🏡 PROJECT REPORT: HOUSE PRICE PREDICTION

---

1. Introduction

Predicting house prices is a classic regression problem where various features such as the number of bedrooms, bathrooms, square footage, and more influence the selling price of a house. This project focuses on creating an optimized **machine learning model** that accurately predicts house prices.

We implemented several models and compared their performances, ultimately selecting **LightGBM** for its superior speed and accuracy.

---

2. Objective

To build a regression model that:

1. Accurately predicts house prices based on given features.
2. Provides real-time predictions through an interactive web application.

---

3. Dataset

The dataset contains the following key features:

- **Numerical Features**: sqft_living, sqft_lot, sqft_above, sqft_basement, etc.
- **Categorical Features**: city, statezip, waterfront, etc.
- **Target Variable**: price (house selling price).

---

4. Why LightGBM?

We tested multiple machine learning models, including **Linear Regression**, **Random Forest**, and **XGBoost**, before selecting **LightGBM** as our final model.

Reasons for Choosing LightGBM

1. **Handles Large Datasets**: LightGBM efficiently handles datasets with many rows and features.
2. **Speed**: It is faster than Random Forest and XGBoost due to its histogram-based approach.
3. **Accuracy**: It achieves high accuracy, especially on structured data.

4. **Supports Feature Importance**: Provides insights into which features most influence predictions.
5. **Robustness**: Handles missing values and categorical data effectively.

---

5. Model Comparison

Model Performance Metrics

We evaluated models based on **Mean Absolute Error (MAE)**, **Root Mean Squared Error (RMSE)**, and **R² Score**.

Model	MAE (↓)	RMSE (↓)	R² Score (↑)	Observations
Linear Regression	$207,555.79	$992,929.75	0.033	Poor fit, assumes linear relationships that don’t exist in the data.
Random Forest	$165,051.21	$988,778.64	0.041	Better than Linear Regression but prone to overfitting with outliers.
XGBoost	$154,132.43	$980,645.02	0.057	Strong performance but slower training compared to LightGBM.
LightGBM	**$98,754.94**	**$163,268.78**	**0.6846**	Fastest and most accurate model, robust to outliers and efficient with memory.

---

6. Feature Importance

LightGBM provides a ranking of feature importance, which helps us understand the factors driving house prices. Below are the most influential features:

1. **sqft_living**: Total living area in square feet.
2. **statezip**: Location of the property.
3. **house_age**: Age of the house in years.
4. **city**: The city in which the property is located.
5. **sqft_above**: Above-ground square footage.

These insights can help prioritize features in future iterations of the model.

---

7. Conclusion

1. **LightGBM is the best model** for this project due to its superior performance across all metrics.
2. It provides the best balance of **speed, accuracy, and robustness** for the given dataset.
3. The **Streamlit app** makes predictions accessible in real-time for users with an intuitive interface.

---

8. Future Work

1. **Hyperparameter Tuning**: Further fine-tune the LightGBM model to improve R^2 score.
2. **Geospatial Features**: Incorporate geographic coordinates for better location-based predictions.
3. **Outlier Detection**: Automatically detect and handle extreme outliers for better generalization.
4. **Enhanced UI**: Add interactive visualizations to the web app for better insights.

---

9. Acknowledgments
