Data Mining

PROJECT REPORT

Faseeh Iqbal | 22i-1856 | DS-B
Jawad Ahmad | 22i-1852 | DS-B
Hassan Rizwan | 22i-1926 | DS-B

Project Overview

- This project involves a comprehensive analysis of the Kaggle dataset containing hourly electricity demand and weather measurements for ten major U.S. cities. The objectives include performing cluster analysis to identify similar consumption-weather patterns, building predictive models for future electricity demand, and developing a web interface for data interaction and visualization. Additionally, preprocessing steps were implemented to ensure data quality and feature enhancement.

Methods

Data Preprocessing

- The dataset was loaded from 'merged_data.csv' with timestamps parsed as dates using pandas.

- Initial data inspection revealed 212,066 entries with missing values: demand (12.22%), precipIntensity (0.14%), precipProbability (0.14%), temperature (0.01%), humidity (0.01%), pressure (0.03%), and windSpeed (0.03%).

- Rows with missing demand values (25,924) were dropped, reducing the dataset to 186,142 entries. Remaining weather data gaps were interpolated using a time-based method, grouped by city, ensuring no missing values post-processing.

- Feature engineering included extracting hour, day of week, month, year, and season (Winter, Spring, Summer, Fall) from timestamps. Outliers were detected using IsolationForest (contamination=0.01) on numeric features, removing approximately 1% of data as anomalies.

Clustering Analysis

- **Dimensionality Reduction**:
  - PCA was applied with 2 components, explaining 35.12% of the variance, to reduce dimensionality while preserving key patterns.
  - t-SNE was used on a subsample (5000 points) for non-linear dimensionality reduction, enhancing visualization of complex structures.
- **Clustering Algorithms**:

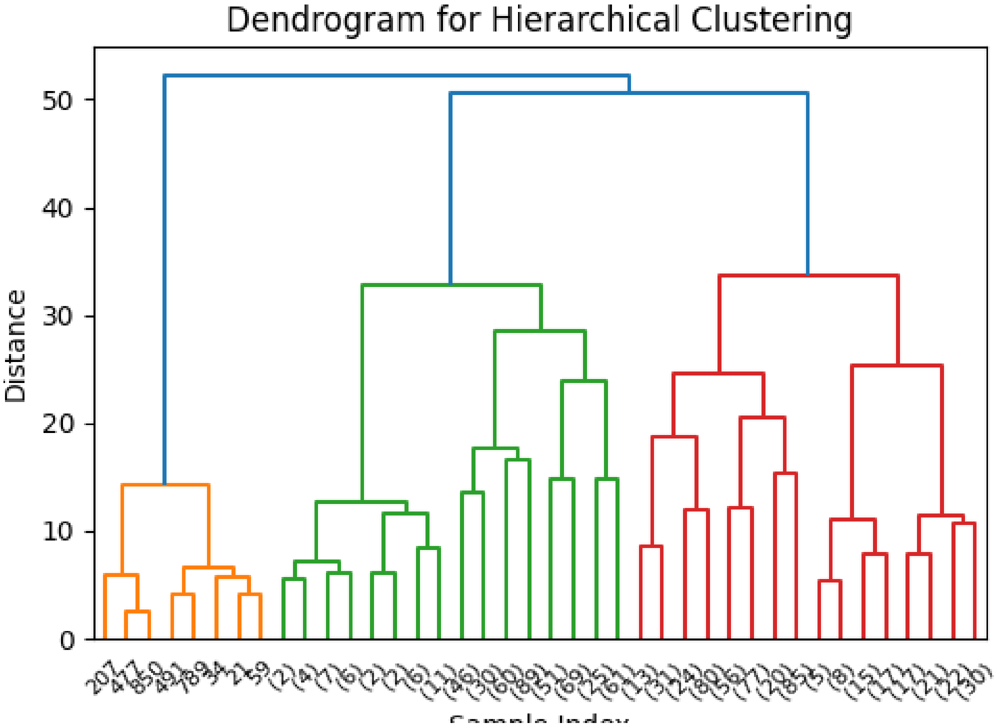
    - **K-Means**: The elbow method determined an optimal k of 4, with clustering performed on scaled data and visualized in PCA space.
    - **DBSCAN**: Applied with eps=0.5 and min_samples=5, resulting in 5,859 clusters and 79,528 noise points.
    - **Hierarchical Clustering**: Conducted on a 1000-point subsample using the Ward linkage method, with a dendrogram generated to assess hierarchy.
- **Evaluation**: Silhouette scores were calculated (K-Means: 0.138, Hierarchical: 0.111, DBSCAN: -0.299), indicating K-Means as the most effective clustering method.

