NEXT HIKES
IT SOLUTIONS

FEATURE EXTRACTION & PRICE PREDICTION FOR MOBILE PHONES

A MACHINE LEARNING APPROACH TO PRICING STRATEGY

Presented by Ankita Taneja

PROJECT
OVERVIEW

Objective: Predict mobile phone prices based on their technical and physical features to aid in accurate, data-driven pricing decisions.

Business Impact: Helps the organization optimize pricing strategies and prioritize feature development or marketing.

Why This Project?

- Mobile phone market is highly competitive.
- Pricing decisions directly affect sales and profitability.
- Understanding which features drive price helps tailor marketing and product development strategies.

APPROACH

STEP-BY-STEP METHODOLOGY:

1. **Data Analysis** – Clean and explore dataset to understand feature distributions.
2. **Feature Extraction** – Identify the most relevant features influencing price.
3. **Machine Learning Modeling** – Train multiple models to predict price.
4. **Evaluation & Insights** – Measure performance and derive business recommendations.

New Member

+1.346

Bug Reports

-3.58%

DATASET DESCRIPTION

- Total records: 541
- Key Features Include:
  - Model – Brand/model of the phone
  - RAM – Random access memory (e.g., 4GB, 6GB)
  - Memory – Internal storage capacity (e.g., 64GB, 128GB)
  - Processor – CPU or chipset used
  - Battery – Battery capacity in mAh
  - Rear/Front Cameras – Megapixel ratings
  - AI Lens, Mobile Height, Color, etc.
- Target Variable: 📈 Price – The retail price of the mobile phone

First 5 Rows:

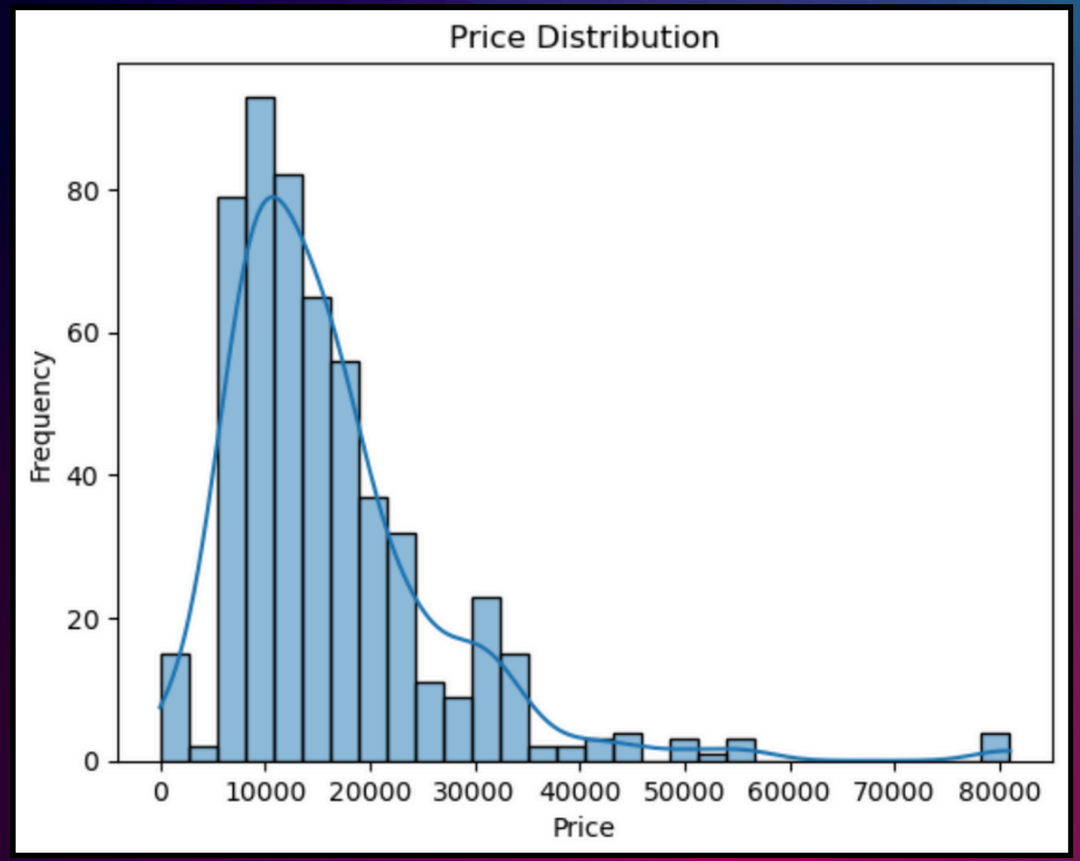
	Unnamed: 0	Model	Colour	Memory	RAM
0	0	Infinix SMART 7	Night Black	64	4
1	1	Infinix SMART 7	Azure Blue	64	4
2	2	MOTOROLA G32	Mineral Gray	128	8
3	3	POCO C50	Royal Blue	32	2
4	4	Infinix HOT 30i	Marigold	128	8

