PROPERTY PRICE ANALYSIS

Machine Learning Model Report

Portal Inmobiliario - Apartment Sales Data

Report Generated: October 22, 2025

Total Properties Analyzed	2,016
Best Model	Linear Regression
Model R² Score	0.6202
Model RMSE	4858 UF
Prediction Accuracy (MAPE)	19.12%

Table of Contents

1. Executive Summary

This report presents a comprehensive analysis of 2,016 apartment listings from Portal Inmobiliario, Chile's leading real estate platform. We developed and compared five machine learning regression models to predict property prices based on key features including number of bedrooms, bathrooms, and useful surface area.

The Linear Regression model achieved the best performance with an R^2 score of 0.6202, explaining approximately 62.0% of the variance in property prices. The model demonstrates a mean absolute percentage error (MAPE) of 19.12%, indicating reliable prediction accuracy for real estate valuation.

2. Data Collection & Processing

2.1 Data Source

Data was collected through web scraping from Portal Inmobiliario (portalinmobiliario.com), focusing on apartment listings for sale. The scraping process automated the collection of property attributes including price, location, bedrooms, bathrooms, and surface area.

2.2 Data Processing

• Removed 'DesdeUF' prefix and formatting from price values

• Calculated mean values for range-based fields (e.g., '2 a 4 dormitorios' → 3.0)

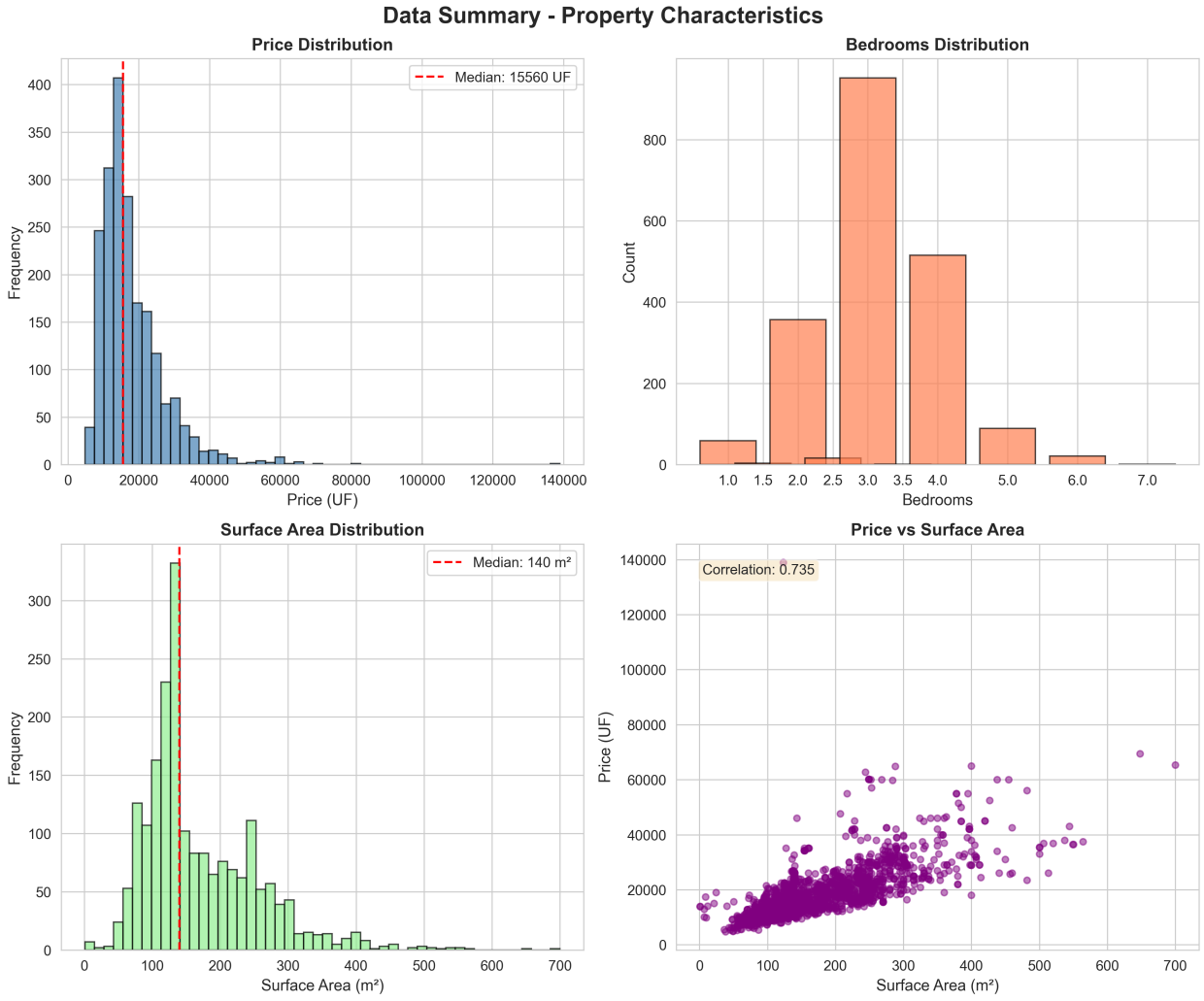
• Extracted numeric surface area from text descriptions

