CAPSTONE PROJECT PROPOSAL

Real-Time Energy Market Intelligence & Anomaly Detection System for the Texas ERCOT Grid

Prepared by: Alina Hasan & Sai Dinesh
Program: M.S. in Data Science

1. Introduction

Texas operates one of the most unique and complex electric grids in the United States through the Electric Reliability Council of Texas (ERCOT). Unlike other states, Texas runs an independent deregulated electricity market, which means:

- Prices are highly sensitive to weather and supply-demand imbalances
- Renewable energy penetration (especially wind and solar) is among the highest in the country
- Extreme weather events (heat waves, storms, polar vortex events) create rapid fluctuations
- Real-time grid reliability depends heavily on accurate short-term forecasting

These characteristics make Texas a high-impact but challenging region for developing real-time energy forecasting and anomaly detection systems.

This capstone project proposes building a Real-Time Energy Market Intelligence Platform specifically for the ERCOT Texas grid, combining live data ingestion, forecasting, anomaly detection, and dashboard visualization.

2. Problem Statement

ERCOT's operational environment presents several challenges: frequent heat waves, winter storms, sudden renewable generation drops, and a fully deregulated wholesale market where small changes can lead to significant price volatility. Current analytical tools depend on offline or historical data and struggle to provide real-time intelligence during critical grid events. As a result, grid operators, analysts, and market participants lack the ability to anticipate short-term stress conditions or identify abnormal behavior as it occurs. This creates operational inefficiencies, delayed decision-making, and heightened financial risk. Therefore, there is a clear need for a system that integrates live ERCOT grid data with Texas-specific weather information to deliver short-term demand forecasts, detect anomalies, and support proactive decision-making.

3. Motivation & Background

Texas offers a uniquely compelling case for this project due to:

- Independent Grid: ERCOT is isolated from the Eastern and Western interconnections. This means Texas must manage almost all supply-demand imbalances internally.
- High Renewable Penetration: Texas leads the nation in wind generation and is rapidly expanding solar capacity, increasing variability.
- Extreme Weather Events: Events like Winter Storm Uri have demonstrated the importance of real-time situational awareness.
- Market Volatility: ERCOT's deregulated market results in high price sensitivity to demand, outages, and weather.

4. Project Objectives

The main objectives of this capstone are:

- Develop a real-time ingestion pipeline for ERCOT grid and Texas weather data
- Build short-term demand forecasting models (1-hour and 24-hour horizons)
- Implement anomaly detection models to identify unusual grid behavior
- Quantify potential operational or financial risks associated with detected anomalies
- Develop an interactive dashboard to visualize load forecasts, anomalies, and regional grid stress
- Implement MLOps-ready components such as containerization and modular pipelines

5. System Architecture

The system will be organized into three layers:

A. Data Pipeline (Backend)

The backend is responsible for continuously ingesting, streaming, and storing real-time data from both the ERCOT grid and Texas weather sources.

1. Ingestion Layer
   The ingestion layer consists of Python-based micro-services running in Docker containers, each dedicated to a specific data source:

   - `ingest-grid-service`: Polls the ERCOT API at regular intervals to retrieve systemwide load, wholesale price (LMP), and transmission outage information.

   - `ingest-weather-service`: Polls NOAA / OpenMeteo APIs to collect weather variables relevant to Texas, including temperature, wind speed, and solar irradiance.

2. Messaging Layer
   All incoming data streams are routed through Apache Kafka, which acts as the central messaging backbone of the system. Kafka enables reliable, high-throughput, real-time data streaming between ingestion services, storage, and downstream modeling components, ensuring that the system can scale and respond to live grid conditions.

3. Storage Layer
   Processed and raw time-series data are stored in TimescaleDB (built on PostgreSQL). TimescaleDB is chosen for its native support for time-series workloads, efficient querying, and ability to retain historical data needed for model training, validation, and analysis.

