YEDİTEPE UNIVERSITY
Following Atatürk's Renaissance

COMPUTER ENGINEERING DEPARTMENT

CSE464 Case Study 1

INTRODUCTION TO DATA SCIENCE & BIG DATA ANALYTICS
(AUTUMN SEMESTER 2024)

ESRA KAYA
20200702007

I. Short Story of your startups Business Idea

- Our startup's business idea is to optimize energy consumption in the energy-intensive steel industry, reducing both costs and environmental impact. To achieve this, we have developed a platform that comprehensively analyzes, predicts, and provides actionable recommendations for improving energy efficiency. The platform collects energy consumption data from industrial facilities, uses advanced analytics to process the data, and applies machine learning to accurately predict future energy demands. Additionally, it offers customized recommendations to enhance energy efficiency, helping steel manufacturers lower their energy costs while transitioning toward more sustainable production practices.

II. Problem Summary/Definition

- The steel industry is an energy-intensive sector, where high energy consumption leads to both increased costs and environmental issues. Optimizing energy consumption in this sector is complex and dependent on numerous factors. These factors include production processes, equipment used, weather conditions, reactive power requirements, and even the day of the week. Steel manufacturers are unable to accurately manage their energy consumption due to a lack of understanding of the interactions between these factors, resulting in unnecessary energy waste and high costs. Existing energy management systems are often based on retrospective analyses and are not capable of predicting future energy needs with sufficient accuracy. This leads to operational inefficiencies and difficulty in cost control.

III. Justification of the Problem (SWOT Analysis)

Strengths:

- The dataset includes comprehensive records of energy consumption, reactive power, and CO_2 emissions, providing a solid foundation for analysis.
- Advanced visualization tools and machine learning models enhance the capability to extract actionable insights.
- The steel industry’s focus on energy optimization aligns with sustainability and cost-efficiency goals, creating an impetus for data-driven decision-making.

Weaknesses:

- High energy dependency makes the industry vulnerable to fluctuations in energy prices and supply.
- Limited granularity in certain data points, such as detailed operational schedules, could impact model precision.
- Potential challenges in real-time data integration and predictive model deployment.

Opportunities:

- Implementing predictive analytics can significantly improve energy usage forecasting, reducing costs and environmental impact.
- Insights from the analysis can guide the adoption of renewable energy sources or more efficient energy technologies.
- Enhanced operational efficiency can strengthen the company's market competitiveness.

Threats:

- Regulatory pressures and environmental compliance requirements may increase operational costs.
- External factors, such as global energy crises or supply chain disruptions, pose risks to energy availability and cost stability.
- Over-reliance on historical data may lead to inaccuracies if operational dynamics change significantly.

IV. Alternative Solution/ Recommendations/Decisions

1. **Implement Real-time Energy Monitoring Systems:**
    - Deploy sensors and IoT devices to capture real-time energy consumption data.
    - Integrate real-time data into predictive models to improve accuracy and responsiveness.

2. **Adopt Renewable Energy Sources:**
    - Explore options for solar, wind, or other renewable energy sources to supplement traditional energy usage.
    - Utilize the insights from energy patterns to optimize renewable energy integration.
3. **Introduce Dynamic Energy Pricing Strategies:**
    - Use predictive insights to negotiate better energy contracts or implement dynamic pricing during low-demand periods.
    - Optimize operations during periods of low energy costs to reduce expenses.
4. **Improve Operational Efficiency:**
    - Schedule energy-intensive tasks during off-peak hours based on energy consumption patterns.
    - Regularly evaluate and maintain equipment to reduce energy waste.
5. **Employee Training and Awareness:**
    - Conduct training sessions to educate employees on energy-saving practices.
    - Foster a culture of sustainability and accountability within the organization.
6. **Invest in Energy Storage Solutions:**
    - Deploy battery storage systems to manage peak loads and store excess energy during low-demand periods.
    - Use storage solutions to balance renewable energy variability.
7. **Compliance with Environmental Regulations:**
    - Regularly update energy usage practices to align with evolving environmental standards.
    - Leverage energy analytics to document compliance and gain certifications.

V. Implementation Plan/Strategy

Step 1: Data Loading and Preprocessing

- **Load Data:** Import the dataset using pandas.read_csv(). Ensure that the date column is correctly parsed.
- **Initial Inspection:** Check the data types and get a quick overview of the dataset with .dtypes and .describe().
- **Handle Missing Values:** Identify missing values with data.isna().
- **Feature Engineering:** Extract features such as day, month, year, and time from the date column using pd.to_datetime(). This allows us to capture temporal patterns.

