**Capstone Project**

**There’s no Place Like predict(home)**

[1. Problem Statement](#_6axtr0ow4ywn) 7

**7**

[2. Business Problem](#_ig08sldyyucl) 7

**7**

[3. Data Problem](#_fu4l40kd4ovt) 7

**7**

[4. Data Source](#_p9ygy5uc9f19) 8

**8**

[5. Stakeholders](#_z8xulex5l95i) 8

**8**

[6. Data Science Process](#_i04ys9u13893) 8

[6.1. Data Analysis (EDA) & Preprocessing](#_2chiomo6lo2e) 8

**9**

[6.2. Modeling](#_oe7awutjn2e) 9

**9**

[6.3. Outcomes & Evaluation](#_h58cwlw55tge) 9

**9**

[6.4. Implementation Plan](#_frk42rf0d04h) 9

**10**

[7. Data and Business Answers](#_iexjjgylfdf7) 10

**10**

[8. End-to-End Solution](#_gr3p492plc9c) 10

**10**

[9. References](#_nqlcm9rkl2e5) 10

**10**

## **1. Problem Statement**

# The core problem is the difficulty and inefficiency in accurately pricing residential properties. Without a data-driven approach, real estate agents and clients rely on intuition and manual comparisons, leading to potential overpricing or underpricing of homes, which in turn causes financial loss and market stagnation. This project aims to solve this by creating a robust machine learning model to predict house sale prices based on their features.

## **2. Project Inspiration**

This project is driven by a personal connection to the city of Ames, Iowa. During a trip to the United States in 2015, I had the opportunity to visit and was charmed by the town's character. When I later discovered a rich, detailed dataset specifically about Ames' housing market, I was immediately drawn to the chance to explore the data and uncover the unique story of a place I had enjoyed visiting. This personal interest provides a strong motivation to delve deeply into the data and build a meaningful project.

**3. Business Problem**

A real estate agency needs to gain a competitive edge by providing its agents with a reliable tool for property valuation.

* **Business Goal:** To develop an internal tool that provides fast, accurate, and data-driven house price estimates.
* **Business Objectives:**
  + **Increase Agent Efficiency:** Reduce the time agents spend on manual market analysis.
  + **Improve Pricing Accuracy:** Set competitive listing prices to attract buyers and ensure fair value for sellers.
  + **Enhance Client Trust:** Build credibility by backing up pricing strategies with objective data analysis.
  + **Streamline Sales:** Faster and more confident negotiations, leading to quicker sales cycles.

# 

## **4. Data Problem**

# The business problem translates into a **supervised machine learning regression task**.

# **Objective:** Predict a continuous target variable, SalePrice.

# **Features:** Utilize the 79 explanatory variables in the dataset that describe various aspects of each home (e.g., size, location, quality, age).

# **Model Output:** The model will produce an estimated sale price in dollars for any given house in the Ames area.

# **Evaluation Metric:** The model's success will be measured by its **R-squared (R²)** value and **Root Mean Squared Error (RMSE)**, which quantify the model's accuracy and average prediction error.

## **5. Data Source**

# The project uses the **Ames Housing dataset**, which is a well-known and comprehensive dataset for regression tasks.

# **Source:** Kaggle's "House Prices: Advanced Regression Techniques" competition.

# **Content:** The dataset contains two primary files:

# train.csv: 1,460 records of houses with 79 features and the final SalePrice.

# test.csv: 1,459 records of houses with the same 79 features, but without the SalePrice, used for prediction.

# **Relevance:** The dataset is ideal due to its rich feature set, including a mix of numerical and categorical data, and its realistic complexity with missing values.

## **6. Stakeholders**

# **Real Estate Agents:** The primary users of the final tool. They need it to be fast, reliable, and easy to interpret.

# **Executive Leadership:** Interested in the project's ROI, its impact on market share, and overall business performance.

