US ACCIDENTS

TECHNICAL SLIDES BY TEAM 17

PROJECT PIPELINE

1. DATA LOADING

- Imported 7.7M+ rows and 46 columns from Kaggle.
- Included raw, redundant, and missing-value features.

2. DATA PREPROCESSING

- Cleaned and transformed data to 6.1M rows and 41 columns.
- Removed low-quality columns, filtered outliers, binned continuous variables, and engineered new features.

PROJECT PIPELINE

3. EXPLORATORY DATA ANALYSIS (EDA)

- Analyzed 20+ features using PySpark.
- Visualized patterns with Matplotlib, Seaborn, Geopandas, and Plotly.

4. MODEL DEVELOPMENT

- Built classifiers: Naive Bayes (RDD) , Logistic Regression (PySpark), Random Forest (PySpark)
- Addressed class imbalance with under/oversampling.

PROJECT PIPELINE

5. MODEL EVALUATION & ITERATION

- Assessed model performance and refined approaches.
- Compared results to select the best solution.

6. FUTURE WORK

- Identified areas for further analysis and model improvement.

PREPROCESSING

1. HANDLING MISSING VALUES

- **Analyzed missing data** for each column.
- **Dropped columns** with >25% missing values:
  - End_Lat, End_Lng, Precipitation(in), Wind_Chill(F)

```
=== NULL Count and Percentage by Column ===
End_Lat: 3402762 (44.03%)
End_Lng: 3402762 (44.03%)
Precipitation(in): 2203586 (28.51%)
Wind_Chill(F): 1999019 (25.87%)
```

- **Dropped rows** with any remaining NULLs (as they were present in only a small number of columns).
- **Reason:** prevent bias from incomplete records.
- **Result:**
  - Reduced to 7,051,556 rows, 42 columns with no NULLs

PREPROCESSING

2. DROPPING IRRELEVANT COLUMNS

- **Removed columns** with little analytical value:
  - ID, Country, Source, Turning_Loop
- **Reason:** improve computational efficiency due to less number of columns
  - **ID:** Unique values, not informative for analysis.
  - **Country:** Only one value (US).
  - **Source:** Only three values (Source1, Source2, Source3).
  - **Turning_Loop:** Only one value (False).
- **Result:**
  - Dataset now contains only relevant columns.

Country
US

Source
Source3
Source2
Source1

Turning_Loop
false

PREPROCESSING

3. REMOVING REDUNDANT COLUMNS

- **Dropped redundant twilight columns:**
    - Civil_Twilight, Nautical_Twilight, Astronomical_Twilight (kept Sunrise_Sunset)
- **Checked for high correlation among similar features.**
- **Reason:** Prevents duplication in analysis.
    - Kept continuous variables as they were not highly correlated.
- **Result:**
    - Reduced to **35 columns.**

```
Correlation matrix for continuous variables:
Columns: ['Distance(mi)', 'Temperature(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)']
[[ 1.         -0.05598764  0.01140223 -0.09040334 -0.0395247   0.00874711]
 [-0.05598764  1.         -0.33053564  0.12005678  0.22400662  0.03458547]
 [ 0.01140223 -0.33053564  1.          0.10986754 -0.38682745 -0.17264725]
 [-0.09040334  0.12005678  0.10986754  1.          0.04117808 -0.0222837 ]
 [-0.0395247   0.22400662 -0.38682745  0.04117808  1.          0.01454486]
 [ 0.00874711  0.03458547 -0.17264725 -0.0222837   0.01454486  1.        ]]
```

PREPROCESSING

4. OUTLIER DETECTION & REMOVAL

- **Identified outliers using 2nd and 98th percentiles for continuous variables.**
- **Removed rows outside these ranges.**
- **Reason:** Removes extreme values that could distort analysis and model results.
  - Tried 1st/99th percentiles (data stayed at ~7 million rows).
  - Tried 2.5th/97.5th percentiles (data dropped to ~4 million rows).
  - Chose 2nd/98th percentiles as a balanced approach.
- **Result:**
  - 6,141,325 rows (improved data quality).

```
Distance(mi): 2nd percentile = 0.0, 98th percentile = 4.496
Temperature(F): 2nd percentile = 17.0, 98th percentile = 93.0
Humidity(%): 2nd percentile = 15.0, 98th percentile = 100.0
Pressure(in): 2nd percentile = 25.27, 98th percentile = 30.35
Visibility(mi): 2nd percentile = 1.0, 98th percentile = 10.0
Wind_Speed(mph): 2nd percentile = 0.0, 98th percentile = 20.0
```

PREPROCESSING

5. FEATURE ENGINEERING

- **Binned continuous variables** using equal-width binning.
- **Added discretized versions** of these 6 features to the dataset.
- **Grouped 'Weather_Condition'** from 78 detailed categories into 7 broader, more meaningful groups.
- **Added Urban/Rural classification** using external US Census data.
- **Reason:** Enables better data analysis insights by simplifying complex features.
- **Result:**
    - Dataset now includes binned versions of all main continuous variables, a simplified weather condition feature, and an Urban/Rural indicator.

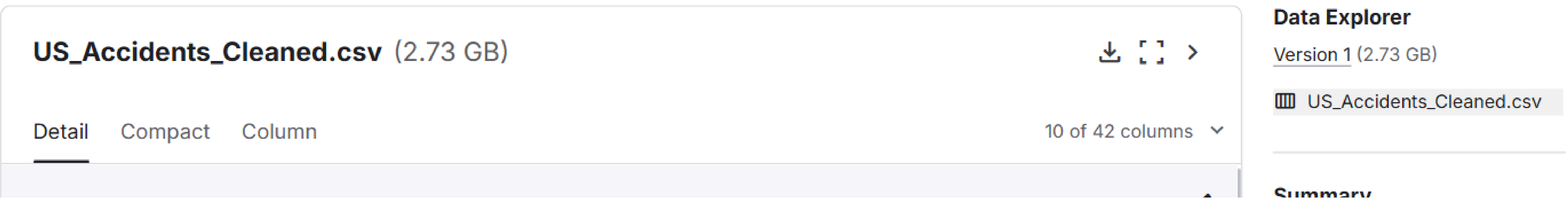
Wind_Speed(mph)	count
0.00-5.00	1672798
10.00-15.00	1229628
15.00-20.00	498514
5.00-10.00	2740385

Weather_Condition
Cloudy
Snowy
Thunderstorm or Hail
Clear
Rainy
Freezing Rain & Ice
Hazy & Dusty

PREPROCESSING

6. FINAL CLEANED DATASET

- **Before:** 7,728,394 rows, 46 columns (raw)
- **After:** 6,140,189 rows, 42 columns (cleaned, binned, grouped, enriched)
- **Exported** cleaned dataset for further analysis and modeling.
- **Reason:** Now much easier to start a project in a Kaggle notebook—just load the cleaned data and begin analysis.



