Capstone Proposal: Predictive Electricity Theft Detection

Business Understanding

Problem & Motivation: Electricity theft costs global utilities $96 billion annually. In emerging markets like Kenya, losses reach 20-40% of revenue, inflating tariffs for honest customers and destabilizing grids. Our team is motivated to apply data science to combat this infrastructure corruption, making essential services more affordable and reliable.

Industry & Audience: The project targets electricity distribution utilities in emerging markets (e.g., Kenya Power, Eskom) and energy regulators.

Impact & Novelty: If deployed, our AI system targets a 30-50% reduction in theft losses. While 47+ academic papers have used our core dataset for load forecasting, none have applied it to electricity theft detection. We bridge this gap by combining this real consumption data with established IEEE (Institute of Electrical and Electronics Engineers) research on synthetic theft patterns.

Data Understanding & Preparation

Primary Dataset: UCI Electricity Load Diagrams (2011-2014), https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams . This provides clean, 15-minute interval consumption data for 370 Portuguese clients over 4 years (approx. 540,570 daily records).

Data Creation & Features: Since real theft labels are unavailable, we will inject realistic synthetic theft into 5% of customers based on IEEE-defined patterns (e.g., meter tampering, cable bypass). We will engineer a targeted set of 25 predictive features from the time-series data, including:

- **Consumption Patterns:** Daily variability (`daily_std`), peak-to-average ratio.
- **Anomaly Metrics:** Statistical Z-scores (`z_score_30d`), sudden consumption drop flags.
- **Theft-Specific Signals:** Benford's Law violation scores- a targeted forensic tool that exploits a universal statistical property of natural numerical data to expose potential tampering, non-linearity checks.
- **Temporal & Comparative:** Weather-normalized consumption, deviation from similar customers.

Preprocessing & Challenges: Key steps include handling European decimal formats, aggregating to daily data, and creating realistic synthetic labels. The main challenge is ensuring the synthetic theft patterns are diverse and credible.

Modeling, Evaluation & Deployment

Modeling Approach: This is a **binary classification** problem with severe class imbalance (5% theft). We will progress from a simple, interpretable baseline (Logistic Regression) to advanced, tree-based ensembles (XGBoost, Random Forest) capable of handling non-linear patterns and imbalance.

Success Metrics: Our primary metric is the **F2-Score**, which emphasizes recall to catch more thieves, with a target > 0.70. We will also track Precision@K and estimate potential annual savings (KES billions).

Deployment Plan: The final model will be deployed via an interactive **Streamlit dashboard**. It will allow utility investigators to view customer risk scores, investigate flagged accounts with explainable AI insights, and simulate the cost-benefit of different intervention strategies.

3-Week Execution Plan

Week	Focus	Key Activities	Deliverable
1	**Foundation & MVP**	Data acquisition, cleaning, injecting basic theft patterns, engineering 10 core features.	A working data pipeline, baseline model (F2 > 0.50), and initial EDA.
2	**Advanced Modeling**	Engineering the full 25+ feature set, model experimentation (XGBoost, ensembles), hyperparameter tuning.	An optimized model achieving F2 > 0.65 with a validated feature set.
3	**Deployment & Polish**	Building the Streamlit dashboard, finalizing business impact analysis, and preparing the presentation.	A presentation-ready project with a live demo dashboard and a model target of F2 > 0.70.

Team & Risk Mitigation

Team Roles (6 Members):

1. Data Engineer (pipeline/database)
2. Feature Specialist (engineering/research)
3. ML Engineer (model development)
4. Validation Lead (metrics/statistics)
5. Dashboard Developer (Streamlit app)
6. Project Manager (coordination/business case)

Key Risks & Mitigation:

- **Class Imbalance:** Use F2-score, class weighting, and anomaly detection techniques.
- **Unrealistic Synthetic Data:** Base patterns strictly on IEEE research and maintain a realistic 5% theft rate.
- **Scope Creep:** Prioritize the core 25-feature model and a functional dashboard over extra complexity.