FusionSite

Driver Safety Model

Trang Hoang and Isabel Arvelo
04-14-25

Agenda

1 Problem Description
2 Data Description
3 Model Development
4 Results
5 Highlights
6 Challenges

What is FusionSite?

FusionSite is a high-growth premier waste management provider with services across 15 states.

Problem Description

PROBLEM

Driver Safety

FusionSite needs proactive driver safety protocols across its fleet operations

Current approach is reactive (waiting for incidents)

Industry leader in portable sanitation, waste management & luxury trailer rentals
Reputation for reliability at stake

SOLUTION

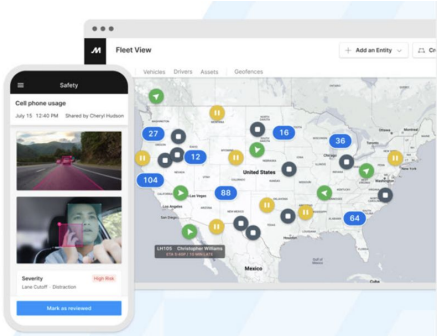
Risk Prediction

Developing a Daily Driver Risk Model Leveraging GoMotive Fleet Management & Driver Safety Platform to analyze driving events data

Integrating driver historical driving patterns and pre-trip inspections

Using location-based data to incorporate state crash data and precipitation data in the radius of each site

Data Sources



Motive API

FusionSite

Internal Company Data

Climate Data

Geospatial Data

FMCSA
Federal Motor Carrier Safety Administration

Transportation Data

Insurance Claims Data

Motive Data Preprocessing

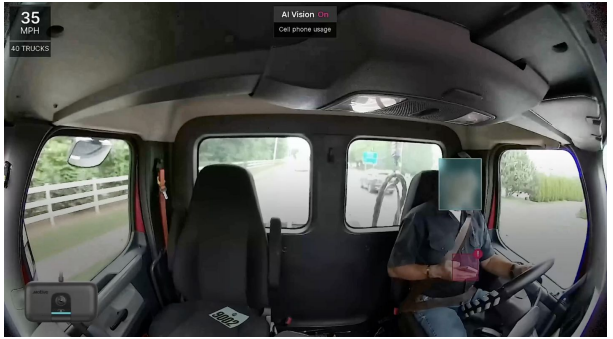
Data about driver and vehicle activity from January 2023 to February 2025:

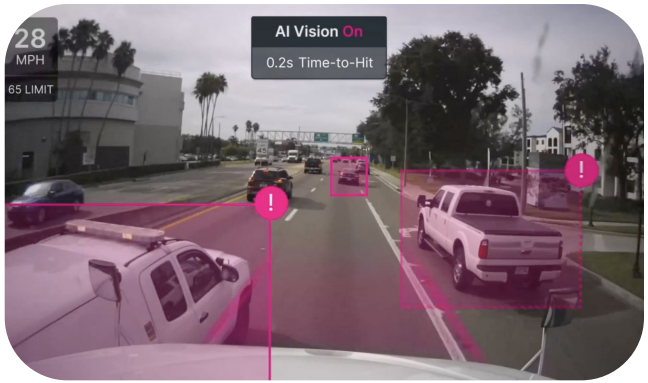
1. **Aggregated Driving Events**
   Captures each driver's behavior, history, and context on a per-trip basis, forming a comprehensive feature set.

2. **Aggregated Combined Events**
   Tracks the volume and types of safety-related events (e.g., seatbelt violations, distractions, speeding), including severity levels of speeding (low/mid/high). It helps identify risky behaviors and monitor unresolved coaching issues.

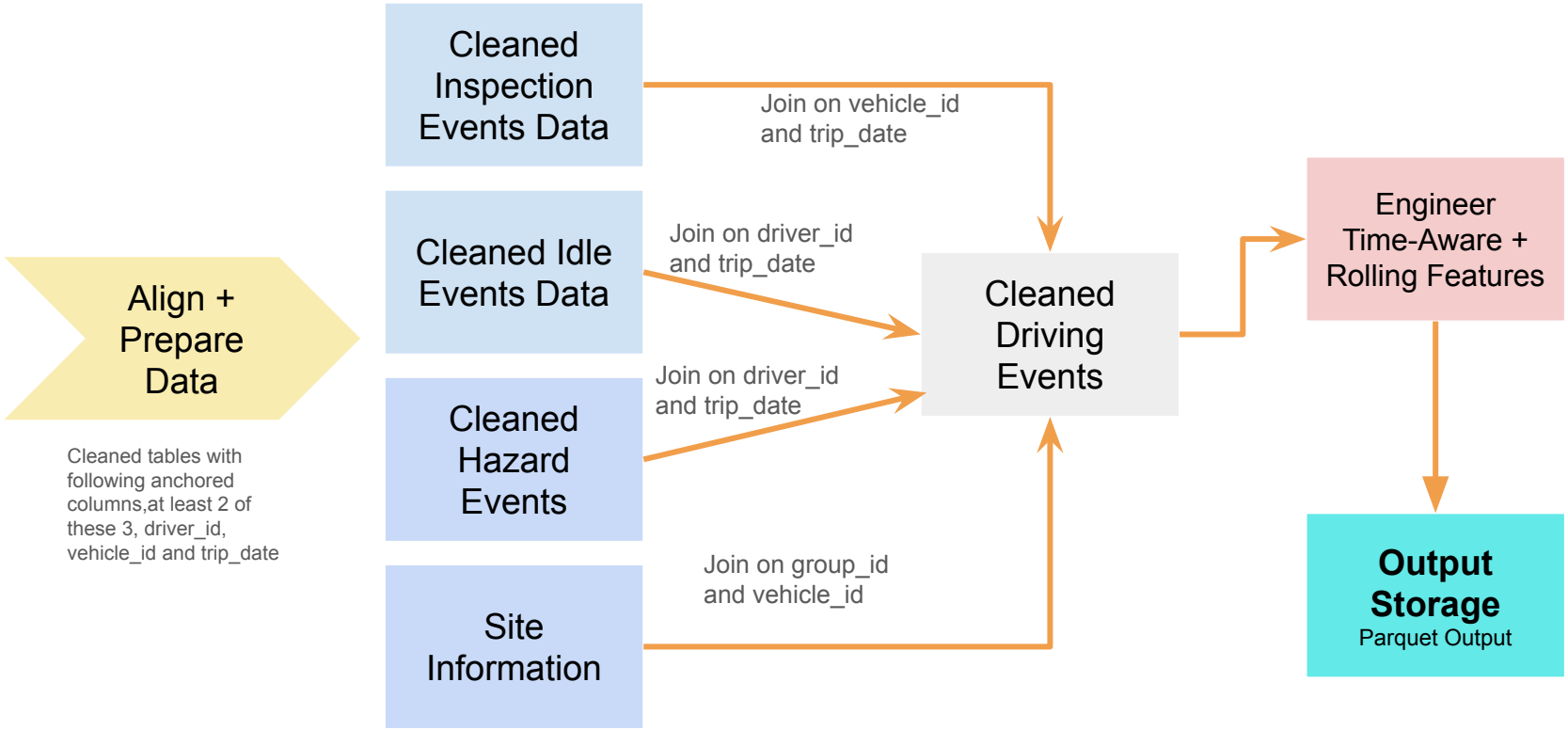
3. **Aggregated Inspection Events**
   Provides a trip-level summary of inspection activities per vehicle by counting total inspections and flagging open or resolved issues.

4. **Aggregated Idle Events**
   Summarizes idle behavior per driver per day to gauge frequency and duration of truck idling, useful for assessing driver efficiency, fuel consumption, and potential safety concerns.