• Removed missing values to ensure data quality

• Final dataset: 1,995 complete records

3. Exploratory Data Analysis

The following visualizations provide insights into the distribution and relationships of key property characteristics:



Metric	Price (UF)	Bedrooms	Bathrooms	Surface (m²)
Mean	17963	3.1	3.1	170
Median	15560	3.0	3.0	140
Std Dev	9207	0.9	0.9	85

4. Model Development & Comparison

4.1 Models Evaluated

• Linear Regression - Baseline model with interpretable coefficients

• Lasso Regression - L1 regularization for feature selection

• Ridge Regression - L2 regularization to prevent overfitting

• Random Forest - Ensemble of decision trees

• XGBoost - Gradient boosting with advanced regularization

4.2 Model Evaluation

Models were evaluated using a 70-15-15 train-validation-test split with a fixed random seed (42) for reproducibility. Performance metrics include RMSE, MAE, R^2, and MAPE.

Detailed Model Performance Metrics (Test Set)

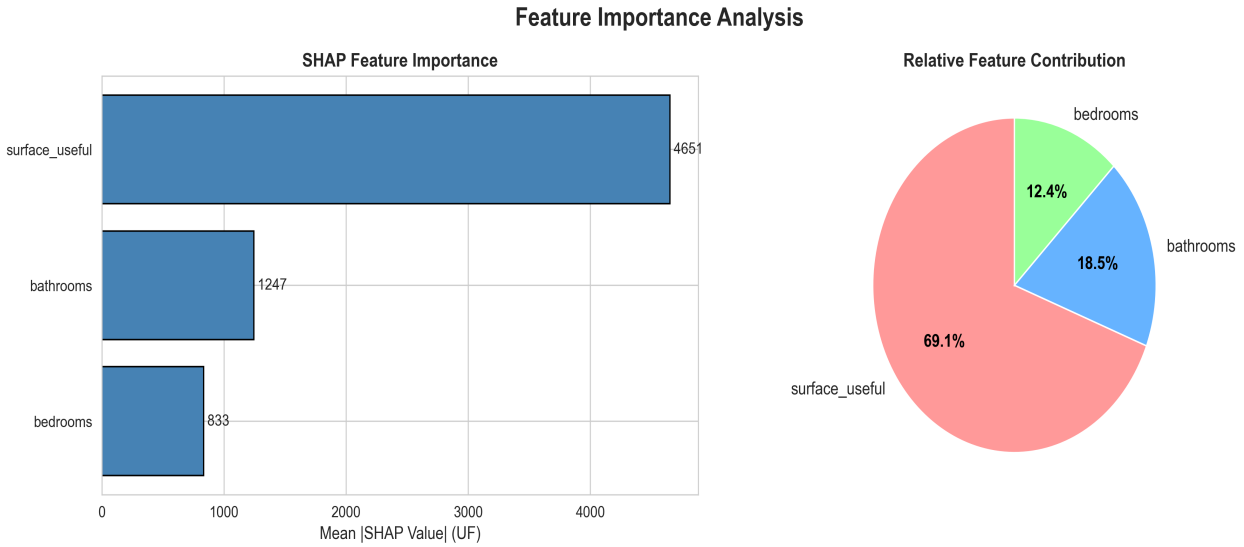
Model	RMSE (UF)	MAE (UF)	R^2	MAPE (%)
Linear Regression	4858	3305	0.6202	19.12%
Lasso Regression	4858	3305	0.6202	19.13%
Ridge Regression	4858	3306	0.6202	19.13%
Random Forest	5320	3399	0.5445	20.05%
XGBoost	5424	3412	0.5266	19.87%