B. Data Science Core (Modeling)

The modeling layer transforms the incoming time-series data into forecasts and anomaly signals.

1. Preprocessing and Feature Engineering
   Incoming data streams are cleaned and aligned through:

   - Handling missing or delayed values

   - Aligning timestamps and time zones

   - Constructing engineered features such as lag variables, rolling window statistics, and calendar-based features (hour of day, day of week, seasonality indicators)

2. Model 1: Demand Forecasting (Regression)
   The first modeling component focuses on short-term demand forecasting:

   - Goal: Predict grid load at 1-hour and 24-hour horizons.

   - Candidate Models:

   - XGBoost regression as an initial baseline model

- LSTM/GRU-based deep learning models to capture complex temporal dependencies and non-linear patterns in ERCOT demand and weather signals.

3. Model 2: Anomaly Detection (Unsupervised)
   The second modeling component focuses on detecting abnormal grid conditions:

- Goal: Identify "abnormal" states such as situations where prices are unusually high given the observed demand, or where grid behavior deviates significantly from historical norms.

- Candidate Algorithms:

- Isolation Forest

- One-Class SVM
  These models compute an anomaly score in real-time, which is later used to trigger alerts in the dashboard.

C. Presentation Layer (Frontend)

The frontend presents real-time insights in a clear, intuitive format for operators, analysts, and other stakeholders.

1. A web-based dashboard will be implemented using Streamlit or React.

2. Key visual components include:

- Live Load vs. Predicted Load plots to show model performance and current demand trajectory.

- Highlighting of regions experiencing potential grid stress or abnormal conditions on a map.

- Real-time alerts that are triggered when anomaly scores exceed predefined thresholds, signaling possible reliability or financial risk.

6. Data Sources (Texas-Focused)

Source	Texas-Specific Metric	Frequency	Purpose
ERCOT API	Texas System-Wide Load	Every 5 minutes	Forecast target variable
ERCOT API	Texas Wholesale LMP Prices	Every 15 minutes	Market & financial insights
ERCOT API	Texas Transmission Outages	Hourly/Daily	Grid stability indicators
NOAA/ OpenMeteo	Texas Temperature, Wind, Irradiance	Hourly	Weather + renewable drivers

7. Methodology

- Collect and ingest real-time grid and weather data
- Perform exploratory analysis to identify key trends
- Engineer features relevant to forecasting and anomaly detection
- Train and validate models using historical + real-time data
- Deploy models within a modular pipeline
- Build dashboard for real-time visualization
- Test forecasting accuracy and anomaly alert responsiveness

8. Expected Outcomes

- A real-time ERCOT data pipeline
- Forecasting models capable of accurate short-term predictions
- Anomaly detection engine highlighting unusual grid behavior
- Interactive dashboard for stakeholders
- A complete end-to-end data science system replicating industry workflow

11. Project Timeline (15 Weeks)

Phase	Weeks	Deliverables
Planning & Setup	Weeks 1–3	Architecture design, pipeline setup
Data Collection & EDA	Weeks 4–6	Real-time data collection, analysis
Modeling	Weeks 7–10	Forecasting + anomaly detection models
Dashboard & Integration	Weeks 11–13	Frontend + backend integration
Finalization	Weeks 14–15	Testing, final report, presentation

15. References

- ERCOT – Electric Reliability Council of Texas.
  "ERCOT Grid and Market Data." Available at: https://www.ercot.com
  (*Primary source for Texas load, price, and outage data.*)

- Hochreiter, S., & Schmidhuber, J.
  "Long Short-Term Memory." *Neural Computation*, 9(8), 1735–1780, 1997.
  (*Foundational paper for LSTM models used in demand forecasting.*)

- Liu, F. T., Ting, K. M., & Zhou, Z.-H.
  "Isolation Forest." *2008 IEEE International Conference on Data Mining*, pp. 413–422. (*Core algorithm for anomaly detection in your project.*)