Outcome: This step ensures the dataset is ready for further analysis and modeling by converting date information and handling any missing values.

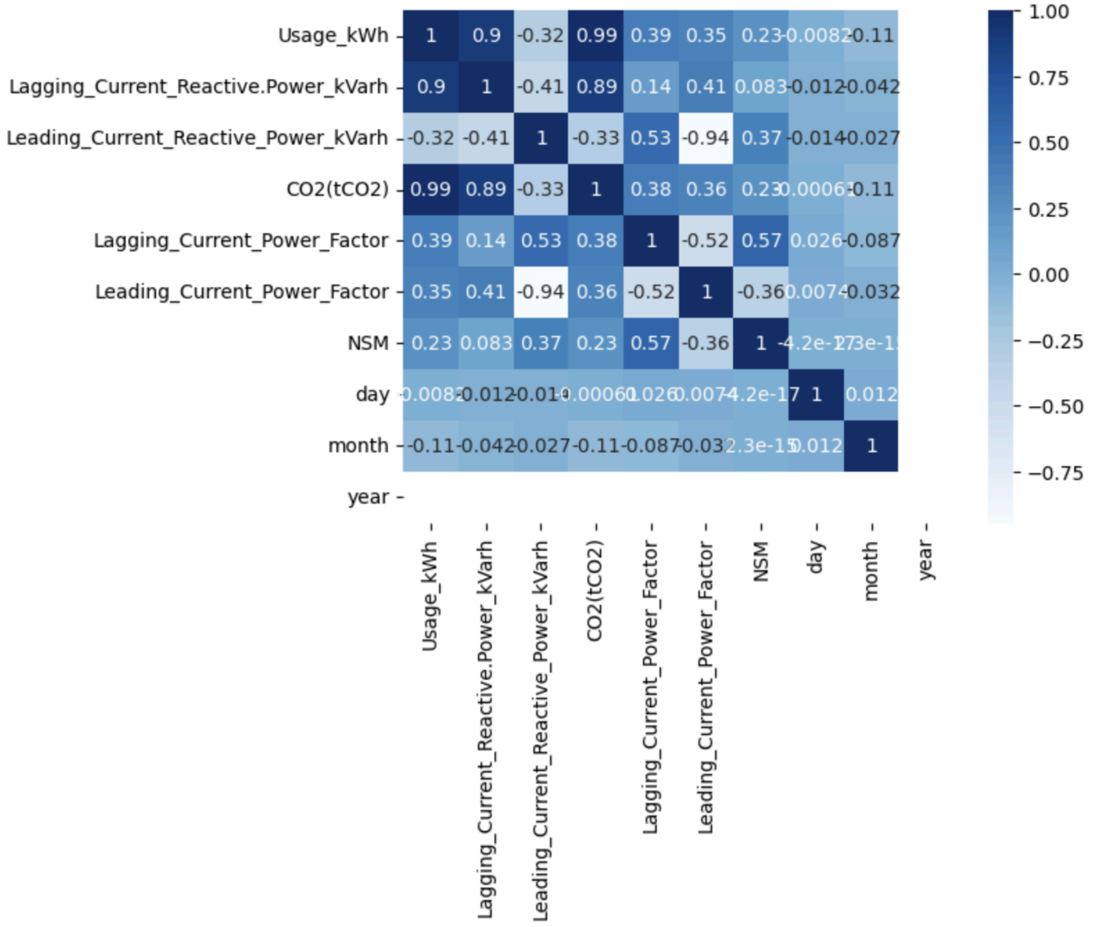
Step 2: Exploratory Data Analysis (EDA)

- **Summarize Usage Data:** Calculate and print basic statistics for Usage_kWh, including average, minimum, and maximum consumption.
- **Aggregate Usage Data by Time Periods:**
  - **Weekly Usage Patterns:** Calculate average energy consumption per week and visualize with bar plots.
  - **Daily Usage Patterns:** Group the data by the day of the week and analyze daily usage with bar charts and pie charts for total consumption by day.
  - **Seasonal Usage Patterns:** Segment the data by season (Winter, Spring, Summer, Autumn) and visualize total electricity consumption by season and load type.

