

Housing Price Classification



Identifying High-Value Investment Properties:
A Data-Driven Approach

Leveraging Machine Learning for Smarter Real Estate
Investments

Spotting Investment Gems:

Our Project Goal



- **The Challenge:** Our real estate investment firm currently relies on experience and manual checks to find properties. This can be slow and sometimes inconsistent.
- **Our Solution:** We're building a "smart predictor" using historical data to quickly identify properties that are likely to be "High Value" investments.
- **Why This Matters:** This helps us find promising properties faster, make more informed decisions, and focus our valuable resources on the best opportunities.
- **Key Question We're Answering:** Can we reliably predict which properties are "High Value" before we invest?

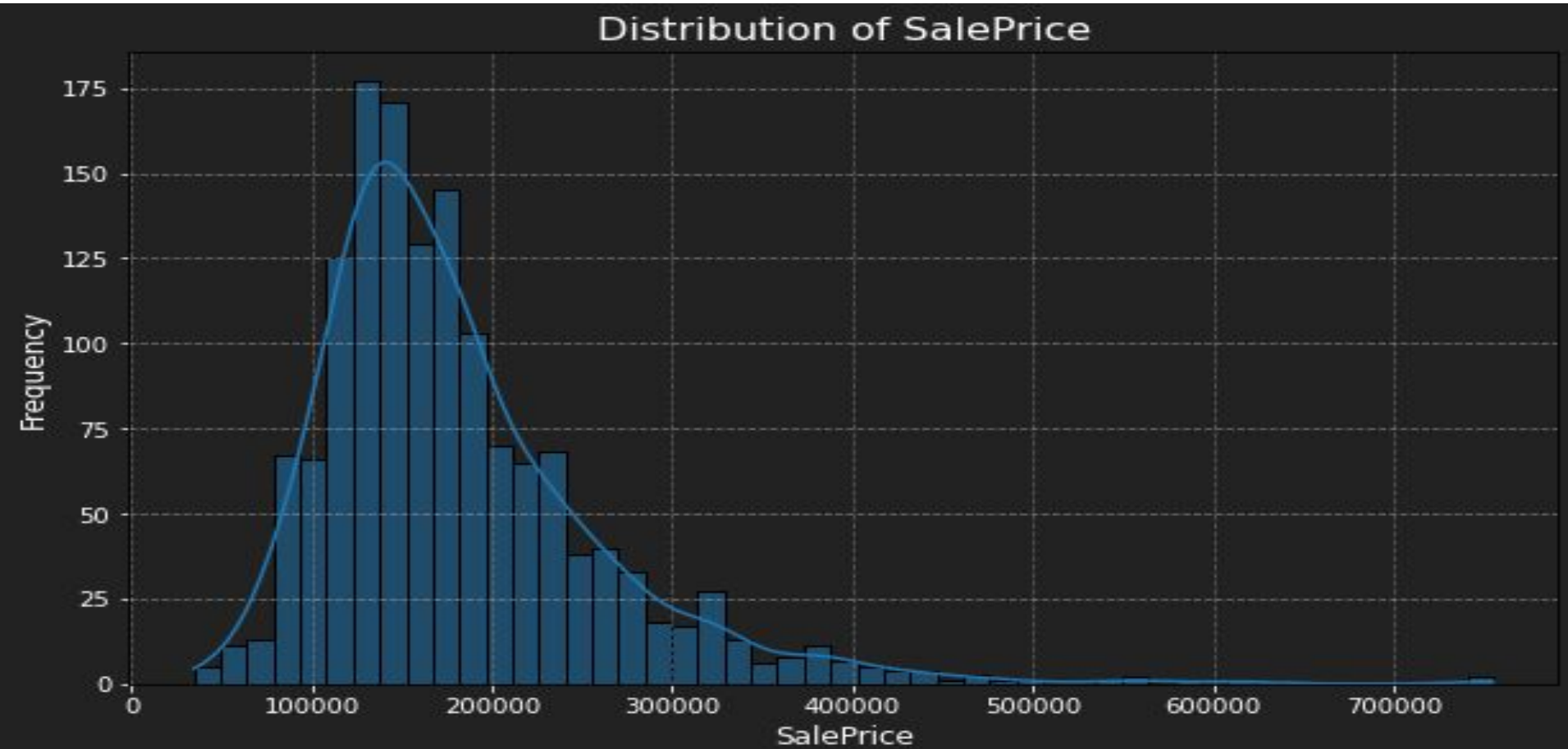
Business & Data Understanding

What Makes a Property "High Value"?

