भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

# Symposium on Business Analytics and Artificial Intelligence - 2026

# BOOK OF ABSTRACTS

# LIST OF PROJECTS

**AI-Based Root Cause Analyzer (RCA) for Predictive Process Disturbance in Refinery Chiller Systems**

**Abstract:** In industrial process plants, particularly in refinery refrigeration systems, alarms are triggered when process variables deviate from safe operating thresholds. However, operators often face difficulty in identifying the *true root cause* of disturbances due to multiple simultaneous influencing factors, leading to delayed response, alarm flooding, and potential production losses. Specifically, for the Low-Level Chiller (LIC1072), crossing the PVLO threshold (28.75) can result in refrigeration failure, hydrate formation, and reduced NGL recovery, significantly impacting plant efficiency.

This project develops an AI-driven Root Cause Analyzer (RCA) that predicts and ranks the most probable causes of process disturbances in real time. The system applies a multi-class probabilistic model along with unsupervised techniques, such as Principal Component Analysis (PCA), to address overlapping root causes and determine their relative contributions. The model trains on three years of high-frequency industrial time-series data, consisting of approximately 1.05 million records at one-minute intervals, sourced from DCS and historian systems. Development utilizes Python (Pandas, Google Colab) and follows a structured pipeline.

For data engineering we filtered steady-state regions, segmented disturbance patterns, and applied correlation-based feature reduction to eliminate multicollinearity. The machine learning approach includes supervised learning, specifically multi-class multi label classification and regression, for root cause identification. Unsupervised learning uses Principal Component Analysis (PCA) for pattern discovery and dimensionality reduction. Deep learning explores artificial neural network (ANN) and recurrent neural network (RNN) architectures to capture complex temporal patterns. The project aims to predict root causes with at least 70% accuracy, output a prioritized list of simultaneous root causes, and establish a proactive warning window of about three minutes before operator intervention becomes necessary.

The RCA system transforms plant operations from reactive to proactive alarm management. By improving root cause visibility, operator decision-making, and situational awareness, this solution reduces production disruptions, mitigates alarm fatigue, enhances overall plant safety, and provides a scalable platform for AI-driven process optimization in industrial automation.

**AI-Driven Crop Yield Prediction for Field Corn Using Phase-Aware Vegetation Index Engineering and Machine Learning**

**Abstract:** Accurate crop yield prediction in field corn (Zea mays) remains a significant challenge due to dynamic environmental interactions, temporal crop variability, and the presence of redundant spectral features in remote sensing datasets. This project aimed to develop a robust AI-driven yield prediction framework for the Kharif 2024 season by optimizing vegetation index (VI) selection and implementing growth-stage-aware feature engineering.

To achieve this, 25 spectral vegetation indices were systematically evaluated using agronomic relevance, remote sensing literature, and redundancy analysis. Based on the evaluation, 12 agronomically relevant indices (5 Core and 7 Conditional) were selected, while redundant and crop-mismatched indices were excluded to improve model efficiency.

The 130-day corn growth cycle was segmented into four analytical phases (G1–G4) to capture critical physiological transitions related to canopy development, nitrogen uptake, water stress, and reproductive-stage sensitivity. Advanced feature engineering techniques, including inter-phase delta analysis, rate-of-change metrics, variability analysis, and phase specific aggregation, were implemented to capture temporal crop dynamics more effectively than conventional static parameter modeling.

The engineered dataset, consisting of satellite-derived spectral data from approximately 297 farms, was used to train and evaluate machine learning models including XGBoost and Random Forest using Python and Scikit-Learn. Dimensionality reduction and correlation analysis were applied to optimize feature selection and reduce noise within the predictive pipeline. The framework achieved a 52% reduction in feature redundancy and identified the Day 70–85 reproductive stage as the most yield-predictive satellite acquisition window for field corn.

The findings demonstrate that phase-aware vegetation index engineering significantly enhances predictive capability and model interpretability by capturing crop stress trajectories and growth-stage transitions. Beyond technical improvements, the proposed solution delivers strong socio-economic impact by enabling early stress detection, improving harvest planning, reducing input wastage, and minimizing financial risk for smallholder farmers. By transforming raw satellite data into actionable agricultural intelligence, the framework supports precision farming, sustainable resource management, stable food security, and long-term agricultural resilience.

