NOVA IMS
Information Management School

DATA-DRIVEN DECISION MAKING

Pricing Regression for Cuckoo Cribs Corporation

Gabriel Calero - 20240357
Tingting Mo - 20241230
Dylan Fourie - 20211651
Gonçalo Duarte - 20241313

EXECUTIVE SUMMARY

This project introduces a smarter, data-driven approach for Cuckoo Cribs to estimate house prices. Instead of relying on gut feelings or guesswork, we use real historical housing data to build a model that can predict prices based on key property features, such as size, condition, location, and number of rooms.

Using the KNIME analytics platform, we will build a pricing model to help the company estimate property values more accurately and consistently. This will allow real estate agents to make faster, better-informed decisions and spot good deals earlier.

The model has been tested and shows strong results—its predictions are very close to actual market prices. With this model, Cuckoo Cribs can reduce pricing mistakes, avoid overpricing or underpricing, and improve overall profits.

We also suggest ways to use this model in daily business, for example, giving agents quick price suggestions during negotiations. As market conditions change, we recommend updating the model regularly to keep it accurate.

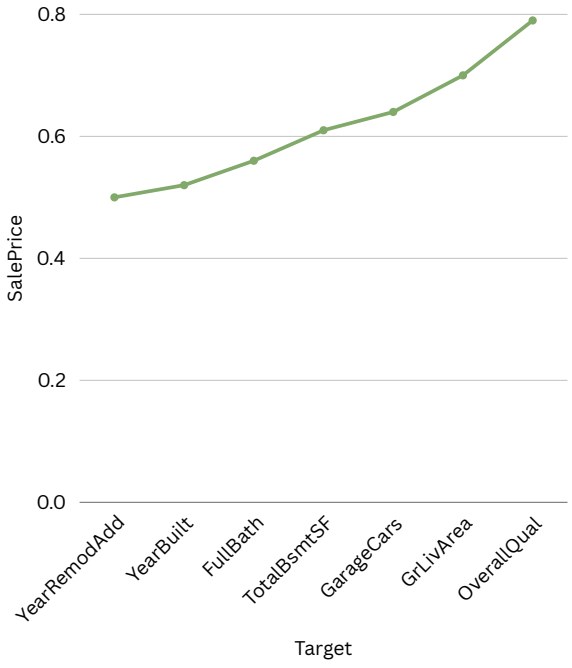
In conclusion, this report shows how using data can help Cuckoo Cribs make smarter decisions, stay competitive, and feel more confident in today's fast-changing real estate market.

AVAILABLE DATA

Feature Engineering and Data Selection

To get the best results, we focused on selecting the most important variables for predicting house prices:

For numerical variables:
We selected those that showed the highest correlation with the target variable, 'SalePrice'. This means we kept the numbers that had the strongest relationship with the final house price.



The correlation between the Features and SalePrice

Target	OverallQual	YearBuilt	YearBuilt	TotalBsmtSF	GrLivArea	FullBath	GarageCars
SalePrice	0.79	0.52	0.5	0.61	0.7	0.56	0.64

For categorical variables:
We removed categories where more than 80% of the data was the same value. These variables added little new information for the model and could reduce prediction quality.

The final set of variables used in the model includes features like the overall quality of the house, year built, size, number of bathrooms, garage size, neighborhood, and other important characteristics that help explain differences in house prices.
The complete list of selected variables can be found in the appendix.

By choosing only the most relevant variables, we made the model simpler, faster, and more accurate.

Model Details and Performance

At Cuckoo Cribs, our goal is to predict property values in a precise and reliable way. To do this, we used an advanced machine learning model called Gradient Boosted Trees. This model is well known in the industry for giving very accurate results.

Reasons for Choosing the Gradient Boosted Trees Model:
- High accuracy: It combines the "opinions" of many small models (trees) to give strong and reliable predictions.

- Proven in the industry: This technique is used by leading companies in real estate and finance for complex estimates.

- Clear and reviewable: Even though it is advanced, its results can be explained in a simple way and checked easily.

Technical results and a simple example
-R^2 of 0.91:
The model can explain 91% of the changes in house prices.
Example: If we have 100 houses, for 91 of them the model predicts a price very close to the real price; only for 9, there will be bigger differences.

- Mean Absolute Error: $14,584
On average, the model's prediction is $14,584 off from the real price.
Example: If a house is worth $200,000, the model will usually predict between $185,416 and $214,584.

RowID	Prediction (SalePrice) Number (double)
R^2	0.91
mean absolute error	14,584.132
mean squared error	498,768,832.204
root mean squared error	22,333.133
mean signed difference	687.134
mean absolute percentage error	0.085
adjusted R^2	0.91

- Mean Absolute Percentage Error: 8.5%
The average error is only 8.5% of the real price, which is low and gives us confidence. Example: If a house costs $300,000, the typical error will be about $25,500 higher or lower.

