Predicting Fault Volumes Using Weather Data

Presented by Pushpa NB
24th May 2024

INTRODUCTION

- To develop a predictive model for daily fault intake based on weather events, particularly rainfall.
- Analysis covers fault and rainfall data over a 3-year period for a specific geographic area.
- The primary goal of this project is to implement predictive modeling techniques to accurately forecast fault volumes, enabling proactive resource planning and optimizing operational efficiency.

Business Problem

Fault volume forecasting has significant repercussions, including inefficient resource allocation, operational disruptions, service delays, and escalated reactive maintenance costs.

Key Considerations:

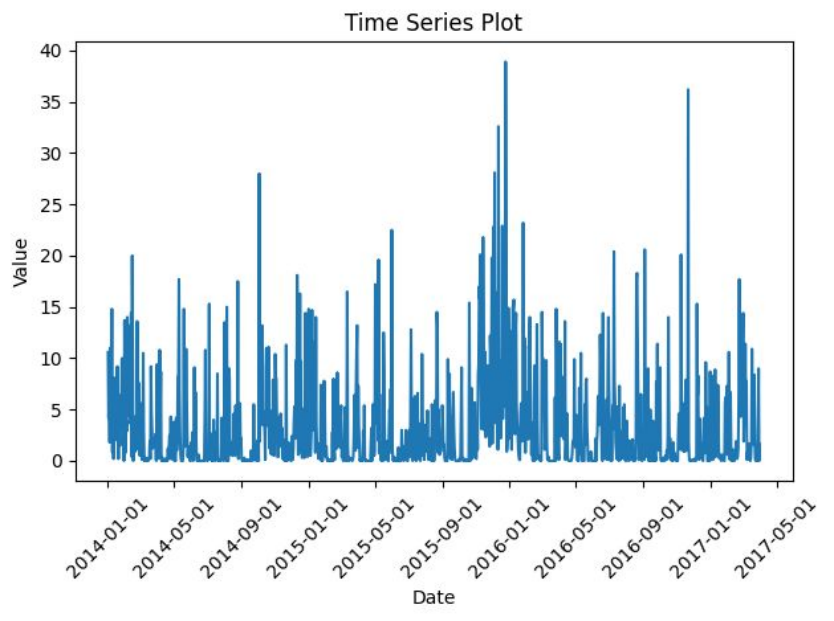
- Service Delivery: This can result in delays in resolving customer issues due to insufficient capacity or unnecessary costs incurred from over-allocation.
- Customer Satisfaction: Dissatisfied customers are more inclined to switch providers, leading to revenue loss and damage to the organization's reputation.

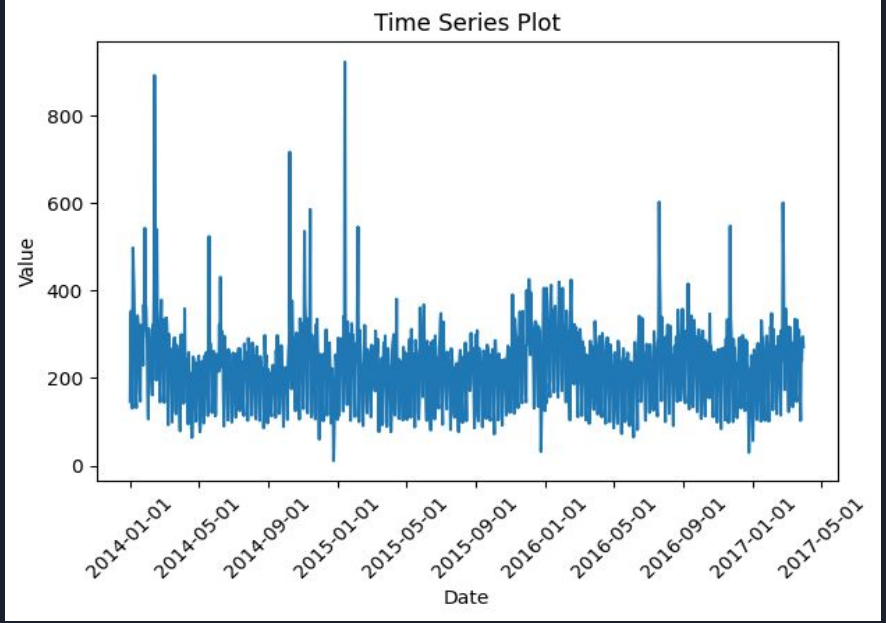
Data Overview

- Datasets Provided
  - Fault Data: Information on individual faults over a 4-year period, including fault types and report dates.
  - Rainfall Data: Daily rainfall measurements for the same period.
  - Calendar Lookup: Details such as day of the week and bank holidays.
- Key Fields:
  - Fault Data: Fault ID, Report Date, Initial MFL (fault location).
  - Rainfall Data: Observation Date, Rainfall (mm).
  - Calendar Data: Actual Date, Day of Week, Day Num Cal Week/Month/Year, Bank Holiday indicator.

