PRICE DISCOVERY AND REGULATION MODEL FOR REAL ESTATE LISTINGS

Building Trust Through Transparent and Accurate Property Pricing

By Leonidas Mugema

CONTENTS

BUSINESS UNDERSTANDING

Company X Platform:

- Operates a nationwide real estate aggregator connecting property owners with buyers through a centralized system.
- Needs accurate and transparent pricing to maintain trust and lead in a competitive real estate market.

Problem Statement:

- Users report significant price variations for similar properties, undermining trust in the platform.
- This inconsistency has led to reduced user trust, higher support costs, and operational challenges.
- Ultimately, it risks damaging Company X's reputation and competitive edge.

PROJECT OBJECTIVES

Develop a fair pricing model that ensures accurate property price estimation based on key attributes like amenities, sub-area, property types, and more.

Build trust by fostering transparency and improving operational efficiency across the platform.

Restore user confidence and position Company X as the leader in transparent, data-driven real estate solutions.

EXPLORATORY DATA ANALYSIS

ABOUT THE DATA

- 200 properties in Pune, Maharashtra with 15+ features each.
- Data includes a mix of numerical, categorical, and text variables.
- Missing Values are present in some columns.

KEY INSIGHTS

- Categorical challenges: Variables are inconsistent and require significant cleaning for usability.
- Numerical challenges: Mixed data types(e.g, strings and numbers) necessitate thorough preprocessing for accurate analysis.

EXAMPLE

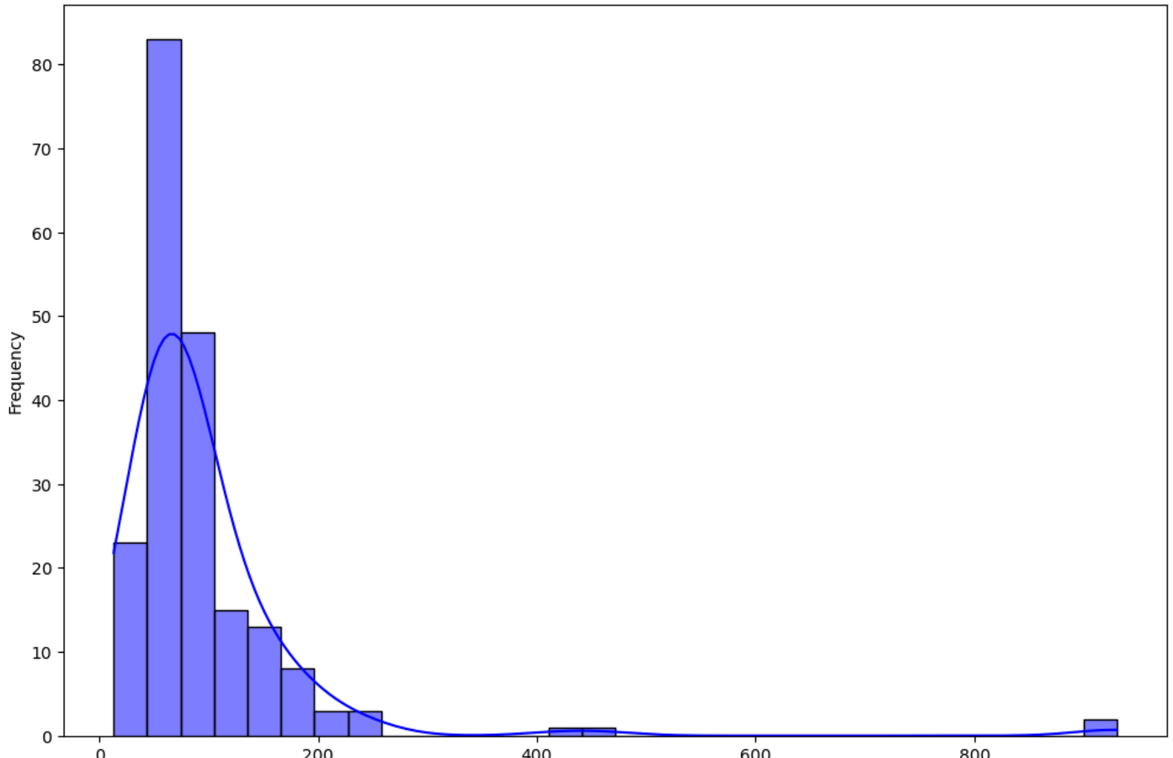
- Unique property types: “1 BHK, 2bhk, 2 BHK, 1BHK, …”