This model lets us move away from guesses or gut feelings, and instead make pricing decisions. With it, Cuckoo Cribs can spot good opportunities, set fair prices, and avoid risks. In short, our “pricing machine” helps make sure every decision is backed by the best possible evidence, just like the world’s most successful companies.

Business Outcome

This project’s predictive house pricing model brings clear and measurable business value to Cuckoo Cribs. Compared to relying on gut feeling, the company can now price properties more accurately and confidently—leading to higher revenue and reduced business risk.

The model achieved a prediction accuracy of 91%, with a Mean Absolute Error (MAE) of $14,584. In other words, the predicted price of most properties differs from the actual sale price by no more than this amount. This gives sales agents a reliable reference point when setting prices, helping them avoid losses from overpricing or underpricing.

Estimated business impact in real-world scenarios:

- Scenario 1: Reducing losses from undervalued sales

If 10 homes are undervalued by $20,000 each, $200,000 in revenue could be lost. The model helps avoid this with fair, data-driven estimates.

- Scenario 2: Faster sales and lower holding costs

More accurate pricing can cut market time by two weeks per property, saving $1,000–$2,000 in marketing and maintenance.

- Scenario 3: Identifying undervalued properties for smarter investments

The model helps identify undervalued properties, supporting better acquisition strategies.

Additionally, based on the five most important variables identified by the model (**Exterior Material, Lot Shape, Roof Style and House Style**) a clear pattern emerges in the average predicted values:

Feature	High-Value Homes	Median Price	Low-Value Homes	Median Price
Exterior Material	Cement Board, Vinyl Siding	$220,000 – $300,000	Stucco, Asbestos Shingles	< $150,000
Lot Shape	IR2, IR3 (Irregular)	$220,000 – $230,000	Regular (Reg)	< $150,000
Roof Style	Hip, Mansard	$180,000 – $220,000	Gambrel	~$110,000
House Style	2Story, 2.5Fin	$190,000 – $200,000+	1.5Unf, SFoyer, 2.5Unf	$100,000 – $150,000

Deployment

To bring the predictive pricing model into real-world use, we recommend integrating it into a simple, user-friendly interface that real estate agents at Cuckoo Cribs can easily access—such as a web-based dashboard or a mobile app.

Here's how agents could use the model in daily work:

Instant Price Estimation Tool

Agents input a property's key details—such as size, location, number of rooms, year built, and overall condition—into the platform. The model instantly generates a reliable price estimate, helping agents make quicker, more confident pricing decisions.

Support During Client Consultations

When meeting with sellers or buyers, agents can use the tool to show data-backed pricing suggestions. This builds trust with clients and improves negotiation outcomes.

Lead Prioritization

The model can help identify properties that are likely undervalued or overvalued. This allows agents to focus on the best opportunities, whether for sales, investments, or faster transactions.

Training and Onboarding

New agents can use the pricing tool as a learning resource, helping them understand how different property characteristics affect value, without needing years of experience.

To ensure adoption, we suggest starting with a pilot in one sales team. Feedback can help refine the interface and improve the model's usability before rolling it out company-wide. The tool should also connect to the company's internal property database to reduce manual data entry.

Monitoring and Retraining

Long term reliability and business value of the predictive pricing model is paramount. To ensure this we propose a comprehensive model monitoring and retraining framework.

Performance Monitoring
Model performance should continuously be tracked because the market is always changing. Error metrics, like RMSE, MAE, and MAPE, should be tracked on new incoming sales. Any changes to key features such as OverallQual, GrLivArea, or Neighbourhood distributions should be monitored using drift indicators or statistical tests.

Retraining
Schedule-based retraining with updated data from Cuckoo Cribs' property transactions should be implemented every 3 to 6 months. When the RMSE exceeds a threshold that was defined by the business or if the MAPE exceeds 10%, the model should be retrained. The retraining pipeline built into the KNIME workflow allows for the model to be refreshed with minimal manual intervention.

Improvements and Next Steps

The current model performs really strongly with a R^2 of 0.91 and a MAE of $14,584, however there is still room for improvement.

Model Comparison
Comparing the current model against others like XGBoost to test for overfitting and gain some explainability trade-offs.

Explainability Tools
By integrating SHAP or LIME, agents will have real-time explanations for why a certain price was predicted. This could improve trust and adoption.

User Feedback Loop
This will allow the agents to flag inaccurate predictions and their feedback can be incorporated into retraining cycles.

Data Ownership
A clear data ownership and quality process should be established to ensure reliable retraining data.