University of Trieste
Architecture and Engineering Department

Master Degree in Computer Science

Automated Property Value Prediction System

Stefano Chen
IN2000246

Introduction

Problem:

- ❑ Accurate property valuation is vital
- ❑ California real estate agency relies on manual time-consuming expert assessment (~20% error margin)

Project Goals:

- ❑ Build an ML-based system to improve valuation accuracy
- ❑ Using MLOps practices: versioning, tracking, deployment, monitoring and alerting

Tools

Jira

- is a project management and issue-tracking software
- supports different Agile Development frameworks (e.g. Karban)

GitHub

- is a web-based platform for version control and collaboration
- supports CI/CD pipelines (GitHub Actions)

Comet ML

- is a platform designed to simplify MLOps workflows
- offers many essential features (e.g. Experiment Tracking, Artifact Versioning and Model Registry)

Streamlit

- is an open-source Python library to build interactive web applications
- easy deployment on Streamlit Community Cloud

MongoDB Atlas

- is a fully managed cloud database service
- deploy, manage, scale and secure MongoDB database in the cloud

Our Approach to solve the Problem

1. Software Requirements

Functional Requirements:

Defines what a product must offer

- web interface for user inputs
- return a predicted value with an associated confidence score
- log all the user input and prediction result

Non-Functional Requirements:

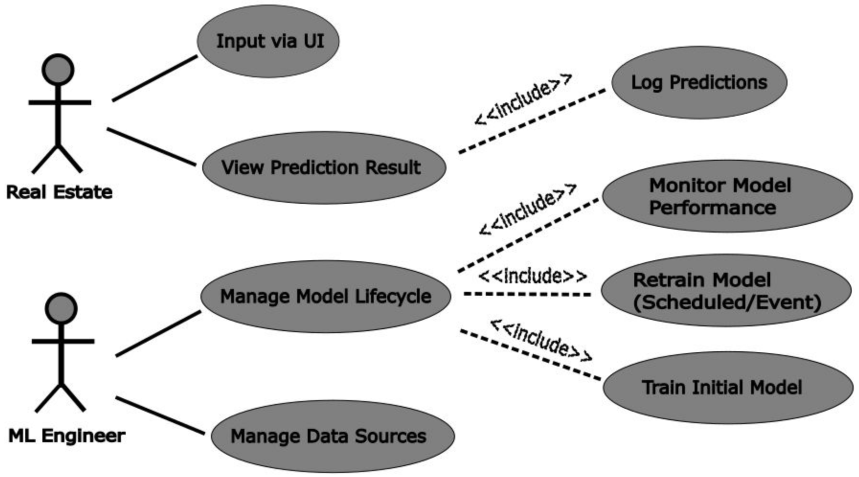
Are the quality constraints that the system must satisfy

- maintain a prediction error rate below 20%
- return prediction results within 2 seconds
- deployed in a secure, scalable cloud environment

