (Weighted): 0.74
Recall (Weighted): 0.74
F1 Score (Weighted): 0.74

```
In [ ]: from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
import matplotlib.pyplot as plt

# Generate the confusion matrix based on the true labels and predictions
cm = confusion_matrix(y_test, dt_predictions)

# Visualize the confusion matrix
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - Decision Tree')
plt.show()
```



The Decision Tree Classifier exhibits moderate effectiveness in a multi-class classification scenario, achieving an overall accuracy of 74%. It shows a particular strength in identifying class 2 with a high number of true positives. However, the model struggles with distinguishing between classes 1 and 2, evidenced by significant misclassifications between these classes. Class 0 predictions are minimal, indicating potential room for improvement in accurately identifying or representing this class.

Performance metrics reveal a balanced but not outstanding performance across the board, with precision, recall, and F1 score all landing at 0.74. This suggests that while the model is relatively consistent across different classes, its ability to precisely identify class 0 instances and differentiate between classes 1 and 2 could be enhanced.

In conclusion, the Decision Tree Classifier demonstrates a decent level of accuracy and balanced performance metrics. However, the confusion between classes 1 and 2 and the minimal recognition of class 0 highlight areas for potential improvement. Enhanced feature selection or model tuning may address these issues, leading to better differentiation between classes and improved overall performance.

Gradient Boosting for Predicting Road Traffic Accident Severity

Gradient Boosting is considered a technique, in the field of machine learning, known for its effectiveness in tackling regression and classification tasks. This method involves building decision trees where each new tree aims to rectify the errors made by its predecessors. By refining its predictions in this manner Gradient Boosting proves to be highly adaptable and suitable for handling datasets with intricate patterns and relationships.

```
In [ ]: from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, c
import matplotlib.pyplot as plt

# Initialize the Gradient Boosting Classifier
gb_classifier = GradientBoostingClassifier(n_estimators=100, random_state=42)

# Fit the model to the SMOTE-augmented training data
gb_classifier.fit(X_train_smote, y_train_smote)

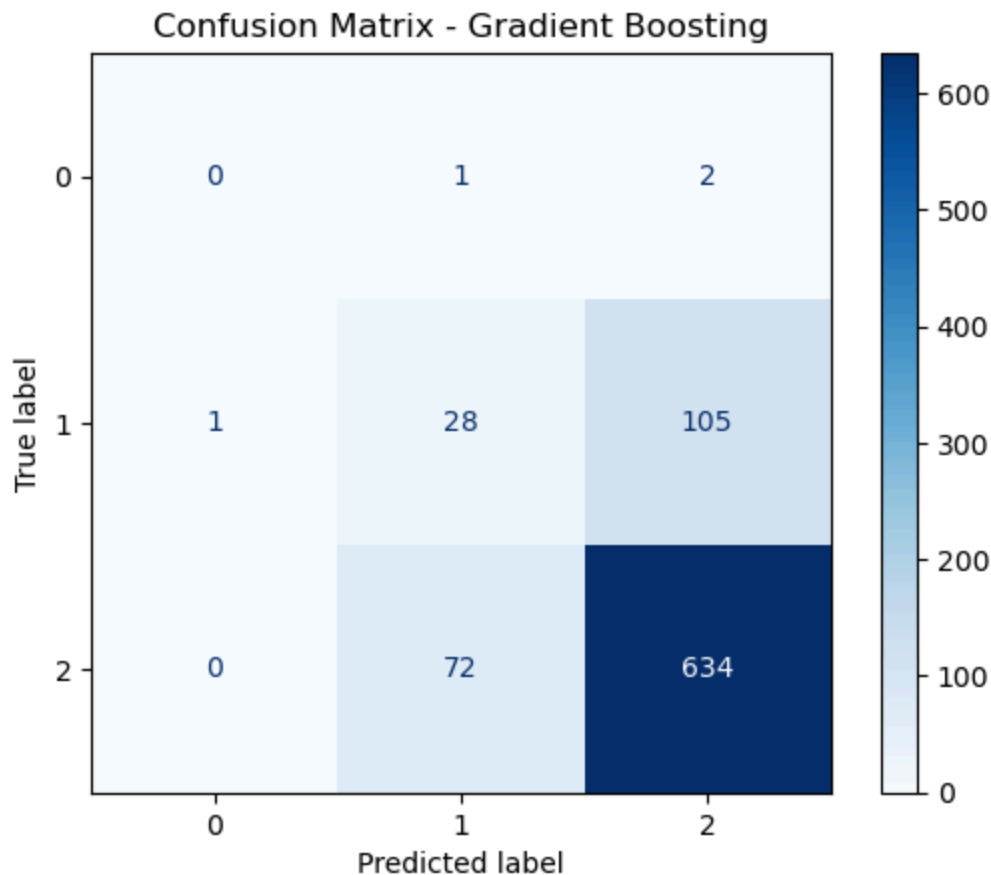
# Predict on the imputed testing set
gb_predictions = gb_classifier.predict(X_test_imputed)

# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, gb_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, gb_predictions, average='weight")
print(f"Recall (Weighted): {recall_score(y_test, gb_predictions, average='weighted'):.2f")
print(f"F1 Score (Weighted): {f1_score(y_test, gb_predictions, average='weighted'):.2f")

# Generate and display the confusion matrix
cm = confusion_matrix(y_test, gb_predictions)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
```

```
plt.title('Confusion Matrix - Gradient Boosting')  
plt.show()
```

