Data Science Mini Project

Sharayu Kalambe
PRN : 22070521056

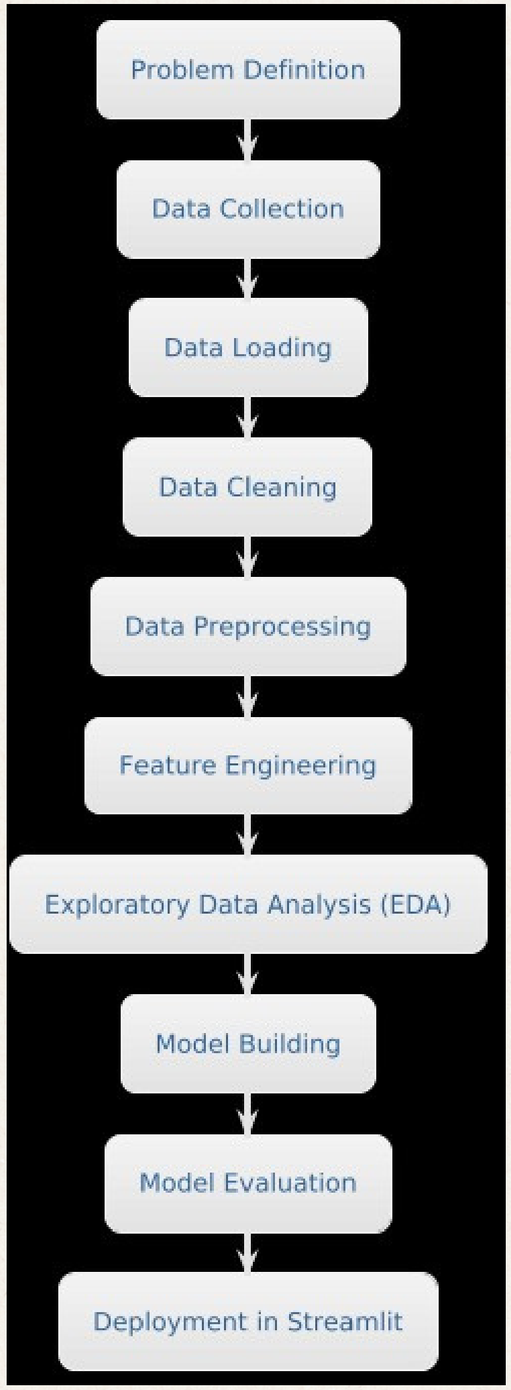
Problem Statement

- Inconsistent coal stock levels across thermal power stations affect power generation reliability.
- Many plants face shortages where actual stock falls below required levels, leading to critical operational risks.
- Supply–demand gaps in coal receipt and consumption impact plant performance and efficiency.
- Variations in indigenous vs. imported coal usage influence cost, availability, and sustainability.
- Low Plant Load Factors (PLF) may be linked to inadequate coal management practices.
- Lack of clear insights into criticality reasons hinders timely corrective action.

About Dataset

- The dataset is taken from the India Data Portal.
- Represents Coal Stocks data collected from thermal power stations across India.
- The data is recorded daily from the year 2006 to 2024.
- Contains coal stock details categorized by state, sector, and individual power station.
- Key data fields include:
- Mode of transport, plant capacity, and sector/utility information.
- Normative and actual coal stocks, daily receipt, and daily consumption.
- Plant Load Factor (PLF) and criticality status of each station.
- Reasons for criticality and other remarks related to coal supply.
- The dataset enables comprehensive analysis of coal availability, stock adequacy, and operational efficiency across India's thermal power plants.

