Advanced House Price Prediction

A Machine Learning Approach Using LightGBM

Presented By : Anirban Biswas

Problem Statement & Objectives

The Challenge

To accurately predict the final sale price of residential homes within King County, USA, addressing the inherent complexity and variability in the real estate market.

Our Objective

Develop a high-performance regression model utilising advanced feature engineering and the powerful LightGBM algorithm for superior predictive capability.

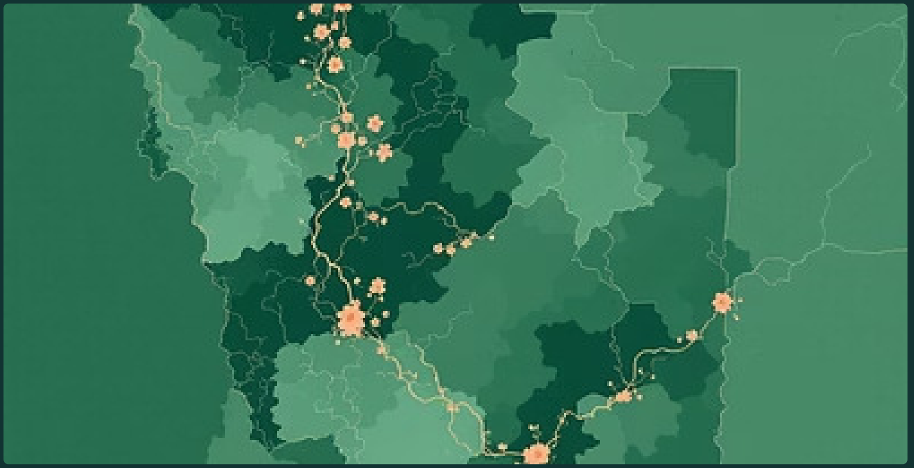
Key Goal

Provide actionable insights into the pivotal factors that significantly influence house prices, thereby aiding stakeholders in informed decision-making.

The Dataset: King County House Data

Our analysis is grounded in the **kc_house_data.csv** dataset, a comprehensive collection of residential property listings from King County, USA.

- Dataset Size: Comprising **21,613 individual property listings**, providing a robust foundation for model training.
- Key Features: Includes essential attributes such as **price, number of bedrooms, living area (sqft_living), construction grade, geographical coordinates (latitude and longitude), and year built (yr_built)**.
- Target Variable: The **price** of the property, which our model aims to accurately predict.



This rich dataset facilitates the exploration of diverse factors impacting residential property valuations.

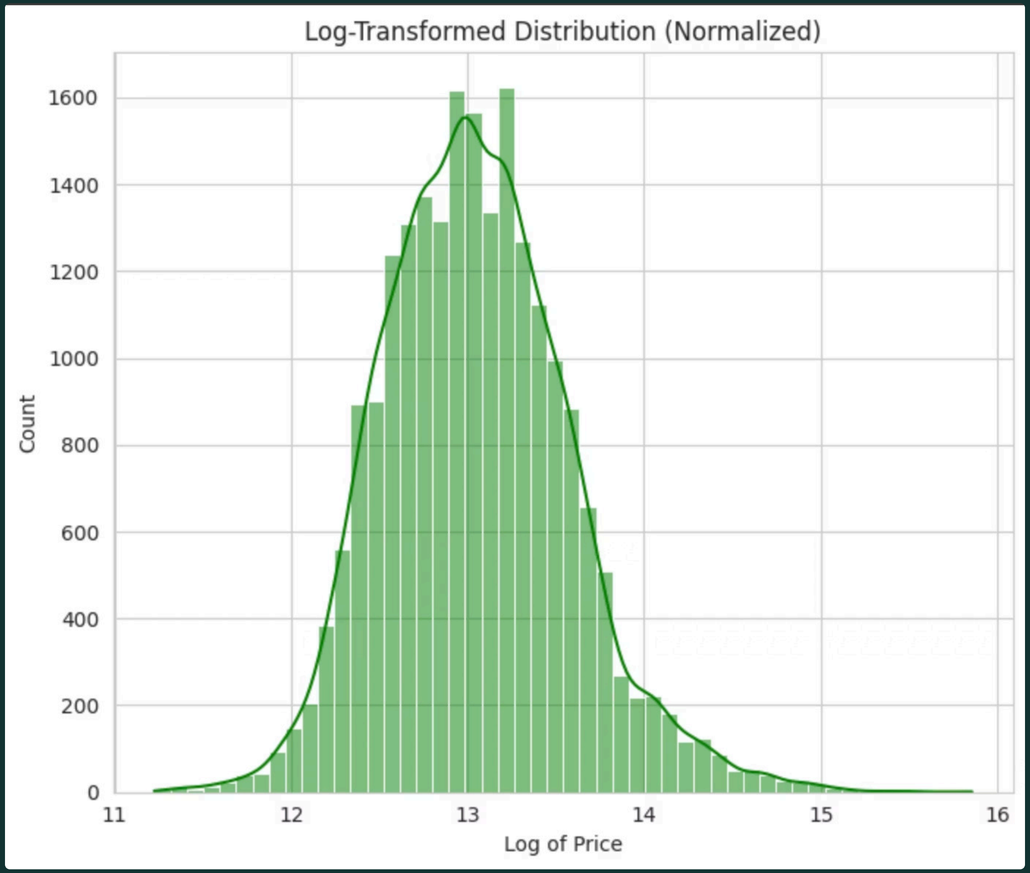
Exploratory Data Analysis (EDA)