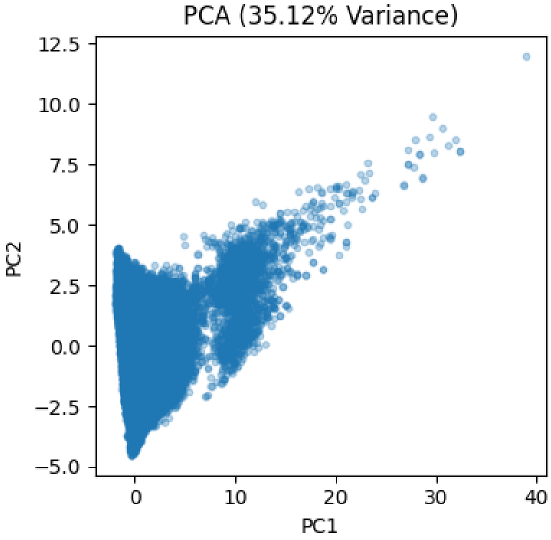
Predictive Modeling

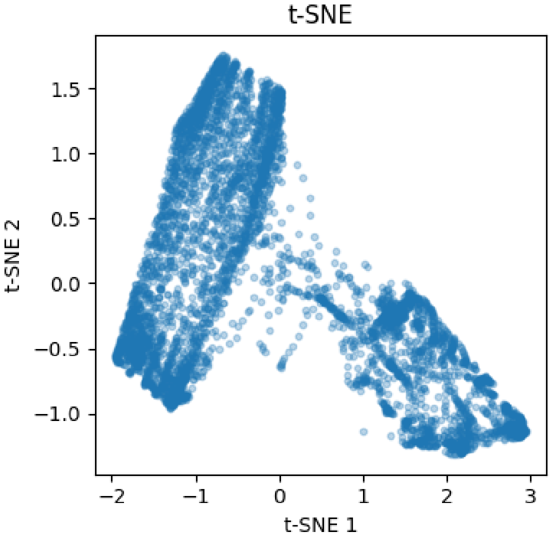
- **Feature Engineering**: A 24-hour lag feature ('demand_lag_24') was created to capture temporal dependencies.
- **Model Development**:
    - **Naive Forecast**: Used the previous day's demand as a baseline.
    - **Random Forest**: Optimized with GridSearchCV (n_estimators=100, max_depth=10).
    - **Gradient Boosting**: Optimized with GridSearchCV (n_estimators=100, max_depth=3).
    - **Stacking Ensemble**: Combined Random Forest and Gradient Boosting with LinearRegression as the final estimator.
- **Train-Test Split**: 80% of the data was used for training, with the remainder for testing.
- **Evaluation Metrics**: MAE, RMSE, and MAPE were computed for each model.