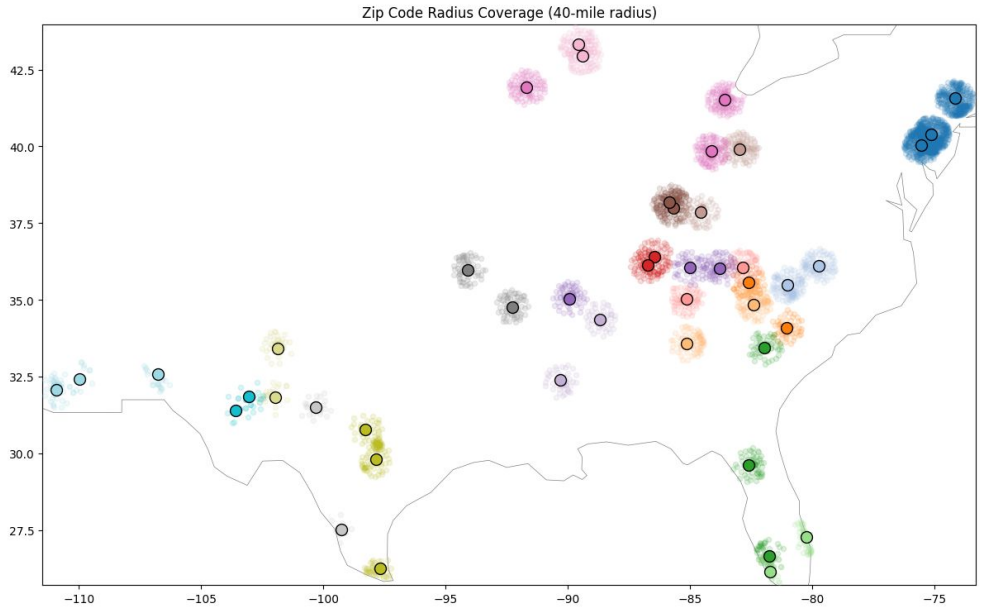




Joining FusionSite trips data

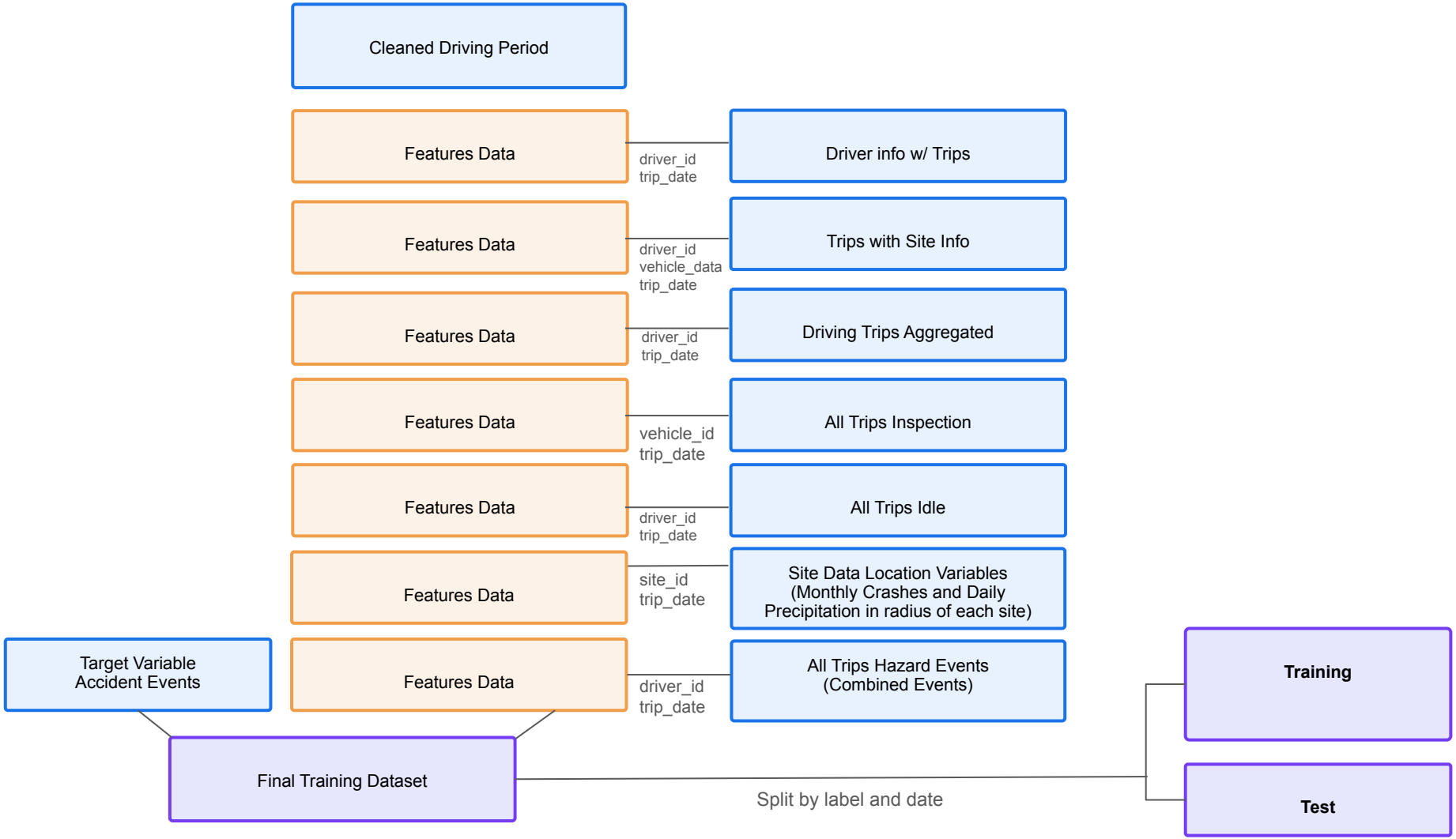


Joining External Data to Site Radius

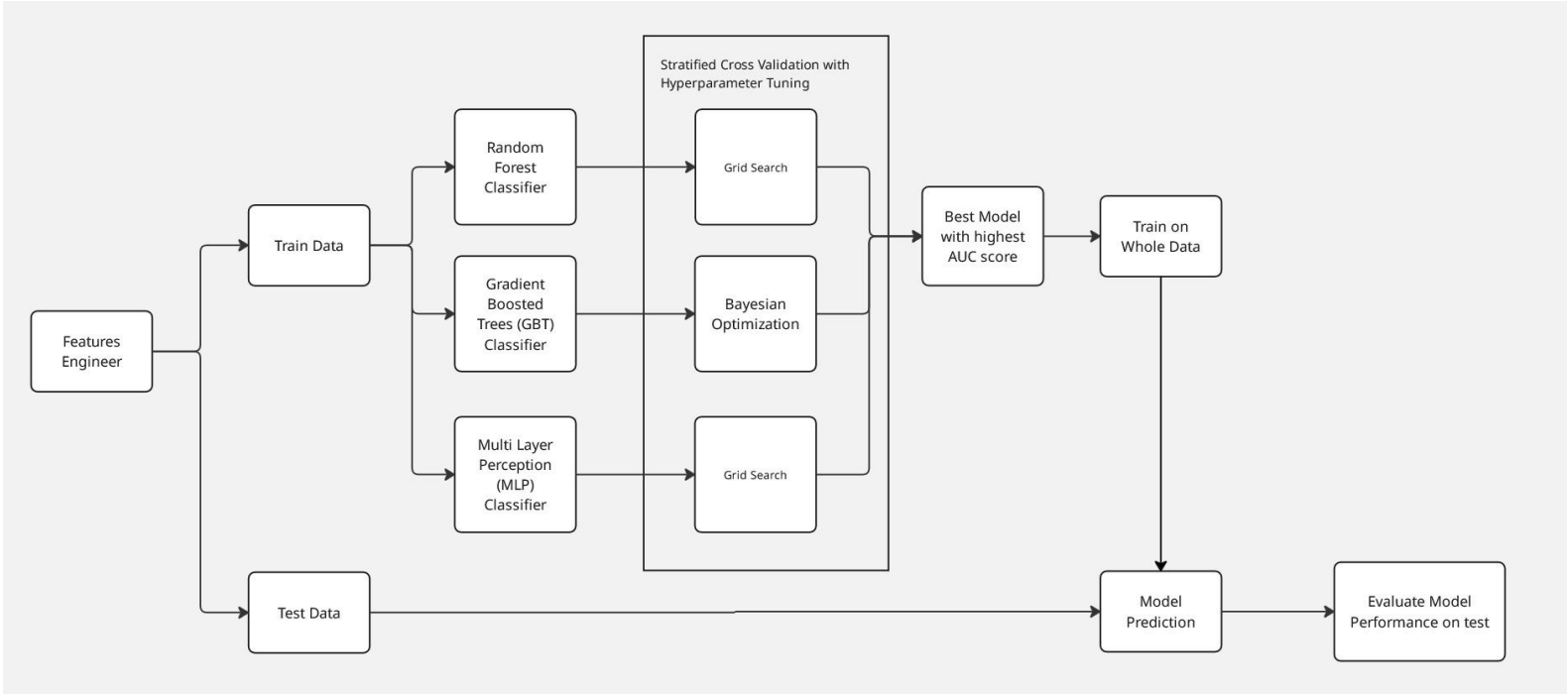


For each site radius we:

- Aggregated details on individual crashes to monthly crash statistics across all counties that intersect with the site's service area zip codes
- Aggregated spatial climate data to summary statistics for daily precipitation in the service area
- Created lookback windows for moving averages of each variable


Cleaned Driving Period
Features Data
driver_id
trip_date
Driver info w/ Trips
Features Data
driver_id
vehicle_data
trip_date
Trips with Site Info
Features Data
driver_id
trip_date
Driving Trips Aggregated
Features Data
vehicle_id
trip_date
All Trips Inspection
Features Data
driver_id
trip_date
All Trips Idle
Features Data
site_id
trip_date
Site Data Location Variables
(Monthly Crashes and Daily
Precipitation in radius of each site)
Features Data
driver_id
trip_date
All Trips Hazard Events
(Combined Events)
Target Variable
Accident Events
Final Training Dataset
Split by label and date
Training
Test

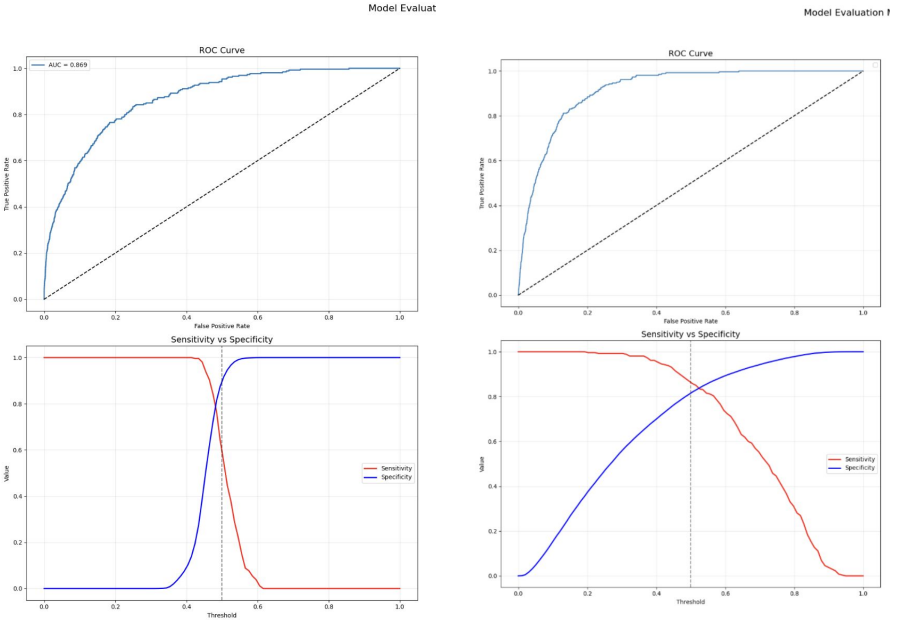
Hyperparameter Tuning



Results

Training Data

- Our models learned something with AUC values > 0.90 after training, but did they learn the relationship between the features and driver accidents?