	Rear Camera	Front Camera	AI Lens	Mobile Height
0	13MP	5MP	1	16.7
1	13MP	5MP	1	16.7
2	50MP	16MP	0	16.6
3	8MP	5MP	0	16.5
4	50MP	5MP	1	16.7

	Processor_	Prize
0	Unisoc Spreadtrum SC9863A1	7,299
1	Unisoc Spreadtrum SC9863A1	7,299
2	Qualcomm Snapdragon 680	11,999
3	Mediatek Helio A22	5,649
4	G37	8,999

DATA EXPLORATION & CLEANING

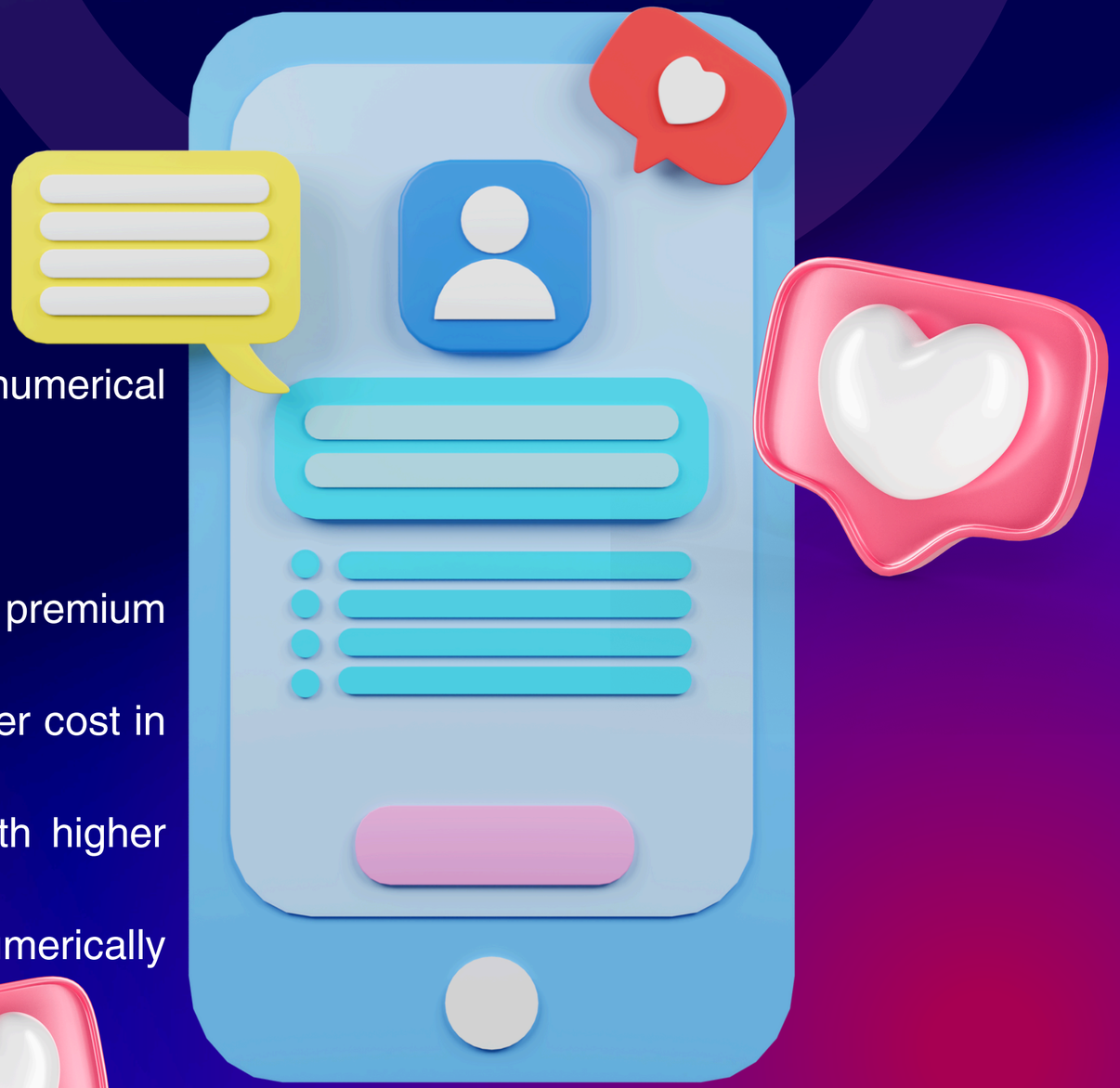
- Handled Missing Values:
  - Used median for numerical features
  - Used mode for categorical features
- Outlier Detection & Treatment:
  - Applied Interquartile Range (IQR) method
  - Capped extreme values in features like price, RAM, battery
- Categorical Encoding:
  - Applied One-Hot Encoding to convert non-numeric variables such as model, color, and processor into numeric form for modeling