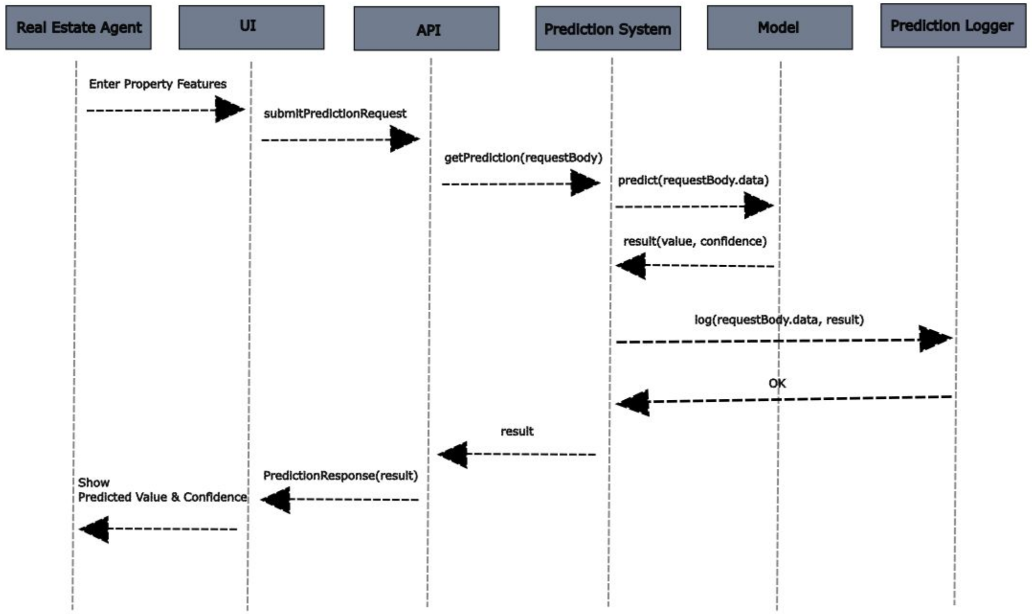
2. System Modeling

Involves creating abstract representation of the system from different perspectives, it usually use UML (Unified Modeling Language).