EXPLORATORY DATA ANALYSIS

DISTRIBUTION OF PROPERTY PRICES (IN LAKHS) BEFORE OUTLIER HANDLING

Key Observations:

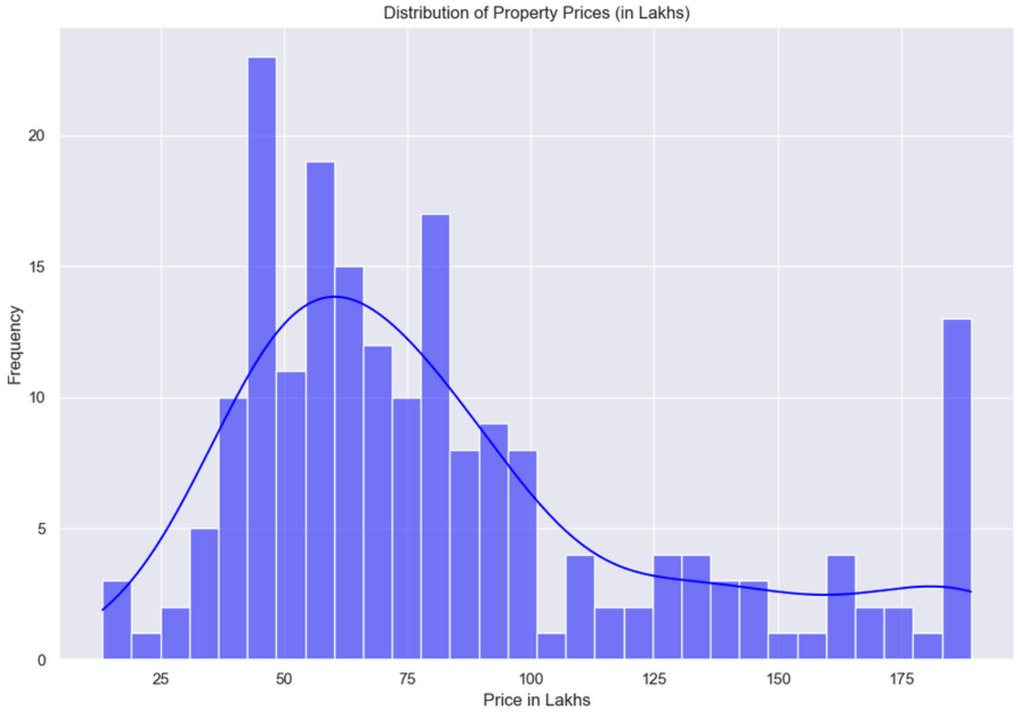
- The distribution is highly skewed, with a long tail extending to the extremely high prices.
- The majority of property prices are concentrated between 50 to 100 lakhs.



DISTRIBUTION OF PROPERTY PRICES (IN LAKHS) AFTER OUTLIER HANDLING

Key Observations:

- After removing outlier, the majority of property prices are clearly concentrated in the 50 to 100 lakhs range.