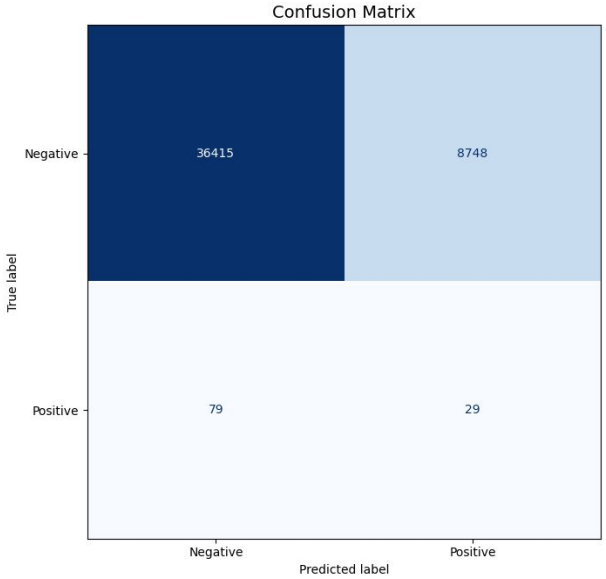
Random Forest

Gradient Boosted Tree

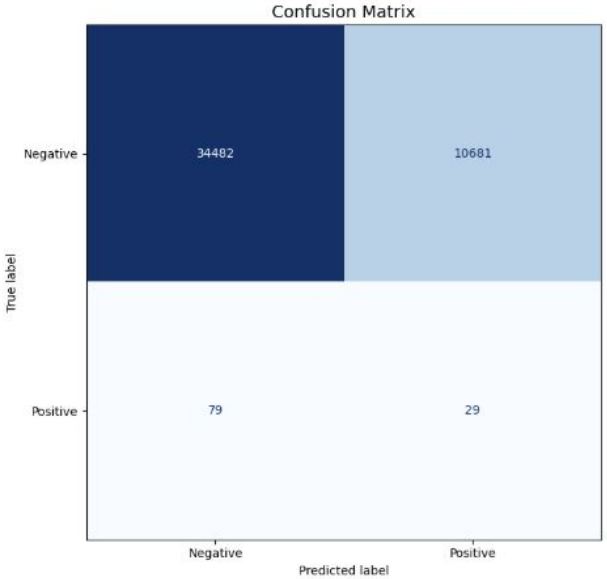
Results

Test Data

- Both models performed just marginally better than random with AUCs of 0.57
- The positive predictive value was extremely low at just 0.32% for RF and 0.30% for GBT
- Less than half a percent of positive predictions were actual accidents.
- Only about 18-28% of true accidents while generating thousands of false positives.



Random Forest



Gradient Boosted Tree

Success

- Served as an informal data audit, uncovering key data inconsistencies
- Provided actionable insights into current data quality and structure
- Helped lay the foundation for future in-house predictive modeling efforts
- Acquired extensive knowledge of the Spark package, recognizing some Spark model limitations like non-stratified cross-validation.
- Fostered smooth collaboration among team members throughout the process.

Challenges

- Date Imputation
- No guarantee drivers will stay within 40 miles of site for any given trip
- Data not missing completely at random (MCAR)
- Data Completeness in insurance claims
- Extreme Class Imbalance.
- Scalability

Where to go from here

Data

A predictive model can only ever be as good as the data that goes into it.

