

Data Analytics On Road Accidents in the US

Capstone Part 4: Non-Technical Presentation

Prepared & Presented By:
Priyanka Maurya

07th April 2025

AGENDA

- 1 *Project Introduction*
- 2 *Data Cleaning*
- 3 *Data Analysis & Visualization*
- 4 *Prediction & Forecasting Model*
- 5 *Conclusion & Q/A*





Purpose

This presentation will analyze accidents trend, severity, hotspots, and Forecast daily accident counts and predict accident severity using Machine Learning models, aiming to provide insights for improving road safety.

3. Data Analysis & Visualization

*Create dashboards & Reports
in Tableau.*

2. Data Cleaning and Processing

*Identify dataset,
perform data clearing
& processing in Python.*

4. Prediction and Forecasting

*Create Machine Learning
Models in Python.*

1. Project Introduction

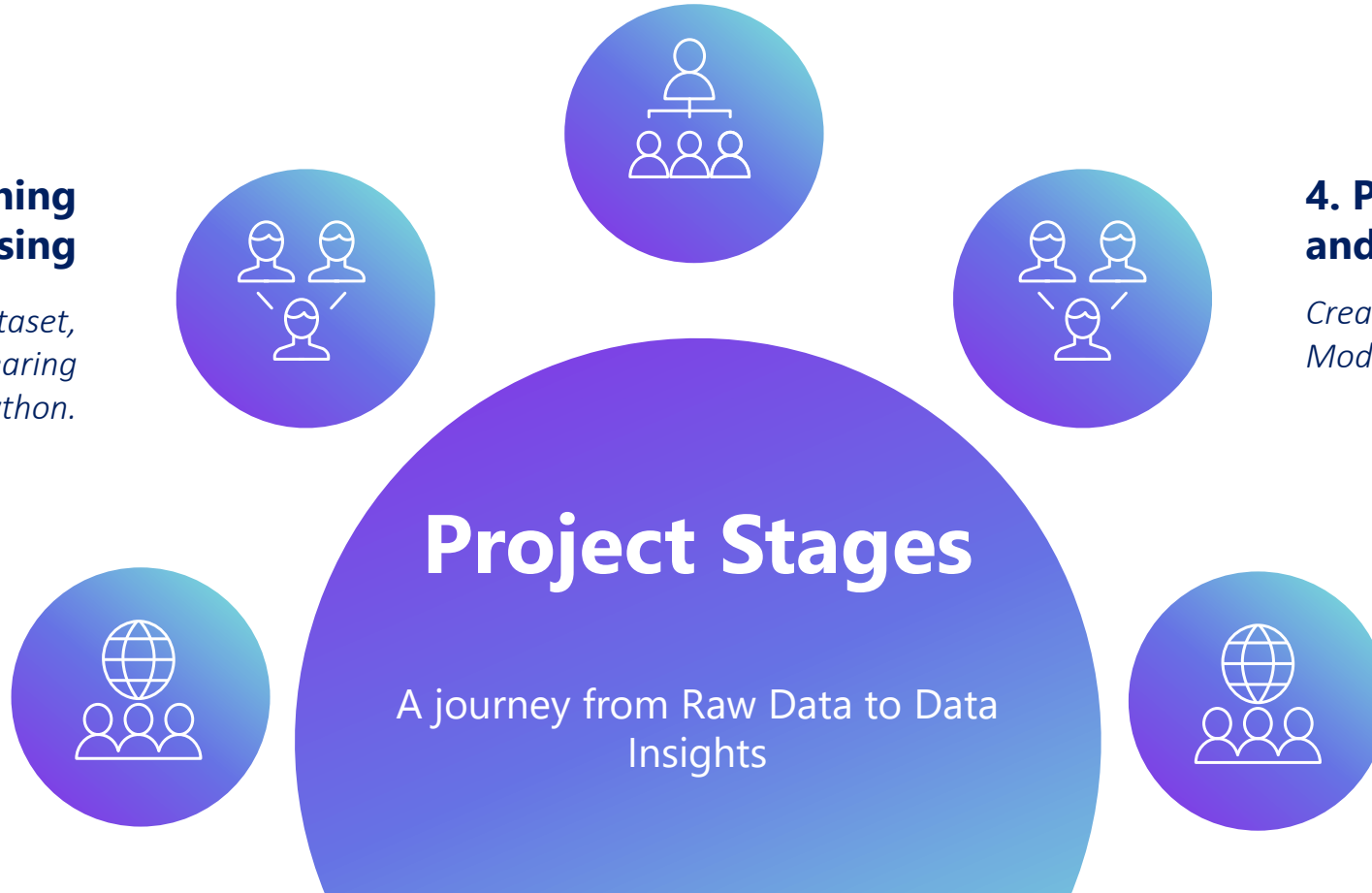
*Create Persona &
Project Objective*

5. Conclusion & Next Step

*Create slides to share
findings & next step in
MS PowerPoint.*

Project Stages

*A journey from Raw Data to Data
Insights*



1

Project Introduction

2

Data Cleaning

3

Data Analysis & Visualization

4

Prediction & Forecasting Model

5

Conclusion & Q/A



Facts & Figures

According to the World Health Organization (WHO), road traffic injuries are a leading global health issue, causing **1.3 million deaths each year worldwide**. In the United States, road traffic fatalities remain a serious concern, with nearly 39,000 deaths each year.

1.19 million deaths annually from crashes

Leading cause of death for ages 5-29

92% of fatalities occur in low-income countries

Over half of deaths are vulnerable road users

Road traffic costs 3% of GDP globally

Source: WHO [\[Link\]](#)





Objective

This analysis aims to provide insights for City Planners and Emergency Services such as Ambulance & Traffic Authorities; by identifying accident hotspots, peak times, and predicting severity. This will enhance road infrastructure, optimize traffic management, and improve resource allocation and emergency response efficiency.



Persona#1: Government City Planners

Pain Point

- *Limited insight into accident hotspots*

Key Performance Indicator

- *Identify top accident-prone state, city, and hotspots*
- *Determine peak accident periods (month, day, time)*
- *Analyze the impact of weather (humidity & temp) on accidents*

Benefits

- *Analyze accident hotspots and peak times to improve safety and optimize traffic management*

Persona#2: Traffic & Ambulance Services

Pain Point

- *Unpredictable accident surges, especially severe ones.*
- *Difficulty in efficient resource allocation for emergencies.*

Key Performance Indicator

- *Forecast daily accident counts and severity for optimized resource planning.*