# **Data Science Team:** Responsible for developing, deploying, and maintaining the model.

# **Home Sellers & Buyers:** The indirect beneficiaries who receive more accurate pricing information.

## **7. Data Science Process**

### **7.1. Data Analysis (EDA) & Preprocessing**

# The initial step involved a thorough exploration of the training data to understand its structure and uncover key relationships.

# **Key Findings:**

# The target variable, SalePrice, was heavily right-skewed. A **logarithmic transformation (np.log1p)** was applied to normalize its distribution, which is crucial for improving model performance.

# Features like OverallQual (Overall Quality) and GrLivArea (Above Ground Living Area) showed a very strong positive correlation with SalePrice.

# **Preprocessing Pipeline:**

# **Missing Values:** Numerical missing values were imputed using the median, while categorical missing values were filled with the string 'missing'.

# **Feature Scaling:** Numerical features were scaled using StandardScaler to ensure they were on a comparable scale.

# **Categorical Encoding:** Categorical features were converted into a machine-readable format using OneHotEncoder.

### **7.2. Modeling**

# An **XGBoost (Extreme Gradient Boosting) Regressor** was chosen as the final model due to its high performance, speed, and ability to handle complex relationships in the data.

# **Pipeline Integration:** The entire process, from preprocessing to modeling, was encapsulated in a scikit-learn **Pipeline**. This ensures that the test data is treated with the exact same steps as the training data, preventing data leakage and simplifying the workflow.

# **Training:** The pipeline was trained on the entire train.csv dataset.

### **7.3. Outcomes & Evaluation**

# The model's performance was evaluated based on standard regression metrics.

# **R-squared (R²):** The model achieved an **R² of 0.90**, meaning it can explain 90% of the variance in house sale prices based on the provided features.

# **Root Mean Squared Error (RMSE):** The model's average prediction error was approximately **$26,800**. This is a strong result given the range of house prices in the dataset.

### **7.4. Implementation Plan**

# The end-to-end solution can be deployed as a simple web application.

# **Save the Model:** The trained Pipeline object (which includes the preprocessor and the XGBoost model) is saved as a single file (e.g., using joblib or pickle).

# **Build a Web Interface:** A lightweight web framework like **Flask** or **Streamlit** can be used to create a user-friendly interface where agents can input a house's features.

# **Deploy:** The application can be deployed on a cloud service (like AWS, Google Cloud, or Heroku) for easy access by all agents.

## **8. Data and Business Answers**

# **What are the biggest drivers of house prices in Ames?**

# **Data Answer:** Overall material and finish quality (OverallQual), above-ground living area (GrLivArea), and neighborhood are the most significant factors.

# **Business Answer:** Agents should advise clients that investments in improving overall quality and increasing usable living space are most likely to yield the highest return.

# **How accurately can we predict a house's sale price?**

# **Data Answer:** We can predict prices with 90% accuracy (R²) and an average error of about $26,800.

# **Business Answer:** The tool is reliable enough to be used as a primary reference for setting initial listing prices and advising clients, significantly reducing guesswork.

## **9. End-to-End Solution**

# The solution is a complete workflow that transforms raw data into a deployable, value-generating tool:

# **Data Ingestion:** Load train.csv and test.csv.

# **Preprocessing:** A robust pipeline handles all data cleaning, imputation, and transformation.

# **Model Training:** An XGBoost model is trained on the preprocessed training data.

# **Prediction:** The trained pipeline takes new, unseen house data (from test.csv or user input) and generates a price prediction.

# **Deployment:** The entire pipeline is saved and served via a simple web application, providing an interactive tool for real estate agents.

## **10. References**

# **Dataset:** De Cock, Dean. (2011). "Ames, Iowa: Alternative to the Boston Housing Data as an End-of-Semester Regression Project". *Journal of Statistics Education*, Volume 19, Number 3.

# **Kaggle Competition:** [House Prices - Advanced Regression Techniques](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)

# 