EDA

1. FEATURE ENGINEERING WITH PYSPARK

- **Created new features:**
  - Accident duration (in minutes) using unix_timestamp on Start_Time and End_Time.
  - Time-of-day segments (Morning, Afternoon, Evening, Night) using the hour function.
  - Combined road features (e.g., FeatureCombo for presence of multiple road objects).
- **Aggregated data:**
  - Grouped by state, weather, time, road features, and urban/rural classification for multi-level insights.
- **Why:** Enabled multi-dimensional grouping and efficient feature creation on millions of records.

EDA

2. STATISTICAL ANALYSIS & PATTERN DISCOVERY WITH PYSPARK

- **Summary statistics:**
  - Calculated mean, median, and count for severity, duration, and accident frequency across states, weather, and road features, etc..
- **Correlation & interaction analysis:**
  - Used .corr() and groupby to explore relationships (e.g., severity vs. weather, road features, time).
  - Performed chi-square tests for categorical relationships (e.g., weather vs. urban/rural).
- **Pattern discovery:**
  - Identified trends such as higher severity at certain times, under specific weather, or with certain road features.
  - Used FP-Growth for frequent feature pattern mining.
- **Why:** Directly uncovered actionable patterns and relationships in the cleaned dataset.

EDA

3. VISUALIZATION WITH PYTHON LIBRARIES

- **Matplotlib & Seaborn:**
  - Grouped bar charts (e.g., severity by time of day), heatmaps (e.g., severity by weather and state), pairplots (e.g., duration vs. severity).
- **Plotly:**
  - Interactive 3D scatter plots (e.g., accident duration by severity, visibility, temperature, wind).
  - Stacked bar charts for accident frequency by weather and urban/rural.
- **GeoPandas:**
  - Plotted accident locations and clusters on US maps, visualized top accident-prone states.
- **Why:** These advanced visualizations are not possible in PySpark alone.

MODEL TRAINING

OBJECTIVE

Classify accidents into binary severity levels (Severity = 4 vs. Severity ≠ 4) using Logistic Regression, Random Forest, and Naive Bayes.

HYPERPARAMETER TUNING:

Logistic Regression: Regularization strength (regParam), elastic net mixing (elasticNetParam), max iterations (maxIter).

PREPROCESSING:

Dropped redundant columns:Distance(mi), Temperature(°C),Humidity(%),Pressure(in), Visibility(mi),Wind_Speed(mph),Severity, Weather_Timestamp,Description,End_Time, Distance(mi)_cont.

MAPREDUCE STEPS FOR NAIVE BAYES IMPLEMENTATION

- **Data Preparation:** Convert DataFrame to RDD as (label, features) tuples; split into train/test RDDs.
- **Map Step 1:** Map (label, 1) to compute class counts.
- **Reduce Step 1:** Reduce by key to get total counts per class and compute priors.
- **Map Step 2:** Map (label, (features, features^2, 1)) to compute sums for means and variances.
- **Reduce Step 2:** Reduce by key to aggregate sums, squares, and counts per class.
- **Final Step:** Calculate feature means and variances per class using aggregated sums.

HANDLING DATA IMBALANCE

IMBALANCE INSIGHT:

- Majority class (Severity ≠ 4): 5,955,314 records.
- Minority class (Severity = 4): 142,813 records.
- Imbalance ratio: 41:1

METHODS TESTED:

- More Feature Exploration
- Downsampling majority class with varying factors:
  - 10x more majority samples than minority.
  - 5x more majority samples than minority.
- Upsampling

FURTHER FEATURE EXPLORATION

EXPLORATION:

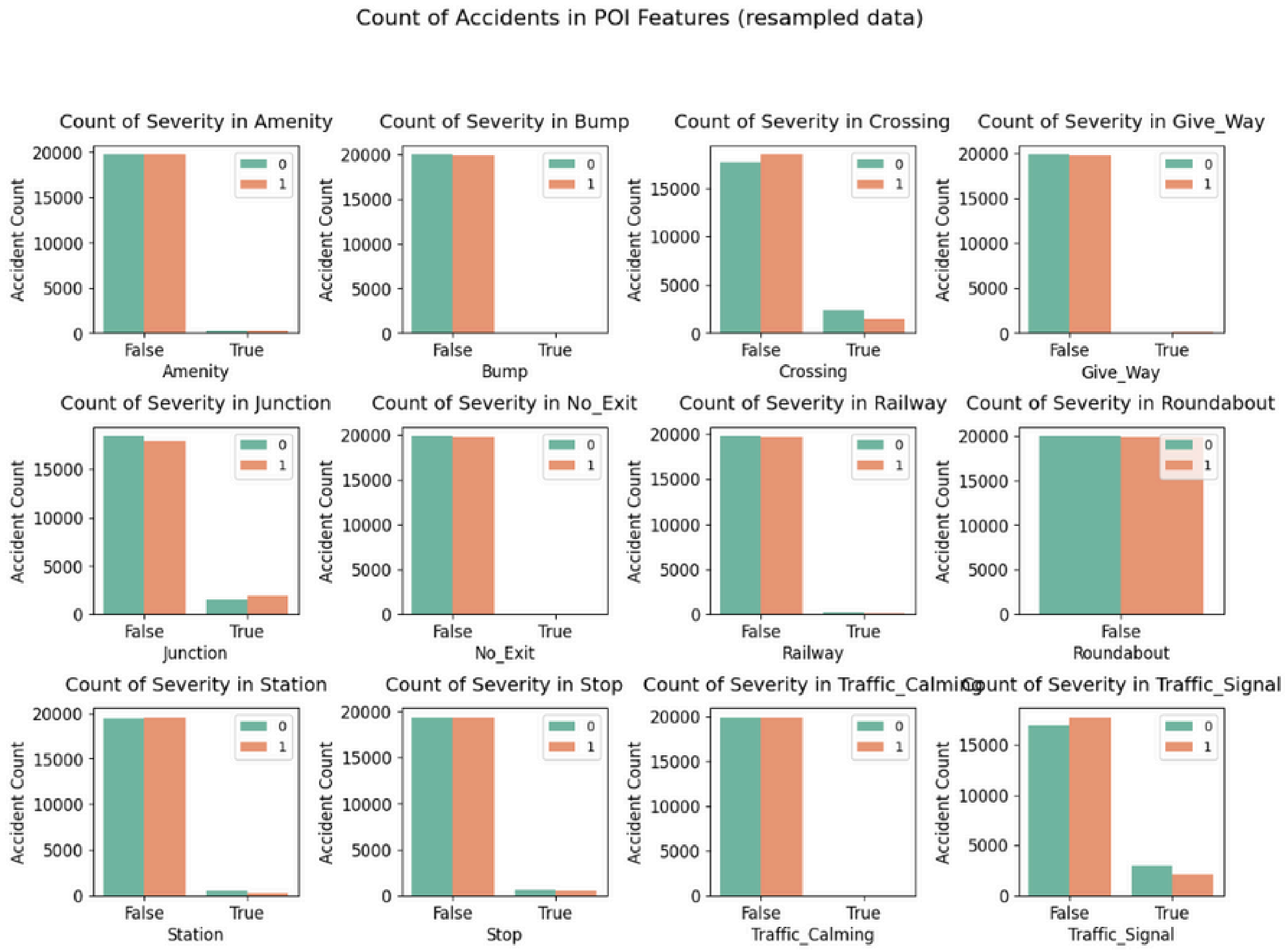
- Resampled 20,000 rows to analyze Points of Interest (POIs) and accident severity.