Benefits

- *Better resource allocation for emergencies.*
- *Enhanced response efficiency and optimized traffic management.*

1

Project Introduction

2

Data Cleaning and Processing

3

Data Analysis & Visualization

4

Prediction & Forecasting Model

5

Conclusion & Q/A



This project began with a massive dataset of over 5 million US road accident records spanning 2016 to 2023. Dataset had 46 columns; covering everything from accident time and location to weather and road conditions.







Step # 1

Data Cleaning: Bringing Order to Chaos

The raw data had inconsistencies like missing timestamps, outliers in distance/weather values, and nulls in key columns.

I performed:







-  **Converted time** — Datetime conversion helped break accident timestamps into useful parts (like hour, day, and month)
-  **Handled odd values** — Outlier capping using IQR removed extreme values so they don't confuse the model.
-  **Filled in the blanks** — Imputation added reasonable values where data was missing (both numbers and categories).
-  **Fixed impossible values** — Negative value correction ensured fields like distance or temperature didn't have invalid negatives.

 **Why it was crucial:**

Without this, models would be misled by noise or invalid values, resulting in unreliable predictions.

This project began with a massive dataset of over 5 million US road accident records spanning 2016 to 2023. Dataset had 46 columns; covering everything from accident time and location to weather and road conditions.

*I explored how accident **severity** relates to factors like **weather, temperature, and time of day**, then engineered meaningful features for the models.*

-  **Spotted patterns** — Explored trends, correlations (how things relate), and category distributions.
-  **Figured out what matters most** — Ranked features using Random Forest Importance.
-  **Broke down time** — Extracted Hour, DayOfWeek, and Month from Start_Time.
-  **Gave categories a number** — Encoded categorical variables into numeric format for the model.
-  **Removed extra baggage** — Dropped redundant or low-impact columns (like Wind_Chill(F) when Temperature(F) was kept).
-  **Prepared Zipcodes** — Created Zipcode_5 for easier regional analysis.

Why it was crucial:

*This process reduced noise, **highlighted impactful variables**, and ensured the model had **cleaner, more meaningful input**, leading to better performance.*



Step # 2

**EDA & Feature Engineering:
Understanding the Story
Behind the Data**

This project began with a massive dataset of over 5 million US road accident records spanning 2016 to 2023. Dataset had 46 columns; covering everything from accident time and location to weather and road conditions.







