PropWise: Data-Driven Property Price Prediction and Market Insights

Project Context: E.D.G.E. Competition Submission

Group & Course Details

Class	Date
TY BSC Data Science and Big Data Analytics	7th November 2025

Group Members	PRN
Neha Jadhav (leader)	1132230022
Sneha Patil	1132230347

1 Our Challenge: Problem, Vision, and Value Proposition

The complexity of real estate pricing demands a modern, data-driven approach. **PropWise** is our solution, transforming opaque valuation processes into transparent, actionable insights.

1.1 The Core Problem & Solution (Dual-Lens Integration)

The Problem Statement	PropWise Solution (The Dual-Lens Approach)
Traditional property valuation lacks consistency and fails to incorporate broader macroeconomic shifts. Stakeholders lack a unified tool that connects raw data, model output, and business context.	**Lens 1: Predictive Engine (Property Level)** Uses advanced ML (XGBoost) trained on property features to deliver highly accurate **Price per Unit Area** forecasts. **Lens 2: Insight Engine (Market Level)** Analyzes macroeconomic time-series data to track market shifts and generates interactive dashboards for **optimal buying times** and **undervalued zones**.

1.2 Unique Selling Proposition (USP)

PropWise uniquely blends data science, machine learning, and business analytics to not just predict property prices, but also deliver actionable market insights through dashboards and prescriptive analytics—making it a complete decision-support system.

1.3 Key Project Summary, Goals, and Outcomes

This section summarizes the measurable success criteria, the core deliverables, and the technical approach.

1.3.1 Key Performance Indicators (KPIs)

These metrics define the quantitative success of our predictive engine:

- **Prediction Accuracy (R^2):** Goal: **$R^2 \geq 0.90$**. (Achieve high prediction accuracy and reliability in property valuation.)
- **Model Reliability:** Goal: **P_{95}** Confidence Interval $\pm$**5%**. (The predicted price should be within a low margin of error for high confidence.)

1.3.2 Deliverables and Scope Summary (What We Will Build)

These are the tangible items delivered by Phase I of the project:

- **Modeling:** Three rigorously tested models (LR, RF, XGBoost) and a **Feature Importance Ranking** of the top 3 drivers.
- **Data:** Two clean, integrated, and ready-to-use data tables (Property and Market Trend data).
- **Analysis:** An **EDA Report** with visual charts showing price drivers (e.g., furnishing, status).
- **Future Roadmap:** A clear **Deployment Plan** for Phase II, including **Streamlit** UI and advanced **ARIMA/Prophet** forecasting.

1.3.3 Methodology Summary (How We Will Build It)

This outlines the tools and outputs of our technical approach:

- **Technical Stack:** Utilizes Python, `scikit-learn`, **XGBoost**, and visualization tools.
- **Output Artefacts:** The final Predictive Engine (`.pkl` file) and an initial set of **Interactive Dashboard Mockups**.

2   Project Scope: The What and The Phases

Our project is divided into two clear phases: the core deliverable scope (this submission) and future enhancements. This defines the **WHAT** the PropWise tool will achieve.

2.1   Phase I: Core Deliverables (For E.D.G.E. Competition Submission)

This phase focuses on the complete ML pipeline and foundational market analysis:

- **Data Preparedness:** Two integrated, clean, and preprocessed datasets (Property and Market Trend data).

- **Model Selection:** Development, training, and rigorous testing of three competing regression models (**LR, Random Forest, XGBoost**).

- **Performance:** The best model, confirmed to meet the **$R^2 \geq 0.90$** KPI.

- **Insight Generation:** Initial **EDA Report** and identification of the **Top 3 Feature Drivers**.

2.2   Phase II: Future Enhancements (Post-Competition Roadmap)

Our vision extends beyond the submission. Future work will focus on:

- **Full Deployment:** Implementing a user-friendly front-end interface (e.g., **Streamlit UI**) for public consumption.

- **Macro-Aware Forecasting:** Integrating advanced **ARIMA/Prophet** time-series models to predict future market shifts, making the price prediction truly macro-aware.

- **Geo-Local Integration:** Incorporating geographical data (City/Locality) for hyper-local price accuracy.

3   Methodology: The How and The Tools

This section details the technical **HOW** behind PropWise, outlining the specific tools, design principles, and data flow used to achieve the project goals.

3.1   Technical Toolkit (Tools of the Trade)

We are leveraging the standard Data Science toolkit: Python for our language, with `pandas` for data handling, `scikit-learn` and the high-performance **XGBoost** for predictive modeling, and `matplotlib/seaborn` for visualization. The final, best-performing model will be saved using `joblib`.

3.2 Predictive Engine Design (The Assembly Line)

The entire valuation process is managed through a **Scikit-learn Pipeline** to guarantee consistency and prevent data leakage:

1. **Preprocessing:** A **ColumnTransformer** handles data types (e.g., converting categorical text features into numerical format via **OneHotEncoder**).

2. **Training & Prediction:** The clean, transformed data is passed to the **XGBoost** model to output the final accurate price prediction.

3.3 Market Trend Data Flow

We analyze historical economic data (`QINN368BIS.csv`) to track market movements over time. This data acts as a "pulse check" on the market. In future iterations, the forecasted results from this data will be integrated as a powerful new feature into the prediction pipeline, creating a complete "Macro-Aware" valuation system.