Workflow

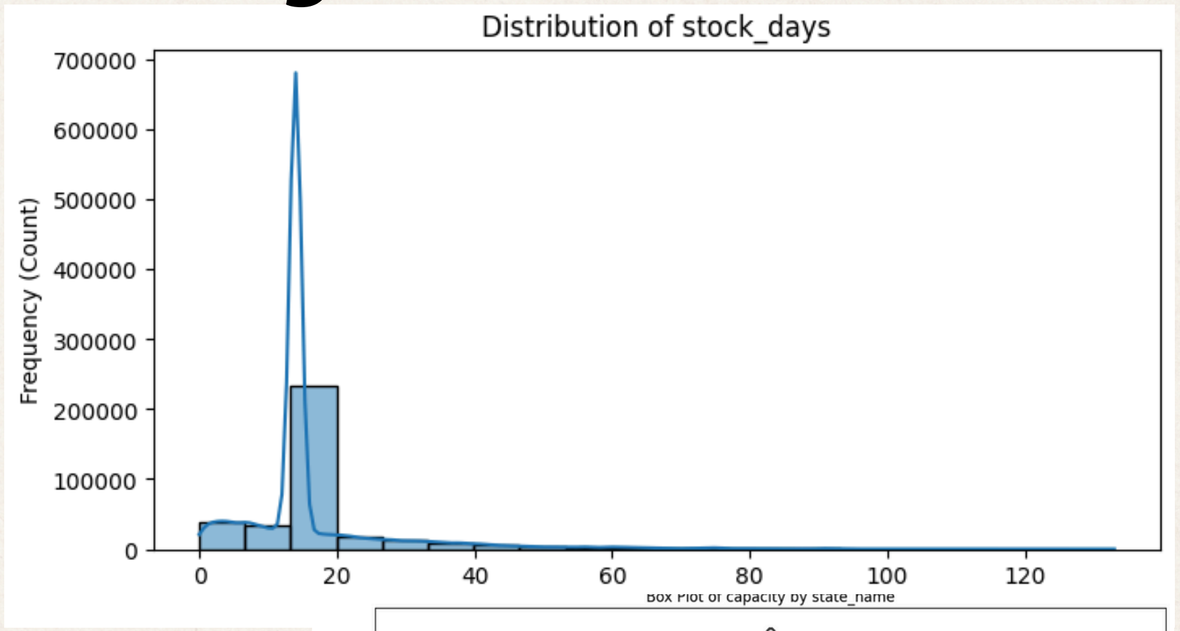

Problem Definition
Data Collection
Data Loading
Data Cleaning
Data Preprocessing
Feature Engineering
Exploratory Data Analysis (EDA)
Model Building
Model Evaluation
Deployment in Streamlit

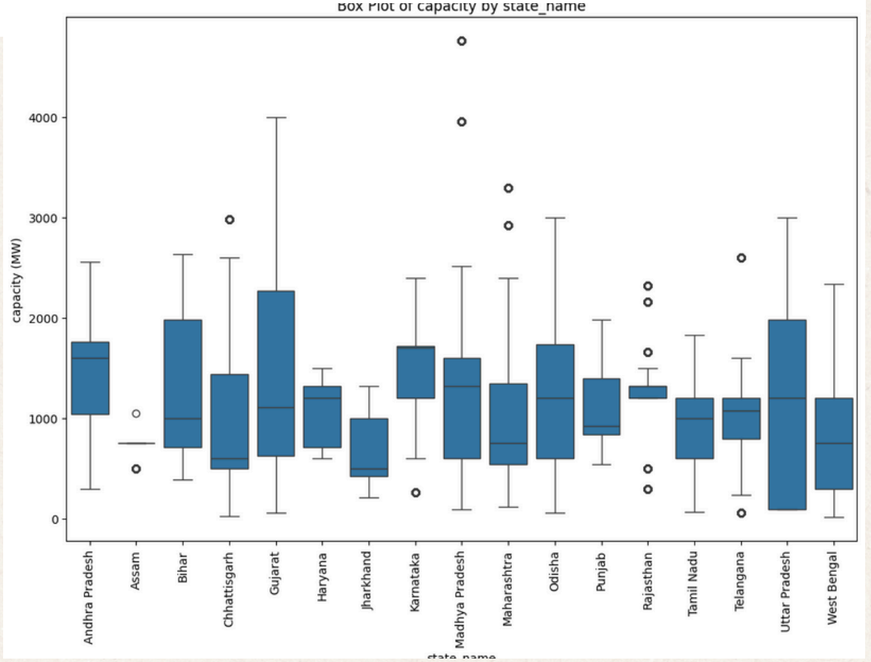
Preprocessing

- Performed data cleaning by checking and handling missing values across all columns.
- Dropped rows with missing entries in key fields like capacity, daily requirement, daily receipt, and daily consumption.
- Converted the ‘date’ column to proper datetime format and standardized it to YYYY-MM-DD.
- Removed duplicates and corrected data inconsistencies for accuracy.
- Prepared and saved the cleaned file as cleaned_daily-coal-stocks.csv for further EDA and modeling.