COLLECTION

Focus on systematic data collection protocols.

START SIMPLE

Start with simple model based on known safety hazards.

ITERATE

Analyze patterns, profiles and behaviors over time.

Appendix

Motive Data Preprocessing - Data Driving Events

Description: The final table generate a comprehensive feature set that captures each driver's behavior, history, and context on a per-trip basis, The date range of this data is from Jan 2023 to Feb 2025.

Raw Features

- Id
- Event_id
- driver_id
- driver_first_name
- driver_last_name
- vehicle_id
- start_date
- end_date
- driving_distance
- Driving_period_type
- driver_company_id
- minutes_driving
- created_at/updated_at
- number
- status-2
- make
- model

Aggregate by driver and trip date

Derived Features

- driver_id
- Trip Date
- driver_total_trip_count
- first_driving_date
- driver_log_trip_count
- rolling_xday_trip_count
- previous_trip_date
- days_since_last_trip
- driving_year_since_first_trip

Motive Data Preprocessing - Data Combined Events

Description: This aggregation tables table captures the volume and types of safety-related events (e.g., crashes, distractions, speeding), their severity in term of speeding (low/mid/high). It helps assess driving behavior, identify risky patterns, and monitor unresolved coaching interventions

Raw Features

- Id
- Event_id
- driver_id
- driver_first_name
- driver_last_name
- vehicle_id
- coaching_status
- start_date
- severity
- group_id
- group_name
- month
- created_at
- updated_at
- max_over_speed_in_kph
- max_over_speed_in_mph

Aggregate by driver and start/trip date

Derived Features

- driver_id
- Trip Date
- Total_events_per_trip
- total_<event_type>_count
- speeding_bin
- cnt_max_over_speed_in_mph_per_speed_bin
- pending_review_count

Motive Data Preprocessing - Data Vehicle Inspection Events

Description: This aggregation creates a trip-level summary of inspection activity per vehicle by counting total inspections and identifying trips with open or resolved issues.

Raw Features

- id
- event_id
- driver_id
- driver_first_name
- driver_last_name
- vehicle_id
- start_date
- end_date
- driving_distance
- driving_period_type
- driver_company_id
- minutes_driving

Aggregate by vehicle and trip date

Derived Features

- vehicle_id
- trip Date
- num_inspections_per_trip
- num_issues_per_trip
- inspection_date
- last_inspection_date
- last_inspection_status

Motive Data Preprocessing - Data Idle Events

Description: This aggregation summarizes idle behavior per driver per day. It helps quantify how often and how long a truck is idling — key indicators of driver efficiency, fuel usage, and potential safety concerns.

Raw Features

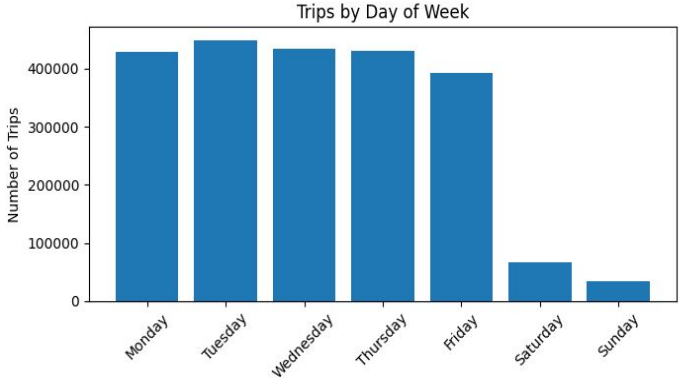
- Id
- Event_id
- driver_id
- vehicle_id
- start_time
- end_time
- driver_company_id
- minutes_idling

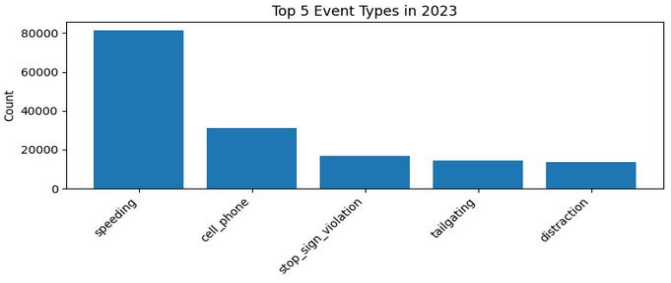
Aggregate by driver and trip date

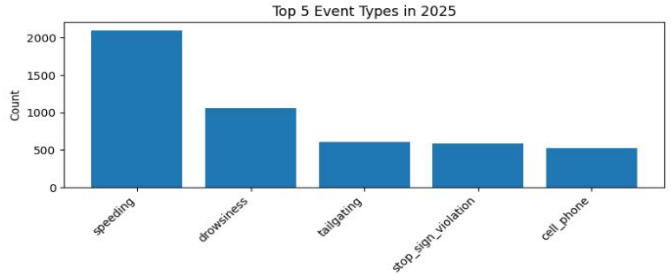
Derived Features

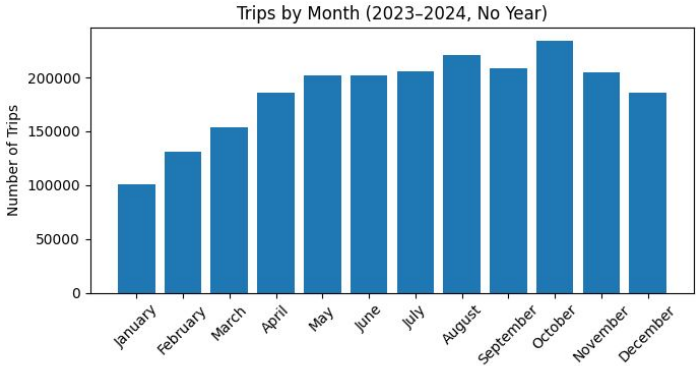
- driver_id
- idle_date
- idle_event_count_per_trip
- avg_idle_duration_per_trip
- total_idle_minutes_per_trip

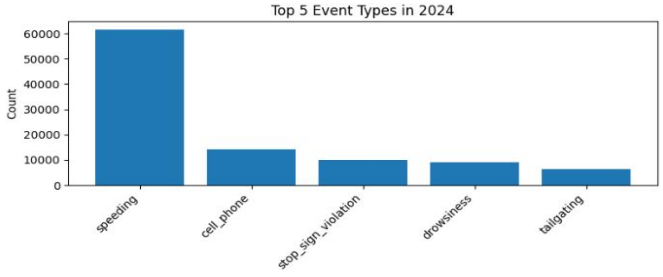
Motive EDA


Trips by Day of Week
Number of Trips
0
100000
200000
300000
400000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

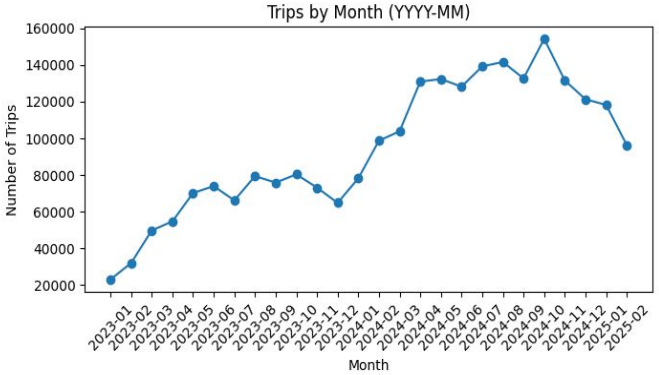

Top 5 Event Types in 2023
Count
0
20000
40000
60000
80000
speeding
cell_phone
stop_sign_violation
tailgating
distraction