EXPLORATORY DATA ANALYSIS

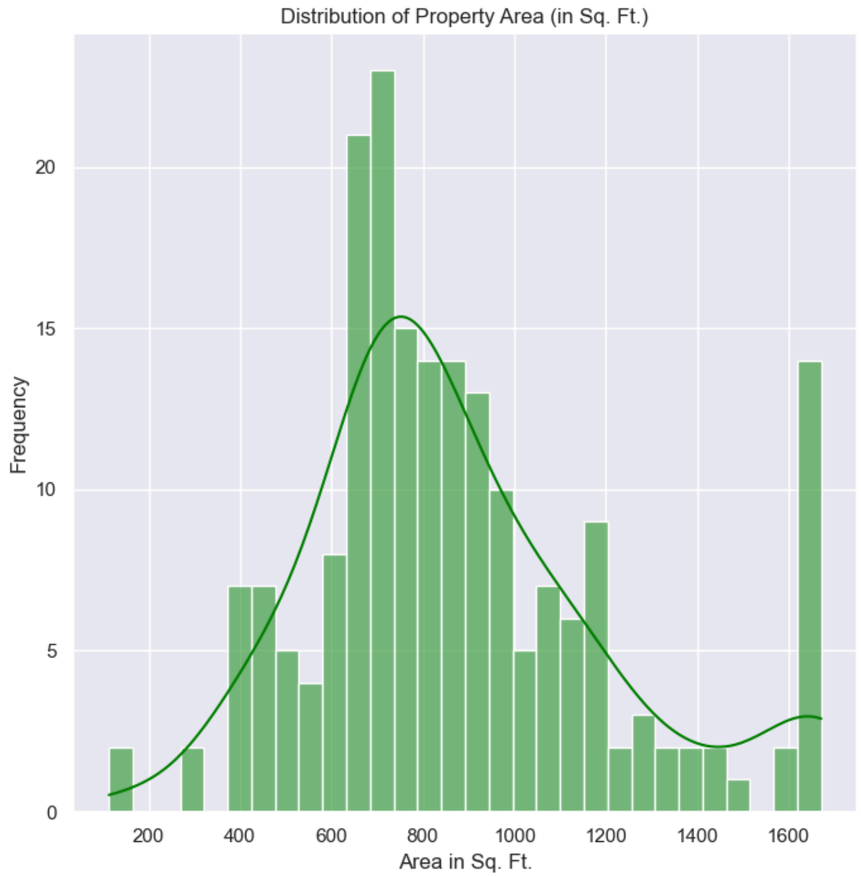
PROPERTY AREA DISTRIBUTION

Key Observations:

- Majority of properties fall within 600-1000 sq. ft., indicating standard property sizes
- Smaller clusters below 400 sq. ft. and above 1200 sq. ft. suggest outliers such as large luxury homes.

Impact:

- The distribution aligns with pricing patterns, making area an essential feature for accurate predicitions in the pricing model.