Step # 3

Model Creation & Refinement: Training Smart Models for Smarter Decisions

Two paths were taken:

- *Time Series (SARIMAX): Forecasted daily accident volume*
- *Classification Models (Random Forest, XGBoost): Predicted severity levels*

Enhanced using:

-  **Prioritized smart inputs** — Feature importance ranking helped identify which inputs were most useful.
-  **Tuned for better results** — Hyperparameter tuning adjusted the model settings for improved accuracy.
-  **Double-checked reliability** — Cross-validation tested the model multiple times to ensure it's not just lucky.
-  **Handled imbalance fairly** — Class balancing with weights made sure rare accident types weren't ignored.

 **Why it was crucial:**

Proper model building and tuning improved accuracy and made the results actionable and trustworthy.

1

Project Introduction

2

Data Cleaning and Processing

3

Data Analysis & Visualization

4

Prediction & Forecasting Model

5

Conclusion & Q/A



KPIs

1. Identify top accident-prone state, city, and hotspots
2. Determine peak accident periods (month, day, time)
3. Analyze the impact of weather (humidity & temp) on accidents

Findings

01

- California is the most accident-prone state
- Los Angeles is the most accident-prone city in California
- 91706 is the top accident hotspot in California, Baldwin Park

02

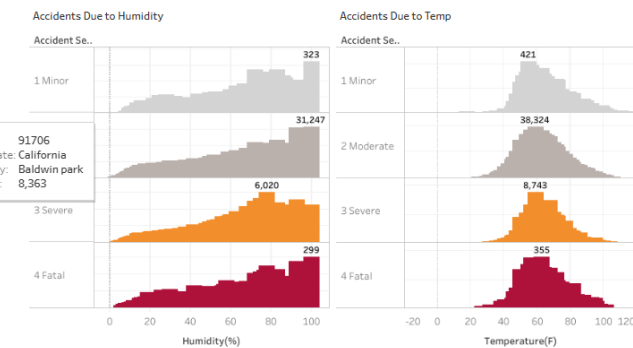
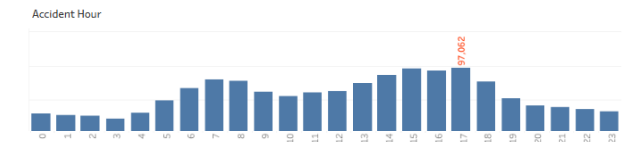
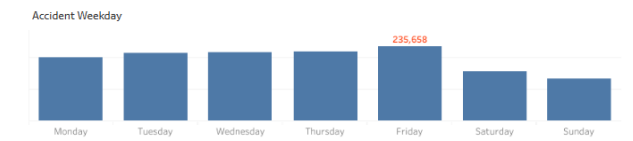
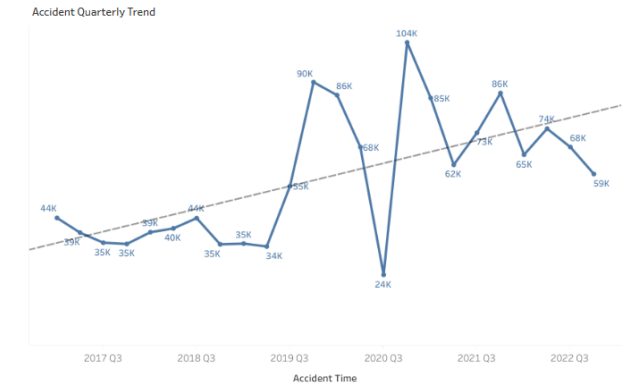
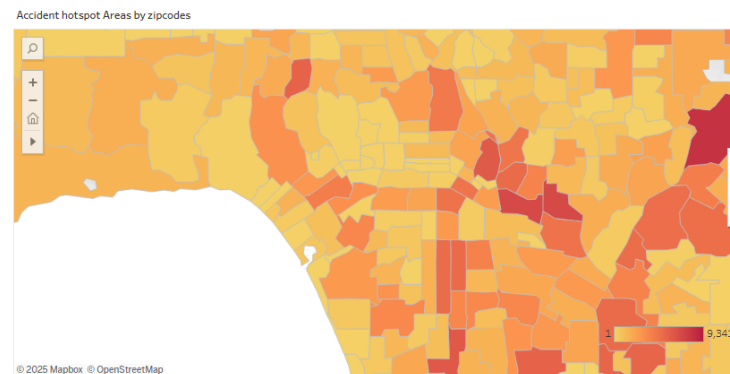
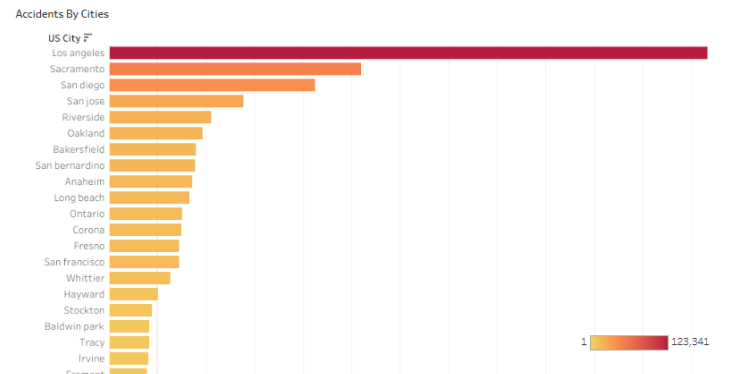
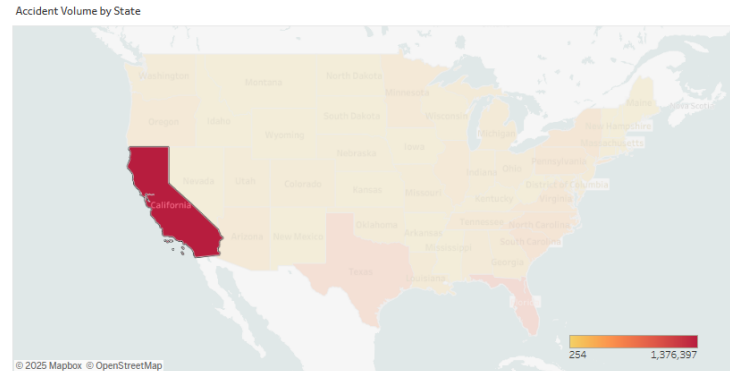
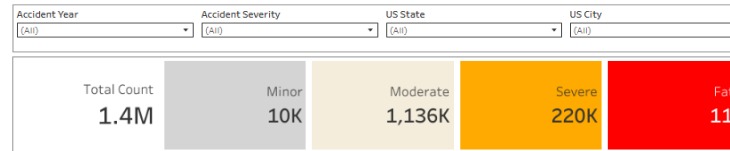
- Peak accident month is **December**
- Peak accident day is **Friday**
- Peak accident hour is **1700 Hours**.