Methodology

- Data Preprocessing:
  - Cleaning and preparing the data for analysis.
  - Merging fault data with rainfall and calendar data.
- Feature Engineering:
  - Creating features to capture short-term and long-term rainfall impact.
  - Categorizing faults based on location (e.g., home, network, exchange).
- Model Selection:
  - Choosing appropriate models (e.g., regression, time series) to predict fault volumes.
  - Justifying model choices based on data characteristics.



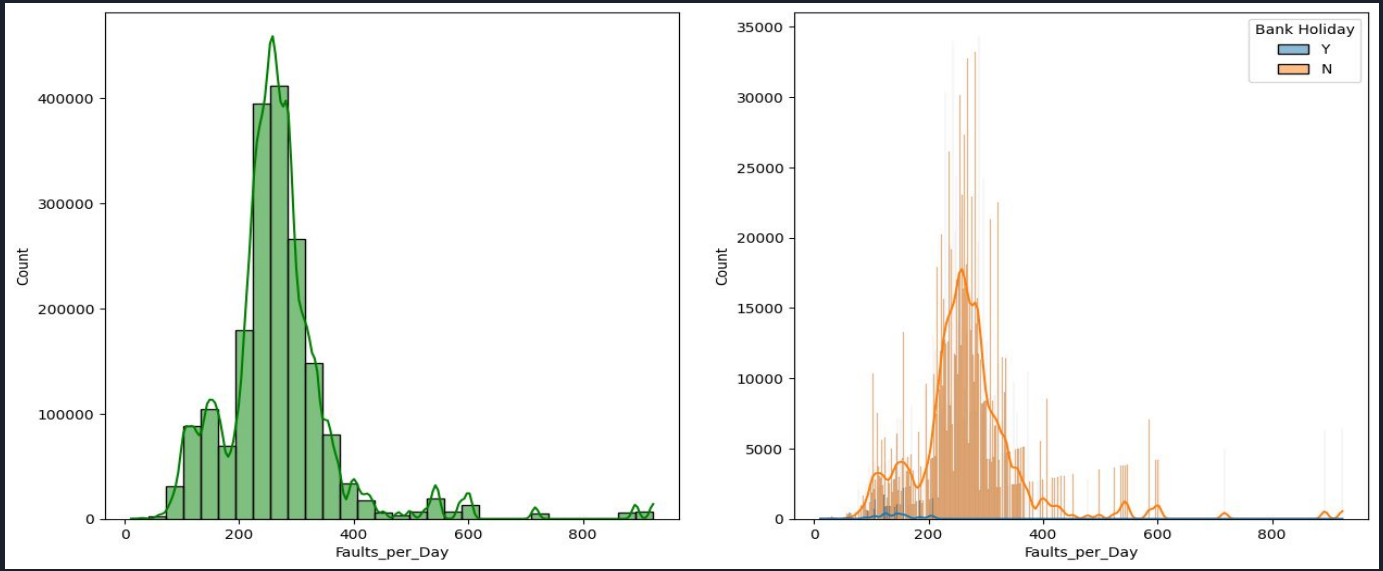


Rainfall Impact Analysis:

- Examined the distribution of rainfall depths over the period.
- Found that rainfall depth typically ranges from 0-15mm, with occasional heavy rainfall events ranging from above 15mm to nearly 40mm.