**Airline Crew Scheduling: FAA-Compliant Crew Pairing & Rostering Optimization**

**Abstract:** Airline crew scheduling demands strict Federal Aviation Administration [FAA] Part 121 compliance — covering Flight Duty Period limits, rest periods, and role-qualification requirements for crew. Legacy rule-based systems yield suboptimal rosters, inflated costs, and crew fatigue, with no real-time what-if analysis or cost-versus-wellbeing trade off capability.

This research proposes an AI-Driven Crew Pairing and Rostering Optimization System that generates FAA-compliant, cost-efficient, and crew-welfare-sensitive duty rosters for cockpit crew. The system constructs legally valid flight-duty sequences and assigns them to qualified individuals through a multi-objective framework minimizing duty cost while considering workload imbalance, with cost-versus-wellbeing trade-off controls. The model initially assigns compliant duty start and end times with qualified PIC and SIC per sector, then extends to incorporate deadheading, hotel stays, crew positioning costs. It further integrates individual leave, holidays to produce fully personalized, operationally realistic rosters capable to scale further for commercial crew.

The optimization engine uses Google's OR-Tools CP-SAT [Constraint Programming – Satisfiability] a hybrid Constraint Programming and SAT-based solver for large-scale scheduling. Hard constraints enforce FAA Part 121 rules: Served via FastAPI, operations teams can retrieve optimized rosters, compliance audit trails through a REST interface. Performance is assessed across six indicators: Flight Coverage Rate (≥98% of ~325 flights assigned PIC and SIC); Solver Runtime (feasible solution within 10-15 mins for ~120 crew); Total Crew Duty Cost (minimized objective across all assignments); Crew Utilization (≤±15% duty-hour variance across the crew pool); Regulatory Compliance (100% hard-constraint satisfaction); and Solution Feasibility (valid complete roster generated for the full operational scenario).

For crew planners, the system eliminates manual constraint checking, reducing planning cycle time with a machine generated FAA compliance traces viz justification, violations. Operations controllers gain real-time what-if capability during irregular operations. Financially, optimized crew utilization reduces deadhead, overtime, and accommodation costs; utilization improvement on a ~300-flight operation translates to millions of dollars in annual cost avoidance

**Cost Optimization for Power Procurement and Scheduling**

**Abstract:** GRIDCO, the bulk power provider for Odisha, faces significant operational hurdles in bulk power management, primarily driven by a demand forecasting gap at the 15-minute block level. The lack of precise forecasting, combined with manual supply optimization complexity and limited cost-benefit visibility, makes it difficult to balance contracted generation, power surrenders, and open market procurement efficiently. This project aimed to bridge the forecasting gap and develop an automated decision-support system that optimizes the power procurement portfolio to minimize costs and maximize revenue impact.

This project utilizes a supervised time-series regression approach to predict continuous power demand values. After evaluating multiple architectures—including SARIMAX, Prophet, and LSTM, LightGBM (Light Gradient Boosting Machine) was selected as the finalized model due to its superior balance of predictive accuracy and computational efficiency. To address the end-to-end procurement cycle, the project is developing an optimization engine using linear/mixed integer programming for supply planning and a scenario simulation module for strategic financial modelling.

To achieve this, a solution was developed using Python and trained on four years of historical demand data (2021–2025) at a granularity of 96 data points per day (15-minute intervals). Key features integrated into the model include autocorrelation lags (Lag_1, Lag_96, etc.), weather signals (average temperature, precipitation), and holiday/weekend indicators. The project will also involve building a pre-market supply optimization engine to determine the most cost-effective supply mix and a post-market re-optimization module to minimize imbalance penalties. These modules will integrate real-time operational conditions and market results into the decision-making workflow.