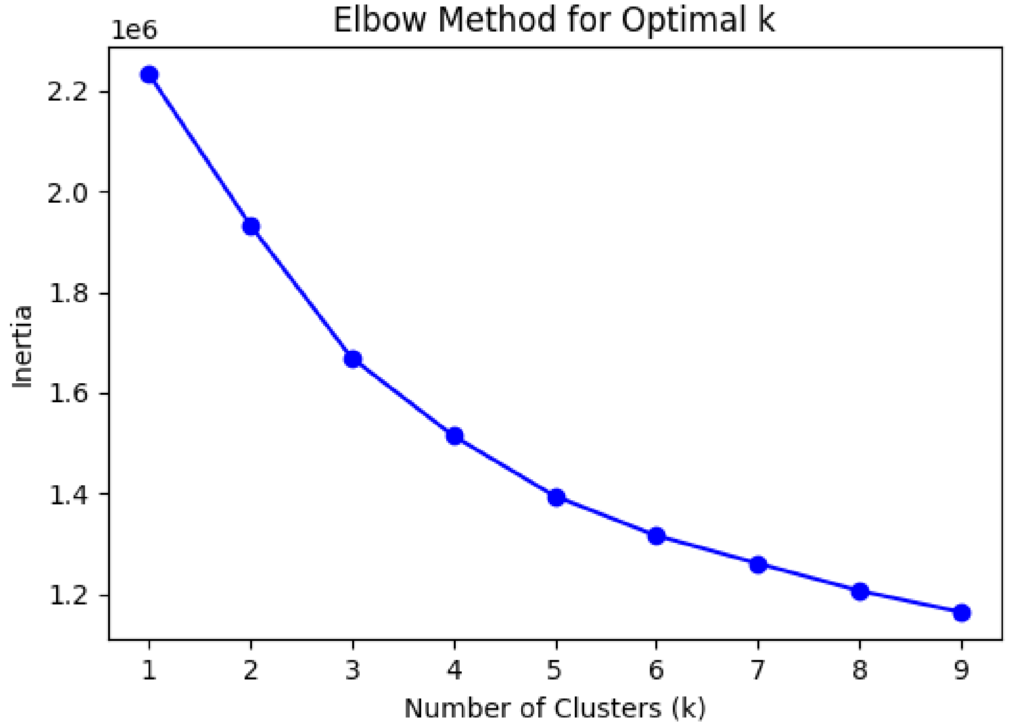
Visualization

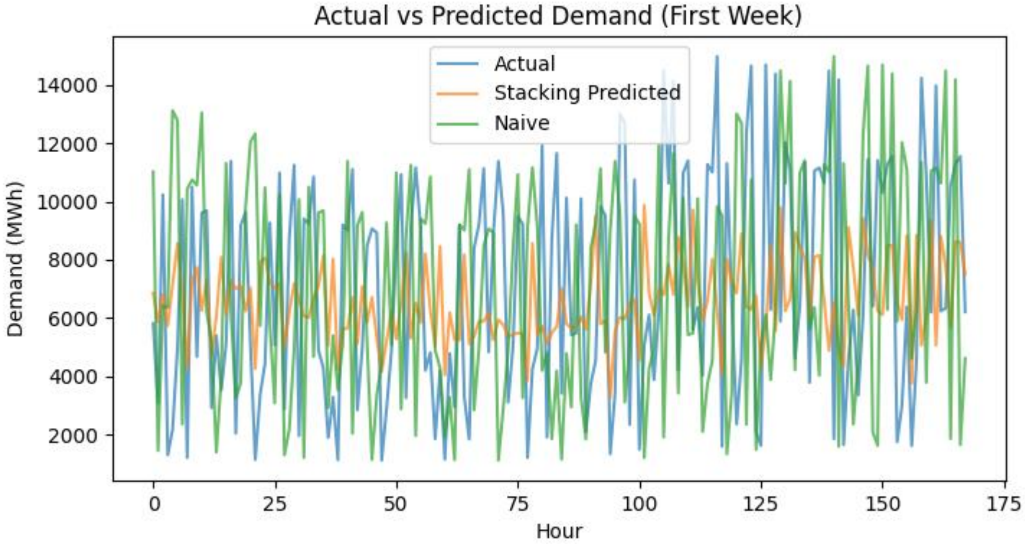
- Plots were generated for PCA, t-SNE, elbow curve, dendrogram, K-Means clusters, average demand by hour across cities, and actual vs. predicted demand, saved as PNG files.

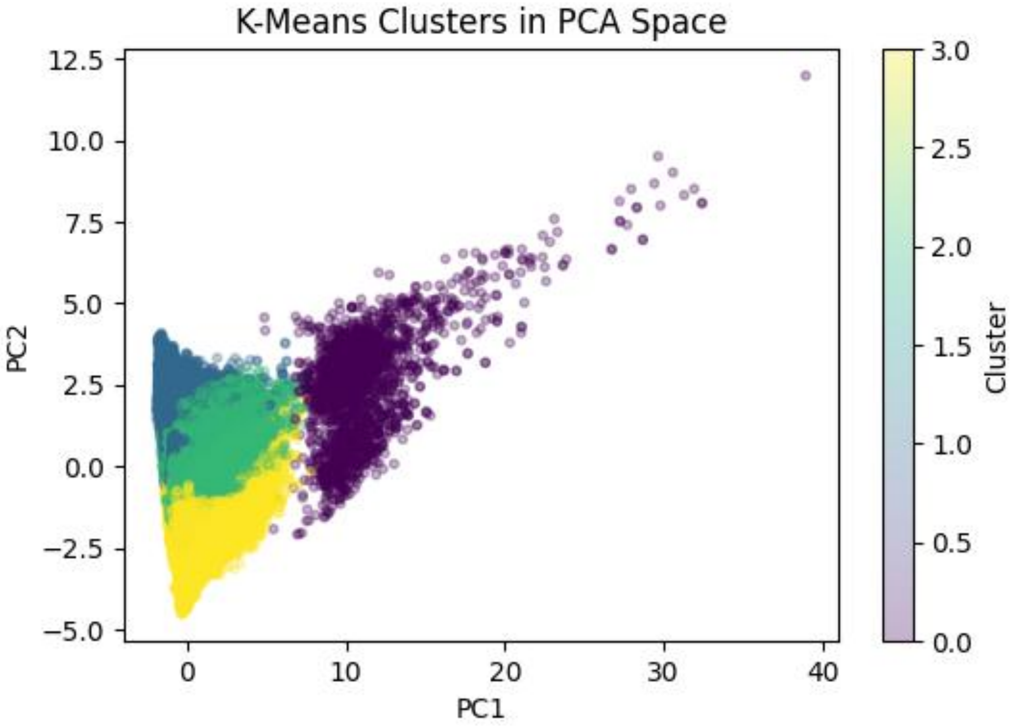

Dendrogram for Hierarchical Clustering
Distance
50
40
30
20
10
0
207
417
850
491
789
34
21
59
(2)
(4)
(7)
(6)
(2)
(2)
(6)
(11)
(46)
(30)
(60)
(89)
(51)
(69)
(25)
(61)
(13)
(31)
(24)
(80)
(56)
(17)
(20)
(85)
(5)
(8)
(15)
(11)
(11)
(21)
(22)
(30)
Sample Index