EXPLORATORY DATA ANALYSIS

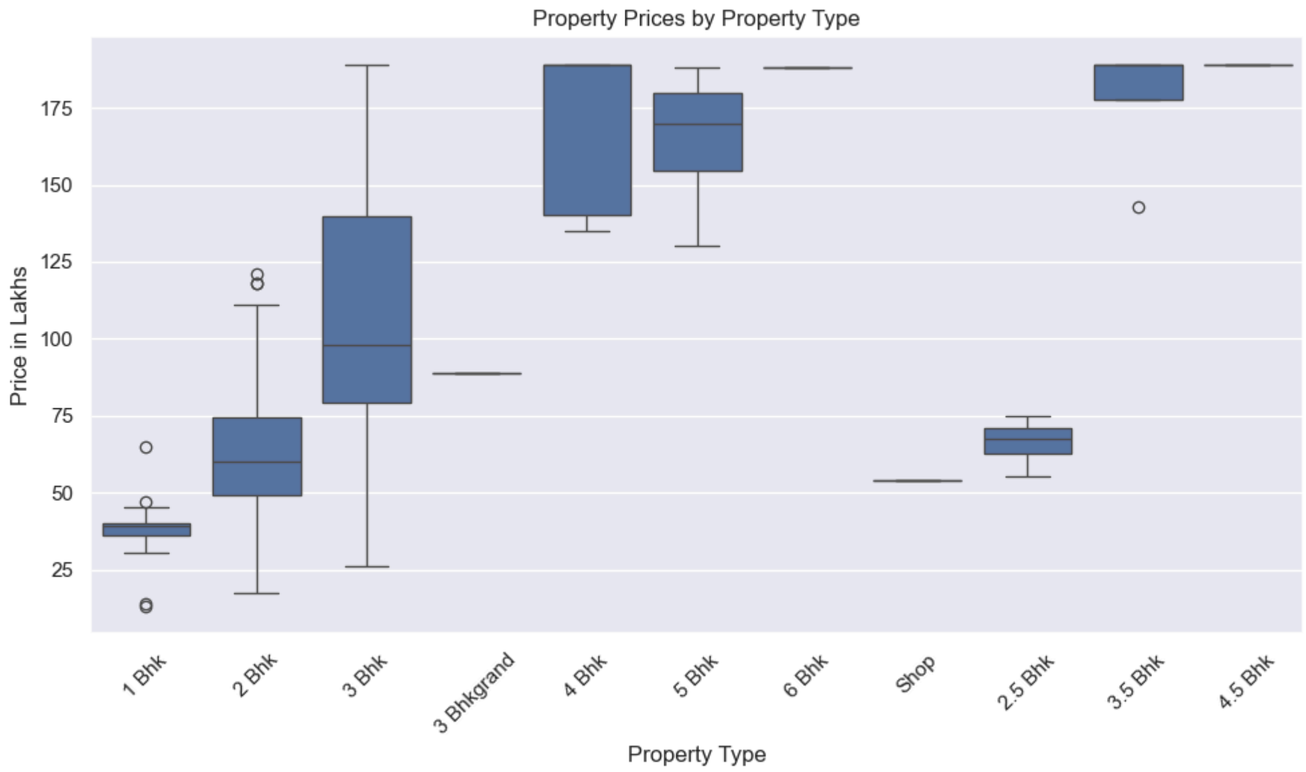
KEY PRICING TRENDS AND INSIGHTS

Key observations

- Prices generally increase with the number of bedrooms, but variability exists.
- Outliers reflect unique factors like luxury feaures or premium areas.

Impact:

- Property type Significantly influences pricing and must be carefully modeled.



EXPLORATORY DATA ANALYSIS

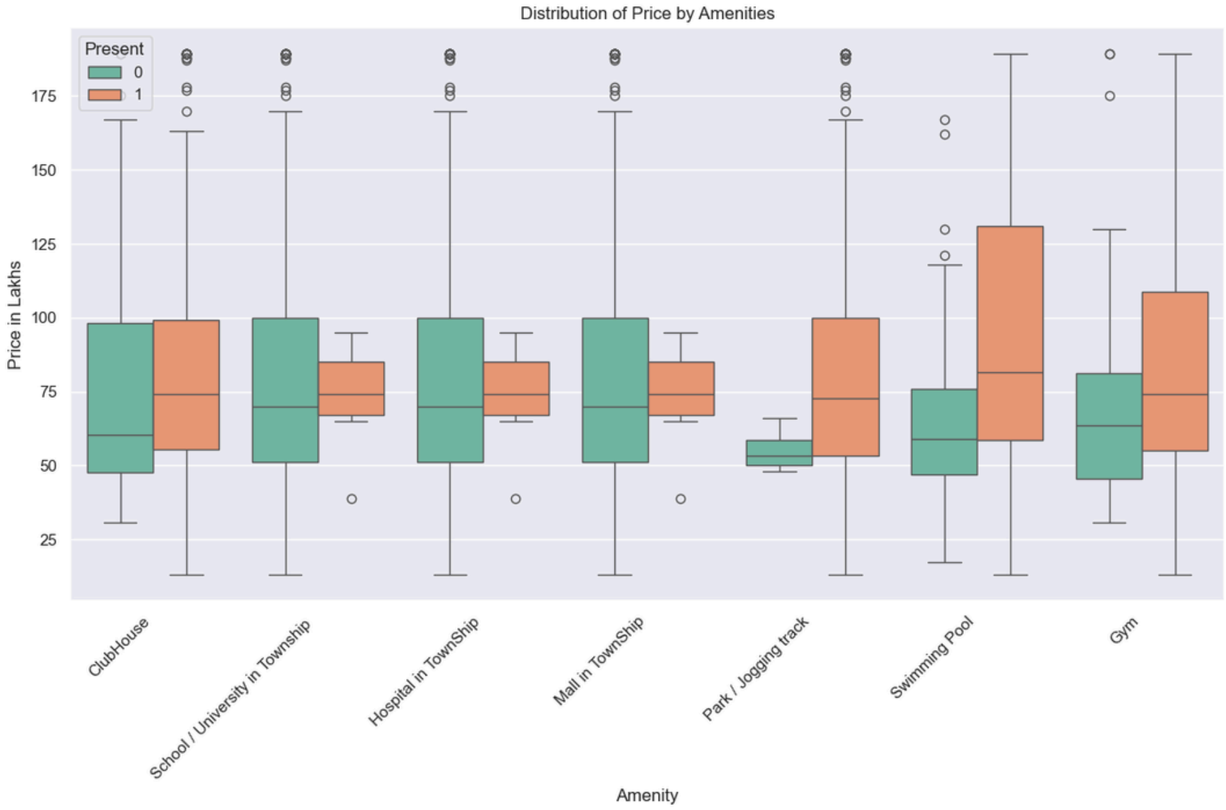
IMPACT OF AMENITIES ON PROPERTY PRICES

Key Trends:

- Premium amenities result in higher prices and narrower variability.
- Missing amenities often lead to lower but more unpredictable pricing.