Exploratory Data Analysis(1/2)

- Analyzed descriptive statistics (mean, median, range) for key variables like daily requirement, consumption, and total stock.
- Visualized coal stock trends over time to identify fluctuations in availability.
- Compared actual vs. normative stock levels to detect shortage or surplus patterns.
- Examined state-wise and sector-wise variations in coal stock and consumption.
- Analyzed mode of transport (rail, road, mixed) and its influence on coal receipt efficiency.

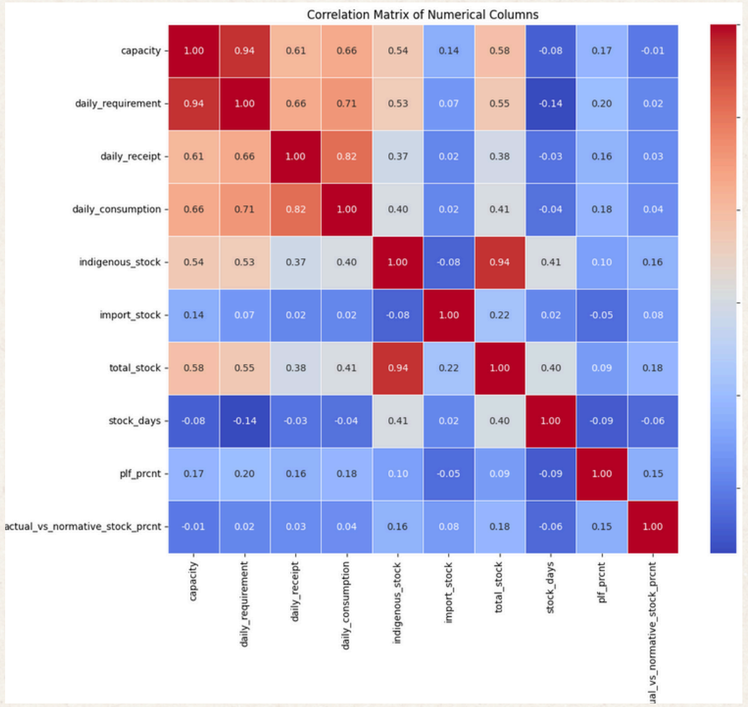


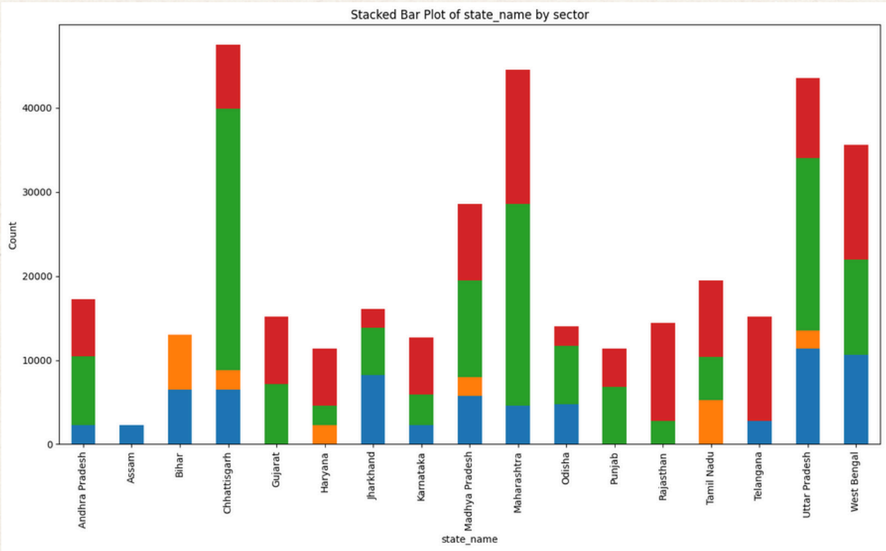


Exploratory Data Analysis(2/2)

Insights and Observations:

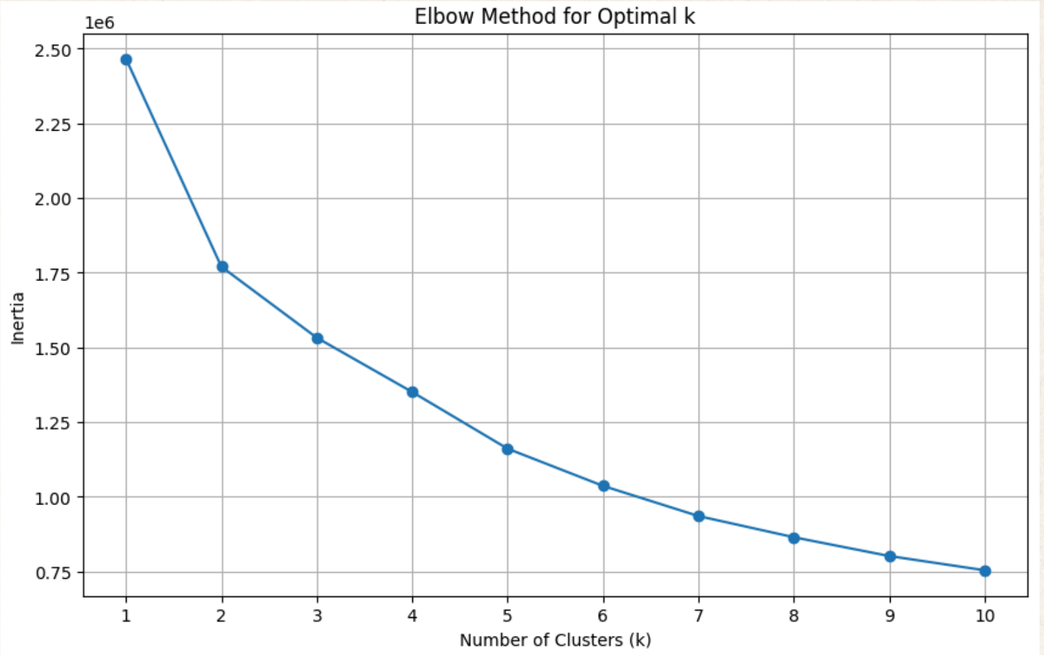
- Several plants frequently operate below normative stock levels, indicating supply delays.
- States with higher daily consumption often face greater stock volatility.
- Indigenous vs. imported coal proportions vary significantly, affecting reliability and cost.
- Plants with low Plant Load Factor (PLF) often correspond to critical stock conditions.
- Criticality remarks provide useful context on supply constraints and operational risks.
- These insights help identify inefficiencies in coal management and guide predictive modeling for stock adequacy.