During our Exploratory Data Analysis, a critical observation was made regarding the distribution of the target variable.

Original Price Distribution

The original 'price' variable exhibited a pronounced right-skewed distribution, indicating a higher frequency of lower-priced homes and a long tail of very expensive properties. This non-normal distribution can negatively impact model performance.

Log-Transformed Price Distribution



To address this, a log-transformation (specifically, `log1p`) was applied. This operation effectively normalized the distribution, leading to a more symmetric, Gaussian-like shape, which is often preferred for linear models and gradient boosting algorithms.

This transformation is crucial for stabilising variance and improving the model's ability to learn relationships between features and the target.

Data Preprocessing & Feature Engineering

Rigorous preprocessing and strategic feature engineering were undertaken to enhance the dataset's predictive power.

Data Cleaning

The non-predictive `id` column was meticulously dropped to remove irrelevant identifiers and streamline the dataset for modelling.

Feature Creation

New, more informative features were engineered to capture nuances in the data, improving the model's contextual understanding.

Temporal Aspects

Extracted `sale_month` and `sale_year` from the `date` feature to capture seasonal and annual market trends.

Property Attributes

Derived `house_age` from `yr_built` and `was_renovated` from `yr_renovated` to reflect property lifecycle impacts.

Model Selection: LightGBM Regressor

LightGBM was selected as the core regression algorithm due to its superior characteristics in handling complex datasets.

High Performance

As a gradient boosting framework, LightGBM is renowned for its exceptional speed and accuracy, delivering state-of-the-art results for regression tasks.

Optimised Efficiency

It efficiently handles large datasets with minimal computational resources, thanks to its leaf-wise tree growth algorithm and optimised data structures.

Enhanced Robustness

LightGBM exhibits strong generalisation capabilities and is less susceptible to overfitting, especially when integrated with techniques like early stopping, ensuring reliable predictions.

Model Training & Evaluation Strategy

A meticulous strategy was employed for model training and subsequent performance evaluation.

01

Data Partitioning

The dataset was rigorously split, with **80% allocated for training** the LightGBM model to learn underlying patterns.

02

Held-Out Validation

The remaining **20% was reserved as a held-out validation set** to impartially assess the model's generalisation capabilities.