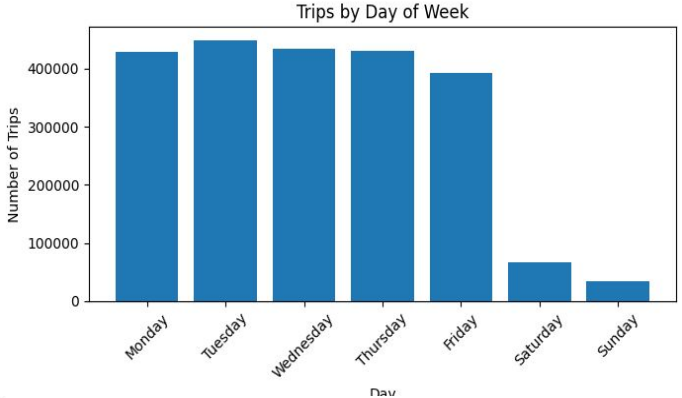

Top 5 Event Types in 2025
Count
0
500
1000
1500
2000
speeding
drowsiness
tailgating
stop_sign_violation
cell_phone

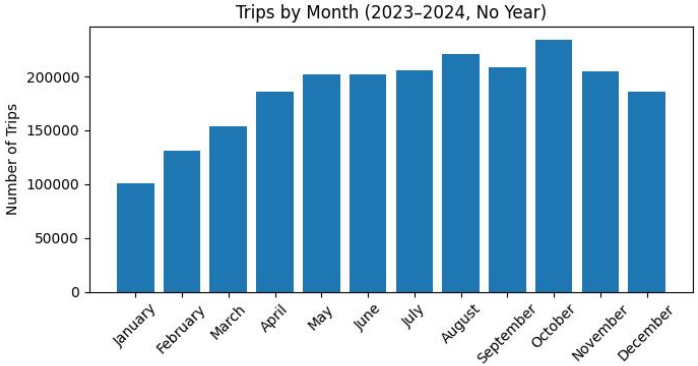

Trips by Month (2023–2024, No Year)
Number of Trips
0
50000
100000
150000
200000
January
February
March
April
May
June
July
August
September
October
November
December


Top 5 Event Types in 2024
Count
0
10000
20000
30000
40000
50000
60000
speeding
cell_phone
stop_sign_violation
drowsiness
tailgating

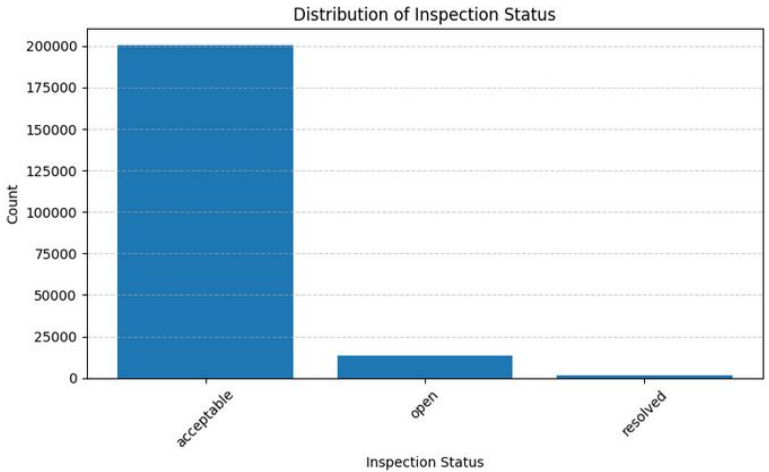
Motive Data Preprocessing - Data Driving Events

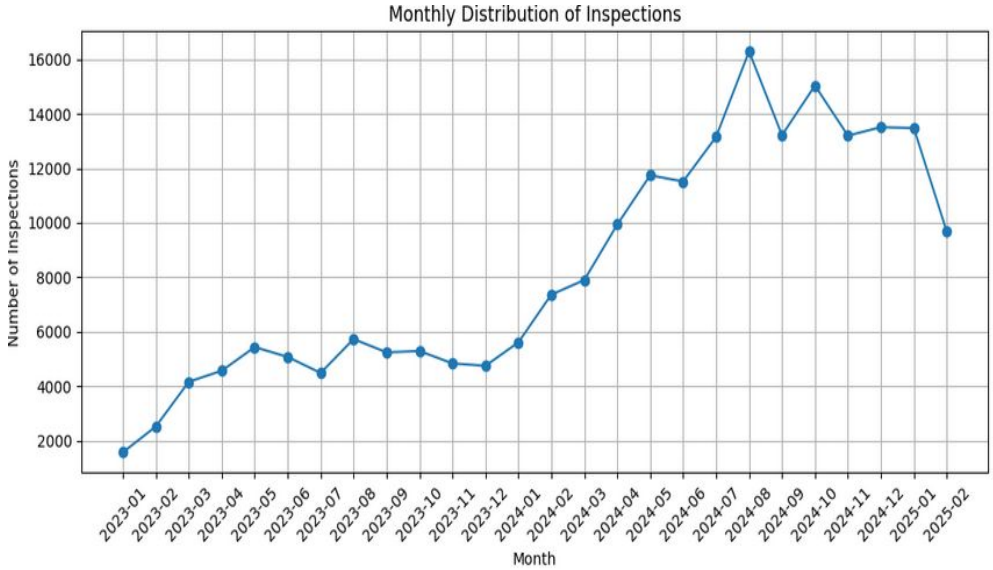




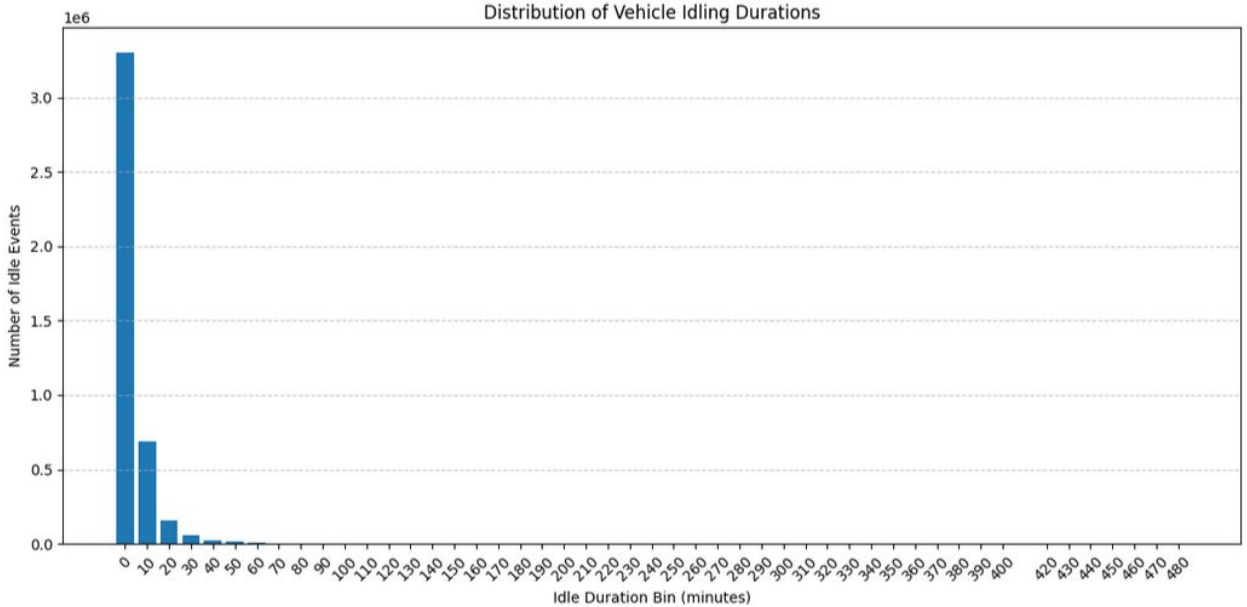


Motive Data Preprocessing - Data Vehicle Inspection Events





Motive Data Preprocessing - Data Idle Events



summary	minutes_idling
count	4264634
mean	8.2234
stddev	8.981
min	2.0
max	485.2

Site Radius Data

Raw Features

- ID
- Brand
- State
- City
- Latitude
- Longitude
- County
- Zip

US Census Zip Code Data

- Zip
- Latitude
- Longitude

- Cross join site and target ZIP codes to create all possible combinations
- Add target coordinates (latitude/longitude) as columns to the DataFrame
- Calculate distances between locations using the Haversine distance
- Filter results to only include locations within 40 miles

Derived Features

- One DataFrame for each site that contains
  - Zip code
  - Distance in miles
- For all zip codes within a 40 mile radius of each site

State Crash Data

Description: This aggregation summarizes provides a monthly summary of crash-related incidents around each site to help encode external risk factors for accidents based on where drivers will spend most of time on the road.