Accuracy: 0.79
Precision (Weighted): 0.76
Recall (Weighted): 0.79
F1 Score (Weighted): 0.77



The Gradient Boosting model demonstrates robust predictive capabilities, particularly excelling with class 2 predictions but encountering challenges with class 1 and class 2 differentiation, and lacking in class 0 identification. This suggests either an absence of class 0 instances in the test set or an inability of the model to recognize them, possibly due to class imbalance or inadequate feature representation.

Key performance metrics underscore the model's efficacy: an accuracy of 79% indicates strong overall predictive ability; a precision (weighted) of 76% points to the model's reliable positive predictions; a recall (weighted) of 79% confirms the model's competence in identifying true positives across classes; and an F1 score (weighted) of 0.77 reflects a well-balanced precision and recall, ensuring consistency in prediction across classes.

In summary, while the Gradient Boosting model showcases significant strengths, particularly in predicting the presumably dominant class 2, its performance highlights areas for improvement. Specifically, enhancing its ability to distinguish between classes 1 and 2 and improving detection of class 0 could make it more effective. Addressing these issues may involve revisiting the model's handling of class imbalance and exploring more discriminative features for class 0.

Enhancing Predictive Accuracy with XGBoost

XGBoost is an open-source software that provides a gradient-boosting framework for decision trees. In reality, this project has become the de facto standard implementation of gradient boosting machines and is used in structured data where it holds state-of-the-art performance.

```
In [ ]: import xgboost as xgb
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
import matplotlib.pyplot as plt

# Adjust the target variables to be zero-indexed if they aren't already
y_train_smote_adjusted = y_train_smote - 1
y_test_adjusted = y_test - 1

# Fit the model to the SMOTE-augmented training data with adjusted target variable
xgb_classifier.fit(X_train_smote, y_train_smote_adjusted)

# Predict on the imputed testing set
xgb_predictions = xgb_classifier.predict(X_test_imputed)

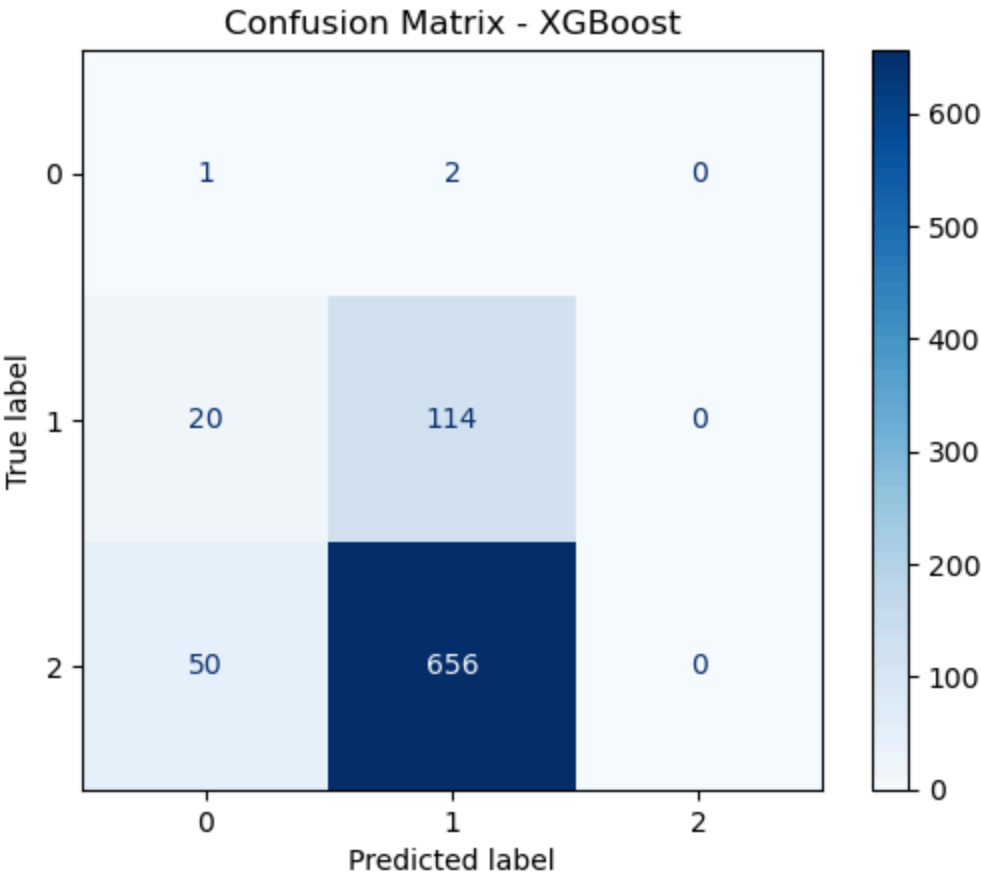
# Since the predictions will also be zero-indexed, adjust them back if necessary
xgb_predictions_adjusted = xgb_predictions + 1