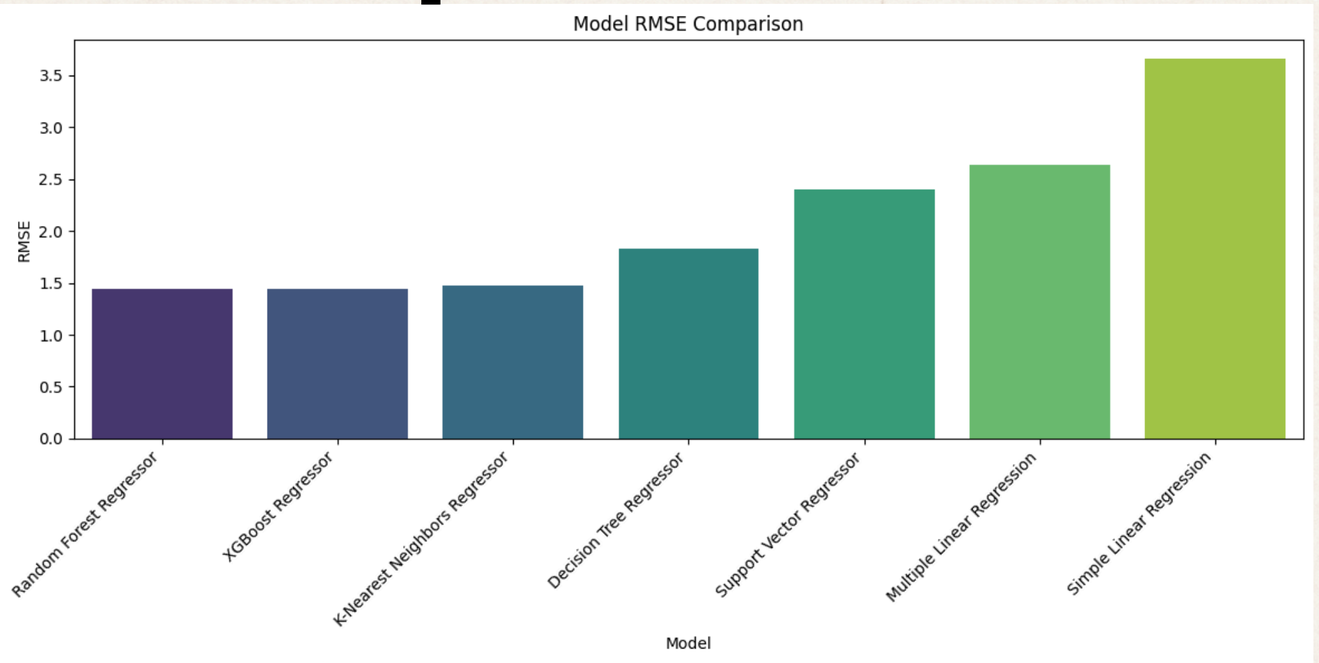


Model Implementations (1/2)

- Implemented multiple machine learning models to predict coal stock adequacy and daily stock requirements.
- Used Python with key libraries such as pandas, NumPy, matplotlib, and scikit-learn.
- The dataset was split into training and testing sets to ensure unbiased evaluation.
- Applied different regression-based algorithms including:
- Linear Regression – to analyze relationships between coal receipt, consumption, and stock.
- Decision Tree & Random Forest – to capture complex, nonlinear supply patterns.
- XGBoost Regressor – for improved predictive accuracy and efficiency.
- Support Vector Regressor (SVR) – to handle variations in stock adequacy across plants.
- Models were compared using performance metrics such as R^2 Score, MAE, and RMSE to determine the most accurate approach.



Model Implementations (2/2)