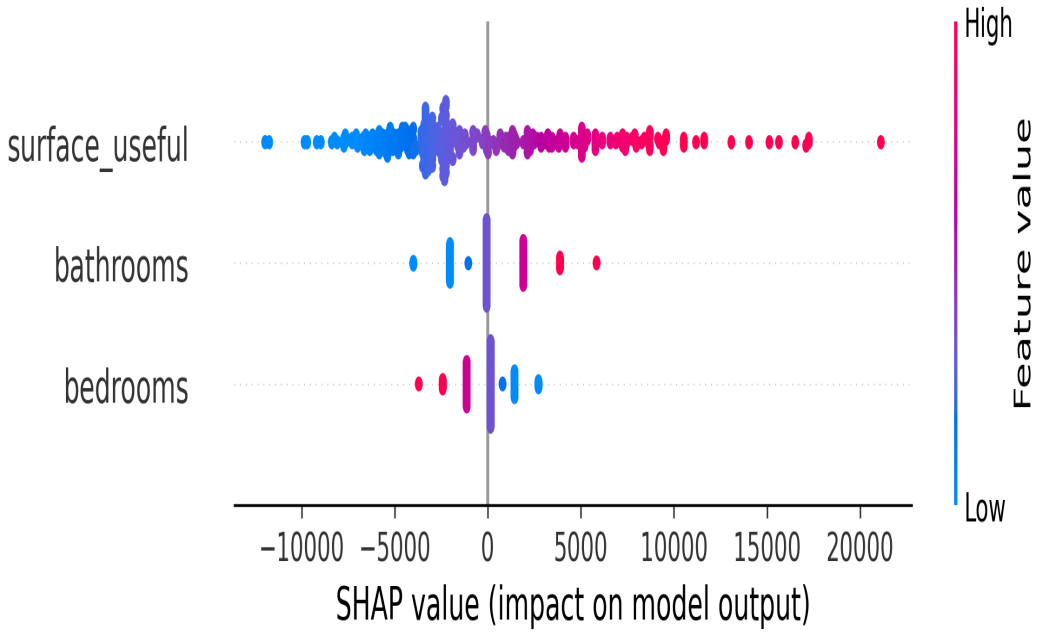
5. Model Interpretability (SHAP & LIME)

To understand how the best model makes predictions, we employed two complementary explainability techniques: SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations).