03

- Higher **fatal** accidents when **Humidity** is between **93% - 100%**
- Higher **Severe** accidents when **Humidity** is between **75% - 80%**
- Higher accidents when **Temp** is between **59° - 65° Fahrenheit**

Refer Tableau Dashboard for Detailed Analysis - [Link](#)

USA Road Accident Analysis (2017 - 2022)



1

Project Introduction

2

Data Cleaning and Processing

3

Data Analysis & Visualization

4

Prediction & Forecasting Model

5

Conclusion & Q/A



Time Series Forecasting - SARIMAX Model

RMSE

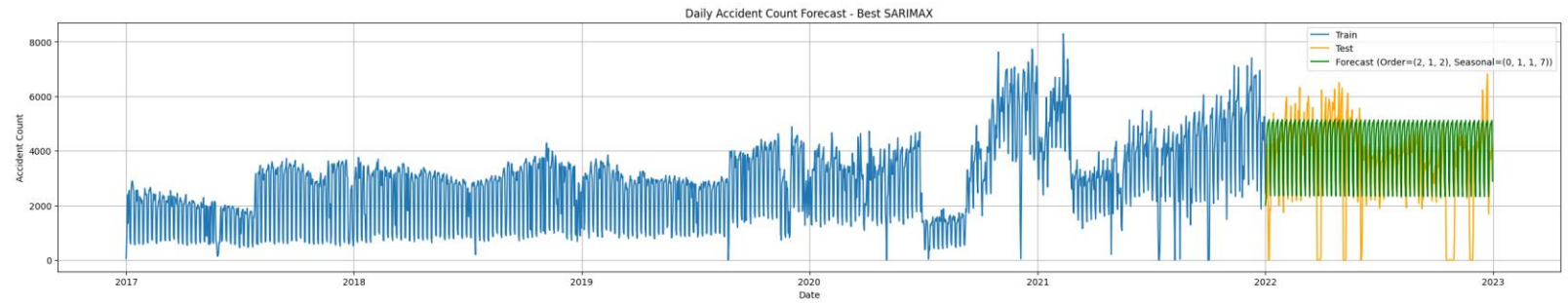
1597

(Root Mean Squared Error – lower is better)

What this model does: The SARIMAX model is a forecasting tool used to predict future accident counts by learning from past patterns in the data. It takes into account seasonal trends (like weekly or monthly cycles) and other time-based behaviors to give accurate short-term forecasts.