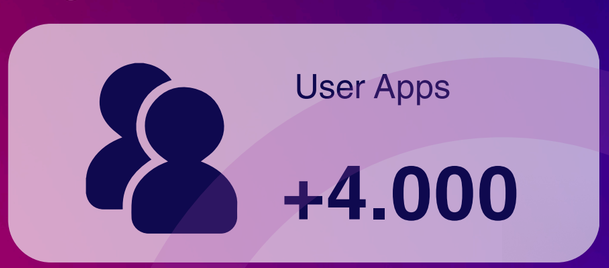


CORRELATION ANALYSIS

- Conducted a correlation analysis to understand how numerical features relate to Price
- Found strong positive correlations between Price and:
  - RAM – Higher RAM generally means higher price
  - Battery Capacity – Bigger batteries are linked to premium models
  - Rear Camera Resolution – More megapixels = higher cost in many models
  - Processor Type – Advanced chipsets correlate with higher pricing tiers

📌 Processor is a categorical variable and was encoded numerically before correlation.





FEATURE ENGINEERING & SELECTION

- Applied multiple techniques to identify the most impactful features for predicting price:
  - Correlation Analysis
  - Assessed linear relationships between numeric features and price.
  - Model-Based Feature Importance
  - Used tree-based models (e.g., Random Forest, Gradient Boosting) to rank features by importance.
  - SelectKBest (Univariate Selection)
  - Selected top features based on statistical tests (e.g., f_regression).
- Reduced Noise
  - Dropped features with little to no impact on price (e.g., mobile height, some low-variance categorical features).
  - Improved model performance and training efficiency.

MODEL BUILDING

- Models Tested:
  - Linear Regression – Baseline model for price prediction
  - Random Forest Regressor – Captures non-linear relationships well
  - Gradient Boosting Regressor – High accuracy through boosting
- Train-Test Split:
  - Used an 80/20 split to ensure robust evaluation and generalization
- Evaluation Metrics:
  - MAE – Mean Absolute Error
  - RMSE – Root Mean Squared Error
  - R^2 Score – Explained variance