- **Definition:** We've defined "High Value" properties as those in the top 25% of sale prices within our historical dataset. All others are considered "Standard Value." This helps us turn a complex price prediction into a clear "yes/no" question.
- **Data Source:** We're using a comprehensive dataset of past home sales, which includes details like size, location, quality, and age of properties.
- **How much data?** We're analyzing thousands of past home sales to learn patterns.



Sale Price Distribution with Threshold



Visualization: Sale Price Distribution with Threshold




- **Type:** Histogram of **SalePrice** with a clear vertical line indicating the 75th percentile, visually separating "Standard" from "High Value" properties.
- **Purpose:** To clearly illustrate how the "High Value" target was defined from the raw sale prices.

This visual clarity helps us grasp the foundational step in converting a regression problem into a classification problem, directly supporting the rationale for using a classification model and setting the stage for understanding the model's categorical output.

Initial Data Insights

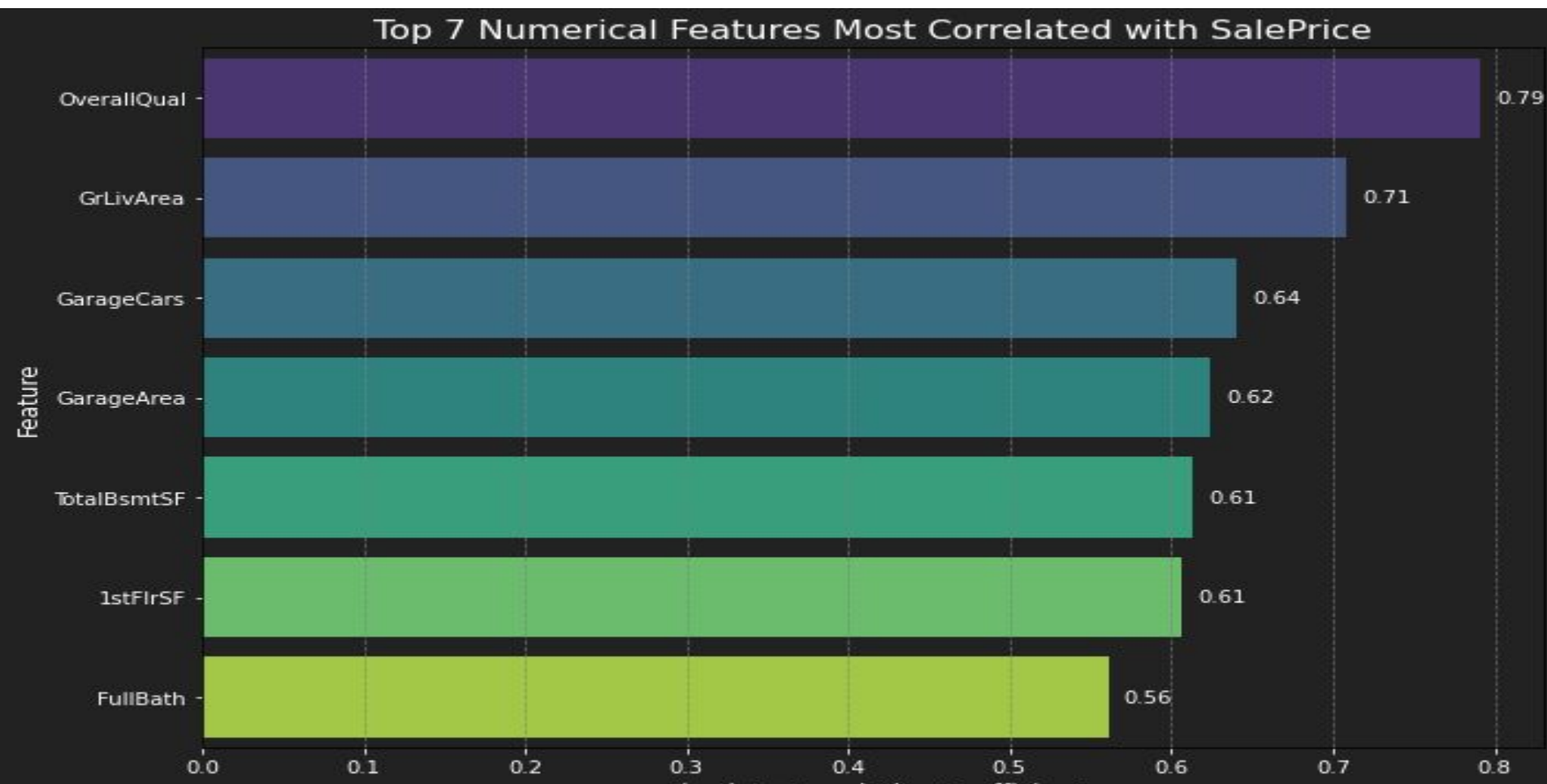
What Drives Property Value?