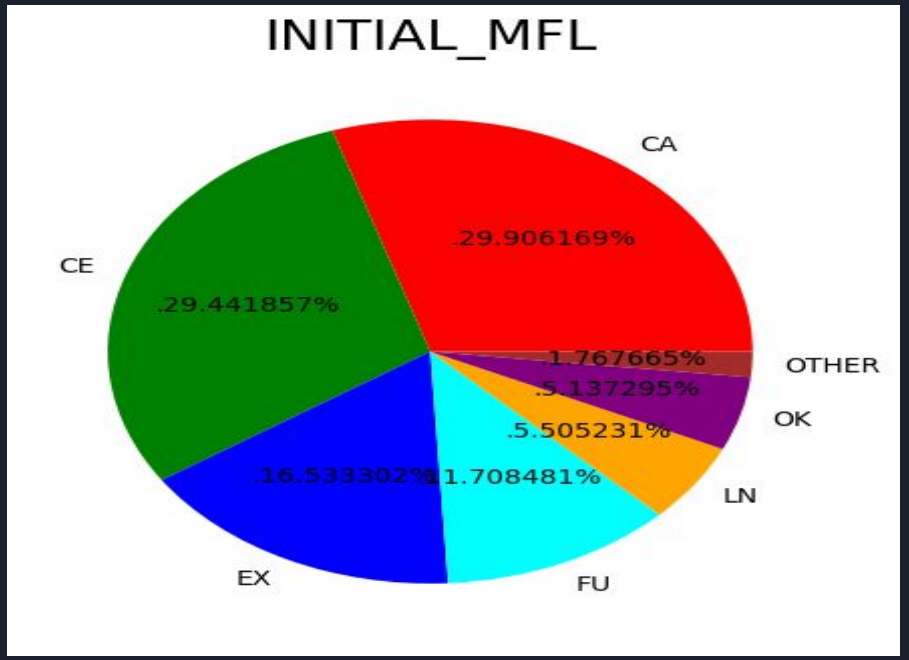
Faults Count Analysis:

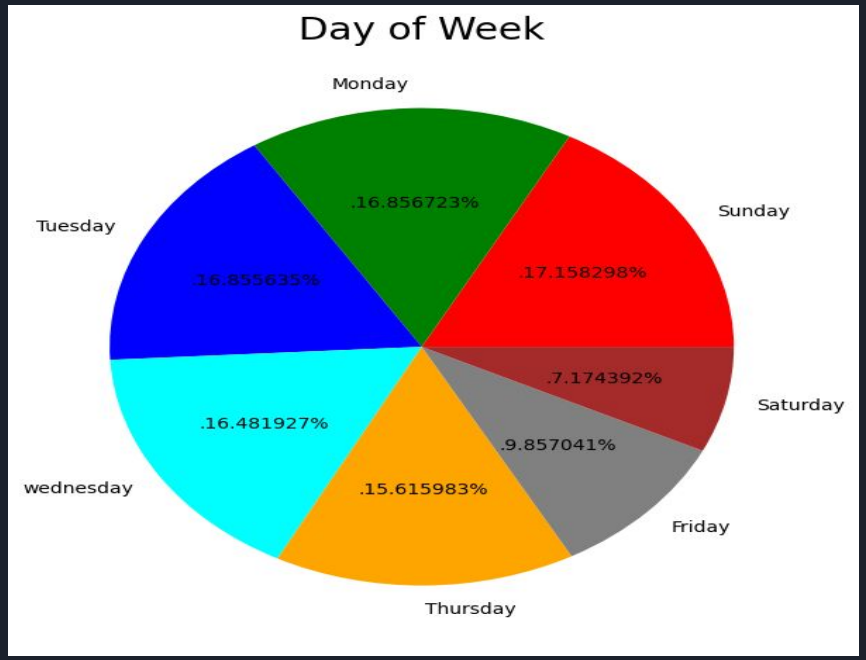
- Daily Counts: Analyzed daily fault counts, noting a typical range of 100-400 faults, with occasional spikes exceeding 600 and even 800.



Bank Holidays Impact:

- Fault Trends: Analyzed fault occurrence on bank holidays versus non-bank holidays.
- Observation: Higher fault counts on non-bank holidays, indicating increased issues when regular maintenance might be less available.