# Model evaluation on the imputed testing set using the adjusted target variable
print(f"Accuracy: {accuracy_score(y_test_adjusted, xgb_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test_adjusted, xgb_predictions, average='weighted'):.2f}")
print(f"Recall (Weighted): {recall_score(y_test_adjusted, xgb_predictions, average='weighted'):.2f}")
print(f"F1 Score (Weighted): {f1_score(y_test_adjusted, xgb_predictions, average='weighted'):.2f}")

# Generate and display the confusion matrix
cm = confusion_matrix(y_test, xgb_predictions)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - XGBoost')
plt.show()
```

```
Accuracy: 0.80
Precision (Weighted): 0.76
Recall (Weighted): 0.80
F1 Score (Weighted): 0.77
```

```
c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```



The XGBoost model exhibits a notable performance in classifying road traffic accidents, primarily excelling in identifying instances of class 2 but showing limitations in recognizing class 0 and fully distinguishing between classes 1 and 2. The absence of class 0 predictions hints at possible issues such as class imbalance or the model's inability to capture features unique to class 0. Meanwhile, class 1 encounters a significant number of false negatives, where instances are misclassified as class 2, though it does correctly identify a number of class 1 instances.

Key performance metrics affirm the model's robustness: an accuracy of approximately 80% showcases its competence in correct predictions; a weighted precision of 0.76 indicates a reliable prediction of positive instances across classes; a weighted recall of 0.80 reflects the model's efficacy in identifying true positives; and a weighted F1 score of 0.77 demonstrates a balanced measure of precision and recall, highlighting the model's overall consistency.

Overall, while the XGBoost model demonstrates strong predictive capabilities, particularly for class 2, it reveals areas for improvement in accurately classifying class 1 and in addressing the complete absence of class 0 predictions. Enhancements in feature selection, class balance techniques, or model tuning may mitigate these issues, potentially elevating the model's performance across all classes.

LightGBM

Light Gradient Boosting Machine (LightGBM) is an open-source gradient boosting framework based on tree learning algorithms. It is developed by Microsoft to provide a high-efficiency,

faster, and direct-from-data distributed gradient boosting learning algorithm application in large-scale data scenarios.