KEY OBSERVATIONS:

- Accidents near traffic signals and crossings: Less likely to be serious (drivers slow down).
- Accidents near junctions: More likely to be serious (speed as a critical factor).
- Some POI features (e.g., Bump, Give_Way) too unbalanced to contribute meaningfully.

ACTION

- Dropped uninformative features:
  - Bump, Give_Way, No_Exit, Roundabout, Traffic_Calming.

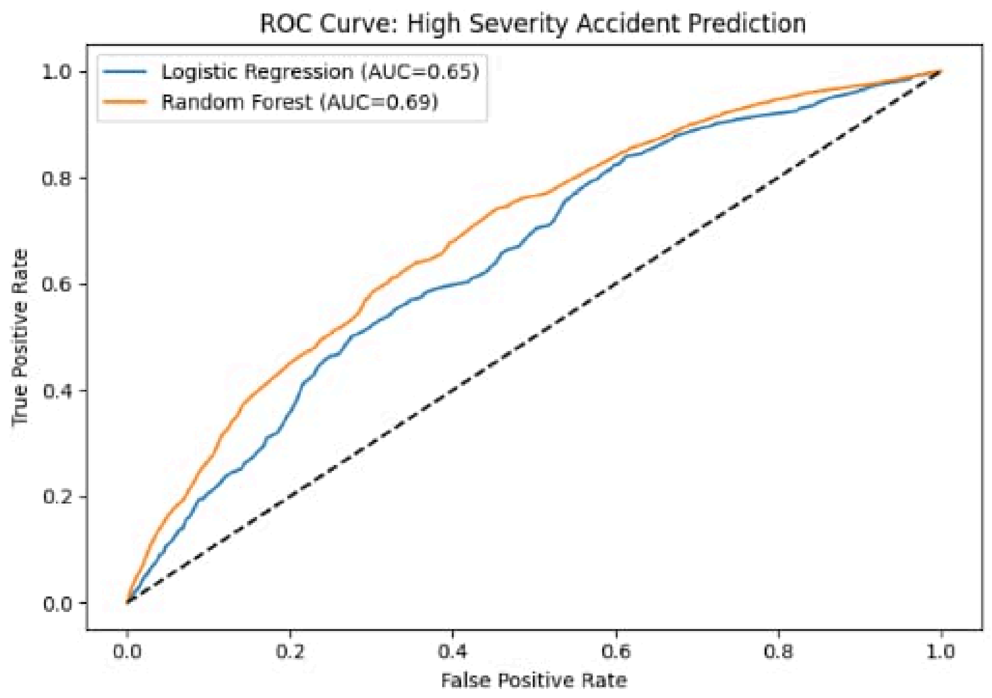


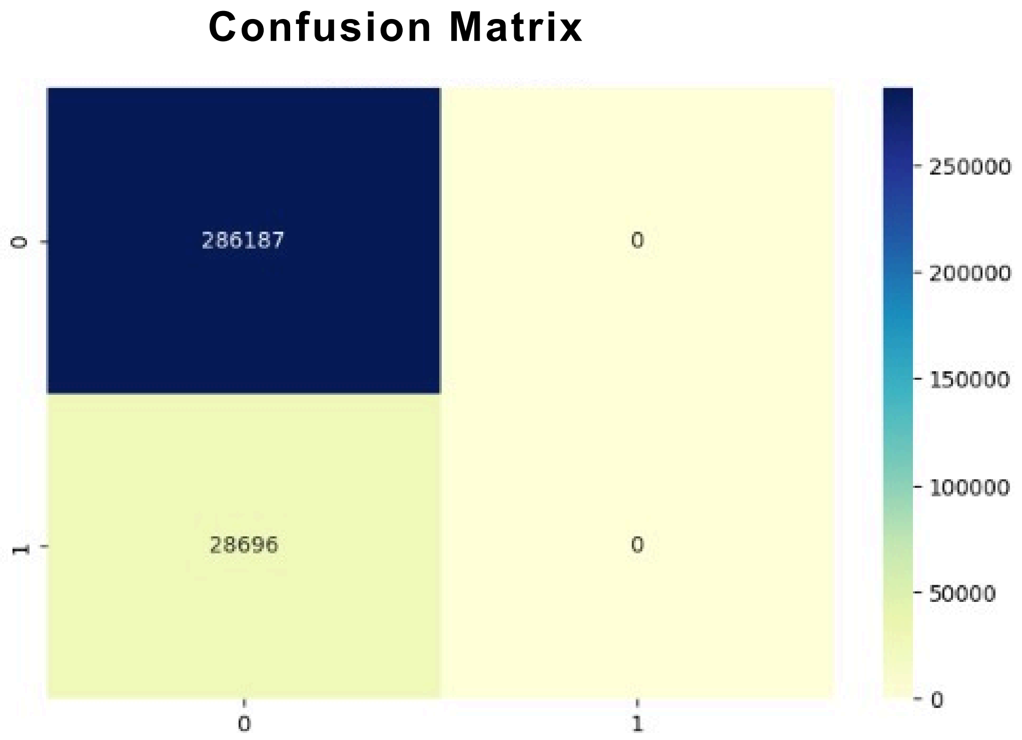
RESULTS OF DOWNSAMPLING

FINDINGS

- Higher sampling ratios (10x, 5x) improved overall accuracy.
- However, models overfit on the majority class (Severity ≠ 4).
- Reduced ratio to 3x to improve precision and recall for the minority class.