03

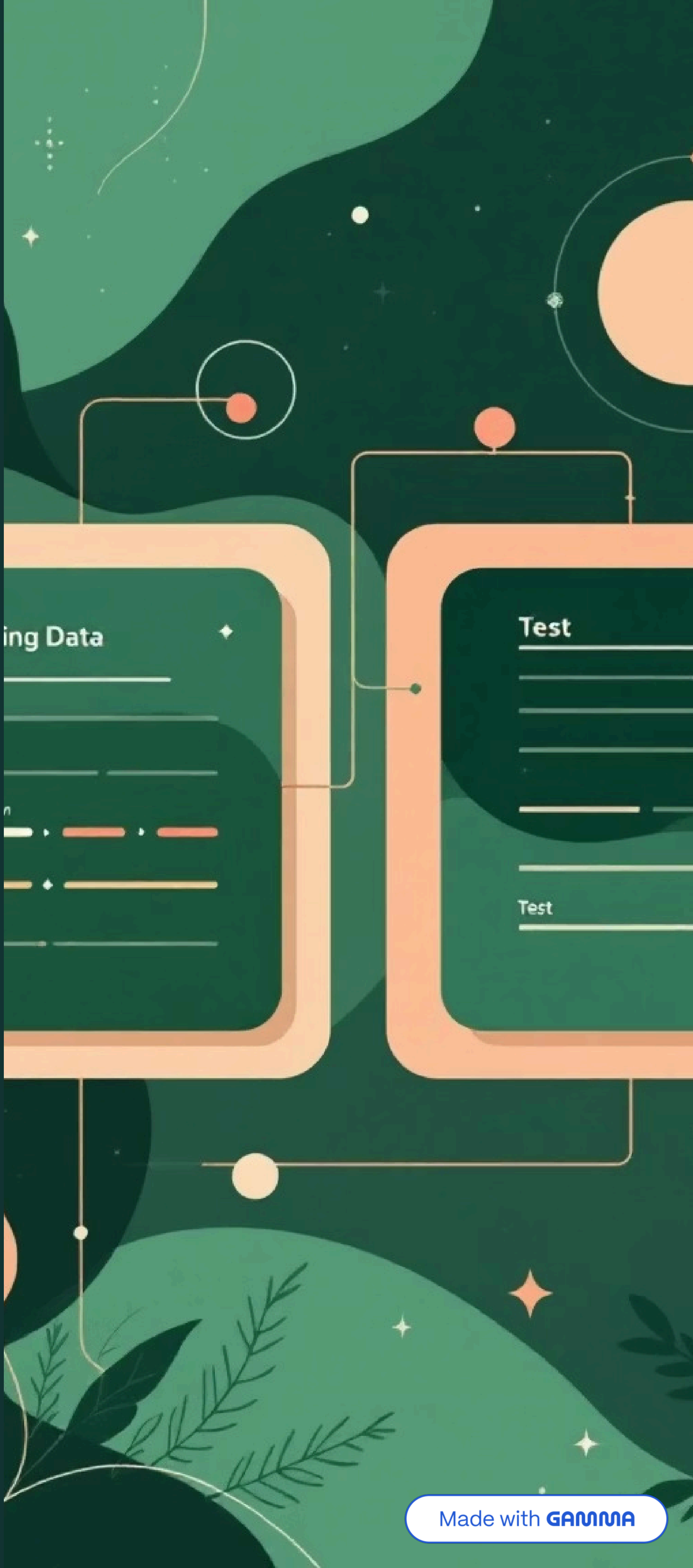
Overfitting Prevention

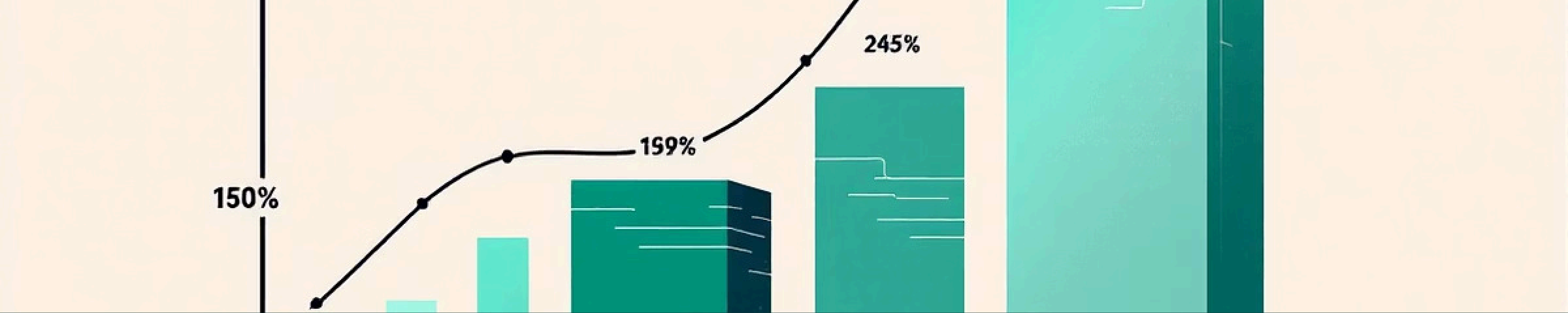
Early stopping was implemented to prevent overfitting; training ceased when validation performance no longer showed improvement.