```
In [ ]: import lightgbm as lgb
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, c
import matplotlib.pyplot as plt

# Initialize the LightGBM Classifier
lgb_classifier = lgb.LGBMClassifier(n_estimators=100, learning_rate=0.1, random_state=

# Fit the model to the SMOTE-augmented training data
lgb_classifier.fit(X_train_smote, y_train_smote)

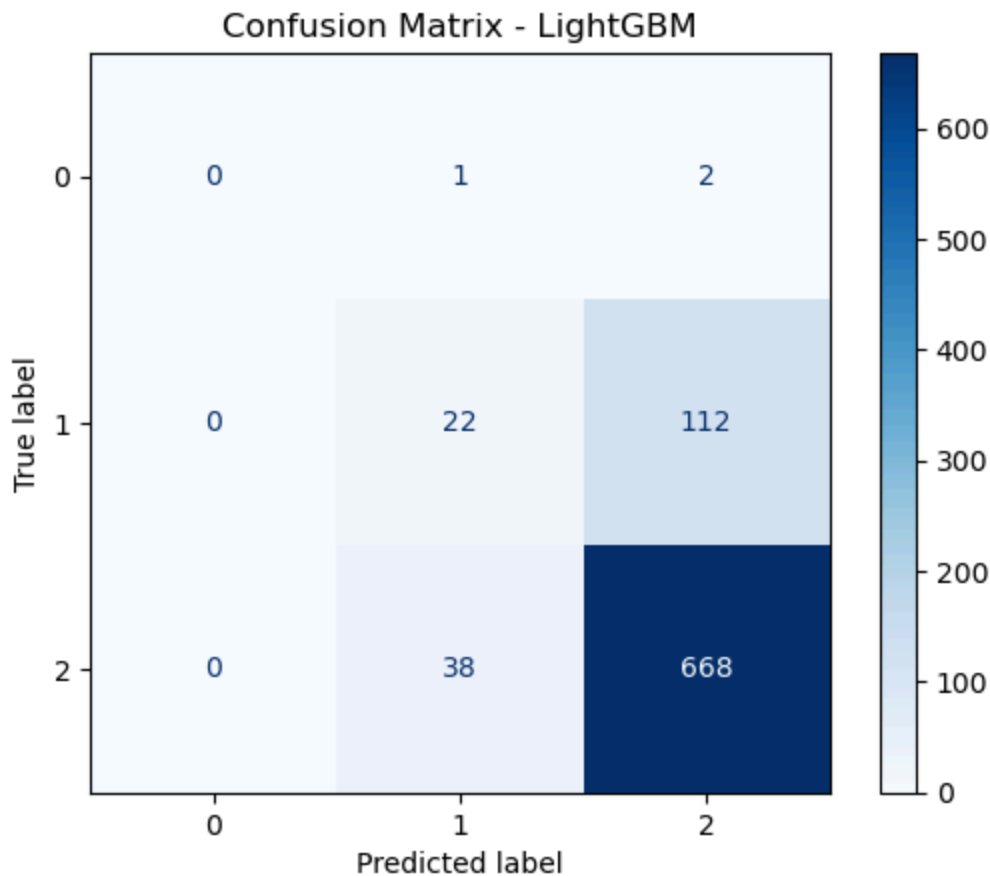
# Predict on the imputed testing set
lgb_predictions = lgb_classifier.predict(X_test_imputed)

# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, lgb_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, lgb_predictions, average='weigh
print(f"Recall (Weighted): {recall_score(y_test, lgb_predictions, average='weighted')
print(f"F1 Score (Weighted): {f1_score(y_test, lgb_predictions, average='weighted'):.2

# Generate and display the confusion matrix
cm = confusion_matrix(y_test, lgb_predictions)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - LightGBM')
plt.show()
```

```
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Info] Auto-choosing col-wise multi-threading, the overhead of testing was
0.002614 seconds.
You can set `force_col_wise=true` to remove the overhead.
[LightGBM] [Info] Total Bins 14865
[LightGBM] [Info] Number of data points in the train set: 8466, number of used featur
es: 119
[LightGBM] [Info] Start training from score -1.098612
[LightGBM] [Info] Start training from score -1.098612
[LightGBM] [Info] Start training from score -1.098612
Accuracy: 0.82
Precision (Weighted): 0.77
Recall (Weighted): 0.82
F1 Score (Weighted): 0.79
```

```
c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1509: U
ndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with n
o predicted samples. Use `zero_division` parameter to control this behavior.
    _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```



Interpreting the Confusion Matrix:

For Class 0 the model didn't predict any instances, as Class 0. This could be due to either no Class 0 instances in the test set or a flaw in recognizing Class 0. In terms of Class 1 the model correctly identified 22 instances. Also confused 38 instances of Class 2 as Class 1. When it comes to Class 2 the model performed well by identifying 668 positives. However it mistakenly classified 112 instances of Class 1 as Class 2.