Use Case Diagram

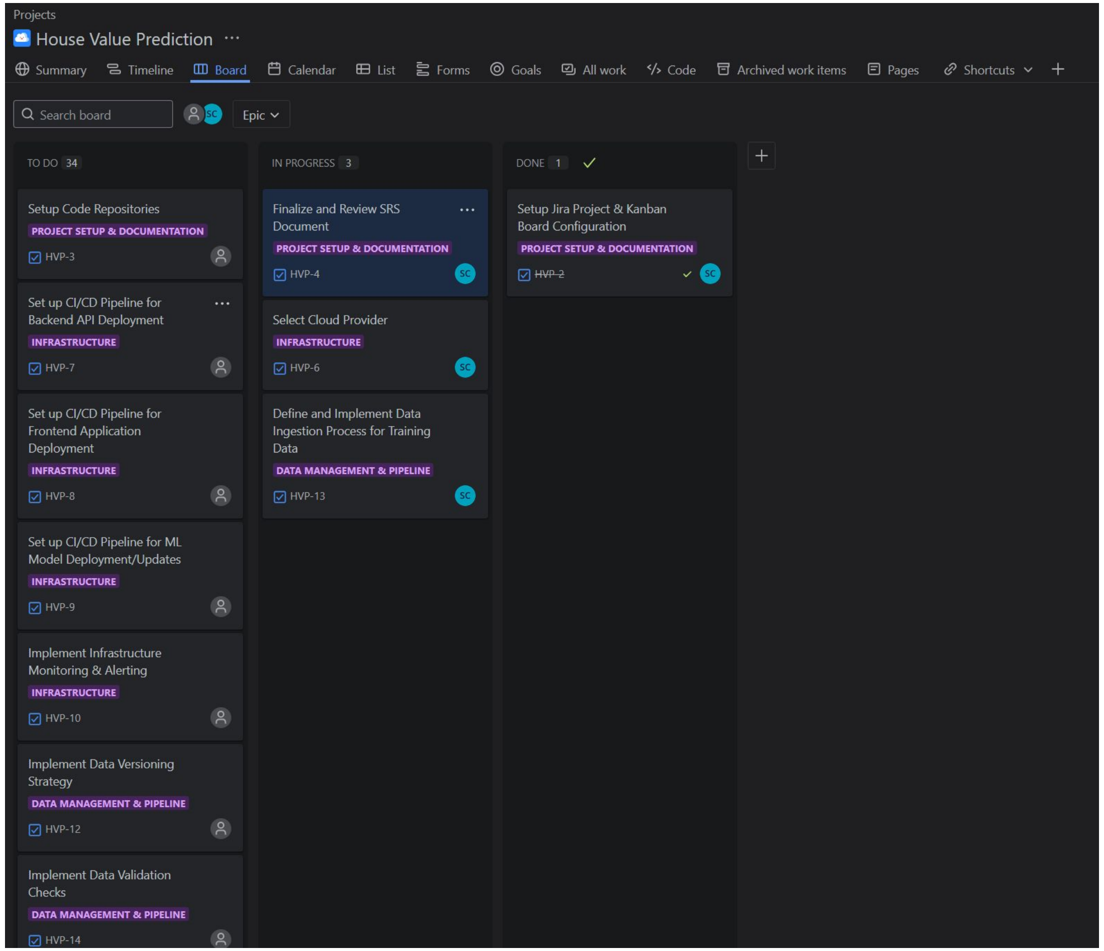


Sequence Diagram

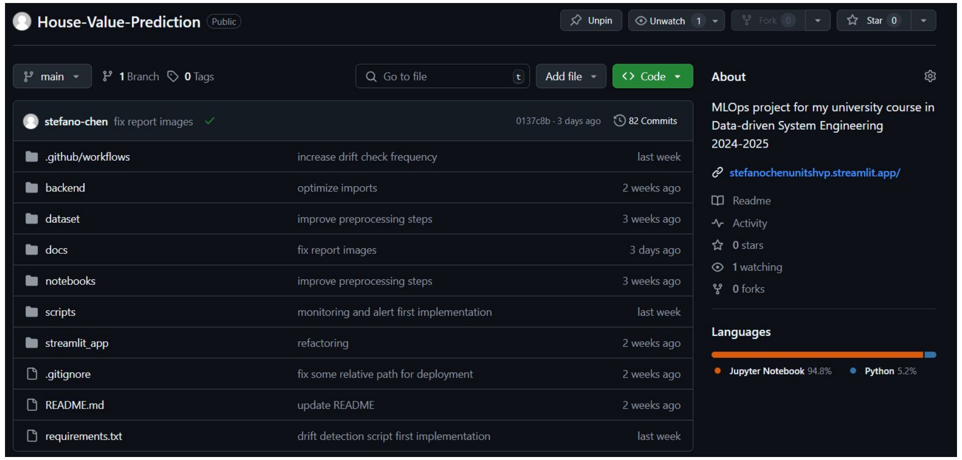


3. Setup Tools

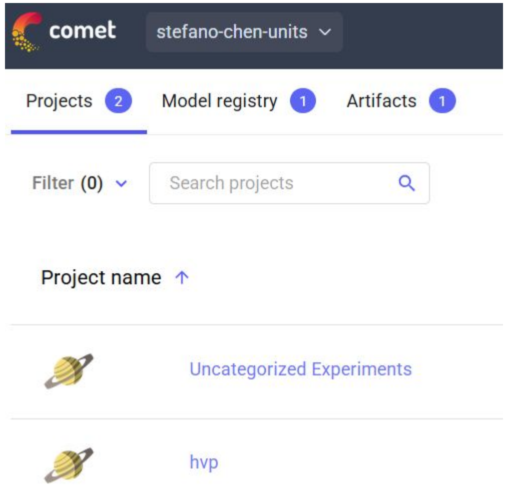
Jira



GitHub



Comet ML



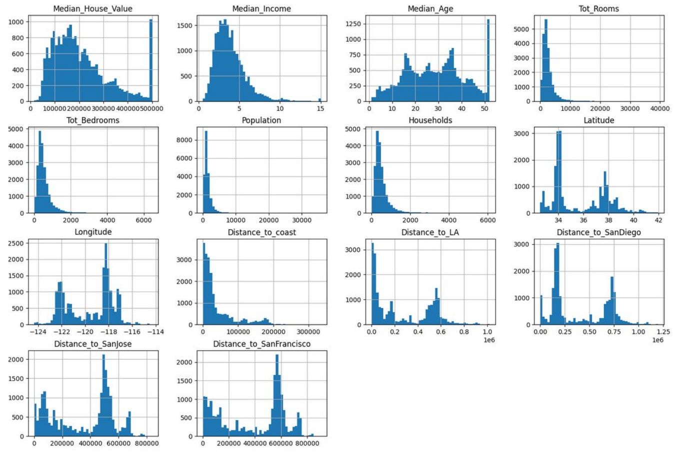
4. Exploratory Data Analysis (EDA)