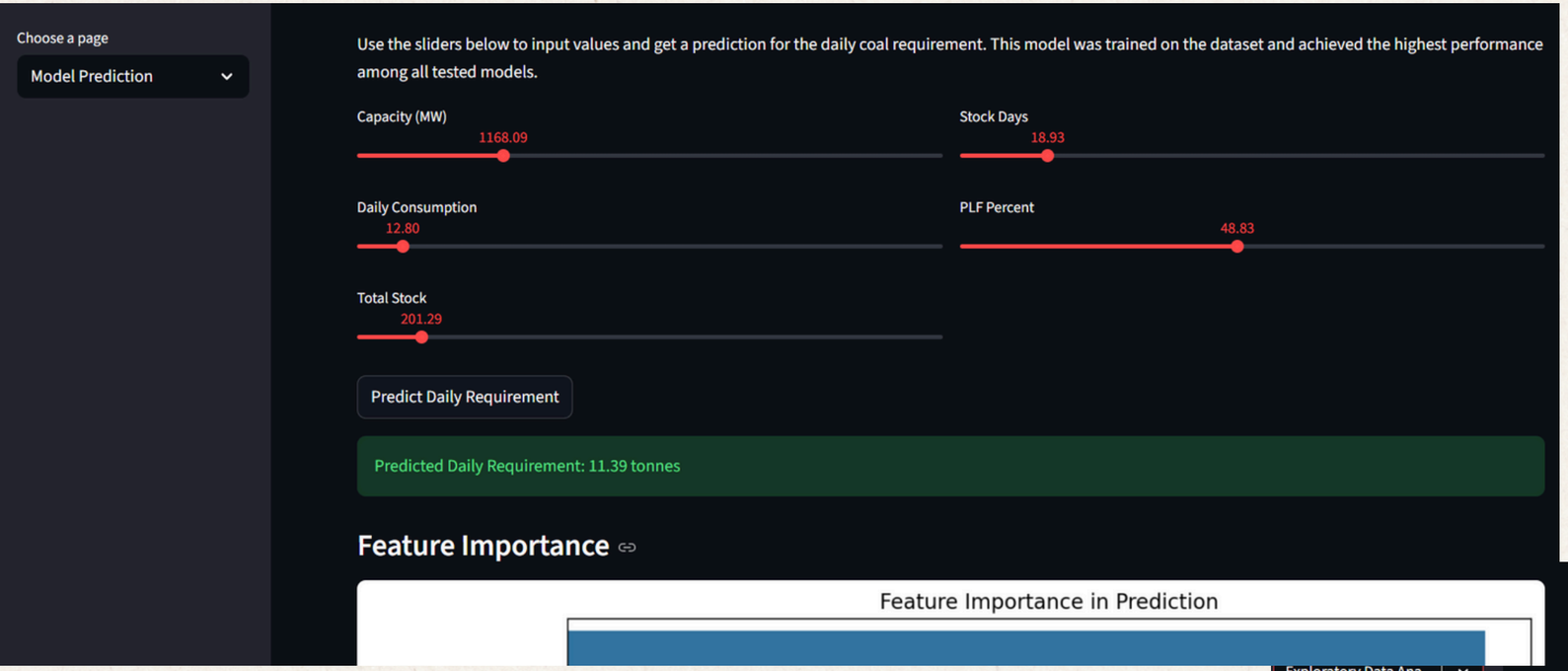


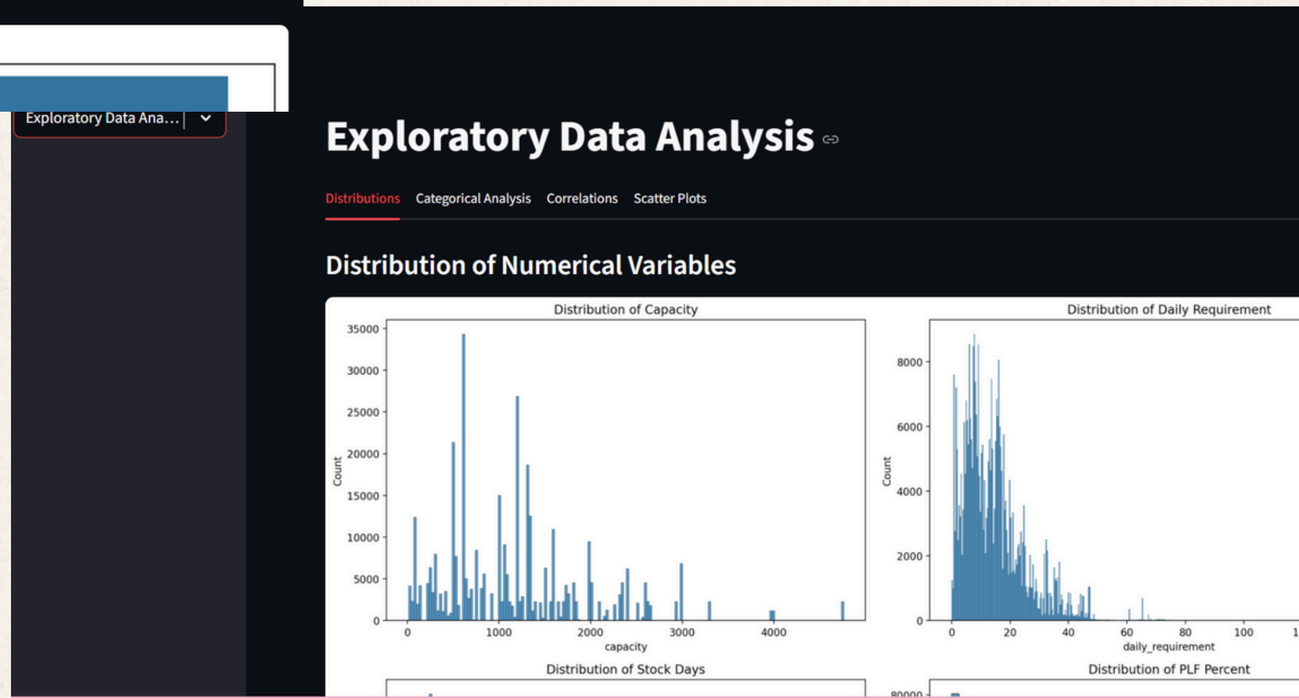
- All models were trained and tested to evaluate their ability to predict coal stock adequacy accurately.
- Performance was measured using R^2 Score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).
- Linear Regression provided a baseline with moderate accuracy.
- Decision Tree improved results by capturing non-linear relations between stock and operational factors.
- Random Forest and XGBoost achieved higher accuracy and lower error rates, showing strong generalization.
- Among all, XGBoost delivered the best performance, indicating its effectiveness for forecasting coal stock adequacy and shortages.

Streamlit Application(1/2)

- Built an interactive Streamlit app for visualization and prediction.
- Includes three sections:
- Dataset Overview: View and explore coal stock data (2006–2024).
- EDA Section: Interactive charts for distributions, correlations, and trends.
- Model Prediction: Input parameters (capacity, PLF, stock days, etc.) to predict daily coal requirement.
- Uses the XGBoost model for accurate predictions.
- Helps in real-time coal stock analysis and decision-making.

Streamlit Application(2/2)





Thankyou!!