We tested this on Logistic Regression and Random Forest, and it yielded the same results, while we did not proceed with Naive Bayes.





	precision	recall	f1-score	support
0	0.909	1.000	0.952	286187
1	0.000	0.000	0.000	143117
accuracy			0.909	314883
macro avg	0.454	0.500	0.476	314883
weighted avg	0.826	0.909	0.909	314883

RESULTS OF DOWNSAMPLING

FINDINGS

- Higher sampling ratios (3x) make the overall accuracy worse.
- Improve precision and recall for the minority class.
- We used 18 features instead of 24 as the models couldn't fit these all data

Logistic Regression Metrics:

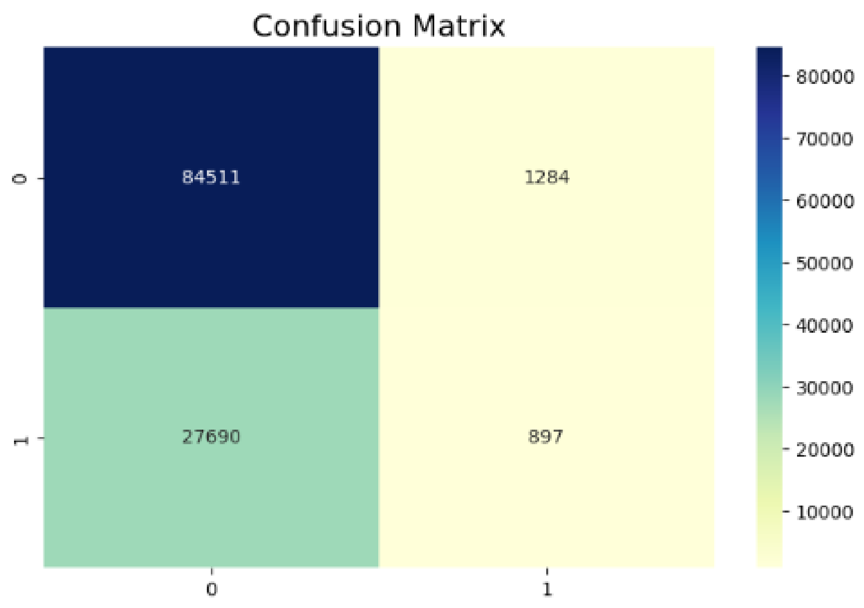
	precision	recall	f1-score	support
0	0.753	0.985	0.854	85795
1	0.411	0.031	0.058	28587
accuracy			0.747	114382
macro avg	0.582	0.508	0.456	114382
weighted avg	0.668	0.747	0.655	114382

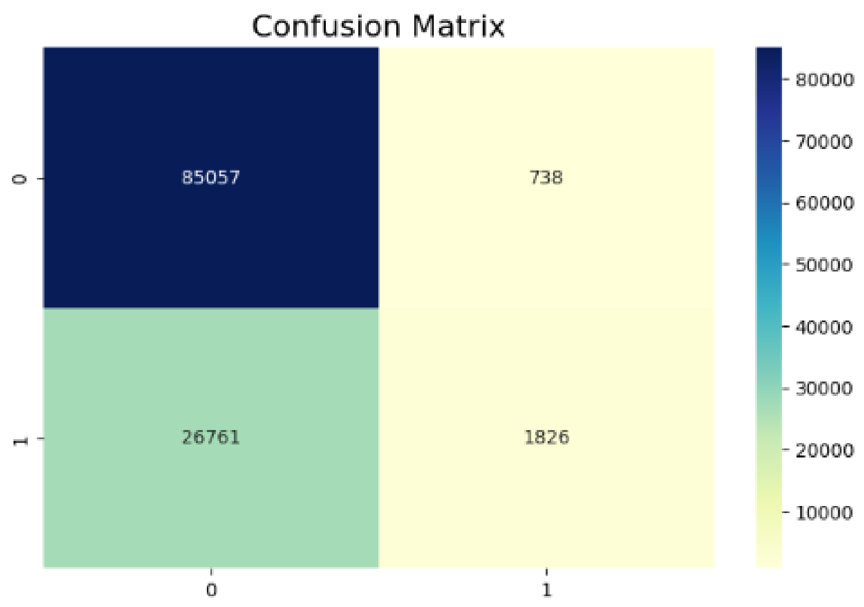
Random Forest Metrics:

	precision	recall	f1-score	support
0	0.761	0.991	0.861	85795
1	0.712	0.064	0.117	28587
accuracy			0.760	114382
macro avg	0.736	0.528	0.489	114382
weighted avg	0.749	0.760	0.675	114382