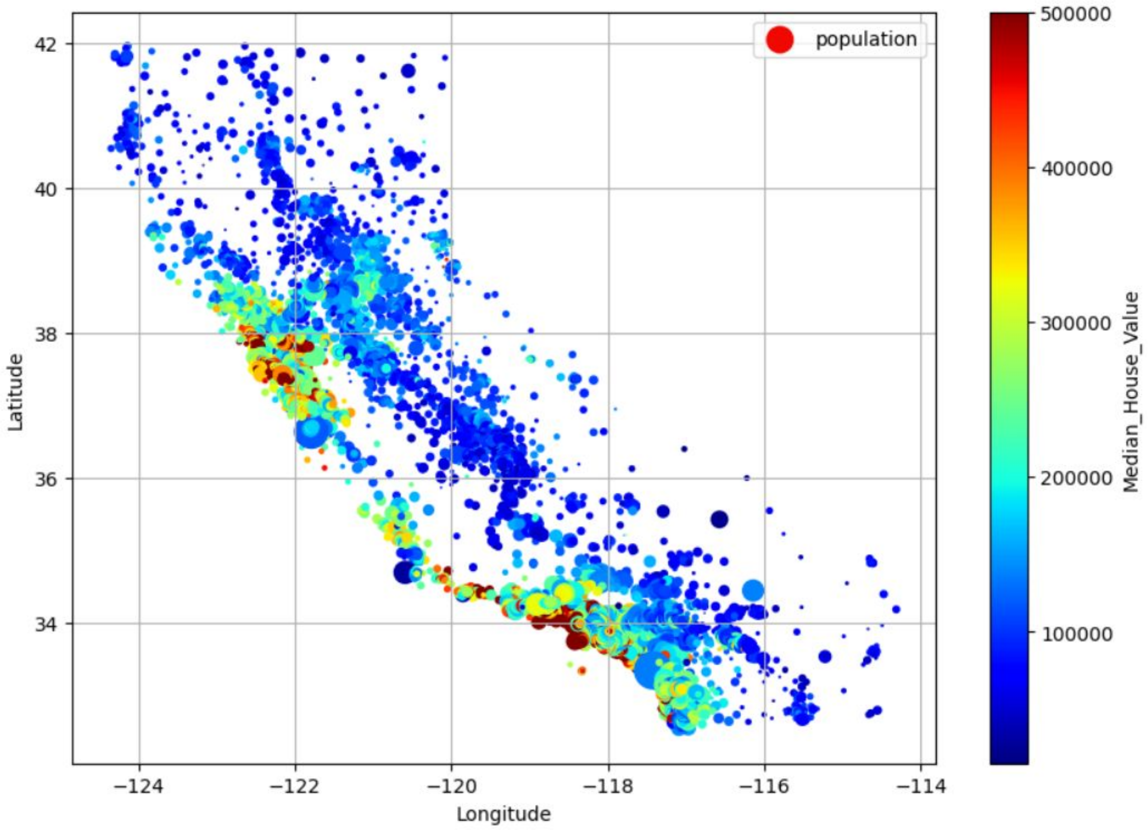
Dataset

20640 entries with 14 columns

Feauture Name	Description
Median_House_Value	Median house value for household within a block (measured in USD) [prediction target]
Median_Income	Median income for households within a block of houses (measured in tens of thousands of USD) [10k$]
Median_Age	Median age of a house within a block; a lower number is a newer building [years]
Tot_Rooms	Total number of rooms within a block
Tot_Bedrooms	Total number of bedrooms within a block
Population	Total number of people residing within a block
Households	Total number of households, a group of people residing within a home unit, for a block
Latitude	A measure of how far north a house is; a higher value is farther north [°]
Longitude	A measure of how far west a house is; a higher value is farther west [°]
Distance_to_coast	Distance to the nearest coast point [m]
Distance_to_LA	Distance to the centre of Los Angeles [m]
Distance_to_SanDiego	Distance to the centre of San Diego [m]
Distance_to_SanJose	Distance to the centre of San Jose [m]
Distance_to_SanFrancisco	Distance to the centre of San Francisco [m]

Feature Distribution





5. Feature Engineering

Correlation Analysis (Before)

Median_House_Value	1.000000
Median_Income	0.688075
Tot_Rooms	0.134153
Median_Age	0.105623
Households	0.065843
Tot_Bedrooms	0.050594
Population	-0.024650
Distance_to_SanFrancisco	-0.030559
Distance_to_SanJose	-0.041590
Longitude	-0.045967
Distance_to_SanDiego	-0.092510
Distance_to_LA	-0.130678
Latitude	-0.144160
Distance_to_coast	-0.469350

