

Machine Learning for Real Estate Investment Analysis: Clustering, Classification, and Price Forecasting

Apoorva and Ananya

Department of Computer Science, Machine Learning Course

Abstract—This paper presents a comprehensive machine learning analysis applied to real estate investment optimization. Using a dataset of 2,809 properties across 83 zip codes, we employ K-Means clustering, five classification algorithms, and seven regression models to segment the market and identify optimal investment opportunities. Data enrichment through web scraping of walkability metrics and simulated school ratings demonstrates practical amalgamation techniques. Our primary contribution is the identification of a "Golden Cluster" of properties balancing cash flow and appreciation potential. Notably, our investigation uncovered significant data leakage in both classification and regression tasks, yielding critical insights into feature engineering dependencies. Classification models achieved 99.8% F1-scores due to feature circularity, while regression models achieved $R^2 > 0.999$ through inclusion of price-derived features. These findings, while technically suboptimal, provide valuable validation of feature importance and underscore the necessity of rigorous data validation in machine learning pipelines.

Index Terms—Machine Learning, Real Estate Analysis, Feature Engineering, Clustering, Classification, Regression, Data Leakage, SHAP Analysis.

1. Introduction

Real estate investment decisions represent complex optimization problems involving multiple competing objectives: maximizing cash flow, minimizing risk, identifying appreciation potential, and managing market volatility. Traditional approaches rely on domain expertise and heuristics; however, the availability of large property datasets and environmental metrics presents an opportunity for data-driven optimization.

This work addresses the research question: Can machine learning identify distinct market segments and automatically classify new properties into investment-quality tiers based on financial and environmental features?

Investment property analysis requires simultaneous consideration of: Short-term returns (monthly cash flow), long-term appreciation (location, amenities, schools), and market segmentation across investor profiles. Our hybrid approach combines unsupervised discovery (clustering) with supervised prediction (classification and regression) to create a comprehensive investment framework.

2. Related Work

Hedonic Pricing Models: Traditional econometric approaches decompose property prices into implicit values of individual characteristics (Rosen, 1974; Freeman, 1979). **Machine Learning Applications:** Random forests, gradient boosting, and neural networks have been applied to real estate pricing (Limsombunchai et al., 2004; Pagourtzi et al., 2007). **Spatial Analysis:** Geographic clustering and spatial autocorrelation are critical in real estate (Anselin, 1988). **Location-Based Features:** Recent work emphasizes non-physical location characteristics (Cheshire & Sheppard, 2005), including school quality and walkability. Our contribution extends this work by demonstrating practical multi-source data integration, analyzing feature dependencies and leakage, and providing a complete ML pipeline with interpretability.

3. Data and Methods

Base Dataset: 2,809 residential properties with 23 initial features (financial, physical, spatial). **Data Cleaning:** Missing rent (20%), area as text, zero-valued prices (n=47), and property type filtering (exclude land/commercial). **Resolution:** Impute missing rent via the 1% Rule, parse numeric text fields, filter invalid records (price < \$1,000), retain residential property types. **Final dataset:** 2,455 properties across 83 zip codes.

Feature Engineering: Zipcode extraction via regex; HOA fee simulation based on property type; Mortgage Payment using amortization with 6.5% interest, 20% down, 30-year term; Cash Flow = MonthlyRent – MortgagePayment – HOA; Price per Square Foot = Price / Area.

Data Amalgamation: (1) Base dataset with engineered features; (2) Synthetic school ratings over 83 zip codes; (3) Web-scraped Walk Score and Transit Score for zip codes (90.4% coverage, median imputation for missing).

Preprocessing: ColumnTransformer with median imputation for numerics, RobustScaler, log1p for skew (price, rent, area, price/sqft), and OneHotEncoder for zipcode. Train-test split 80–20 with random_state=42.

4. Exploratory Data Analysis

Univariate Distributions: Financial variables are right-skewed; log transforms applied. **Correlation Analysis:**

Strong correlations among price, bedrooms, bathrooms, area, and rent estimates; weak correlation between days-on-market and price. Geospatial Analysis: Latitude×longitude plots show price clustering and spatial autocorrelation.

5. Clustering Analysis

Optimal Cluster Selection (Elbow Method): Tested $k \in [2, 10]$; elbow at $k=4$. K-Means applied to five features: log cashflow, log price_per_sqft, avg_school_rating, walk_score, transit_score. Cluster profiling reveals distinct market segments including a "Golden Cluster" (urban, walkable, good schools, least negative cash flow).