Outcome: Gain insights into energy consumption patterns across time (daily, weekly, seasonal).

Step 3: Correlation Analysis

- **Correlation Heatmap:** Visualize correlations between numerical variables, particularly focusing on Usage_kWh and other metrics like CO2 emissions, Lagging Reactive Power, etc.
- **Scatter Plots:** Generate scatter plots to visually assess the relationships between energy consumption and other factors such as Lagging_Current_Reactive.Power_kVarh, CO2 emissions, and NSM.



Outcome: Identify variables strongly correlated with energy consumption, particularly those influencing energy usage like CO² emissions and reactive power.

Step 4: Feature Selection

- **Select Relevant Features:** Use SelectKBest with f_regression to select the top 3 features most predictive of energy consumption.
- **Standardize Data:** Scale features using StandardScaler to ensure consistency and improve model performance.

Outcome: Prepare a clean, standardized dataset with selected features that will be used for model training.

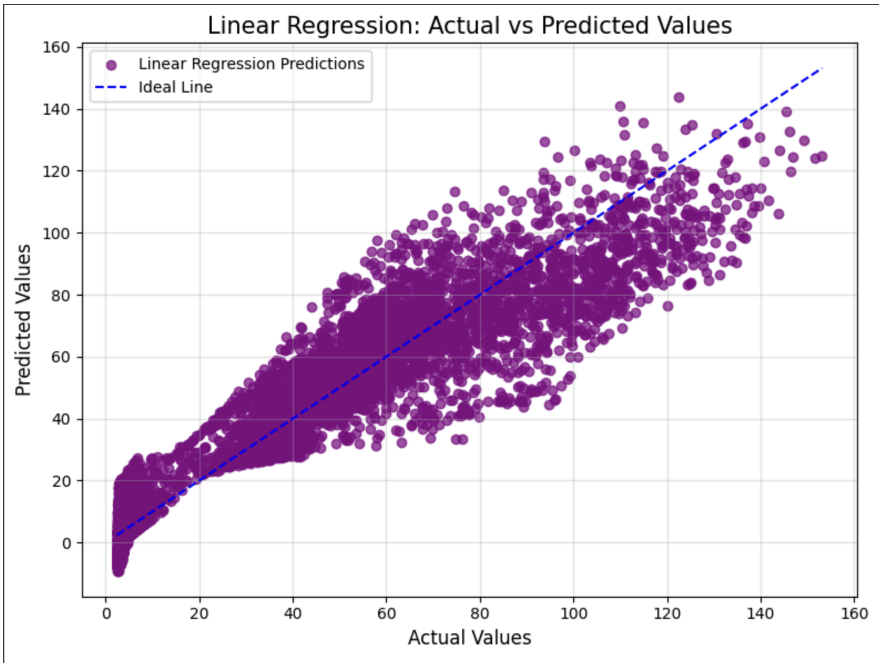
Step 5: Model Training

- **Model Selection:** Train and evaluate the following regression models:
    - Linear Regression
    - Ridge Regression
    - Lasso Regression
    - ElasticNet
    - Random Forest Regressor
- **Split Data:** Divide the dataset into training and testing sets using train_test_split.
- **Fit Models:** Train the models on the scaled training data.

Outcome: Establish a baseline performance using different models for energy consumption prediction.

Step 6: Model Evaluation

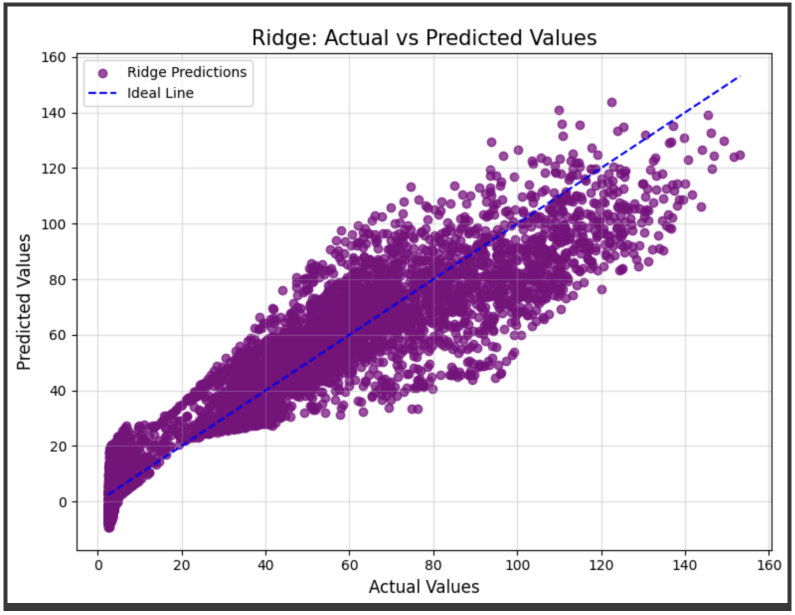
For Linear Regression :