Impact:

- Amenities are critical features for price prediction and should be weighted appropriately.



DATA PREPARATION

FEATURE SELECTION & CLEANING

- Removed redundant columns (Sr. No., Price in Millions) to simplify the dataset.
- Handled missing values to maintain data quality.
- Standardized text data to ensure consistency across categories

FEATURE ENGINEERING

- Created "BHK Number" feature to numerically represent bedroom count.
- Generated area-based price features to capture pricing trends.
- Conducted text analysis of property description column using NLP for qualitative insights.
- Encoded categorical variables into numerical values for model compatibility

	Sub-Area	Propert Type	Company Name	TownShip Name/ Society Name
0	Bavdhan	1 BHK	Shapoorji Paloonji	Vanaha
1	Bavdhan	2 BHK	Shapoorji Paloonji	Vanaha
2	Bavdhan	3 BHK	Shapoorji Paloonji	Vanaha
3	Bavdhan	3 BHK Grand	Shapoorji Paloonji	Vanaha
4	Mahalunge	2BHK	Godrej Properties	Godrej Hills retreat

MODELS TESTED FOR PROPERTY PRICE FORECASTING

Gradient Boosting Models:

- XGBoost: Efficient and accurate for structured data.
- LightGBM: Fast and memory-efficient, ideal for large datasets with categorical features.
- CatBoost: Optimized for categorical data with minimal preprocessing.

Linear Models:

- Lasso Regression: Simplifies models by penalizing less important features.
- Ridge Regression: Reduces overfitting by penalizing large coefficients.

Non-linear Models:

- MLP (Neural Network): Captures complex non-linear relationships between features and target prices.

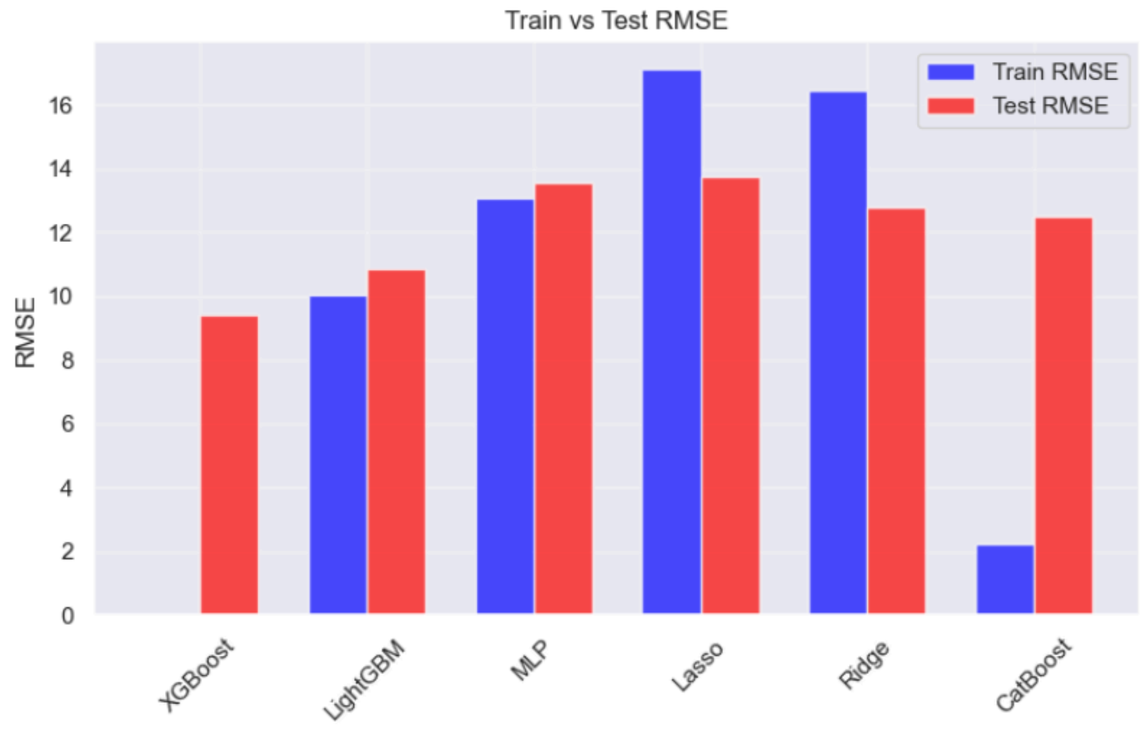
MODEL EVALUATION AND BEST MODEL SELECTION

Evaluation Metrics:

- Models were compared using MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and R^2 (Coefficient of Determination) for both training and test sets.
- Lower RMSE and MAE, combined with higher R^2, indicate better performance.