As a result, the finalized LightGBM model achieved an average MAPE of 3.76% for next-day forecasts and 5.14% for next-7-day forecasts (based on ~6 months testing data). Furthermore, it demonstrated high industrial viability by reducing training time to 2–3 minutes, compared to 8 to 12 hours for statistical models. Ongoing development aims to quantify success through KPIs such as total procurement cost savings and revenue impact per simulated scenario.
These findings suggest that by bridging the forecasting gap and automating optimized supply planning, with high-accuracy and real-time demand insights enables GRIDCO to minimize power acquisition costs and maximize revenue.

**District-Level Retail Expansion for A leading eyewear retail chain: A Data-Driven Multi-Criteria Decision Framework Using AHP and Predictive Analytics**

**Abstract:** India's organized optical retail sector remains significantly underpenetrated in Tier II and Tier III regions, where expansion decisions are often driven by heuristic judgment rather than structured data analysis. A leading eyewear retail chain faces a critical challenge in systematically identifying high-potential districts for expansion across India, resulting in potential inefficiencies in capital allocation and missed growth opportunities.

This project aims to develop a scalable, data-driven decision framework to prioritize districts for retail expansion. The proposed solution leverages a Multi-Criteria Decision Analysis (MCDA) approach using the Analytic Hierarchy Process (AHP), integrating five key dimensions—Demographics, Economic Potential, Eye Health Demand, Retail Ecosystem, and Competition Intensity. Additionally, a gradient-boosted regression model is proposed to predict potential store revenue, enabling translation of composite district scores into actionable financial insights.

The methodology involves constructing a comprehensive dataset comprising multiple socioeconomic and market variables across all Indian districts, sourced from platforms such as NDAP and CMIE Consumer Pyramids. A Python-based analytical pipeline is proposed for data processing, scoring, and modelling. Competitive intensity will be quantified through district-level mapping approaches. The proposed framework is designed to achieve full district coverage and establish a structured scoring mechanism supported by predictive modelling. The expected output is a ranked and segmented list of districts, enabling prioritization of expansion opportunities based on quantified potential.

This framework demonstrates how a structured, analytics-driven approach can enhance the efficiency and effectiveness of retail expansion strategies. By enabling data-backed decision-making, the proposed methodology aims to support optimized capital deployment, reduced decision latency, and improved identification of high-return markets for a leading eyewear retail chain

**Effort-Based Forecasting and Right-Sizing: An AI-Driven Approach to Intelligent Staffing for Enterprise Support Operations**

**Abstract:** A global enterprise automation software company currently rely on volume and time-to-resolution (TTR) metrics to plan workforce capacity across three shifts spanning five global regions (AMER, EMEA, APAC, India, and Japan). This volume-driven approach fails to account for the true complexity of support tickets, process-level gaps, and the specialization requirements of individual engineers, which can lead to operational inefficiencies, increased workload variability, and service delivery challenges during periods of rapid growth or product release cycles

This project builds an effort-based forecasting system that transforms capacity planning from volume-driven to effort-driven, enabling precise, transparent, and continuously updated staffing recommendations per shift, geography, account tier, and specialization topic. Ticket complexity is quantified using a multi-signal clustering engine combining K-Means clustering on behavioural features such as reopen rate, escalation flags, email chain depth, and time at support. With NLP-derived embedding clusters on ticket descriptions and solutions, fractal dimensionality scoring LID-MLE. Incoming tickets are classified in real time against this model and assigned to Very Low, Low, Medium, or High complexity clusters.

Ticket volume is forecasted at the segment level using a Prophet and SARIMAX ensemble with hierarchical MinT reconciliation across all five regions, ensuring coherence between GEO level and total level forecasts. The full pipeline is trained on historical support ticket data using Python, Prophet, PyTorch Forecasting.

The system targets a forecast MAPE of ≤ 5% per segment, maintains engineer occupancy within the 75–85% optimal band, and ensures a backlog flow ratio (closures to arrivals) of ≥ 1.0. Complexity heatmaps expose which GEOs, tiers, and specializations are experiencing disproportionate high-complexity load, and right-sized staffing recommendations are generated per shift with full model-based transparency.