Train set balance

is_high_severity	count
0	345033
1	114897





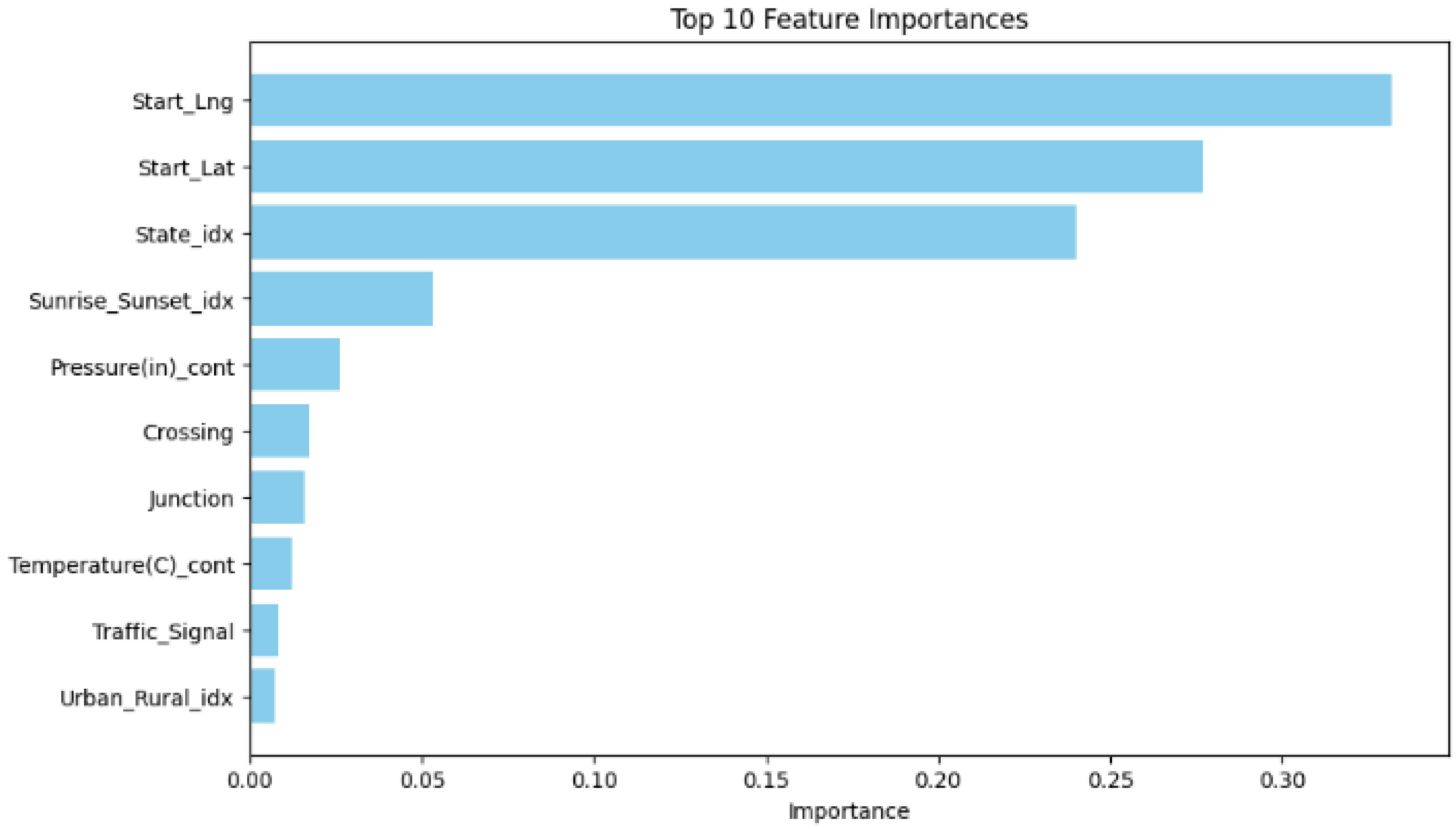
RESULTS OF DOWNSAMPLING

FEATURE IMPORTANCE FOR RANDOM FOREST MODEL

- sampling ratios (3x)
- The feature importance plot indicates that high-resolution spatio-temporal accident patterns are the most predictive features for severity, followed by pressure, population, and road type as other key factors.

Train set balance

is_high_severity	count
0	345033
1	114897

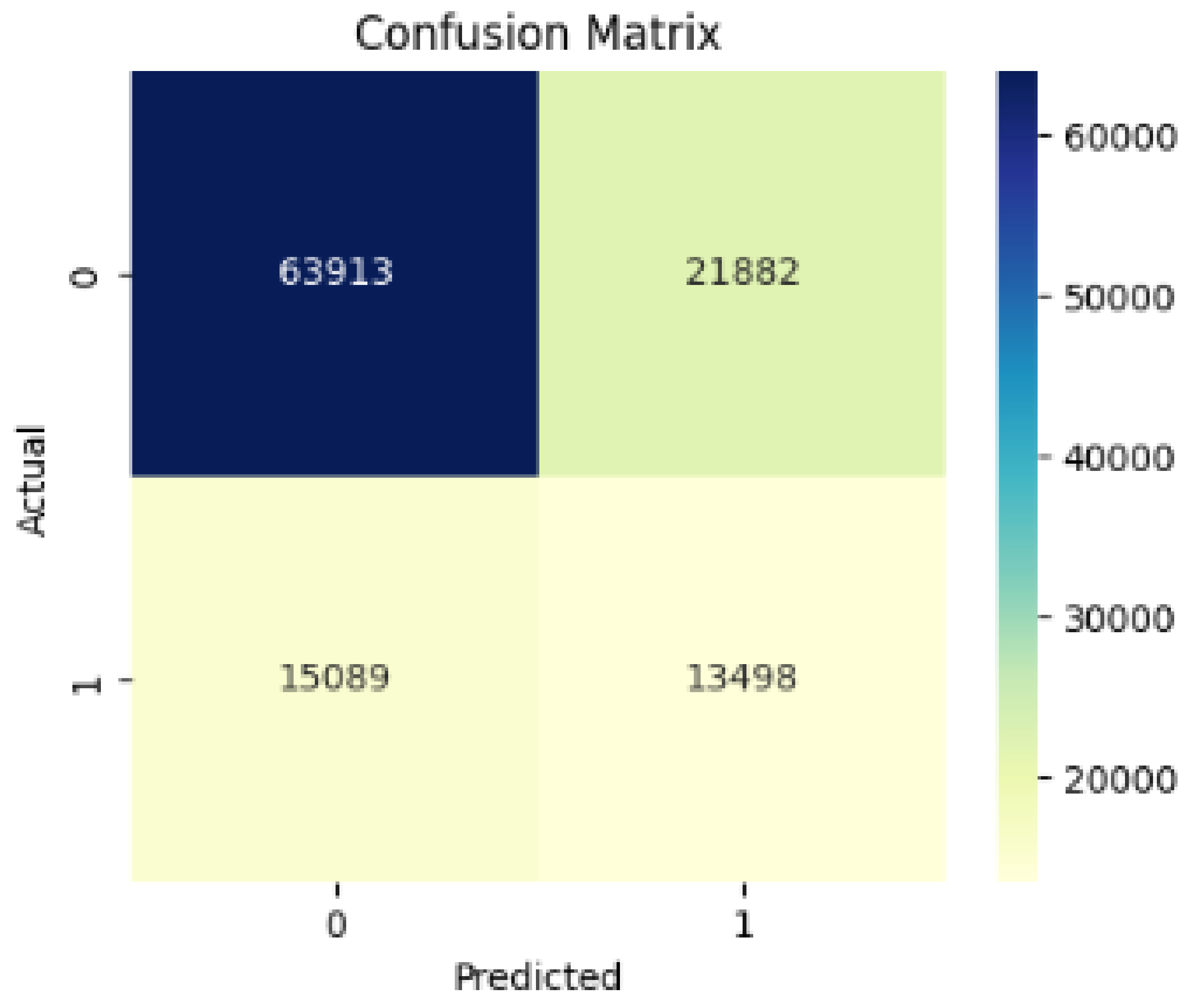


RESULTS OF DOWNSAMPLING

FINDINGS

Gaussian Naive Bias(Map-Reduce)

- sampling ratios (3x)
- Results:
  - Precision: 0.3815
  - Recall: 0.4722
  - F1 Score: 0.4220
  - Accuracy: 0.6768