04

Performance Metrics

Model efficacy was quantified using **Root Mean Squared Error (RMSE)** for error magnitude and **R-squared (R^2)** for variance explanation.





Results: Model Performance

The LightGBM model demonstrated commendable performance in predicting house prices.

0.88

R-squared (R^2)

The model achieved an R^2 score of **0.88**, indicating a very strong predictive fit.

88%

Interpretation

Our model successfully explains **88% of the variance** in house prices.

0.1656

Root Mean Squared Error (RMSE)

The average prediction error on the log-transformed price scale is approximately **0.1656**.

This high predictive capability makes the model a valuable tool for real estate analysis.

Key Insights: What Drives House Prices?

Our model identified several influential factors dictating property values, offering crucial market understanding.

Living Space (sqft_living)

The overall living area of a property consistently emerged as the single most dominant predictor of price.

Construction Grade (grade)

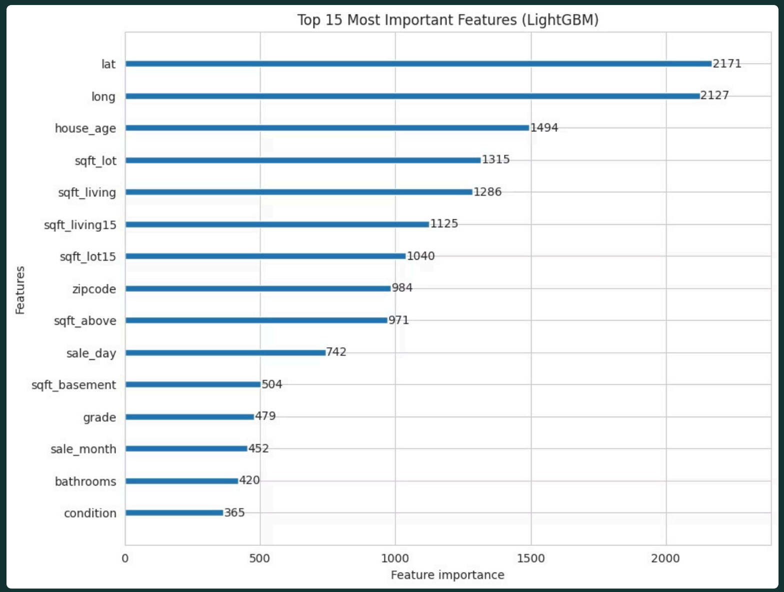
The quality of construction and design, represented by the grade feature, significantly impacts property valuation.

Geographical Location (lat & long)

The precise latitude and longitude of a property are vital, capturing the influence of neighbourhood, amenities, and school districts.

Property Age (house_age)

The age of the house, derived during feature engineering, also plays a notable role, reflecting wear-and-tear or historical value.



This visual representation highlights the relative importance of each feature in the model's predictive process.

Conclusion & Next Steps

We have successfully developed and validated a robust house price prediction model, providing a strong foundation for future advancements.

Project Conclusion

A high-accuracy LightGBM model has been successfully constructed, reliably predicting house prices and discerning key market drivers within King County.

External Data Integration

Incorporate additional external datasets, such as school ratings, crime rates, and local economic indicators, to enrich predictive power.

Model Experimentation

Explore and evaluate other advanced machine learning models, including CatBoost, XGBoost, or deep learning architectures, for potential performance gains.

Thank you for your attention. Questions & Discussion