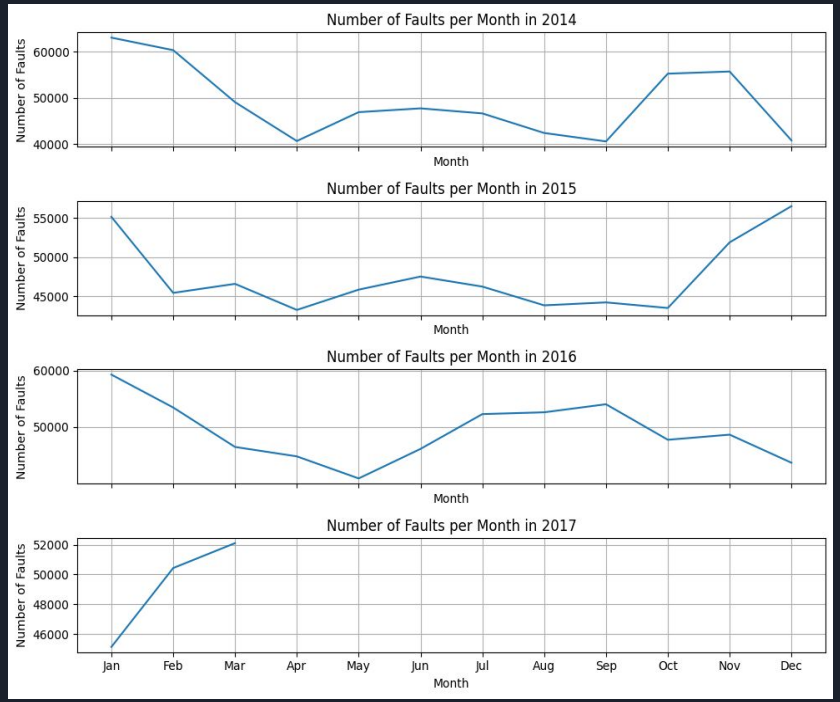


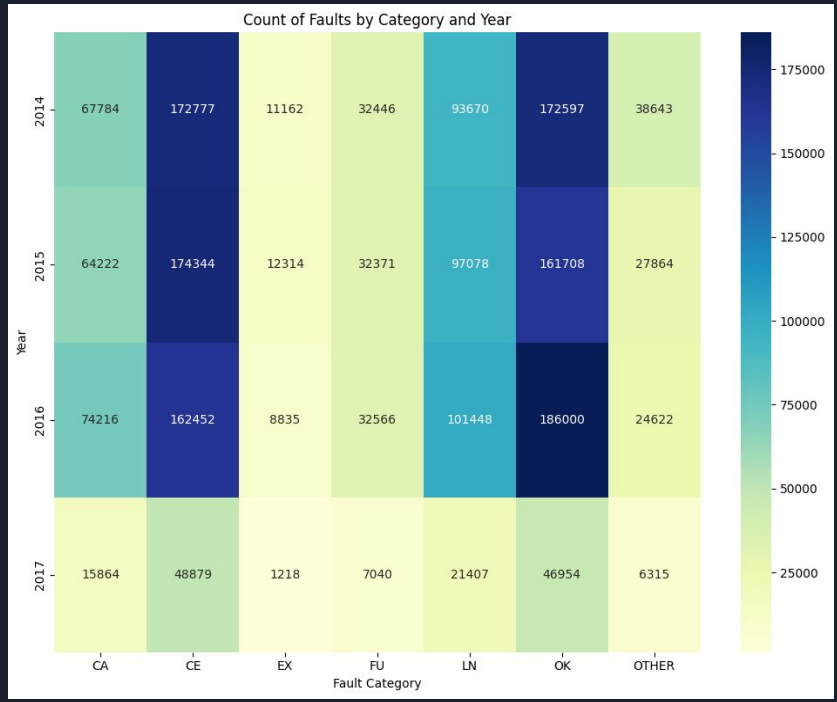
Location-Based Fault Analysis:

- Showed distribution of faults across different locations (CA, CE, EX, FU, LN, OK, OTHERS).
- Higher fault occurrences in CA and CE locations, with fewer faults in other areas.

Weekly Fault Distribution:

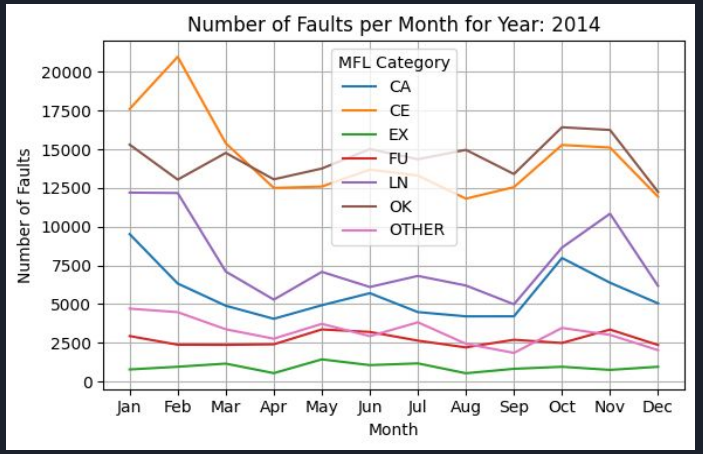
- Day-of-Week Analysis: Investigated fault counts across different days of the week.
- Illustrated that faults are more frequent from Sunday to Thursday, with a noticeable drop on Fridays and Saturdays.