Performance Highlights:

- XGBoost: Best training performance (R^2 = 1.000) but showed slight overfitting due to a larger gap between train and test RMSE.
- LightGBM: Best overall test performance with balanced RMSE (10.84 lakhs) and R^2 (0.9458), making it the most robust model.
- CatBoost: High R^2 (0.9284) but slightly higher test RMSE (12.46 lakhs).
- Linear models (Lasso, Ridge) and MLP showed comparatively weaker performance with higher RMSE and MAE.



HYPERPARAMETER TUNING

Overview:

- Conducted extensive hyperparameter tuning to optimize the performance of the LightGBM model.
- Parameters like learning rate, max depth, number of leaves, and boosting type were fine-tuned.

Key Finding:

- Despite tuning, the optimized model failed to outperform the initial LightGBM configuration in terms of RMSE and R^2 on the test set.

Conclusion:

- The default LightGBM configuration remains the best-performing model due to its balance of simplicity and effectiveness.

Next Steps:

- Leverage the initial LightGBM for deployment and revisit tuning with a larger dataset.

SAMPLE PREDICTIONS WITH CONFIDENCE INTERVALS(PRICE RANGE)

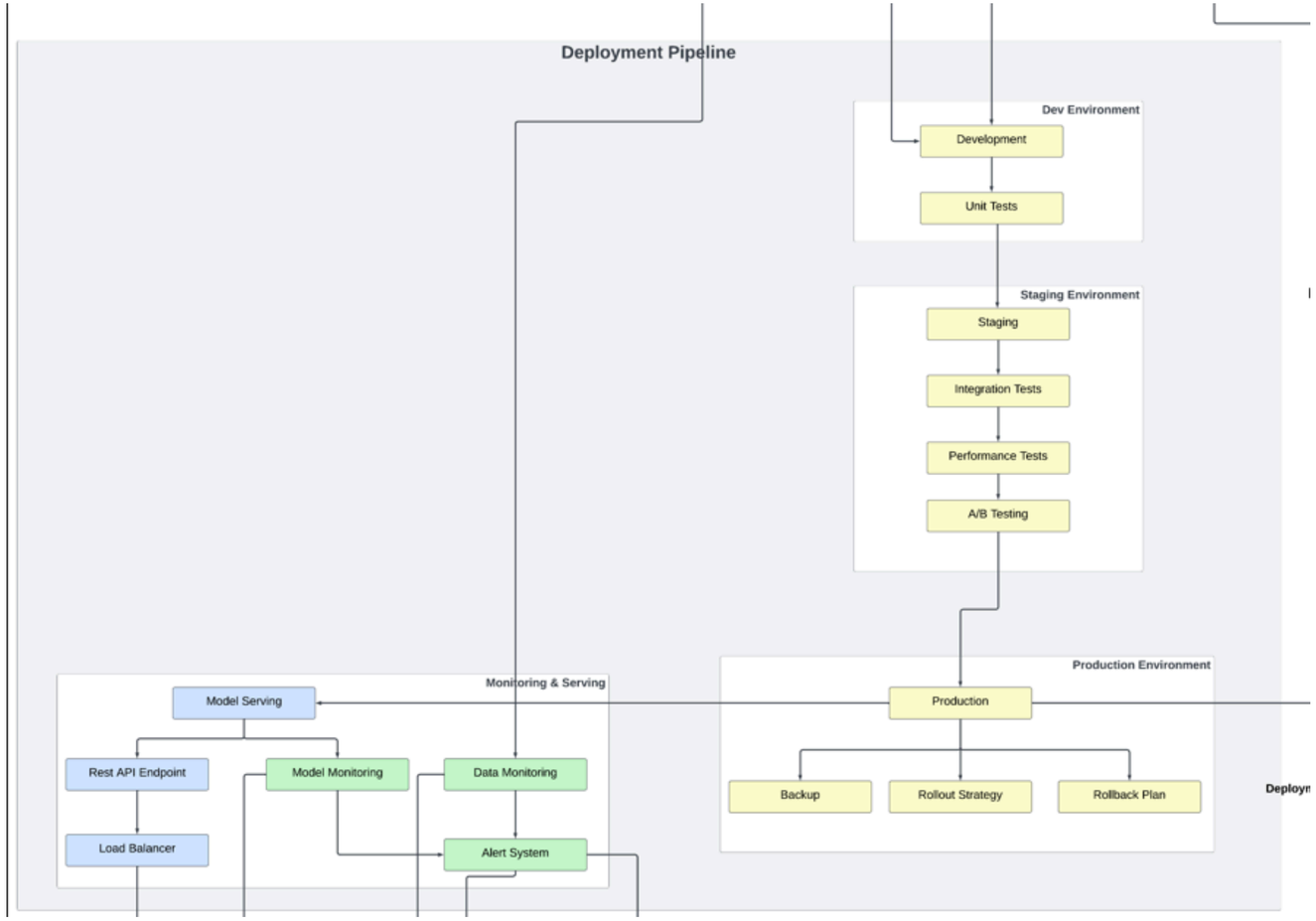
Key Insights:

1. Close Predictions:
   - The model demonstrates strong predictive power, with predicted prices closely aligning with actual values (e.g., row 15: predicted = 39.35 vs. actual = 40).
2. Confidence Intervals:
   - The confidence intervals provide an added layer of transparency, showing the range of uncertainty in the predictions.
   - For example, the actual price of 187 falls within the interval [158.25, 200.68], validating the model's reliability.

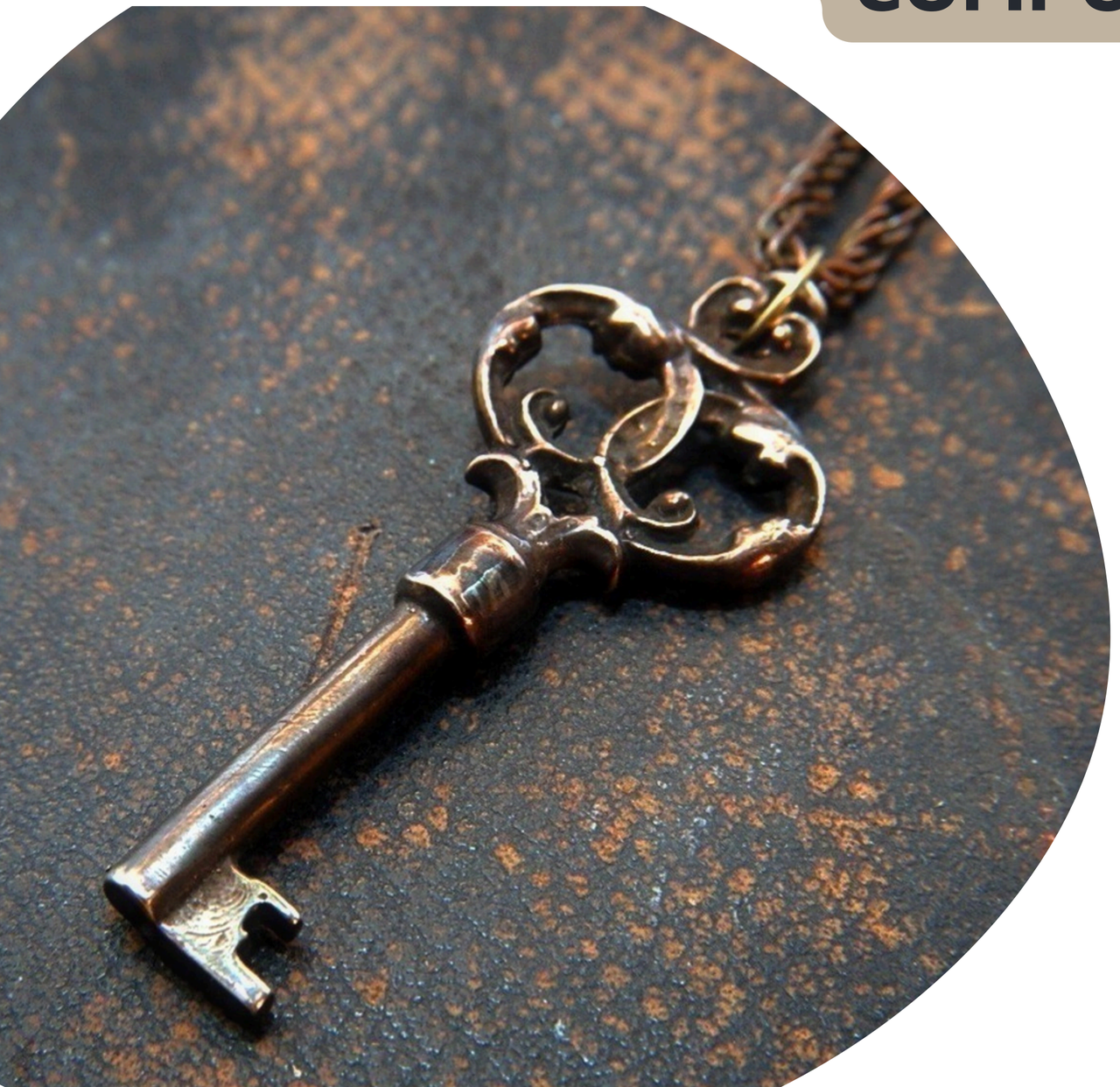
Index	Actual	Predicted	Lower Bound	Upper Bound
95	187.00	179.464995	158.247669	200.682321
15	40.00	39.350772	18.133446	60.568098
30	73.00	72.516065	51.298739	93.733391
158	87.00	83.424238	62.206912	104.641564
128	177.00	145.450604	124.233278	166.667930