Performance Metrics :

MSE: 103.9892
MAE: 7.1535
RMSE: 10.1975
R2: 0.9073
Adjusted R2: 0.9073

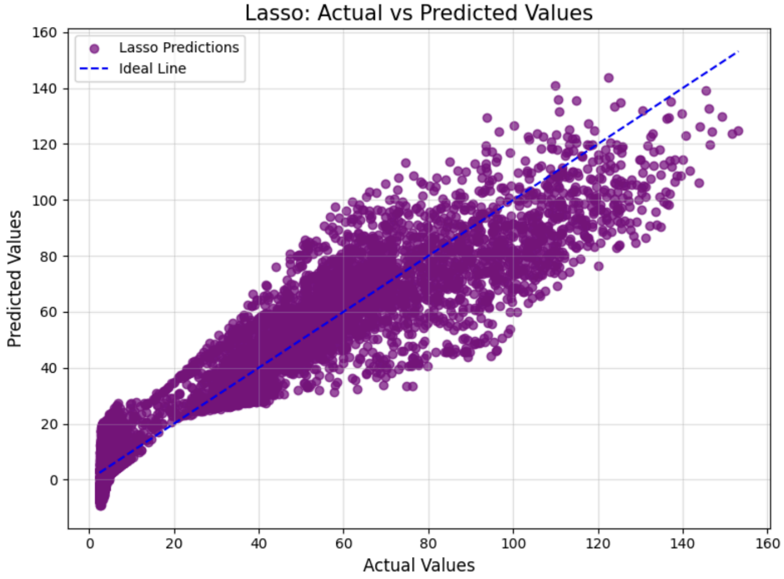
For Ridge Regression :



Performance Metrics :

MSE: 103.9890
MAE: 7.1533
RMSE: 10.1975
R2: 0.9073
Adjusted R2: 0.9073

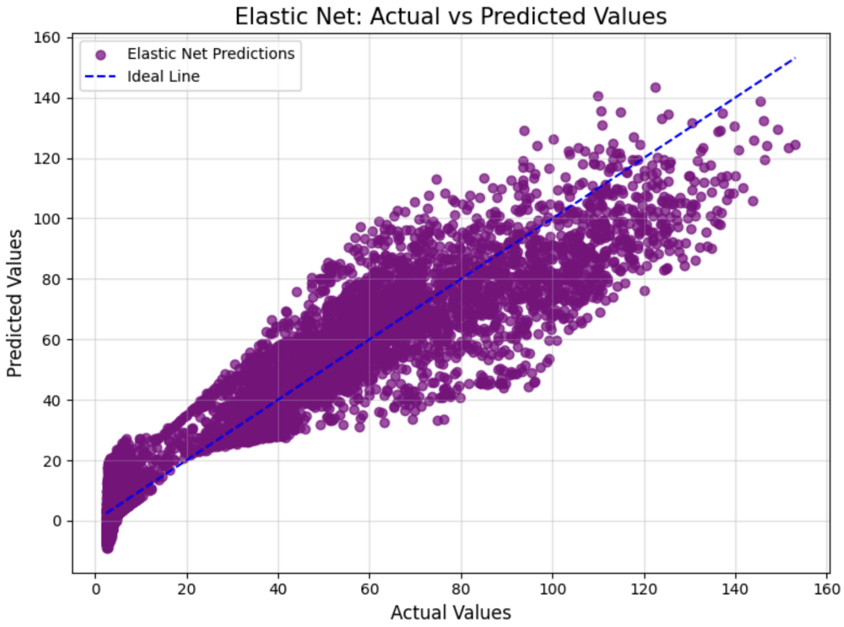
For Lasso Regression:



Performance Metrics :