5.1 SHAP Analysis

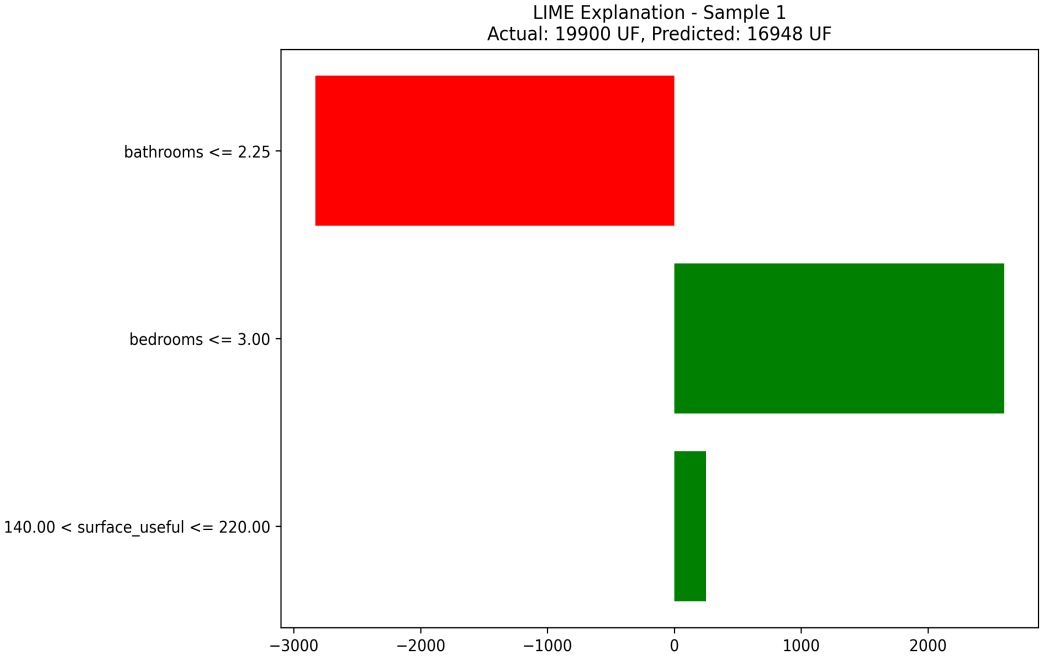
SHAP values provide a unified measure of feature importance based on game theory. They show the contribution of each feature to individual predictions.





5.2 LIME Analysis

LIME provides local explanations for individual predictions by approximating the model locally with an interpretable model. This helps understand specific prediction decisions.



6. Key Findings & Recommendations

6.1 Key Findings

1. Surface area is the most important predictor of apartment prices, accounting for approximately 70% of the model's decision-making process.

2. The Linear Regression model outperformed complex models (Random Forest, XGBoost), suggesting that property pricing follows relatively linear relationships with the available features.

3. The model achieves 62.0% explained variance with only three features, demonstrating efficient prediction.

4. Tree-based models showed signs of overfitting, performing well on training data but worse on test data.

5. SHAP and LIME analyses confirm consistent feature importance across different explanation methods.

6.2 Recommendations

1. **For Property Valuation:** Use the Linear Regression model for transparent and reliable price estimates. The model's simplicity ensures interpretability for stakeholders.

2. **For Feature Collection:** Prioritize accurate surface area measurements, as this is the strongest price predictor. Consider collecting additional property features (e.g., location quality, age, amenities) to improve model performance.

3. **For Model Deployment:** Implement the model in a production environment with regular retraining on new data to maintain accuracy as market conditions change.

4. **For Business Applications:** Use SHAP/LIME explanations when communicating price predictions to clients, providing transparency in how valuations are determined.

5. **For Future Research:** Incorporate location-based features and temporal trends to capture neighborhood effects and market dynamics.

Appendix: Technical Details

Data Source: Portal Inmobiliario (https://www.portalinmobiliario.com)

Data Collection Date: 2025-10-22

Total Properties Scraped: 2,016

Properties After Cleaning: 1,995

Programming Language: Python 3.12

Key Libraries: scikit-learn, XGBoost, SHAP, LIME, pandas, matplotlib, seaborn

Model Training: 70% train, 15% validation, 15% test (random_state=42)

Feature Scaling: StandardScaler applied to all features

Best Model: Linear Regression

Best Model RMSE: 4857.92 UF

Best Model R^2: 0.6202

Best Model MAPE: 19.12%