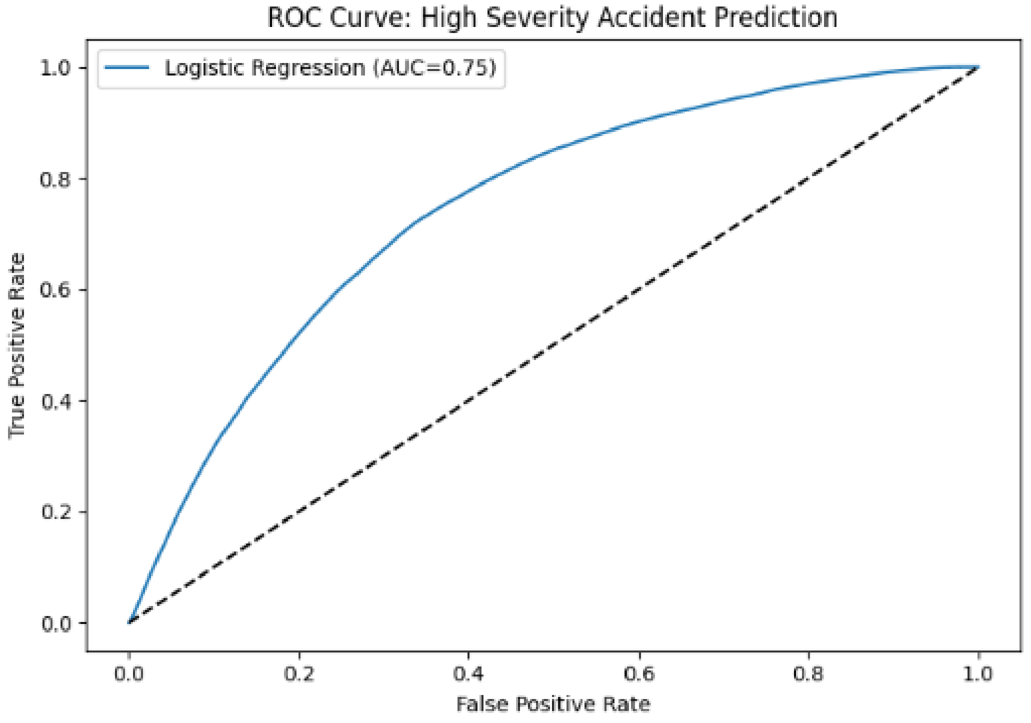
RESULTS OF UPSAMPLING

UPSAMPLING APPROACH:

- Applied 1:1 ratio for majority and minority classes.
- Generated around 700,000 records per class using 24 features.
- Added small noise to numerical features for realistic synthetic minority data.
- We used logistic regression only as random forest couldn't fit all these data.

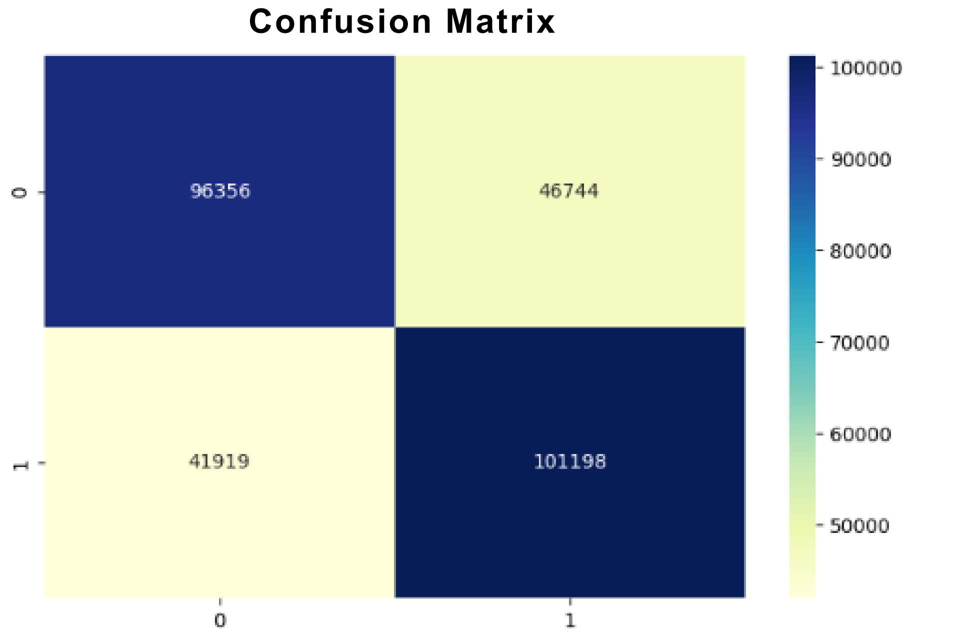
Train set balance

is_high_severity	count
0	574639
1	574179



logistic regression

	precision	recall	f1-score	support
0	0.697	0.673	0.685	143100
1	0.684	0.707	0.695	143117
accuracy			0.690	286217
macro avg	0.690	0.690	0.690	286217
weighted avg	0.690	0.690	0.690	286217



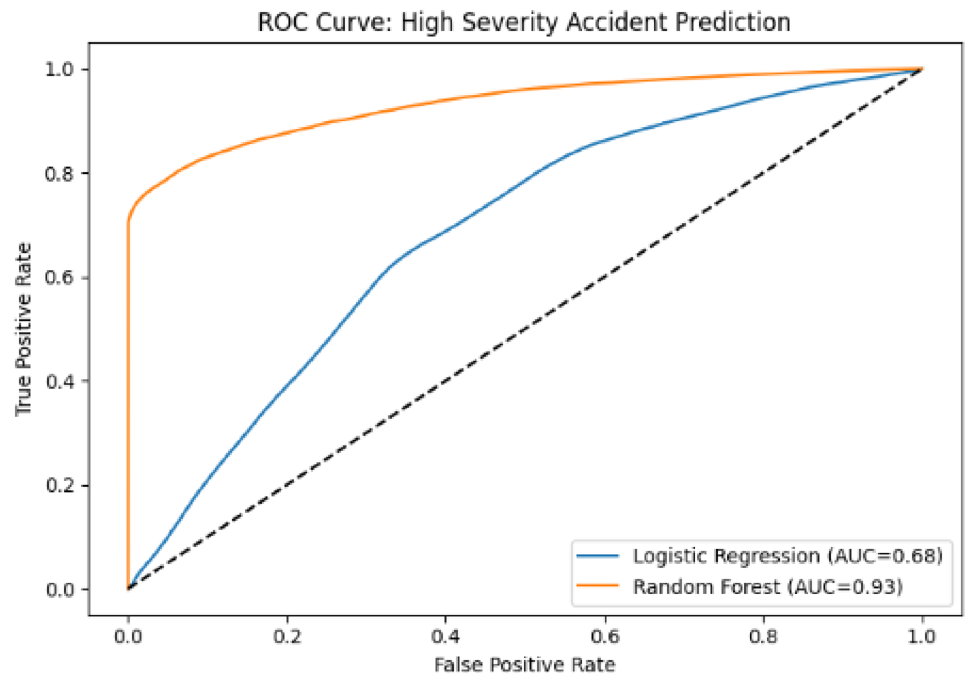
RESULTS OF UPSAMPLING

UPSAMPLING APPROACH:

- Applied 1:1 ratio for majority and minority classes.
- Generated around 700,000 records per class using 18 features instead of 24.