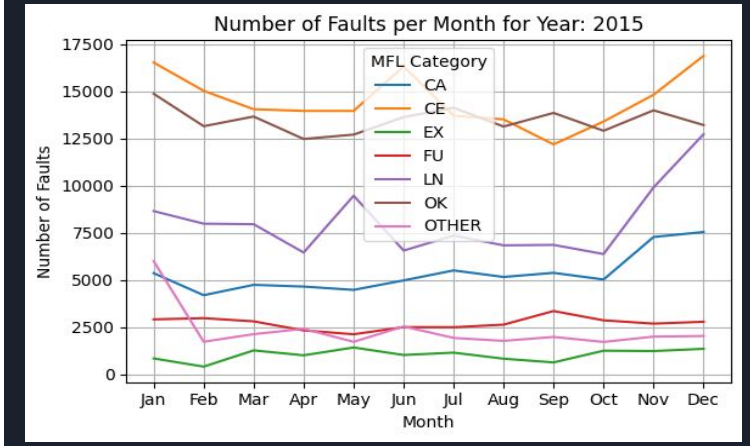


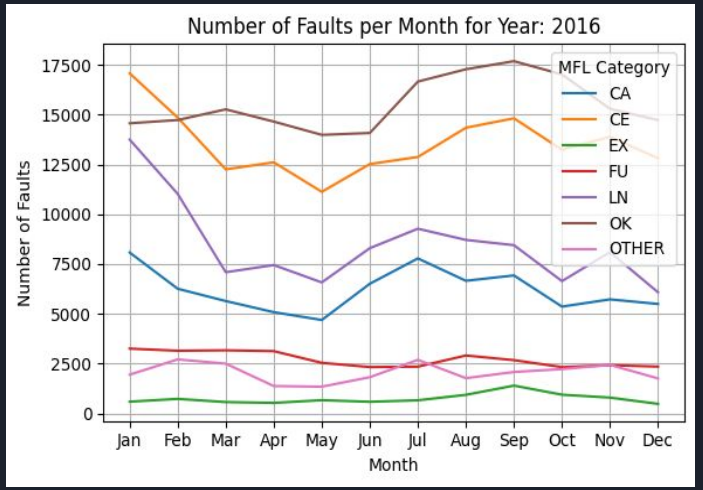


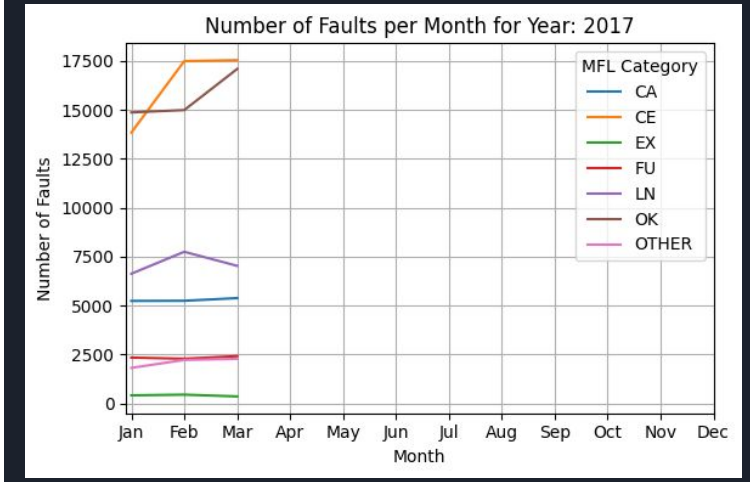
Monthly Fault Distribution:

- Monthly Trends: Analyzed fault counts across different months of the year.
- Yearly Comparison: Found higher fault counts at the beginning and end of each year, with some annual variations (e.g., 2016 showing different patterns).









Yearly Location-Based Fault Analysis:

- Analyzed the number of faults by location over the years.
- Found that CE (D-side overhead network) and OK (Line tested OK) locations consistently had the highest fault counts across multiple years.
- Presented a line graph to illustrate the fault trends by location over the years, highlighting the persistently high fault counts in CE and OK locations.