DEPLOYMENT AND MONITORING PLAN FOR REAL ESTATE MODEL FOCUS ON DATABRICKS

- Established a comprehensive Plan pipeline within Databricks to address price inconsistencies
- This plan enables smooth integration of machine learning models into production
- Allows for ongoing training, automated validation, and monitoring of models
- Boosts scalability and minimizes manual efforts in model management
- Ensures model accuracy over time
- Aims to rebuild user trust by providing consistent and transparent property pricing
- Full image link can be found in Appendix



DEPLOYMENT AND MONITORING STRATEGY: KEY COMPONENTS



- Data Processing & Storage
- Model Development Pipeline
- Deployment Workflow
- Monitoring Systems
- Alert Management

DEPLOYMENT AND MONITORING STRATEGY

Data Sources:

- Property listing
- Historical transactions
- Market indicators

Feature Store:

- Centralized feature repository
- Real-time feature serving
- Version control and tracking
- Feature reuse across teams

Development Environment:

- Databricks notebooks
- Collaborative development
- Automated experimentation

MLflow Integration:

- Experiment tracking
- Model versioning
- Performance monitoring
- Easy deployment

DEPLOYMENT AND MONITORING STRATEGY

Deployment Pipeline

1. **Three-Stage Deployment:**
- Development
- Initial testing
- Unit tests

2. **Staging**:
- Integration testing
- A/B testing
- Performance validation

3. **Production**:
- Gradual rollout
- Monitoring
- Automated rollback

Monitoring and Alerts

Comprehensive Monitoring:
- Model performance metrics
- Data quality checks
- System health
- Business KPIs

Alert System:
- Real-time notifications
- Multi-channel alerts
- Automated responses
- Incident tracking

CI/CD Implementation

Automated Pipeline:
- Code quality checks
- Automated testing
- Model validation
- Deployment automation

GitHub Actions:
- Continuous integration
- Automated workflows
- Quality gates
- Release management

RECOMMENDATIONS

- Create interactive dashboards to track key metrics, such as data drift and prediction consistency for real time insights and stakeholder transparency
- Schedule concise , data-driven presentation every two weeks to update stakeholder on model performance, detected anomalies, and impact on user trust
- Allow users to rate pricing recommendations directly on the platform, creating a continuous feedback loop for improvement.
- Collecting more data before making high-stakes deployment decisions
- Scaling the dataset, especially by integrating external sources

THANK YOU!

Any Questions?

APPENDIX

- Full Image of Deployment and Monitoring Strategy Plan for the Model : https://lucid.app/lucidchart/38f0b041-f3ce-4f94-b6ae-b70f2b12bb62/edit?viewport_loc=-3812%2C1150%2C13875%2C6413%2C0eqBfotfVGlV&invitationId=inv_28fe668a-a005-4a4b-a89d-09092cdd0d84
- Github link : https://github.com/Mugema123/AI-Powered-Price-Discovery-and-Regulation-for-Real-Estate-Listings