MSE: 103.9958
MAE: 7.1487
RMSE: 10.1978
R2: 0.9073
Adjusted R2: 0.9073

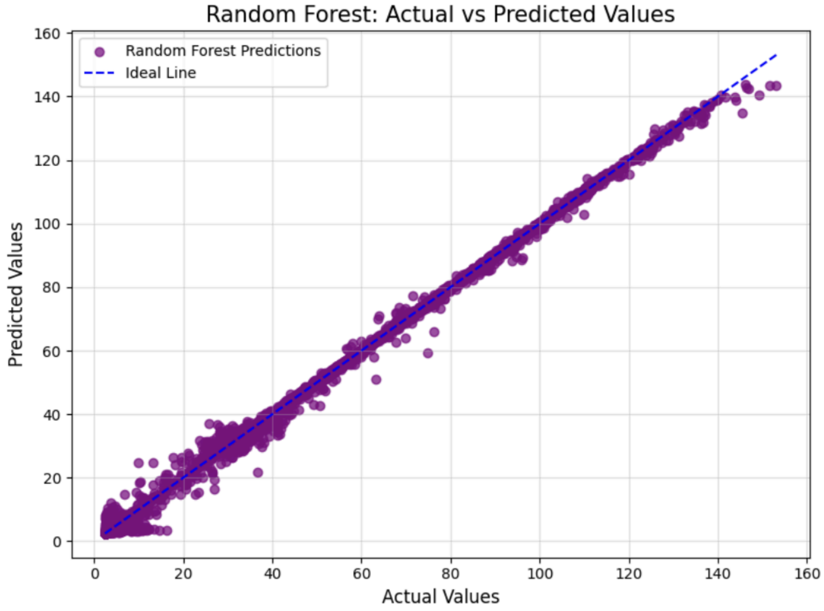
For Elastic Net :



Performance Metrics :

MSE: 103.9987
MAE: 7.1270
RMSE: 10.1980
R2: 0.9073
Adjusted R2: 0.9073

For Random Forest :



Performance Metrics :

MSE: 1.2054
MAE: 0.4572
RMSE: 1.0979
R2: 0.9989
Adjusted R2: 0.9989

- **Evaluate Performance Metrics:** For each model, calculate the following metrics:
  - **MSE (Mean Squared Error)**
  - **MAE (Mean Absolute Error)**
  - **RMSE (Root Mean Squared Error)**
  - **R2 and Adjusted R2**
- **Plot Predictions vs. Actual Values:** Visualize the predicted vs. actual energy usage values for each model to assess how well each model performs.

Outcome: Understand which model is best suited for predicting energy consumption based on performance metrics.

Step 7: Model Comparison and Final Selection

- **Compare Models:** Review the results for each model, comparing R2 scores, MSE, MAE, and RMSE. Pay particular attention to the Random Forest model, which is expected to perform better due to its ability to handle complex patterns.
- **Select Final Model:** Based on evaluation metrics, choose the best-performing model for deployment (likely Random Forest due to its superior R2 score).

Outcome: Select the optimal model to predict energy consumption, which will likely be Random Forest given its superior performance.

Step 8: Conclusion and Future Work

- **Insights:** Conclude the findings, particularly highlighting the impact of variables like CO^2 emissions and lagging reactive power on energy consumption.
- **Model Improvement:** Suggest potential ways to improve the model, such as hyperparameter tuning for the Random Forest model or incorporating additional features.

Outcome: A complete, effective model for predicting energy consumption in the steel industry, with clear next steps for optimization.

VI. Follow Up & Evaluation Plan

1. **Implementation and Training**
   - Collaborate with pilot facilities to integrate the platform and train staff for effective use.
2. **Performance Monitoring**
   - Track key metrics such as energy savings, cost reductions, and carbon footprint impact.
3. **Customer Support and Feedback**
   - Provide ongoing support, deliver quarterly reports, and collect user feedback for improvements.
4. **Scalability Testing**
   - Expand to additional facilities, evaluating adaptability across various operations.
5. **Impact Assessment**
   - Periodically measure long-term results and refine the platform to maximize efficiency and value.

VII. References

CSE464 Classroom Notes 1-10.
Dataset as "`Steel_industry_data.csv`":
Steel Industry Energy Consumption - UCI Machine Learning Repository

Colab Link :
https://colab.research.google.com/drive/15Y7Dnk2gnWEI7HMhc0OpfQ3zaBddC8rV?usp=sharing