This solution directly addresses the capacity planning gap by replacing reactive, volume-based headcount decisions with proactive, effort-calibrated recommendations that account for ticket complexity, engineer specialization, and account growth trajectories thus enabling support the leadership to precisely staff where complexity rises, reduce SLA risk, and make resource allocation backed by data driven decisions.

**GenAI Powered Virtual SOC Analyst: Autonomous Security Alert Triage & Threat Intelligence**

**Abstract:** A Security Operations Centre (SOC) is a centralized team within an organization responsible for continuously monitoring and responding to cybersecurity threats. Millions of daily alerts overwhelm modern Security Operations Centres, causing analyst fatigue, slow response times, and high false-positive rates. Manual triage is infeasible at this scale and current solutions mostly work on keyword-based approach and single event tracking.

The project aims to build an agentic solution by combining volume anomaly detection, classification model, agent-based deep-dive, and mitigation planning. The system escalates when volume crosses a tolerance threshold, does severity classification, and the agent generates structured document with step-by-step remediation.

Team addressed the issue of data availability by using sample records to generate synthetic data for different security alerts. The team then used this data for anomaly detection and running the classification model. The team passed the model output and additional context from frameworks/guidelines to an agentic workflow created using the Langraph library in Python. Additionally, the team used pandas/scikit-learn/xgboost for data processing and classification. Compared to a business (Sigma) rule-based approach, the classification model reduces false negative rates by ~8-10 percentage points for specific threat type, strengthening overall security. False positive rates drop were ~30-40%, significantly lowering analyst fatigue by filtering out scenarios that basic rule-based systems cannot distinguish. Additionally, the agentic workflow automates alert handling and remediation.

The platform frees analysts from repetitive triage, dramatically reduces mean-time-to-detect and mean-time-to-respond, and scales across new client environments. It also provides proof of concept for an agentic SOC solution in cybersecurity and opens a pathway into managed SecOps.

**Intelligent In-Patient Estimate Calculator for Predicting Hospital Stay Costs**

**Abstract**: Manipal Hospitals face a recurring gap between initial treatment cost estimates and final billed amounts at discharge, with many cases exceeding the acceptable ±5% variance. This affects patient trust, financial planning, and operational predictability. The project proposes a system-driven Intelligent In-Patient Estimate Calculator that generates accurate cost forecasts at admission and continuously updates them during the patient's stay.

The solution integrates multi-source healthcare data, including electronic health records, historical billing patterns, diagnosis-related groups, length of stay, ICU utilization, treatments, and pharmacy consumption. It identifies key drivers of billing variance and uses two machine learning models at the unit and department levels to predict a dynamic cost range instead of a fixed estimate. This approach enables early identification of cases likely to exceed the defined variance threshold.

The system incorporates structured and unstructured data by extracting meaningful signals from clinical narratives such as past-history, history of present illness, diagnosis, and treatment progression. It uses BERT-based natural language processing to convert free-text clinical notes, prescriptions, and treatment records into standardized, billable entities, including medications, lab tests, and procedures. These derived features enhance the predictive accuracy of the models.

The solution delivers role-specific outputs across stakeholders. Patients receive transparent cost ranges, while Patient Care Advisors gain visibility into key cost drivers and receive alerts for potential escalations. Finance teams benefit from improved revenue predictability and early variance detection, and operations teams use forecasts to optimize ICU utilization and length of stay. Clinical leaders gain insights into cost-quality relationships, supporting value-based care decisions.

The system reduces billing variance, improves financial transparency, and strengthens resource planning. It also establishes a scalable foundation for predictive analytics and clinical decision support across the hospital.

**Intelligent Lead Scoring for a Solar Company**

**Abstract:** Freyr Energy, operating across both B2B and B2C segments in the solar rooftop space, faces acritical challenge: leads generated from different channels have lower-than-expected conversion to orders. Lower conversion rates lead to inefficient utilisation of their resources and higher customer acquisition costs. Freyr aims to increase this conversion rate with data analytics and AI.

This project aimed to transform the existing manual lead-handling process into a data-driven, optimized system to maximize conversions. To achieve this, a two-stage machine-learning framework for lead prioritisation was designed. In stage 1, leads are scored and ranked to guide tele callers on which prospects to engage first. In stage 2, a refined prioritisation model was designed to identify high-intent leads for the sales team, ensuring focused effort on prospects most likely to convert.