Model Development and Evaluation

Model Training:

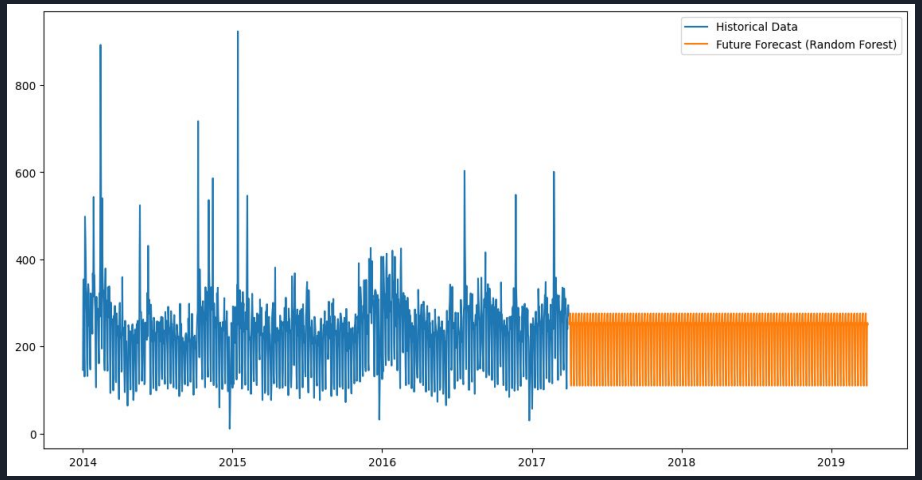
- Employed a range of regression models (Linear Regression, Ridge Regression, Decision Tree, Random Forest Regressor, XGBRegressor) alongside time series models like ARIMA and SARIMA.
- Achieved a moderate accuracy rate, signifying the models' robust predictive capability.

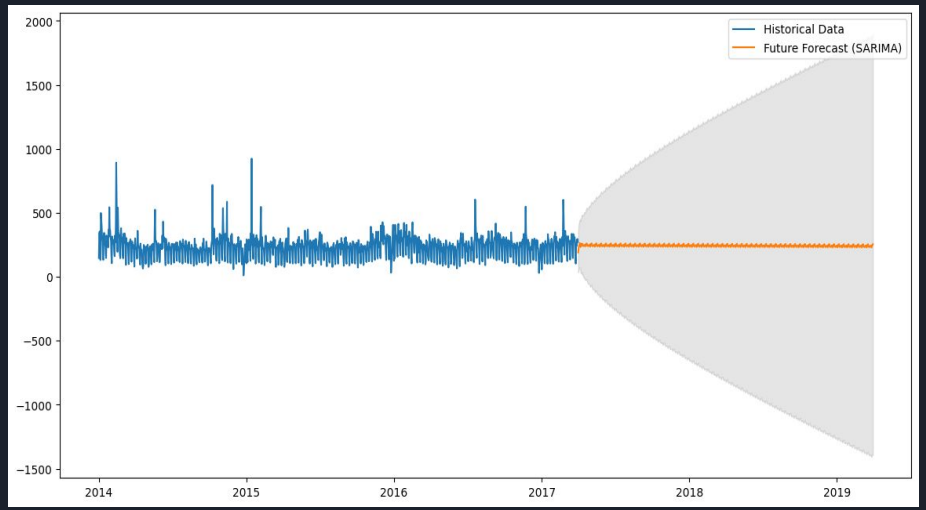
Feature Importance:

- Analyzed the significance of features such as rainfall data, fault locations, day of the week, and bank holidays.
- Identified the relative importance of each feature in predicting fault volumes.

Short-Term Effects Consideration:

- Incorporated short-term effects by leveraging features like Rainfall_7day_avg, Rainfall_14day_avg, Rainfall_30day_avg, Rainfall_1day_lag, and Rainfall_7day_lag.
- Examined how recent rainfall trends influence fault volumes across varying timeframes.





- Employed ARIMA and SARIMA models to forecast the results for the upcoming year based on our existing dataset.
- Attached a graphical representation illustrating the forecasted outcomes.
- Leveraged our current dataset to train the models and make predictions.
- Noted a decrease in error counts, indicating improved accuracy and efficacy of the models in capturing the underlying patterns and trends within the data.

Conclusion and Recommendations

Conclusion:

- Leveraging our predictive model to forecast fault occurrences based on historical data provides valuable insights for planning ahead.
- Analysis reveals distinct patterns, with faults being more frequent during the initial and final months of the year and lower counts observed on Fridays and Saturdays.
- Weekdays, especially Sunday to Thursday, exhibit higher fault occurrences.
- The high accuracy of our predictive model underscores its effectiveness for resourcing planning.

Recommendations:

- Implement the predictive model to enhance the accuracy of resourcing plans.
- Make operational adjustments to accommodate higher fault counts on non-bank holidays and weekdays when planning maintenance and resource allocation.

Thank You