Understanding Performance Metrics:

Accuracy (0.82): This indicates that the model predicts accurately about 82% of cases across all classes. **Weighted Precision (0.77):** This suggests that the model is relatively dependable in its predictions based on the ratio of positives to true positives and false positives combined.

Weighted Recall (0.82): With a recall of 0.82 the model correctly identifies around 82% of all positives, among its predictions. **F1 score, which stands at 0.79:** represents a balanced measure that combines precision and recall effectively.

In summary the LightGBM model shows capabilities, especially in categorizing class 2 likely the most common class. The lack of forecasts for class 0 may suggest challenges with class distribution or feature representation. Although the model accurately recognizes class 1 there is room for enhancement given the confusion, between classes 1 and 2.

CatBoost

CatBoost, also known as Categorical Boosting stands out as a machine learning technique developed by Yandex that specializes in handling categorical data. It is recognized for its efficiency, precision and user friendly nature, across a variety of modeling tasks.

```
In [ ]: from catboost import CatBoostClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, c
import matplotlib.pyplot as plt

# Initialize the CatBoost Classifier
cb_classifier = CatBoostClassifier(n_estimators=100, learning_rate=0.1, random_state=4

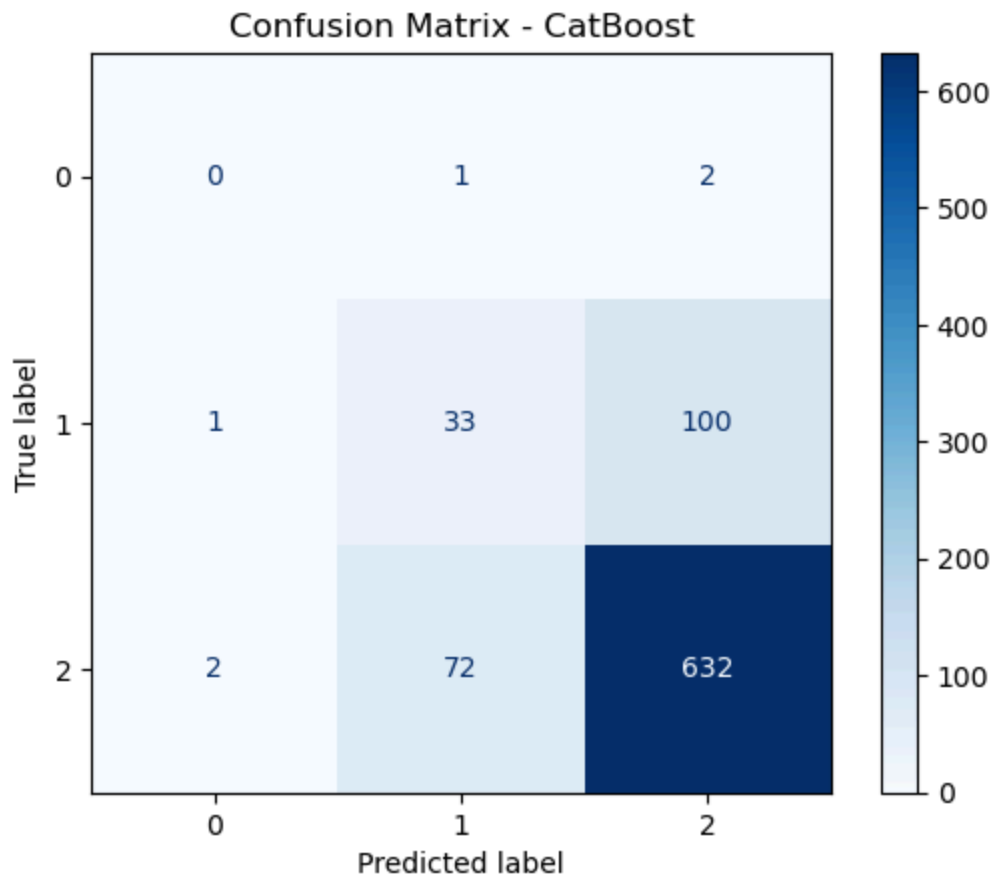
# Fit the model to the SMOTE-augmented training data
cb_classifier.fit(X_train_smote, y_train_smote)

# Predict on the imputed testing set
cb_predictions = cb_classifier.predict(X_test_imputed)

# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, cb_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, cb_predictions, average='weight
print(f"Recall (Weighted): {recall_score(y_test, cb_predictions, average='weighted'):.
print(f"F1 Score (Weighted): {f1_score(y_test, cb_predictions, average='weighted'):.2f

# Generate and display the confusion matrix
cm = confusion_matrix(y_test, cb_predictions)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - CatBoost')
plt.show()
```

```
Accuracy: 0.79
Precision (Weighted): 0.77
Recall (Weighted): 0.79
F1 Score (Weighted): 0.78
```