The ML model was built using 6 months of real business data (April 2026 to October 2026), which was cleaned, processed, and modelled using Logistic Regression & XGBoost in a Python-based environment. Custom features were created based on existing variables, and some variables were transformed. The model's effectiveness was evaluated using gain-based metrics, demonstrating strong predictive power. The models successfully captured 80-90% of high value leads within the top 5 deciles in both stages, significantly improving targeting precision.

By enabling structured prioritisation, the framework ensures that high potential leads are engaged at the right time and with priority, thereby considerably optimising efforts and avoiding missed opportunities. Beyond improving conversion rates, it empowers the business with actionable insights into key drivers for customer decisions, thereby unlocking new opportunities for market expansion.

**Optimizing Sales Performance Through Predictive Lead Analytics, Conversion Intelligence & Effort Efficiency**

**Abstract:** Sales teams at TeleCRM face significant challenges in prioritizing leads due to fragmented CRM data, inconsistent activity tracking across channels, and a severe class imbalance. To address these interconnected gaps, this project applies advanced machine learning and business analytics techniques to predict lead conversion likelihood, conversion speed, and expected time to conversion, while tackling six broader business challenges that collectively impede sales efficiency and revenue growth.

Data is sourced from a PostgreSQL database and processed through a Snowflake-based data lake. The methodology is adopted with rigorous data sanitation — including missing value treatment, removal of leaky features, and class imbalance handling — followed by multiple iterations of feature refinement in close collaboration with our SPOC and Project Guide. EDA conducted across calls, WhatsApp interactions, website visits, and status change logs confirmed that transactional activity variables were the strongest conversion predictors, with converted leads showing 6x greater call durations and 7x more WhatsApp attempts than non-converted leads.

The project employs a multi-model AI/ML architecture tailored to each business question. For Question 1, an XGBoost Classifier was developed and validated, outperforming Logistic Regression, Random Forest, and SVM. Unsupervised clustering and micro-classification models are being built for a more streamlined outcome. Questions 2 through 6 are currently in progress — preliminary modeling indicates Gradient Boosting and Random Forest Classifiers to identify fast-converting leads, regression models to estimate time-to-conversion, Logistic Regression and XGBoost to identify efficiency levers, and a clustering-based framework for lead-to-salesperson matching. Early findings suggest that responding within 10–30 minutes, making 3–5 multi-channel touchpoints, and sustaining calls above 60 seconds are key efficiency thresholds. Together, these six analytical modules aim to form a unified, data-driven sales intelligence system that transforms fragmented CRM activity data into measurable improvements in conversion rates, pipeline velocity, sales effort quality, and resource allocation — ensuring every lead is handled by the right representative at the right time.

**Predicting Delivery Dates & Continuous Adjustments to ETA**

**Abstract:** Getting a truck to arrive exactly when you expect it to in India is, frankly, a nightmare. The Ghats throw steep gradients at you, traffic behaves unpredictably, and cargo weights vary trip to trip - so the classic distance-divided-by-speed formula falls apart almost immediately. For Bosch, a late truck isn't just a scheduling headache. It sets off a chain reaction across production lines and forces warehouses to sit on expensive buffer stock just to hedge against uncertainty. This project started with a simple question: can we replace that guesswork with something that actually works?

Our approach was to understand the journey, not just track it. Rather than feeding the model raw GPS coordinates, we tried to capture the "physics" of what a truck actually experiences on the road. We engineered features like Average Road Gradient and a custom Volume-Terrain Pressure Index - essentially a way to quantify how hard a route is based on cargo weight and terrain steepness combined. One of our biggest wins was a City-level delay feature built from three months of historical data, which let the model learn the specific quirks and bottleneck patterns of each city on the network.