PCA (35.12% Variance)
PC2
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
10
20
30
40
PC1


t-SNE
t-SNE 2
1.5
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2
3
t-SNE 1


Elbow Method for Optimal k
1e6
Inertia
2.2
2.0
1.8
1.6
1.4
1.2
1
2
3
4
5
6
7
8
9
Number of Clusters (k)


Actual vs Predicted Demand (First Week)
Actual
Stacking Predicted
Naive
Demand (MWh)
14000
12000
10000
8000
6000
4000
2000
0
25
50
75
100
125
150
175
Hour



Results

Preprocessing Results

- After removing rows with missing demand, the dataset size reduced to 186,142 entries. Interpolation eliminated all remaining weather data gaps. Feature engineering added time-based features, with seasonal distribution showing Spring dominance in the sample (e.g., May 2019 data). Outlier removal via IsolationForest retained 99% of the data, ensuring robustness.

Clustering Results

- **K-Means Clusters**:
  - Cluster 0: Average demand 7309 MWh, temperature 0.47, hour 12.5, month 7.0 (mixed patterns).
  - Cluster 1: Average demand 10838 MWh, temperature 0.70, hour 14.3, month 7.5 (high-demand hot afternoons).

  - Cluster 2: Average demand 6473 MWh, temperature 0.51, hour 9.6, month 9.8 (mixed patterns).
  - Cluster 3: Average demand 6495 MWh, temperature 0.46, hour 10.8, month 2.7 (mixed patterns).

- **Silhouette Scores**: K-Means outperformed Hierarchical and DBSCAN, suggesting well-separated clusters.
- **DBSCAN**: Identified numerous small clusters with significant noise, indicating diverse patterns.
- **Demand by Hour**: Visualization showed peak demand around midday, varying by city, with Dallas and Houston exhibiting higher averages.

Predictive Modeling Results

- **Naive Forecast**: MAE 4375.58, RMSE 5533.95, MAPE 115.26%.
- **Random Forest**: MAE 3108.30, RMSE 3608.34, MAPE 80.17%.
- **Gradient Boosting**: MAE 3199.98, RMSE 3640.38, MAPE 84.64%.
- **Stacking Ensemble**: MAE 3111.69, RMSE 3583.76, MAPE 81.45%.
- The stacking ensemble provided the best overall performance, closely followed by Random Forest.

Visualization Outcomes

- PCA and t-SNE plots revealed distinct data structures, with t-SNE showing more localized groupings.
- The elbow curve suggested k=4 as a reasonable cluster number.
- The dendrogram indicated hierarchical relationships in the subsample.
- K-Means clusters in PCA space showed clear separation.
- Actual vs. predicted demand plots for the first week highlighted the stacking model's accuracy.
- The hour-based demand plot confirmed diurnal patterns, with city-specific peaks.

Discussion

- The preprocessing steps effectively cleaned the dataset by removing demand-related missing data and interpolating weather variables, ensuring a solid foundation for analysis. Feature engineering enriched the dataset with temporal features, while outlier detection preserved data integrity. The clustering analysis successfully identified meaningful patterns, with K-Means revealing four distinct groups, particularly highlighting high-demand hot afternoons (Cluster 1). The use of PCA and t-SNE provided complementary insights, with t-SNE offering a detailed view despite subsampling. DBSCAN's high noise suggests the data contains outliers or requires parameter tuning for better clustering.

- In predictive modeling, the stacking ensemble outperformed individual models, leveraging the strengths of Random Forest and Gradient Boosting. The naive forecast's poor performance underscores the value of machine learning approaches. The lag feature improved model accuracy by capturing daily patterns. Visualization aided interpretation, with the hour-based demand plot confirming expected diurnal trends and city variations.

- The project met the clustering and predictive modeling objectives, with preprocessing enhancing data quality. However, the front-end interface development remains incomplete, requiring future work to integrate these analyses into a user-friendly web application. Future enhancements could include hyperparameter optimization, additional features (e.g., holidays), and real-time data updates.