Raw Features

- Crash ID
- Crash Year
- Report State
- Fatal Count
- Fatalities
- Injuries
- Vehicles In Accident
- City
- City Code
- State
- Location
- Crash Date
- Year

Aggregate by Year, Month, State, County and count and sum over columns

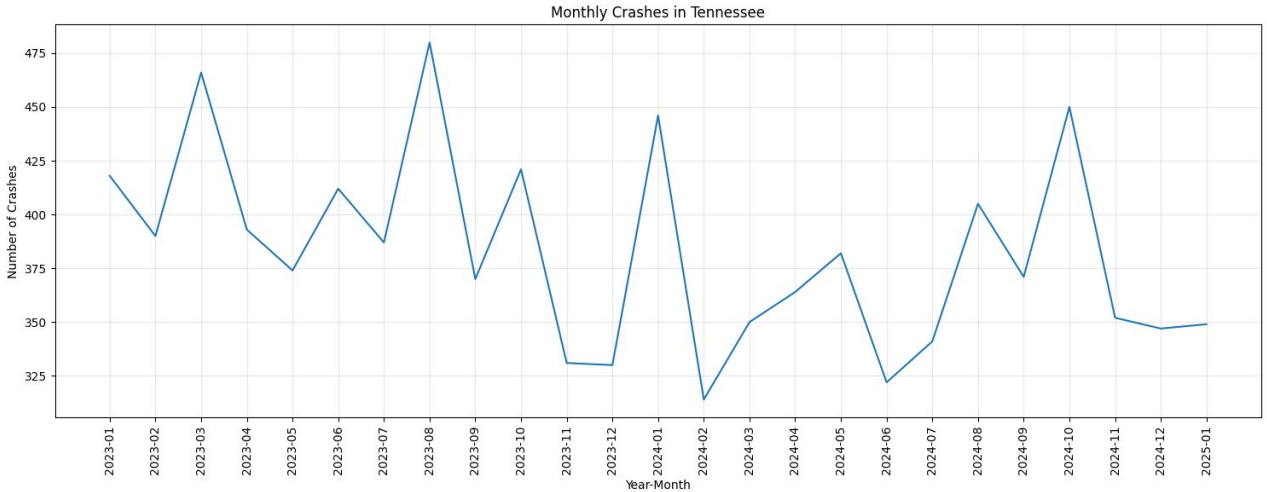
Derived Features

- Crash Year
- Crash Month
- State
- County Code
- Crash Count
- Total Fatalities
- Total Injuries
- Total Vehicles

State Crash Data EDA

Correlation matrix for key metrics:

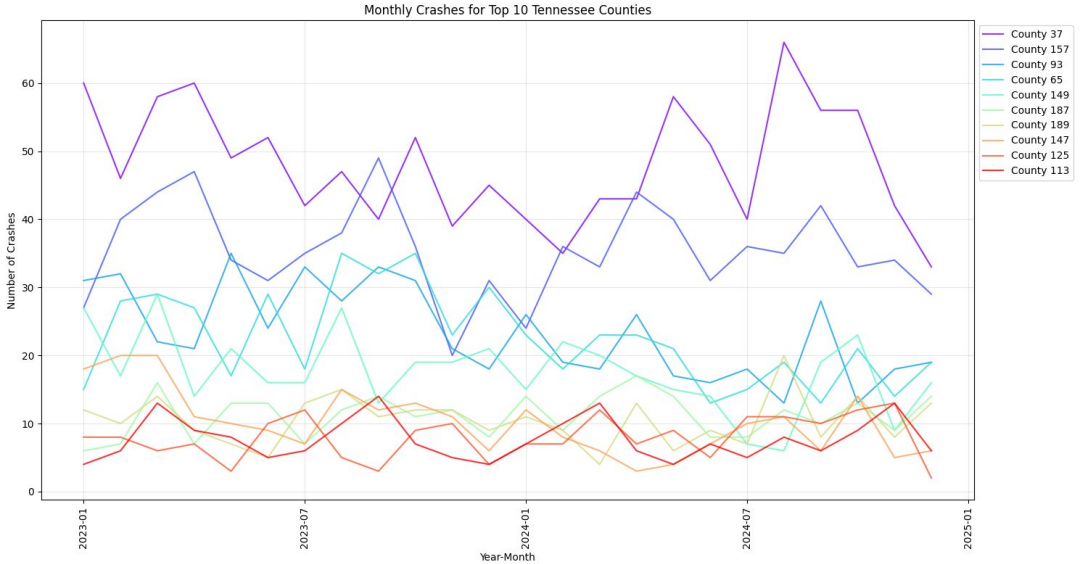
	crash_count	total_fatalities	total_injuries	total_vehicles
crash_count	1.000	0.437	0.851	0.932
total_fatalities	0.437	1.000	0.440	0.481
total_injuries	0.851	0.440	1.000	0.752
total_vehicles	0.932	0.481	0.752	1.000



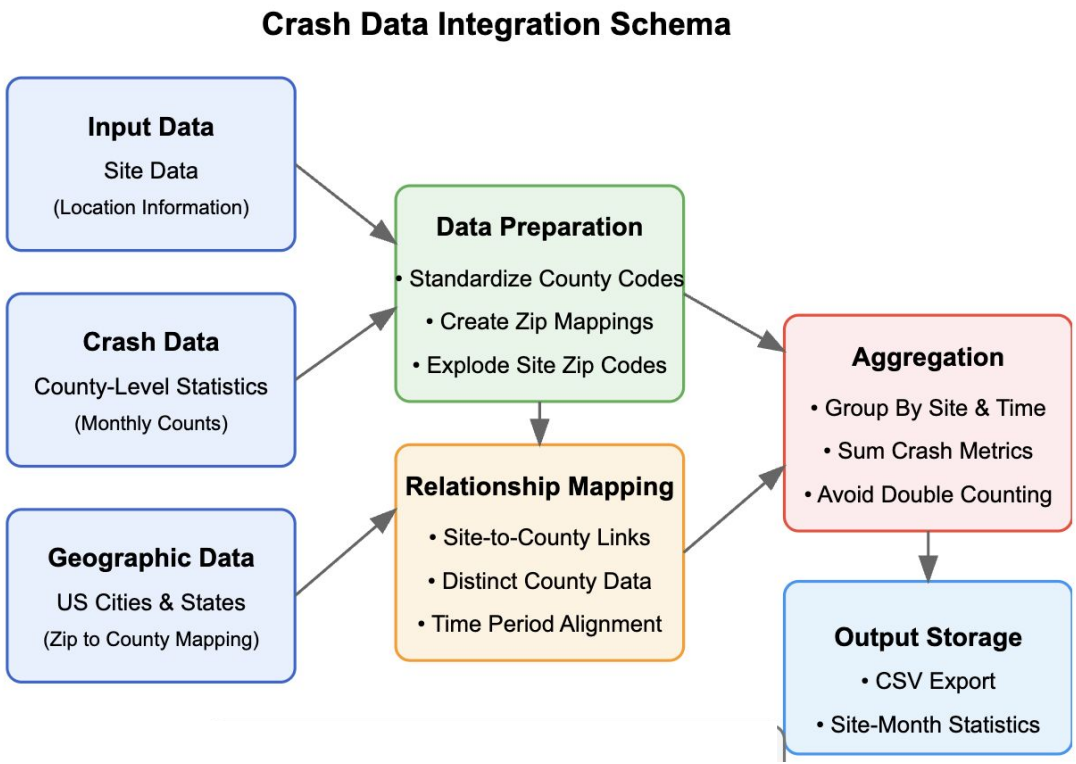
State Crash Data EDA

Correlation matrix for key metrics:

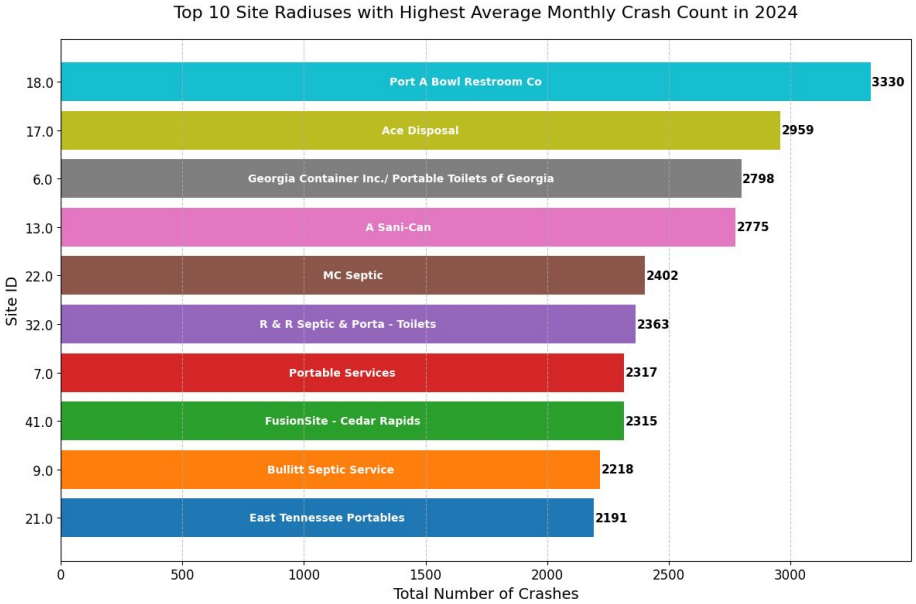
	crash_count	total_fatalities	total_injuries	total_vehicles
crash_count	1.000	0.437	0.851	0.932
total_fatalities	0.437	1.000	0.440	0.481
total_injuries	0.851	0.440	1.000	0.752
total_vehicles	0.932	0.481	0.752	1.000