- 
- Before building our predictor, we looked at general trends in the data.
 - We found that certain features, like overall quality and living area, often have a strong connection to higher sale prices.
 - We also prepared our data for the predictor. Information about a property was frequently missing; for example, the PoolQC (Pool Quality) column was often empty, indicating no pool.
 - For these gaps, we filled them in, such as marking "No Pool" when PoolQC was blank, or using the median value for properties with missing LotFrontage (linear feet of street connected to property) to ensure completeness.



- Additionally, we converted all property details into a numerical format our predictor could easily understand and learn from.
- This included converting descriptive categories like Street ('Pave' or 'Grvl') into distinct numerical representations, or transforming the OverallQual (Overall material and finish quality) scores into standardized values. This meticulous preparation ensured every piece of information was precise and ready for analysis.

Top Correlations with SalePrice





Visualization: Numerical Features most Correlated with SalePrice

Type: Bar chart showing the top 7 numerical features most correlated with **SalePrice**.

Purpose: To reveal the most influential property characteristics that have historically driven sale prices, providing foundational insights for initial investment screening.

Recommendation Connection: This early insight into strong property-value relationships will guide preliminary investment considerations and align with the more precise feature importance identified by our predictive model.

Modeling (Our Predictive Approach)

Classification in Action



Building Our Smart Predictor



- **What is Machine Learning?** Think of it as teaching a computer to find patterns in data so it can make intelligent guesses about new, unseen data.



- **Why "Classification"?** Instead of trying to guess an exact price (like \$250,000), we're teaching our predictor to classify properties into categories: "High Value" or "Standard Value." This is like deciding "Is this property a good fit for our premium portfolio?" (Yes/No). This simplifies decision-making.
- **Our Process:** We started with a basic predictor (like a simple checklist) and then refined it (like adding more detailed criteria and weighting them) to make it even smarter. We built multiple versions to find the best one.

Evaluation (How Well Does Our Predictor Perform?)


How Reliable is Our "High Value" Predictor?



How Reliable is Our "High Value" Predictor?

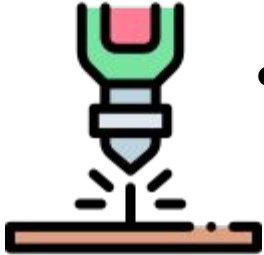


- **Measuring Success:** We used a key metric called ROC-AUC (think of it as a comprehensive score of how good our predictor is at distinguishing between "High Value" and "Standard Value" properties). A score closer to 1.0 is excellent, while 0.5 is like guessing randomly.
Our Results: Our best predictor, the Decision Tree Classifier, achieved an impressive ROC-AUC score of [Insert Tuned Decision Tree ROC-AUC here, e.g., 0.85].
What this Means : This score tells us our predictor is very effective at identifying high-value properties, significantly better than chance. It can help us flag properties with high potential, reducing missed opportunities and wasted effort.



It's also quite good at finding most high-value properties without too many false alarms" succinctly conveys a balance between good recall (finding most high-value properties) and good precision (without too many false alarms).

Precision:



- **Concept:** Think of it as accuracy among the properties your model *flags as high value*. If the model predicts 10 properties are "High Value," precision tells you how many of those 10 were *actually* "High Value" investments.
- **Business Implication:** High precision means fewer "false alarms." The team won't waste time and resources evaluating properties that the model said were great, but turned out to be mediocre. It's about efficiency and avoiding bad investments.