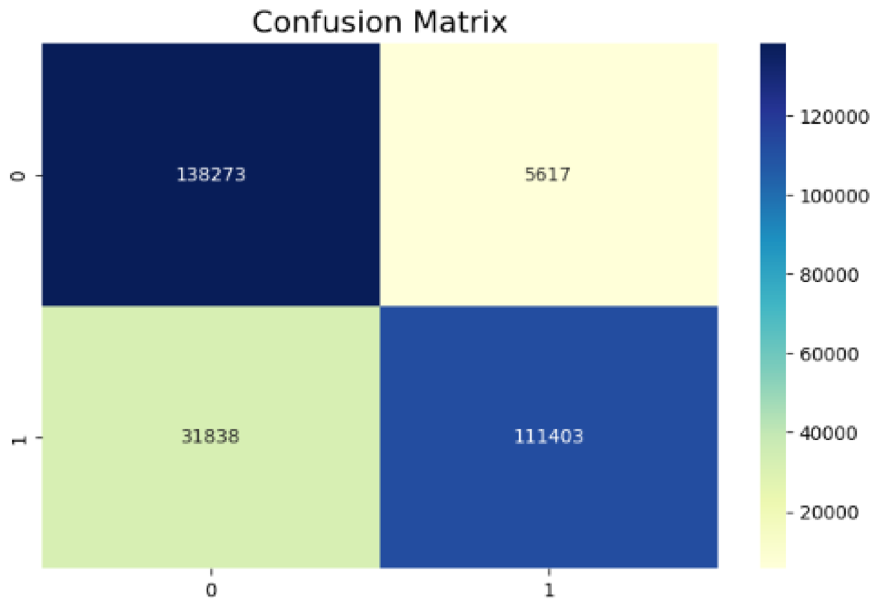
Logistic Regression AUC: 0.684

Random Forest AUC: 0.933



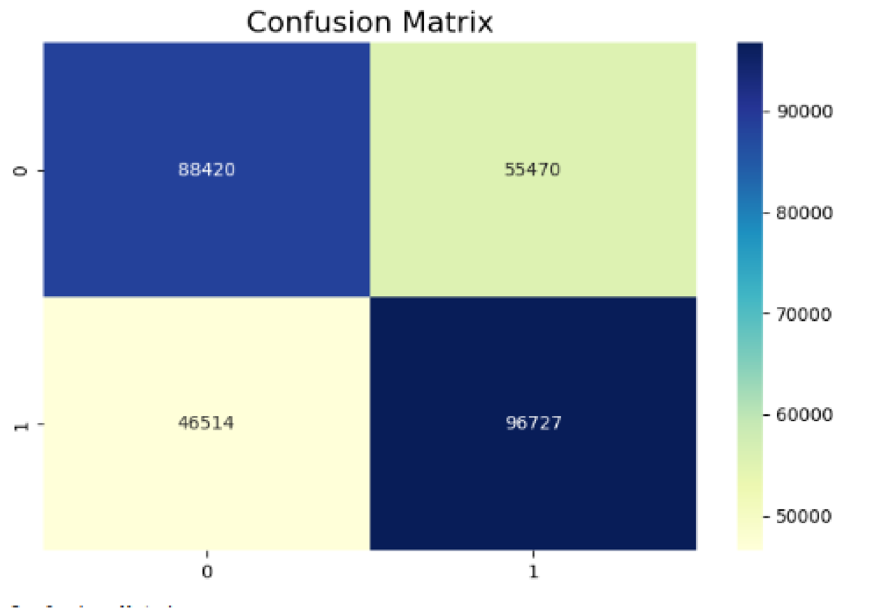
RANDOM FOREST METRICS:

	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.813	0.961	0.881	143890
1	0.952	0.778	0.856	143241
ACCURACY			0.870	287131
MACRO AVG	0.882	0.869	0.868	287131
WEIGHTED AVG	0.882	0.870	0.868	287131



LOGISTIC REGRESSION METRICS:

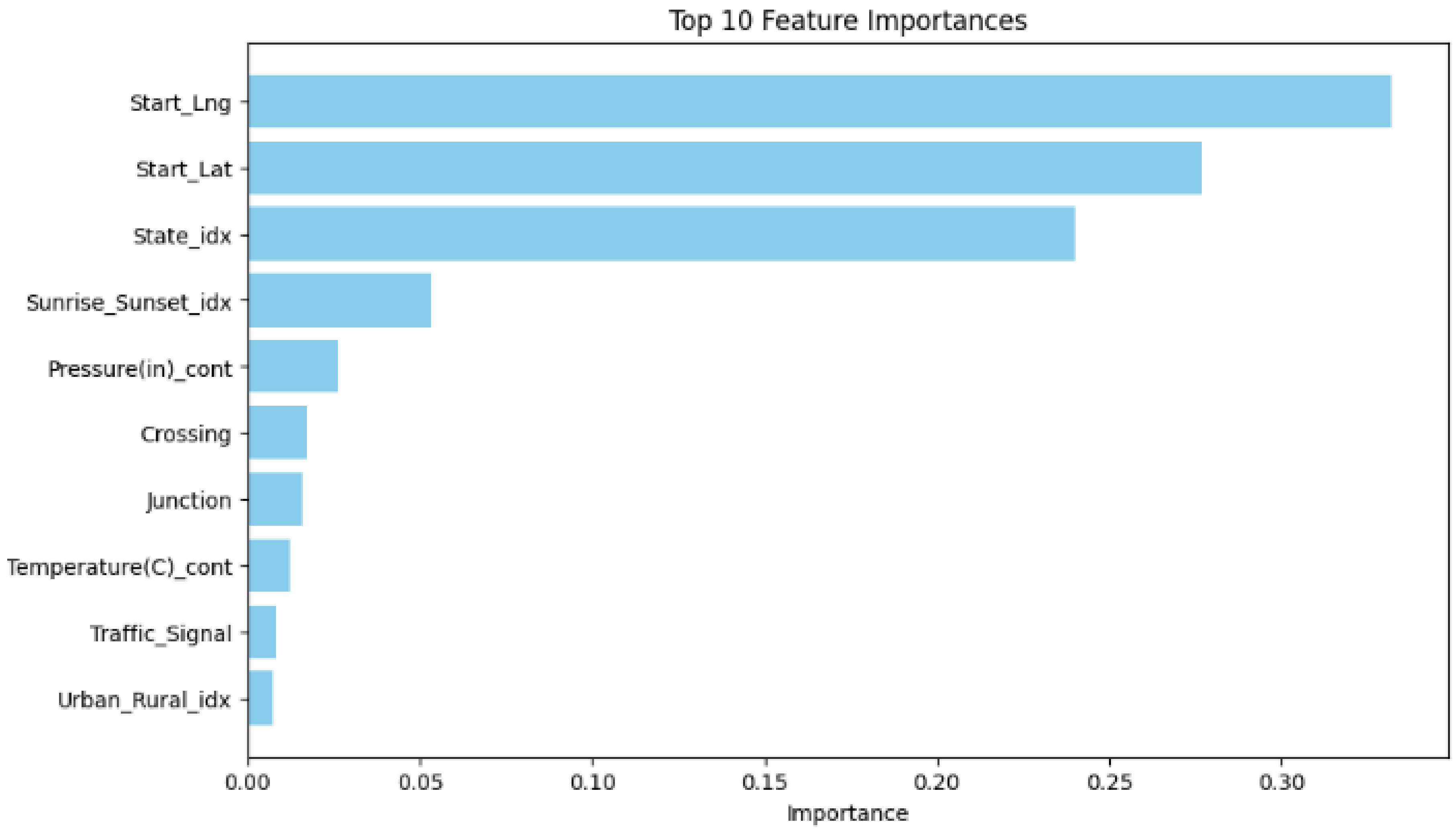
	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.655	0.614	0.634	143890
1	0.636	0.675	0.655	143241
ACCURACY			0.645	287131
MACRO AVG	0.645	0.645	0.645	287131
WEIGHTED AVG	0.645	0.645	0.644	287131