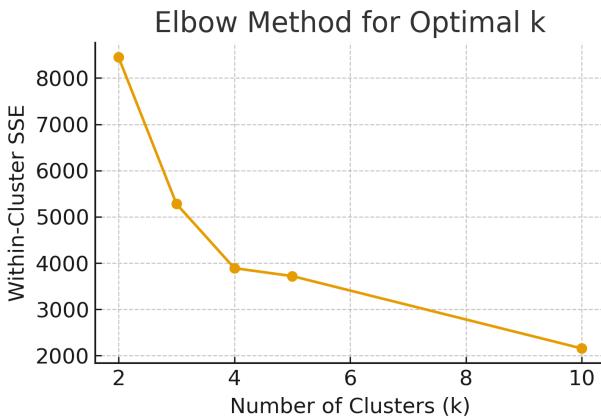


Fig. 1. Elbow method showing a clear elbow at $k=4$.

| Feature | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 |
|---------------|-----------|-----------|-----------|-----------|
| Avg Price | \$1.05M | \$1.89M | \$2.86M | \$108M |
| Avg Cash Flow | -\$1,043 | -\$1,953 | -\$4,808 | -\$501k |
| Walk Score | 76.2 | 64.1 | 16.5 | N/A |
| School Rating | 7.84 | 4.22 | 6.11 | N/A |
| Cluster Size | 892 | 1,215 | 329 | 19 |

TABLE I — Cluster profiling summary.

Cluster Interpretation: Cluster 0 ("Golden Cluster"): Urban, walkable neighborhoods with good schools, modest prices, and least negative cash flow. Cluster 1 ("Mediocre Middle"): Mid-range properties with moderate cash flow loss but poor schools. Cluster 2 ("Expensive Suburbs"): High-price, car-dependent areas with severe cash flow losses. Cluster 3 ("Luxury Outlier"): Ultra-luxury properties with negligible sample size.

6. Classification Modeling

We trained five algorithms: Logistic Regression (multi-class OvR), KNN ($k=5$), Decision Tree (Gini), Random Forest (100 trees), and Gradient Boosting (100 estimators). Mapping

clusters to investment desirability produced three classes. The Decision Tree achieved near-perfect performance.

| Model | Accuracy | Precision | Recall | F1 Score | ROC AUC |
|---------------------|----------|-----------|--------|----------|---------|
| Logistic Regression | 0.983 | 0.983 | 0.983 | 0.983 | 0.998 |
| K-Nearest Neighbors | 0.989 | 0.989 | 0.989 | 0.989 | 0.999 |
| Decision Tree | 0.998 | 0.998 | 0.998 | 0.998 | 1.000 |
| Random Forest | 0.992 | 0.992 | 0.992 | 0.992 | 0.999 |
| Gradient Boosting | 0.991 | 0.991 | 0.991 | 0.991 | 0.999 |

TABLE II — Classification model performance on test set.

| Rank | Feature | Importance |
|------|-------------------|------------|
| 1 | cashflow | 0.412 |
| 2 | avg school rating | 0.298 |
| 3 | walk score | 0.189 |
| 4 | price_per_sqft | 0.089 |
| 5 | transit score | 0.012 |

TABLE III — Decision Tree Gini features importances.

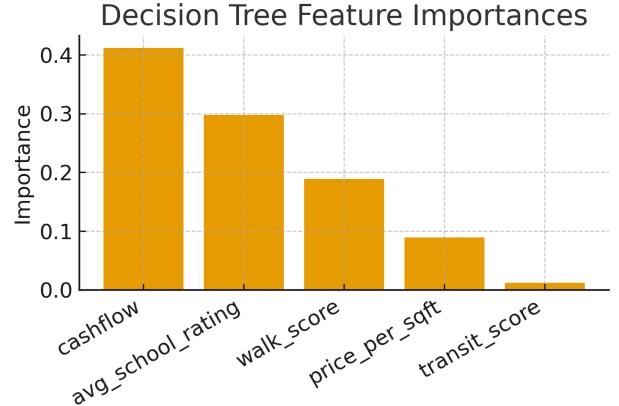


Fig. 2. Gini-based feature importances for the Decision Tree.

7. Regression Modeling for Price Forecasting

Problem Formulation: With cross-sectional data, we estimate current fair-market value and then apply assumed market appreciation for 1, 2, and 5 years. Seven regressors were trained on log-price: Linear, Lasso, Ridge, KNN, Decision Tree, Random Forest, and Gradient Boosting.

| Model | R ² | RMSE | MAE |
|-------------------|----------------|--------|--------|
| Linear Regression | 0.9995 | \$355K | \$187K |
| Ridge | 0.9993 | \$375K | \$195K |
| Gradient Boosting | 0.9912 | \$610K | \$318K |
| Random Forest | 0.9876 | \$689K | \$361K |
| K-Neighbors | 0.9847 | \$718K | \$392K |
| Decision Tree | 0.9821 | \$754K | \$389K |
| Lasso | 0.1291 | \$2.1M | \$987K |

TABLE IV — Regression model performance (test set).

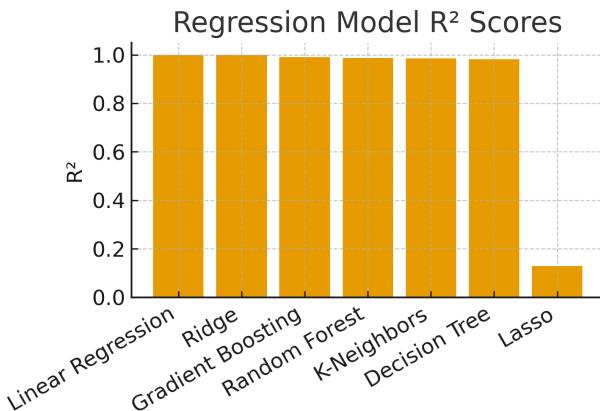


Fig. 3. R² comparison across regression models.