Recall:



- **Concept:** Think of it as the model's ability to find all the actual high-value properties. If there are 100 truly "High Value" properties in the market, recall tells you how many of those 100 your model successfully identified.
- **Business Implication:** High recall means your team won't miss out on good opportunities. You're effectively casting a wide enough net to catch most of the valuable properties. It's about completeness and maximizing opportunities.

Key Drivers of Value

(What Features Matter Most?)



Unlocking Value: The Most Important Property Features



Our "**smart predictor**" doesn't just give us a Yes/No answer; it also tells us *why*.



The model highlights the features that most strongly influence whether a property is "High Value."

Top Influencers: Based on our analysis, the most impactful features for predicting a "High Value" property are:

- Overall Quality: The general condition and finish of the house.
- Above Ground Living Area (Square Feet): The size of the main living space.

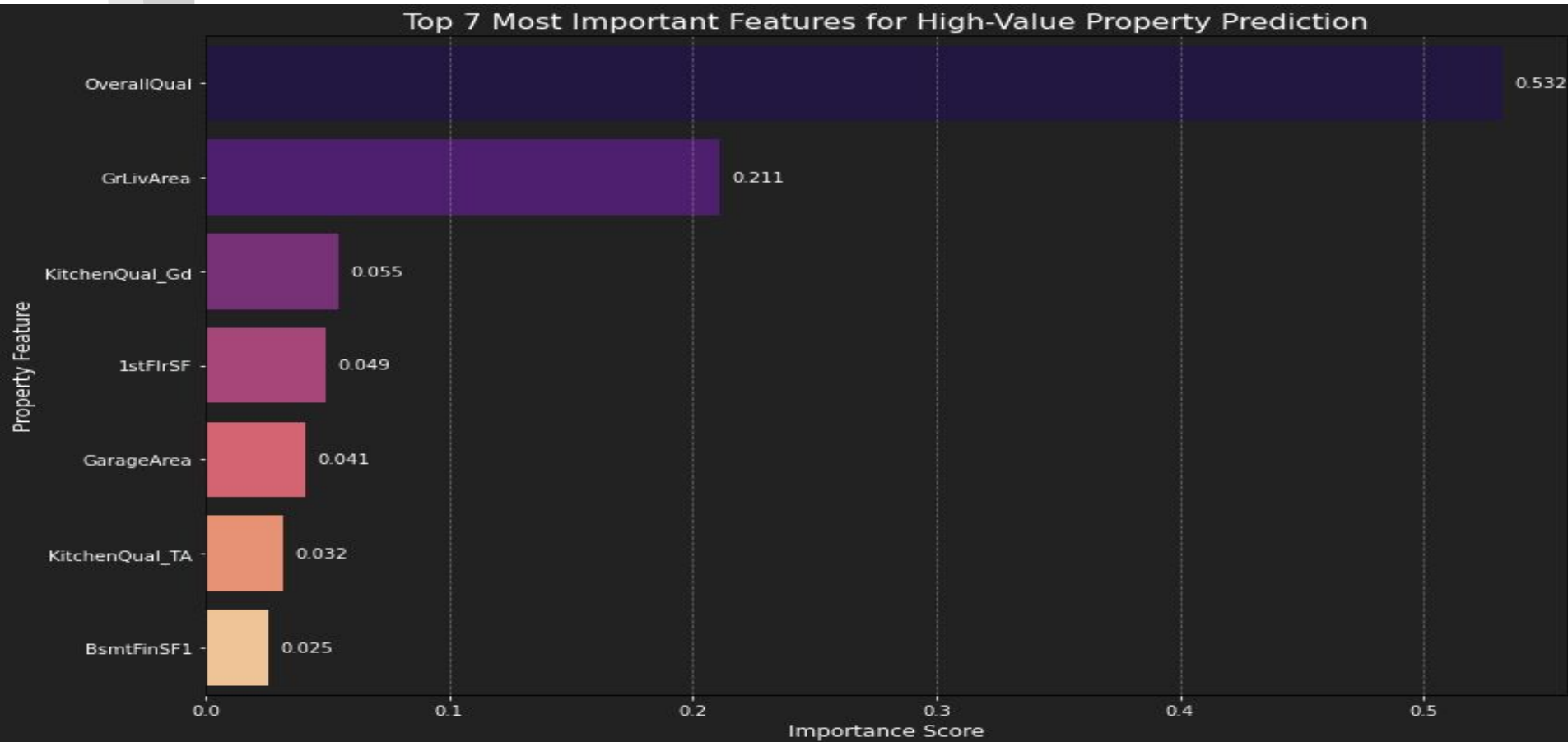


- Garage Capacity (Number of Cars): How many cars the garage can hold.
- Total Basement Square Feet: The size of the basement.
- Year Built: The age of the property.



These are the qualities that our model "learned" are most indicative of a premium investment.

Top 7 Most Important Features for High-Value Property Prediction





Visualization: Top 7 Most Important Features for High-Value Property Prediction

- **Type:** Bar chart showing the top 7 most important features as identified by the final model.
- **Elaboration:** This indicates the visual format: a bar chart.
 - Each bar will represent a specific property characteristic (e.g., "Overall Quality," "Living Area"), and the length of the bar will indicate how "important" that feature was to our predictive model.



- "Important" here means how much that feature helped the model make accurate classifications.

Purpose: To visually highlight the key drivers of property value.

- The core purpose is to provide immediate, actionable insights. By showing which features are most influential, we're giving a clear understanding of what characteristics consistently make a property "High Value" according to historical data. If "Overall Quality" is a top driver, it tells the team that the condition and finishes are paramount for high-value properties.

Recommendation: This visualization directly supports the second recommendation ("Focus on Key Improvement Areas").



- This directly links the visualization to a concrete business action. The second recommendation is about strategically improving properties that might currently be "Standard Value" but have the potential to become "High Value." The bar chart shows "Garage Capacity" as a highly important feature, it directly suggests that investing in adding or improving a garage could be a key strategy to increase a property's value. This connection makes your data analysis highly practical for the investment firm, showing where their renovation efforts could yield the best returns in terms of achieving a "High Value" classification.