RESULTS OF UPSAMPLING

FEATURE IMPORTANCE FOR RANDOM FOREST MODEL

- Applied 1:1 ratio for majority and minority classes.
- Generated around 700,000 records per class using 18 features instead of 24.
- The feature importance plot indicates that high-resolution spatio-temporal accident patterns are the most predictive features for severity, followed by pressure, population, and road type as other key factors.



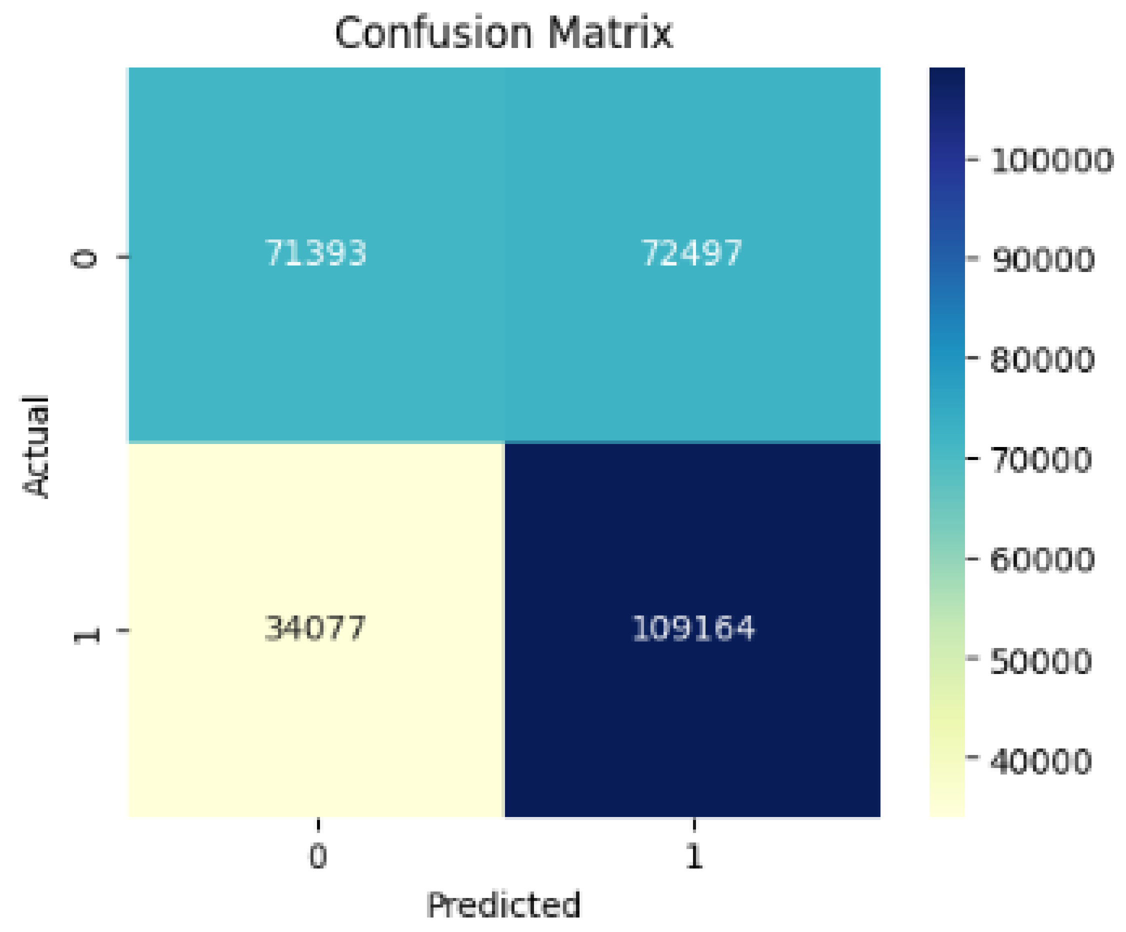
RESULTS OF UPSAMPLING

UPSAMPLING APPROACH:

Gaussian Naive Bias(Map-Reduce)

- Applied 1:1 ratio for majority and minority classes.
- Generated around 700,000 records per class using 18 features instead of 24.
- Results:
  - Precision: 0.6009
  - Recall: 0.7621
  - F1 Score: 0.6720
  - Accuracy: 0.6288

Class priors: {0: 0.5002002057767201, 1: 0.4997997942232799}



ENHANCEMENTS & FUTURE WORK

1. ENHANCED HANDLING OF CLASS IMBALANCE

- Explore advanced resampling techniques and ensemble methods.
- Test additional classifiers such as Gradient Boosting (XGBoost, LightGBM) and deep learning models.

2. EXPANDING DATA SOURCES

- Integrate external datasets (e.g., US Census demographic data) to enrich accident records.
- Link variables like population, commuting patterns, and median income at the county level for deeper analysis.

ENHANCEMENTS & FUTURE WORK

3. REAL-TIME PREDICTION SYSTEM

- Develop a real-time accident risk prediction system.
- Incorporate live traffic, weather, and location data to dynamically assess and visualize accident risk.

4. DATASET LIMITATIONS & QUALITY ISSUES

- Over 1,200 cities reported only one accident, indicating possible underreporting.
- Major cities like New York are missing, despite their high population.
- Data collection stops at March 2023, resulting in incomplete data for the final year.

THANK YOU