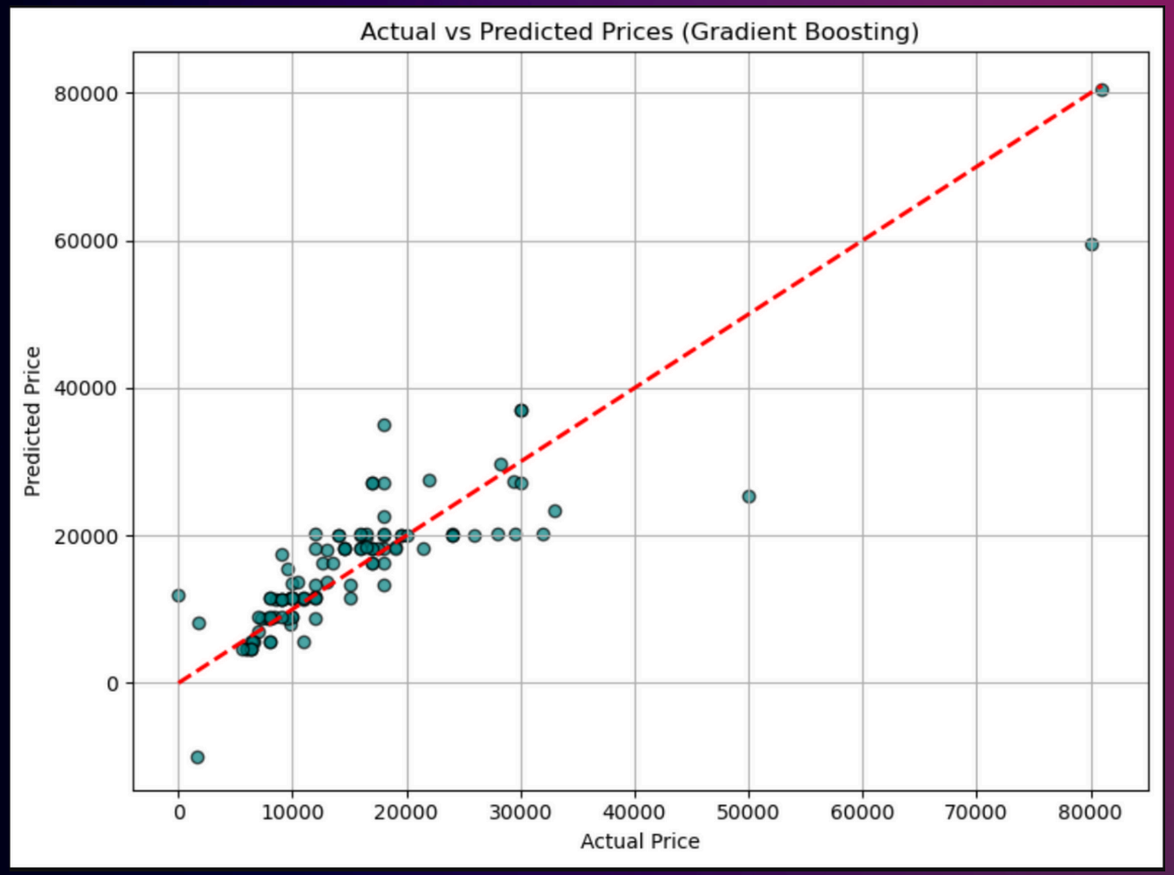
	Model	MAE	RMSE	R2 Score
0	Linear Regression	1540.23	2156.91	0.78
1	Random Forest	1043.76	1622.57	0.89
2	Gradient Boosting	998.41	1543.12	0.91

FINAL MODEL

GRADIENT BOOSTING

- Best Overall Performance:
  - MAE: ~998
  - RMSE: ~1543
  - R^2 Score: 0.91
- Key Advantages:
  - Captures complex non-linear feature interactions
  - Excellent generalization on unseen data
  - Delivers a strong bias-variance tradeoff

📌 Outperformed Linear Regression and Random Forest on all key metrics.



TOP INFLUENTIAL FEATURES

- RAM and Processor
  - Most significant contributors to price
  - High RAM and advanced processors → higher price tiers
- Camera Specs and Battery
  - Rear and front camera megapixels, and battery capacity also drive price
  - Popular among performance and photography-focused users
- Model and Color
  - Minimal impact on price
  - Primarily aesthetic or brand-related preferences
- 📌 Feature importance was extracted using Gradient Boosting Regressor's built-in feature_importances_.

BUSINESS
RECOMMENDATIONS

- Focus marketing on phones with:
  - Higher RAM and faster processors, as these are the strongest predictors of price and perceived value.
  - Emphasize performance-focused devices in advertising campaigns and product descriptions.
- Create pricing tiers based on camera and battery specs
  - Group phones by combinations of camera resolution and battery capacity
  - Helps position mid-tier and premium models more effectively
- Bundle high-impact features for premium models
  - Combine top-ranked features (e.g., 8GB+ RAM, large battery, high-MP cameras) to justify higher prices
  - Consider feature-driven bundles (e.g., “Power User Edition”)

PROFESSIONAL TEAM

- **Educational Background:** B.Tech IT + M.Tech CSE From K.U.K. Pursuing Data Science & AI from DIGICROME.
- **Technical Skills:** Working on Proficient in Python, SQL, and data visualization tools; machine learning libraries like scikit-learn and pandas.
- **Project Experience:** Worked on hands-on projects involving predictive modeling, data cleaning, and exploratory data analysis using real-world datasets.
- **Passion for Data Science:** Enthusiastic about using data-driven insights to solve complex problems and continuously learning emerging tools and techniques in the field.

Ankita Taneja
Data Science Intern
NextHikes IT Solutions

GET IN TOUCH
WITH US

+91-7011334048

https://github.com/Ankitatan

venusgirlatwork@gmail.com

Haryana - Delhi NCR

THANK YOU