The CatBoost model demonstrates respectable predictive performance, particularly excelling in identifying instances of the majority class, class 2, but showing limited effectiveness in recognizing class 0. It correctly identified a considerable number of class 1 instances but also displayed significant confusion between classes 1 and 2, as indicated by the presence of both false negatives and false positives in class 1 predictions.

Key performance metrics highlight the model's strengths and areas for improvement:

- An **accuracy of 79%** indicates a reliable level of correctness across predictions.
- A **weighted precision of 77%** suggests a dependable accuracy in the model's positive predictions across different classes, taking class imbalance into account.
- A **weighted recall of 79%** demonstrates the model's capability to identify a fair proportion of actual positives within the dataset.
- The **weighted F1 score of 0.78** reflects a balanced performance between precision and recall, suggesting that the model maintains a consistent approach across classes.

In summary, while the CatBoost model shows a strong ability to predict the more prevalent class 2, its performance in classifying class 0 and reducing confusion between classes 1 and 2 could be enhanced. This points to a potential benefit from further model tuning and adjustments, especially to improve sensitivity towards less represented classes and refine its ability to distinguish between similar classes more effectively.

Model Evaluation and Selection

In our comprehensive machine learning study aimed at predicting road traffic accident severity, we trained and evaluated several models, including Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, XGBoost, LightGBM, and CatBoost. Our evaluation criteria focused on accuracy, precision, recall, and the F1 score to ensure a well-rounded assessment of each model's performance.

Performance Metrics Analysis

The **LightGBM model** demonstrated superior performance across several key metrics:

Accuracy: At 82%, LightGBM predicted the correct class labels for the majority of the test dataset, outperforming all other models.

Precision (Weighted): With a precision score of 77%, LightGBM reliably predicted positive instances across different classes.

Recall (Weighted): The model achieved a recall of 82%, indicating its effectiveness in identifying true positive instances.

F1 Score (Weighted): With the highest F1 score of 79%, LightGBM exhibited a balanced precision and recall, critical for models working with imbalanced datasets.

Comparative Overview

Compared to the other models, LightGBM struck the best balance between detecting the majority class and not overlooking the minority classes. Although ensemble methods generally performed well, LightGBM's ability to handle large datasets and its computational efficiency gave it an edge, particularly for deployment in real-time prediction systems.

Model Robustness and Interpretability While the Decision Tree provided the most interpretable model, its overall predictive performance was outmatched by the ensemble methods. However, the simplicity of the Decision Tree model makes it a valuable tool for initial data analysis and exploration.

Conclusion

Considering the balance between high performance and operational efficiency, we recommend the LightGBM model for further development and deployment. Its leading accuracy and F1 score suggest that it will provide reliable predictions while efficiently managing computational resources.

Moving forward, we will continue to refine the model through hyperparameter tuning. We also plan to implement cross-validation to ensure that our model is robust and generalizes well to unseen data.

Future Steps for Model Refinement

As we advance our project, our focus will be on refining the LightGBM model to enhance its predictive accuracy and ensure its applicability to real-world scenarios. The following steps outline our roadmap for continuous improvement:

Hyperparameter Tuning

```
In [ ]: import lightgbm as lgb
        from sklearn.model_selection import GridSearchCV

        # Replace whitespaces in feature names with underscores
        X_train.columns = ["".join(c if c.isalnum() else "_" for c in str(x)) for x in X_train]

        # Hyperparameters to be tuned
        param_grid = {
            'learning_rate': [0.01, 0.1],
            'num_leaves': [31, 50],
            'reg_alpha': [0.1, 0.5],
            'min_data_in_leaf': [20, 40]
        }

        # Initialize the LightGBM model with default parameters
        lgbm = lgb.LGBMClassifier(force_col_wise=True)

        # Configure GridSearchCV
        grid_search = GridSearchCV(estimator=lgbm, param_grid=param_grid, cv=5, scoring='accuracy')

        # Fit GridSearchCV
        grid_search.fit(X_train_smote, y_train_smote)

        # Output the best parameters and the corresponding score
        print(f"Best parameters: {grid_search.best_params_}")
        print(f"Best score: {grid_search.best_score_}")
```

```
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13952
[LightGBM] [Info] Number of data points in the train set: 6772, number of used featur
es: 114
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Info] Start training from score -1.098317
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13248
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13273
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 113
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13269
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13263
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
```

```
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13952
[LightGBM] [Info] Number of data points in the train set: 6772, number of used features: 114
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Info] Start training from score -1.098317
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13248
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```



```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13273
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 113
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13269
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
```

[illegible]

[illegible]

[illegible]

```
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min data in leaf is set=20, min child samples=20 will be ignore
```

[illegible]

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13269
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur es: 111
```

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13263
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

43/147

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

53/147

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13211
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13224
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13872
[LightGBM] [Info] Number of data points in the train set: 6772, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Info] Start training from score -1.098317
```

[illegible]

72/147

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13211
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13224
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13269
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13263
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
```

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13248
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13269
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13952
[LightGBM] [Info] Number of data points in the train set: 6772, number of used features: 114
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Info] Start training from score -1.098317
[LightGBM] [Info] Start training from score -1.098760
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=20, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=20
[LightGBM] [Info] Total Bins 13248
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 111
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465

[illegible]

101/147

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13213
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 106
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
```

[illegible]

[illegible]

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13221
[LightGBM] [Info] Number of data points in the train set: 6773, number of used features: 105
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

113/147

[illegible]

[illegible]

[illegible]

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13224
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

file:///C:/Users/glawi/OneDrive - Aston University/Big Data/Individual.html

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13224
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Warning] Found whitespace in feature_names, replace with underlines
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignore
d. Current value: min_data_in_leaf=40
[LightGBM] [Info] Total Bins 13221
[LightGBM] [Info] Number of data points in the train set: 6773, number of used featur
es: 105
[LightGBM] [Info] Start training from score -1.098908
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Info] Start training from score -1.098465
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf
[LightGBM] [Warning] No further splits with positive gain, best gain: -inf

[illegible]

131/147

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[LightGBM] [Warning] min_data_in_leaf is set=40, min_child_samples=20 will be ignored. Current value: min_data_in_leaf=40
[LightGBM] [Warning] min data in leaf is set=20, min child samples=20 will be ignored
```

[illegible]

After doing grid Search we got best parameters so we use those best parameters to fit the model and predict on unseen data

142/147

```
# Fit the model to the SMOTE-augmented training data
lgb_classifier.fit(X_train_smote, y_train_smote)

# Predict on the imputed testing set
lgb_predictions = lgb_classifier.predict(X_test_imputed)

# Model evaluation on the imputed testing set
print(f"Accuracy: {accuracy_score(y_test, lgb_predictions):.2f}")
print(f"Precision (Weighted): {precision_score(y_test, lgb_predictions, average='weighted'):.2f}")
print(f"Recall (Weighted): {recall_score(y_test, lgb_predictions, average='weighted'):.2f}")
print(f"F1 Score (Weighted): {f1_score(y_test, lgb_predictions, average='weighted'):.2f}")

# Generate and display the confusion matrix
cm = confusion_matrix(y_test, lgb_predictions)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix - LightGBM')
plt.show()
```


[illegible]