A turning point in the project came when we caught a subtle but important flaw: we'd been using straight-line (Haversine) distance, which was quietly misleading the model about how long routes actually were. Switching to the Google Distance Matrix API - real road distances, real travel paths - immediately improved how the model interpreted journey complexity. From there, we did a thorough cleanup of the feature set, cutting over 20 redundant variables so the final LightGBM model could focus on what genuinely mattered rath.

After tuning the model using Bayesian Optimization, we landed at a WMAPE of 19% and a Mean Absolute Error of 14 hours. In practical terms, the model now accounts for roughly 75% of the variance in delivery delays - which is a meaningful jump from where we started.

The broader takeaway is that supply chain prediction gets dramatically better when you stop treating roads as straight lines and start treating terrain as a real variable. For Bosch, that translates to tighter scheduling, less money locked up in precautionary stock, and a supply chain that runs with far fewer surprises between factory and customer.

**Social Media Analytics, Customer Segmentation, and Lead Scoring for FacilitEasy — A Data-Driven Approach to Product–Market Fit**

**Abstract:** Kasadara is a technology services company with a growing product portfolio. Its flagship product, FacilitEasy, a facility and asset management platform has an active digital presence across LinkedIn, Facebook, Instagram, YouTube, and a company website. However, the marketing and sales teams lack a data-backed view of what content drives engagement, who the right customers are, and which leads are worth pursuing. This leads to broad campaigns, generic messaging, and low conversions.

This project aimed to build three connected analytics workstreams for FacilitEasy: social media performance analytics, customer segmentation, and predictive lead scoring. To achieve this, we have collected engagement data (impressions, clicks, shares, comments, follower growth) from all social platforms and Google Analytics For segmentation, we plan to apply K-Means Clustering on firmographic and behavioural data from Apollo.io and internal lead records, then train a Random Forest Classifier to predict segment membership for new prospects.

For lead scoring, we plan to build an XGBoost model on historical lead-to-conversion data enriched with engagement and firmographic features, with Logistic Regression as an interpretable baseline.

The tech stack used was Python, Scikit-Learn, XGBoost, Pandas, and power BI for dashboards. As a result, we plan to have social media analysis identify top-performing content formats and posting patterns. Segmentation surface ≥3 actionable audience clusters with distinct profiles.

The lead scoring model achieving ≥80% classification accuracy, the projected improvement in campaign response rate to be 15–20% through prioritized, segment aligned outreach. These approaches suggest that a combined analytics approach across social media, segmentation, and lead scoring can directly improve targeting and conversion for a B2B SaaS product like FacilitEasy — and the framework can be replicated for Kasadara's other products (KIA, Agent KAI, BrikBite) as they scale.

**Value Chain Analytics: From Smart Procurement to Predictive Customer Intelligence**

**Abstract**: This project presents a data-driven approach to improve the overall value chain of a circular economy business of *Saahas Zero Waste (SZW)*. The procurement system has different item categories of waste procured under various business verticals of SZW. The data hasn't been stored appropriately and has been highly fragmented over the cloud and so, SZW always did a random vendor selection at whatever price quoted by the vendor.

This project finds that, over a period of 6 years 2019 to 2025, the vendor composition keeps on changing every year variedly. Hence, the project is aimed to build an effective vendor ranking system using a business weighted Recency-Frequency-Monetary (RFM) metrics. This helps identify top reliable suppliers across the various item categories. It also includes basic risk and churn prediction to avoid supply disruptions. In addition, median price benchmarking is used to compare vendor pricing within each category, helping the company control costs and make better purchasing decisions. All these decisions are captured into an impactful executive dashboard.

The executive dashboard integrates procurement and sales into a unified system enabling appropriate measurement and analysis of revenue performance.

On the sales side, the project helps to forecast the sales of the top business verticals which also, contribute majorly on the procurement side. Thus, this helps in completing the demand supply chain for the same. Along with this, the project also contributes towards forecasting the customer behaviour. All the customers are categorized into clusters, and reinforcement learning is then applied to understand how customers stay, leave, or return over time, which helps estimate Customer Lifetime Value (CLV).

By bringing together supplier insights and customer analytics, this project offers a practical, end-to-end framework to improve profits, manage cash flow better, and support smarter, data-driven decision-making for long-term growth.