Why was this model chosen: SARIMAX is well-suited for time series data with seasonal effects. Unlike basic models like ARIMA, it can handle more complexity such as recurring weekly patterns in accident trends — which are common in real-world traffic scenarios.

How it performed: The model was trained on several years of accident data. Initially, ARIMA was used as a baseline but didn't capture seasonality well. After tuning, SARIMAX with configuration (2,1,2)(0,1,1,7) delivered a better result.



What we did to improve it:

- Explored multiple SARIMAX configurations
- Identified and adjusted seasonal cycles (7-day weekly pattern)
- Removed irregular spikes (outliers) that confused the model
- Ensured clean datetime formatting to help detect patterns

What the result means: An RMSE of 1597 means the model predicts the number of daily accidents with a reasonably small error margin, given the size and spread of the data. This forecasting can be used by traffic authorities or emergency planners to prepare for high-risk days 🚦 📅.

What these models does: Predict the severity level (1 to 4) of an accident using inputs like time, weather, location, and road features.






Why I used these models:

- **Random Forest** is a simple and reliable model giving a good baseline accuracy and is simple to interpret.
- **XGBoost** is a more advanced version, performs better with large datasets, handles imbalanced data better and improves performance on rare classes.
- Both are classification models, useful when predicting categories (like severity levels 1 to 4).

How the models performed:

- The baseline Random Forest model predicted Severity 2 (most common class) with high accuracy, but struggled to detect less frequent severities (like 1 and 4).
- After tuning and using class weights (to give more importance to underrepresented classes), performance improved slightly.
- Final accuracy was around 80%, but detection of Severity 1 remained weak due to its rarity in the dataset.

What was done to improve the results:

-  Applied class weights: This told the model to pay more attention to less frequent severity levels.
-  Performed Grid Search: to fine-tune model settings/finding the best model settings (hyperparameters).
-  Applied Cross-Validation: to validate model stability across different data samples.
-  Trained XGBoost: which improved performance for Severity 3 but still faced challenges with Severity 1.
-  Class imbalance: remained a key challenge despite these efforts.

What the result means: The model is very effective for predicting the most common severity levels (Accuracy: 80%), which are important for traffic planning and emergency services. However, it still misses rare and severe cases, which is a known challenge in imbalanced datasets.

Severity Prediction Models: Random Forest & XGBoost

83%

Overall Accuracy

1

Project Introduction

2

Data Cleaning and Processing

3

Data Analysis & Visualization

4

Prediction & Forecasting Model

5

Conclusion & Q/A



Limitations & Future Steps

What challenges were faced: Class imbalance made it harder to predict rare accident severities like level 1 or 4. Most records were of severity 2, so models tended to focus on that.



Use SMOTE

A technique that creates synthetic examples of rare classes to balance the dataset better.



Bring in external data

like traffic flow, weather alerts, or holidays to provide more context to the model.



More feature tuning

can help uncover patterns that are currently missed.



Deploy as a tool

Which could help local authorities predict both the count and severity of upcoming road incidents.

An illustration on the left side of the slide shows several hands in white and purple tones assembling light blue puzzle pieces. The background is a gradient of teal and blue.

Conclusion

What was built:

⚠️ I developed two powerful models — SARIMAX for predicting how many accidents may happen, and Random Forest/XGBoost for predicting how serious they might be.

What worked well:

✅ SARIMAX gave strong forecasting accuracy ($RMSE \approx 1597$) for daily accident counts.

✅ Random Forest achieved ~80% accuracy and did especially well with Severity 2 (the most common class).

Where there's room to grow:

🔍 Less common severity levels were harder to detect due to limited examples.

🕒 Some improvements (like deeper tuning or external data usage) were limited by time, but could unlock much more accuracy in the future.

Thank You

Let's Begin The Q/A...