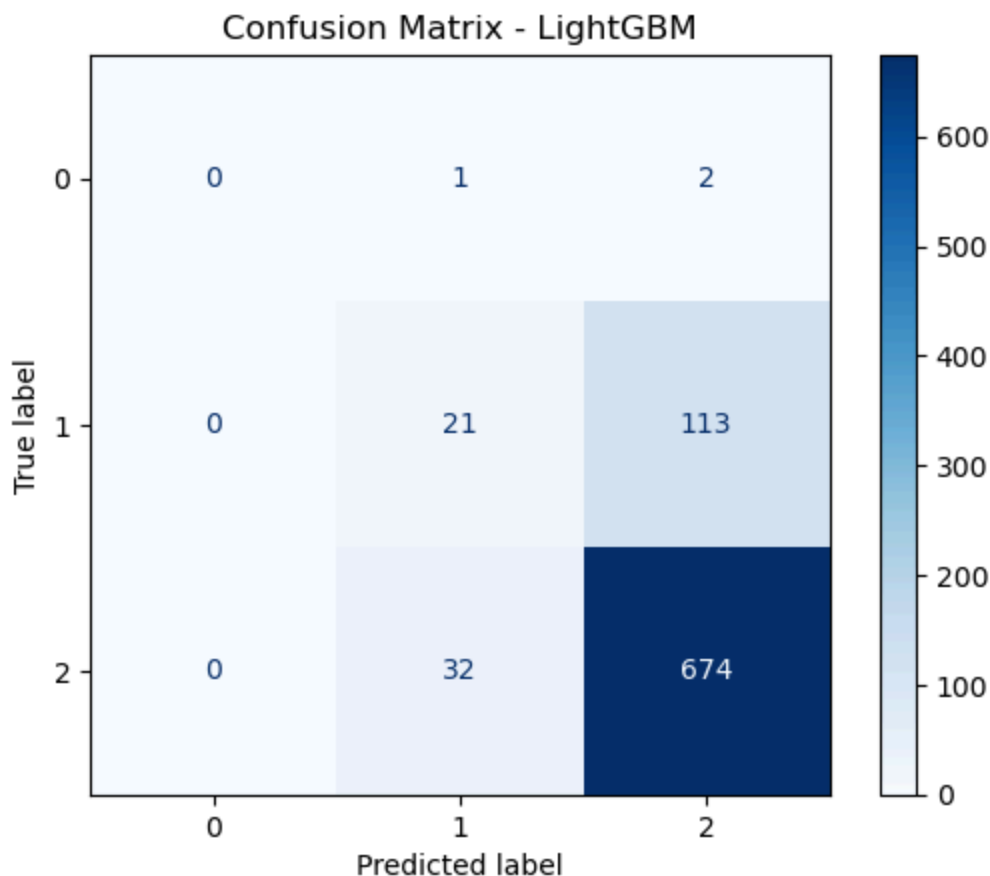
Accuracy: 0.82

```
Precision (Weighted): 0.78
```

Recall (Weighted): 0.82

F1 Score (Weighted): 0.79

```
c:\Users\glawi\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1509: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```



Conclusion

The comprehensive analysis involved preparing and engineering features from a dataset concerning road traffic accidents, followed by the application and evaluation of several machine learning models to predict accident severity. The models evaluated included Logistic Regression, Random Forest, Decision Tree, Gradient Boosting, XGBoost, LightGBM, and CatBoost, each chosen for its particular strengths in handling structured data, computational efficiency, and modeling accuracy.

Key Findings

Model Performance: The LightGBM model outperformed others in accuracy, precision, recall, and F1 score, making it the most suitable for predicting road traffic accident severity. It achieved an accuracy of 82%, a precision (weighted) of 77%, a recall (weighted) of 82%, and an F1 Score (weighted) of 79%.

Model Comparison: Ensemble methods generally showed better performance compared to simpler models like Logistic Regression and Decision Tree. However, the Decision Tree model provided valuable insights due to its interpretability.

Feature Importance: Temporal features (like day of the week and time), environmental conditions, geospatial data, and traffic conditions were identified as significant predictors of

accident severity.

Possible Future Improvements

1. Hyperparameter Tuning: For the LightGBM model, further tuning of hyperparameters through methods such as grid search, random search, and Bayesian optimization could enhance predictive performance.
2. Cross-Validation: Implementing cross-validation techniques would help ensure that the model generalizes well to unseen data and is robust across different data splits.
3. Advanced Models Exploration: While the LightGBM model performed best among the models evaluated, there's room for exploring other advanced machine learning or deep learning models that might offer improvements in predictive accuracy or interpretability.

Recommendation

The analysis recommends the LightGBM model for further development and deployment, given its high performance and operational efficiency. It suggests that this model offers a reliable foundation for a real-time prediction system to assist in road safety measures and decision-making.

Moving forward, the focus will be on refining this model, exploring advanced features, and enhancing its ability to predict with higher accuracy, especially for underrepresented classes in the dataset.