Recommendations (Actionable Insights for Investment)

Action Plan: How to Leverage
Our Predictor

Based on our findings, we recommend the following strategic actions:

1. Prioritize Property Sourcing:



Action: Use this predictor as a first filter for incoming property listings.



Benefit: Quickly identify potential "High Value" properties, allowing our team to focus their deep-dive analysis on the most promising leads, saving significant time.

2. Focus on Key Improvement Areas:

Action: For properties we acquire that are currently "Standard Value" but have growth potential, consider targeted renovations.

Benefit: By enhancing features like "Overall Quality" or expanding "Living Area," we can strategically increase a property's likelihood of becoming "High Value" upon resale.



3. Refine Acquisition Criteria:

Action: Integrate these data-driven insights into our standard property acquisition checklist.

Benefit: Be more strategic in what properties we pursue, ensuring they align with the characteristics of historically "High Value" assets.



Limitations (Important Considerations)

What Our Predictor Can and Cannot Do (Yet)

- **Market Dependency:** The predictor learned from past market data. Major shifts in the economy or local real estate trends could affect its accuracy. We should re-evaluate it periodically.
- **Data Availability:** Some detailed property information that the model uses (like exact quality ratings) might not be instantly available for every new listing.
- **"High Value" Definition:** Our "High Value" threshold (top 25%) is a choice. We can adjust this based on your firm's specific investment goals and risk tolerance.
- **Not a Crystal Ball:** This model is a powerful tool to *guide* decisions, not replace expert human judgment and due diligence.

The Road Ahead: Enhancing Our Data-Driven Strategy



- **Pilot Program:** Test the predictor in a small pilot to compare its recommendations with traditional methods.
- **Explore Smarter Predictors:** Investigate even more advanced machine learning techniques that might offer higher accuracy.
- **Dynamic Goals:** Develop ways to adjust the "High Value" definition based on current market conditions or changing investment strategies.
- **Automate Updates:** Set up a system to automatically feed new property data into the predictor to keep it up-to-date.
- **User-Friendly Tool:** Potentially build a simple online tool where your team can enter property details and get instant "High Value" predictions.

Thank you

Lucinda Wanjiru

