Spencer Polans

June 9, 2025

Pipeline Pitch

1. Clearly articulate the business problem or challenge that your model will address:

The business objective would be to help a real estate firm in New York City accurately estimate housing prices for residential homes based on property features and location. This would help agents set competitive listing prices which would 1) reduce time on the market and 2) maximize the seller’s ROI.

1. Identify the specific stakeholders who would benefit from your solution:

The key stakeholders would be real estate agents, investors, and property owners.

1. Define the primary goals of your project in terms of business outcomes

The goal would be to develop a model that would help estimate housing prices in New York City based on property features and location. The business goal would be to help real estate agents and sellers in NYC set competitive, data backed listing prices.

1. Describe the dataset you have chosen/connect the relevancy of the dataset to your business problem

The dataset I have selected is from Kaggle. It includes 84,548 rows and 22 columns before cleaning. The dataset includes sales from a 12 month period (September 2016-2017). The dataset itself directly addresses the core question of “What drives property prices in NYC?”. It includes variables that directly influence pricing decisions such as size, age, location, building type, etc. It also reflects actual sales transactions (once filtering out gifts, etc). All together this provides a strong foundation for a supervised learning model.

**CAVEAT**: While the dataset reflects sales from 2016-2017, our objective is not to forecast current market prices, but to understand the relationship between property characteristics and value. These relationships, from experience, are relatively stable. With access to more recent data, this model could potentially be re-trained for predictive use.

To strengthen the dataset, I also have included NYC census data for median income by ZIP code as well walk score data by zip code. The API limits the number of calls for the free version, but doing it at ZIP level, should provide some socioeconomic insights and context that may improve predictive performance.

1. Explain why a machine learning approach is appropriate for this problem.

Real estate pricing is influenced by complex, non-linear relationships between location, property attribute, and the market itself. A machine learning model is ideal because it can capture complex patterns and interactions among many features, it scales easily across many records, and it should support improvement as data becomes available.

1. Define how you will measure success from both technical and business perspectives.

Success will be measured technically by RMSE, MAE, and R-Squared.

Business success will be achieved by having a model that is able to make predictions within 10% of the actual sales price for at least 80% of the created test set. I will also attempt to identify the top 5 most influential features (sq ft, borough, unit count, etc.).

Preliminary modeling suggests 80% of predictions within 10% may not be feasible with this dataset. I will, however, benchmark model performance using r-squared and refine the business target as I go through this process.

As I continue, I am thinking about trying to first crack a 60% r-squared and then try to reach 70%. Again, the goal is to try to help stakeholders accurately list home prices in an effort to maximize ROI and reduce time on the market.

Part 2: Problem solving process

1. Data Acquisition and Understanding

I will obtain the dataset from Kaggle (which is a slightly cleaned and polished version from the NYC Department of Finance). I will begin by exploring the dataset. This means looking at the structure, column names, and data types (.info(), .describe(), and .isnull().sum()). This will allow me to assess missing values, any invalid entries (zero or negatives), and see which columns might need to be converted to numeric (SALE PRICE) or one hot encoded. In terms of preliminary visualizations, I will use seaborn and matplotlib to see the initial distribution of prices and correlations. I may also explore price by borough or price vs square footage.

Additionally, I may investigate neighborhood income data, crime data, and geographical city data to help support the model.

1. Data Preparation and Feature Engineering

I will remove columns that are non-informative or redundant (Unnamed: 0 and EASE-MENT). Easement is being dropped because there are no values in the column. I will also clean outliers.

I will feature engineer a few new features to start:

Building age (2017- Year Built)

price\_per\_sqft (Sale Price/Gross Square Feet). Rows that have a value of 0 will also be filtered out.

I also will develop a full address line to do geo-mapping for walkscoreapi.

For pipeline implementation, I will use scikit learn to transform numeric and categorical features.

1. Modeling Strategy

I plan to implement at least 3 different modeling algorithms: Linear Regression (as a baseline model), Random Forest, and XGBoost. This seems like a good blend of simplicity, interpretability, and (hopefully) model performance. The same test/train split will be applied across all models (likely 80/20).

I will also use cross\_val\_score to help prevent overfitting.

For hyperparameter tuning, I will use a GridSearch focusing on n\_estimators and max\_depth. Model performance will be measured by using RMSE, MAE, and R-Squared.

1. Results Interpretation and Communication

How will you translate model results to business insights?

I will attempt to identify most impactful features (by feature importance). These results will be used to explain how certain attributes such as square footage, unit count, building type, etc. influence home values.

Visualizations plans for model performance

For visualizations, I will plot bar plots for feature importance, residual plots, and a distribution of error (histogram). These visualizations will help explore patterns and communicate key insights to non-technical stakeholders.

Strategy for explaining technical concepts to non-technical stakeholders.

I will simplify terms into plain language for stakeholders so they can understand my findings. For example, as I am trying to measure by a 10% error, explaining that a 10% margin on a $500,000 home is anywhere between $450,000 to $550,000.

Conceptual Framework: Flowchart

A diagram of a business process

AI-generated content may be incorrect.

Timeline and Scope (10 days ie. 2 work weeks) \*\*\*subject to change\*\*\*

* **Dataset finalization and problem formulation – Day 1** 
  + Dataset acquisition and initial exploration (Kaggle/NYC Open Data)
  + Begin initial exploration (structure, type, missing values)
  + Business problem statement and success metrics
  + Set up Github
* **Exploratory Data Analysis – Day 2**
  + Comprehensive data profiling (get data distributions and outliers)
  + Statistical analysis of relationships (create visualizations to see)
  + Documentation of insights written out in notebook
* **Data Preprocessing – Days 3-4**
  + Data cleaning implementation (drop redundant columns and figure out what to do with nulls)
  + Feature engineering (see above- continue throughout)
  + Pipeline development (Sckit-learn ColumnTransformer)
  + Data splitting (train/validation/test)
* **Model Development – Day 5**
  + Implementation of baseline models (linear regression)
  + Algorithm comparison (XGBoost/Random Forest)
  + Hyperparameter tuning (GridSearchCV)
  + Cross-validation (cross\_val\_score)
* **Model Evaluation and Refinement – Days 6-7**
  + Final model selection (based on newly created features and incorporated datasets)
  + Performance evaluation on test data
  + Business metric calculation
  + Interpretation of results
  + Continue refinement
* **Documentation and Reporting – Day 8-9**
  + Code commenting and cleanup
  + Technical report writing (summarize methods used, EDA, and results)
  + Executive presentation development (create visuals such as feature importance, residual plots, etc.)
  + Begin building slideshow
* **Final Review and Submission – Day 10**
  + Quality assurance of code (make sure everything works)
  + Video recording (record slides and presentation)
  + Final submission preparation

Potential Risks & changes if needed: If external datasets (Walk Score) take longer to integrate or model tuning reveals poor baseline performance, time can shift from reporting (Day 9) back into modeling (Day 6–7) as needed.