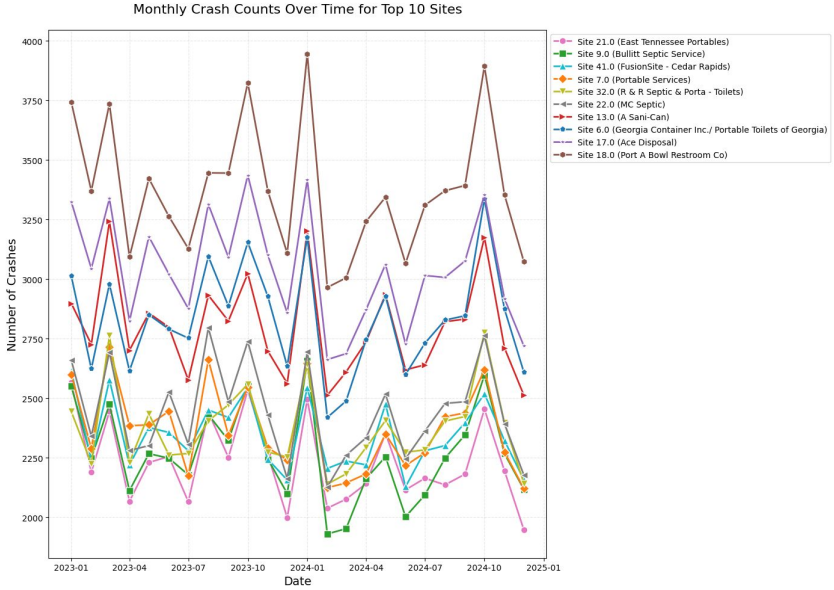


Joining State Crash Data to Site Data

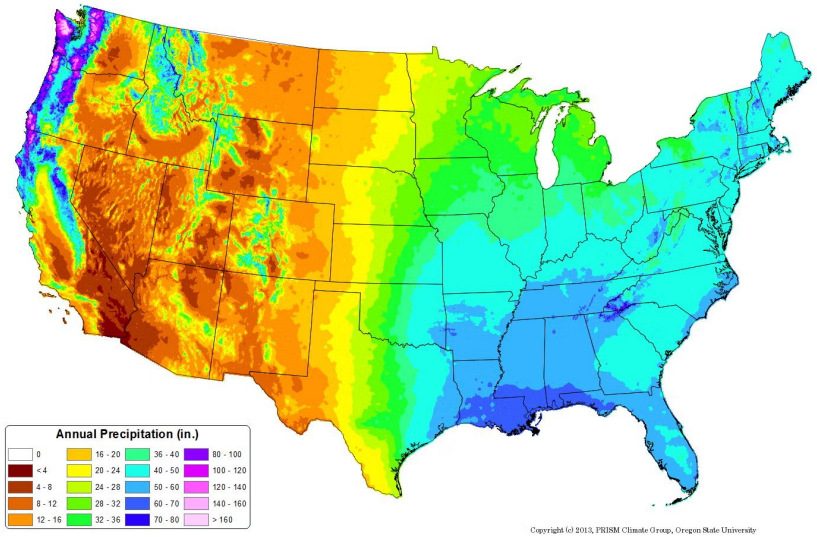


Site Radius Crash Data EDA





Precipitation Data



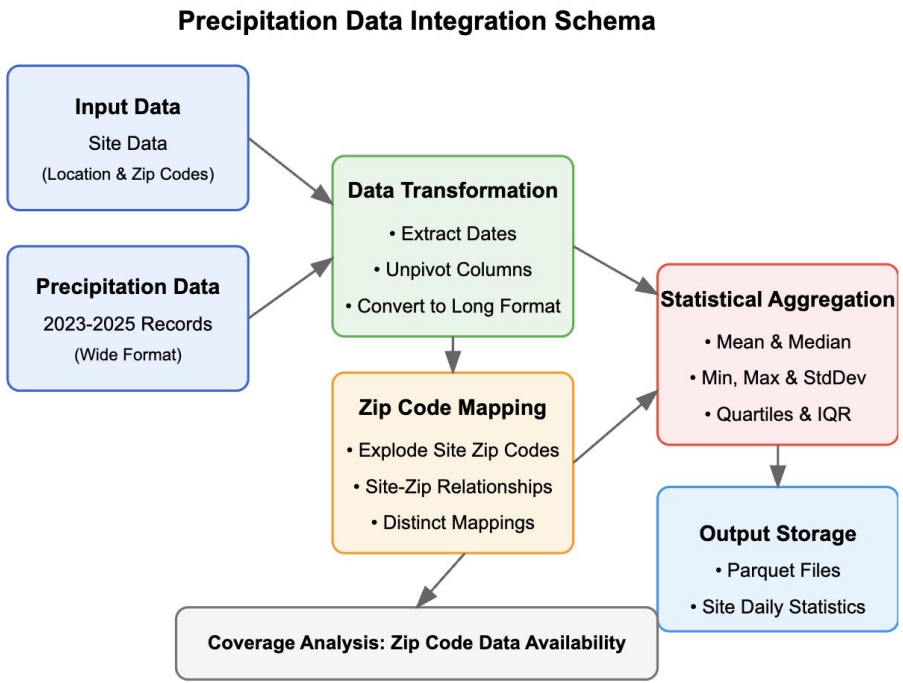
*

Raw Features

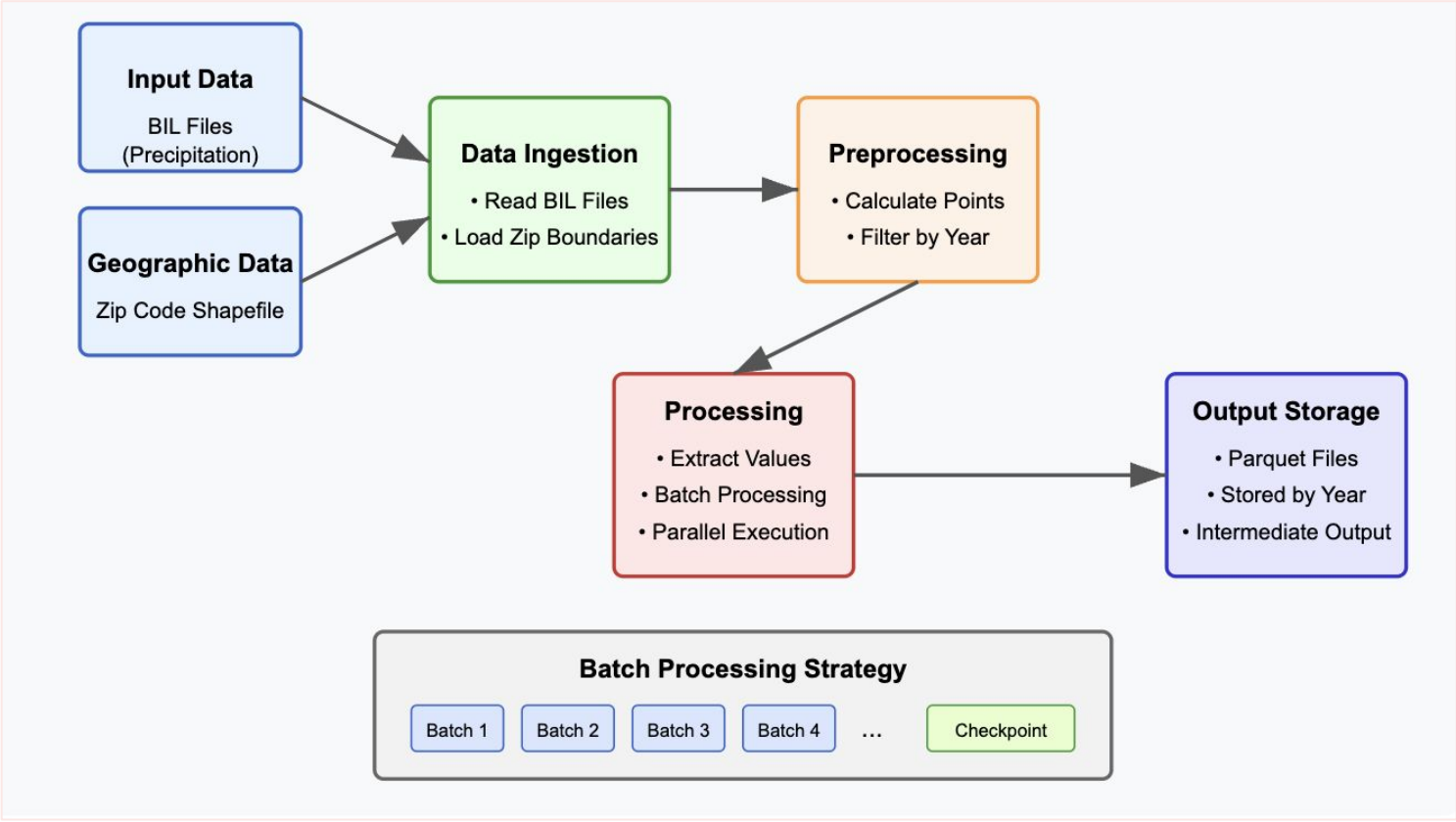
- Zip Code
- Latitude
- Longitude
- Total precipitation (rain+melted snow) for each Date in mm

**See Weather Data Processing for details*

Joining Precipitation Data to Site Data



Weather Data Preprocessing



Weather Data Preprocessing

Precipitation Data (BIL Files)

- Primary raster values
- Raster Metadata
  - Width and height
  - Coordinate reference system (CRS)
  - Transform parameters
  - Bounds
  - Nodata values

Zip Code Shapefile

- Latitude
- Longitude
- Zip Code Identifiers
  - GEOID20
  - ZCTA5CE20

State Crash Data Preprocessing

Preprocessing

- Generate representative points for each zip code polygon
- These points are used as sampling locations where values from the raster data will be extracted
- The code converts zip code polygons to specific point coordinates that fall within each zip code area

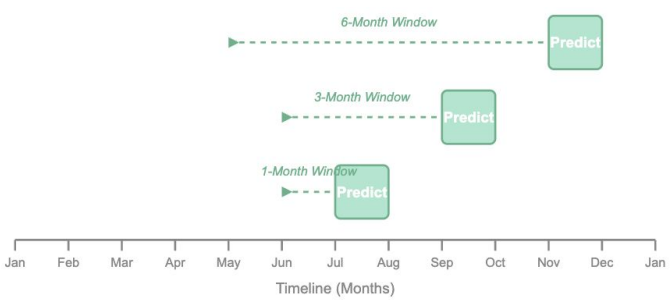
Processing

- Raster values are sampled at the representative points
- Converts geographic coordinates to pixel coordinates in the raster
- Checks if points fall within the raster bounds and handles nodata values appropriately
- Associates specific BIL data values with each zip code location

Location Data Feature Engineering

- Created lookback windows at different time scales for each variable

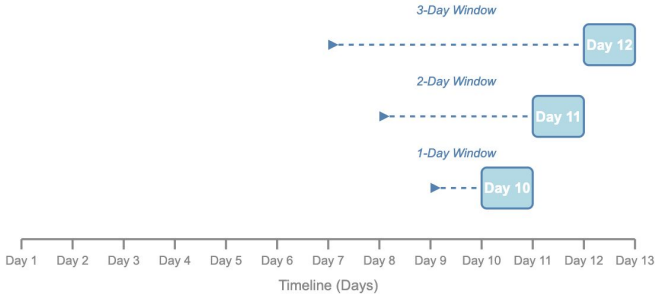
State Crash Data



Accident Data Moving Averages

- Total Monthly Crashes
- Total Monthly Fatalities
- Total Monthly Injuries
- Total Monthly Vehicles

Precipitation Data



Precipitation Data Moving Averages

- Summary statistics (min, max, mean, median, iqr) for rainfall over previous days

Weather Data Preprocessing

Output

- Zip Code
- Latitude
- Longitude
- Total precipitation (rain+melted snow) for each Date in mm

State Crash Data Data Collection

Direct Download from FMCSA

- One Excel File for each State with 19 sheets representing different variables and aggregations

Convert to csv and get variables of interest

- Open each .xlsx, save the data for each year of interest in a csv

State Crash Data

- One csv for each year with raw features for data pipeline

Target Variable

Accident Data

Insurance Claims Data

- Identify common columns
- Concatenate all rows
- Delete any duplicates across the driver id and date columns

All Accidents

- Driver id
- Insured driver name
- Insured driver first name
- Insured driver last name
- Date
- Year