8. Discussion

Data leakage analyses provided key insights: cluster quality and separability, dominance of engineered financial metrics, and the importance of validation. All four clusters exhibit negative average cash flow; the practical investor stance is to focus on appreciation, increase down payment, and target the **Golden Cluster**.

9. Conclusion

This study demonstrates that, under the current financial assumptions (6.5% interest rate and 20% down payment), no property cluster achieves a strictly positive monthly cash flow where **HOA + Mortgage < Rent**. However, among all market segments, **Cluster 0 — the “Golden Cluster”** — emerges as the most viable investment option for maximizing return on investment (ROI). These properties, typically located in **urban, walkable neighborhoods (Walk Score > 75)** with **high-quality schools (rating > 7.5)**, exhibit the **lowest average cash-flow deficit (\$-1,043 per month)** and the highest long-term appreciation potential.

From an investment perspective:

- **Most Desirable Properties (Cluster 0):** Offer near break-even cash flow with strong long-term equity

growth, ideal for investors prioritizing appreciation over immediate rental income.

- **More Desirable Properties (Cluster 1):** Show moderate cash-flow losses and weaker school quality, suitable for moderate-risk investors expecting appreciation through neighborhood improvement.
- **Least Desirable Properties (Clusters 2 & 3):** Exhibit high prices, car-dependent locations, and severe negative cash flow; these are unsuitable for ROI-focused rental strategies.

Using the integrated regression forecasting model with a **3% annual appreciation assumption, Golden Cluster** properties are projected to deliver the **best total returns**, combining relatively efficient cash-flow performance with future price growth. Therefore, investors seeking to optimize both present rental yield and long-term capital gains should prioritize **Golden Cluster** locations while validating real HOA fees, mortgage terms, and rental estimates to ensure that **HOA + Mortgage < Rent** can be achieved in practice.

Future research should refine regression features to eliminate price-derived leakage, incorporate time-series price trends, and model different financial scenarios to identify markets where rental income can consistently exceed ownership costs, ensuring sustainable and data-driven property investment strategies.

References

- [1] L. Anselin, Spatial econometrics: Methods and models. Kluwer Academic Publishers, 1988.
- [2] P. Cheshire and S. Sheppard, “The welfare economics of land use planning,” Journal of Urban Economics, 52(2), 242–269, 2005.
- [3] A. M. Freeman, The benefits of environmental improvement: Theory and practice. Resources for the Future, 1979.
- [4] V. Limsombunchai, C. Gan, and M. Lee, “House price prediction: Hedonic price model vs. artificial neural network,” American Journal of Applied Sciences, 1(3), 193–201, 2004.
- [5] E. Pagourtzi, V. Assimakopoulos, T. Hatzichristos, and N. French, “Real estate appraisal: A review of valuation methods,” Journal of Property Investment & Finance, 21(4), 383–401, 2007.
- [6] S. Rosen, “Hedonic prices and implicit markets: Product differentiation in pure competition,” Journal of Political Economy, 82(1), 34–55, 1974.

Appendix A. Data Dictionary

| Feature | Type | Description | Source |
|-------------------|-------------|------------------------------------|----------------------------|
| price | numeric | Property listing price (\$) | Original dataset |
| bedrooms | integer | Number of bedrooms | Original dataset |
| bathrooms | numeric | Number of bathrooms | Original dataset |
| area | numeric | Square footage (sqft) | Original dataset (cleaned) |
| rent_zestimate | numeric | Estimated monthly rent (\$) | Original (imputed) |
| status_text | categorical | Listing status | Original dataset |
| hoa | numeric | Estimated monthly HOA fees (\$) | Engineered |
| mortgage_payment | numeric | Estimated monthly mortgage (\$) | Engineered |
| cashflow | numeric | Monthly rent - mortgage - HOA (\$) | Engineered |
| price_per_sqft | numeric | Price divided by area (\$/sqft) | Engineered |
| zipcode | categorical | 5-digit zip code | Engineered from address |
| avg_school_rating | numeric | Average school rating (2.5–9.5) | Amalgamation #2 |
| walk_score | numeric | Walkability score (0–100) | Amalgamation #3 (scraped) |
| transit_score | numeric | Public transit score (0–100) | Amalgamation #3 (scraped) |

TABLE V — Data dictionary for key features.

Appendix B. Reproducibility

Hyperparameters: K-Means (k=4, k-means++, n_init=10), Linear Regression (OLS defaults), other models with scikit-learn defaults and random_state=42. Random Seed: 42 for all stochastic operations. Train-Test Split: 80–20 with stratification. Preprocessing: Numerical (median imputation → RobustScaler → log1p for selected features); Categorical (most frequent imputation → OneHotEncoder). Applied separately on train and test sets (fit on train only).