Name: Median_House_Value, dtype: float64

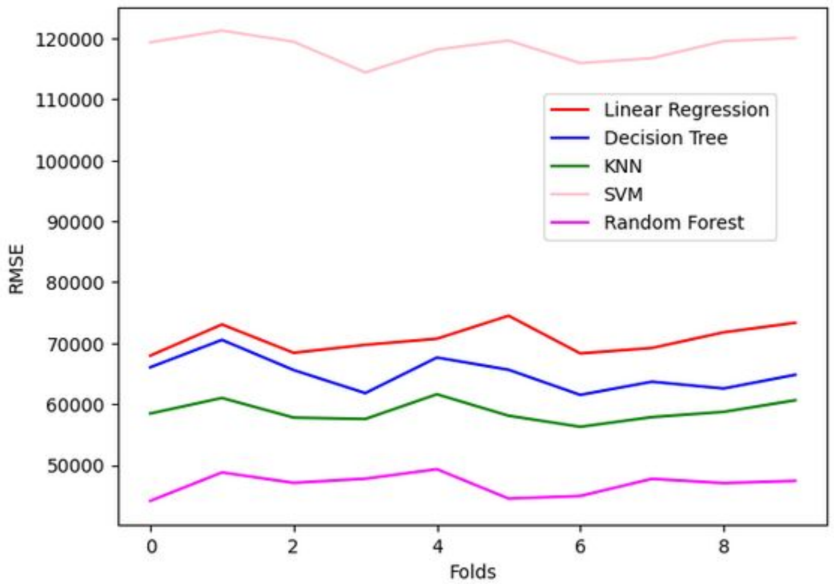
Correlation Analysis (After)

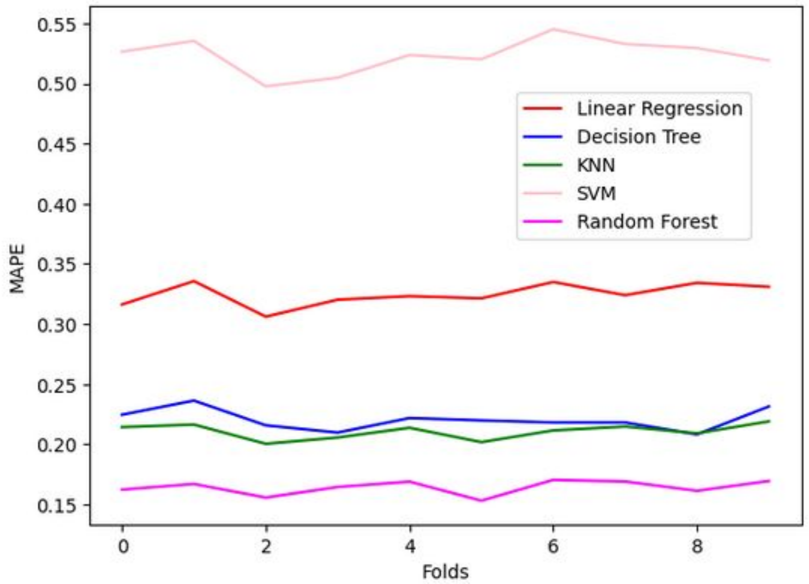
Median_House_Value	1.000000
Median_Income	0.688075
Rooms_Per_House	0.151948
Tot_Rooms	0.134153
Median_Age	0.105623
Households	0.065843
Tot_Bedrooms	0.050594
People_Per_House	-0.023737
Population	-0.024650
Distance_to_SanFrancisco	-0.030559
Distance_to_SanJose	-0.041590
Longitude	-0.045967
Distance_to_SanDiego	-0.092510
Distance_to_LA	-0.130678
Latitude	-0.144160
Bedrooms_Ratio	-0.255624
Distance_to_coast	-0.469350

Name: Median_House_Value, dtype: float64

6. Preprocessing Pipeline

7. Model Selection (10 fold-CV)





8. Hyperparameter Tuning (Optuna)

Search Space:

- n_estimators: 100 - 300
- criterion: [“squared_error”, “absolute_error”, “friedman_mse”, “poisson”]
- min_samples_leaf: 1 - 50

After 100 trials (all tracked in Comet ML)

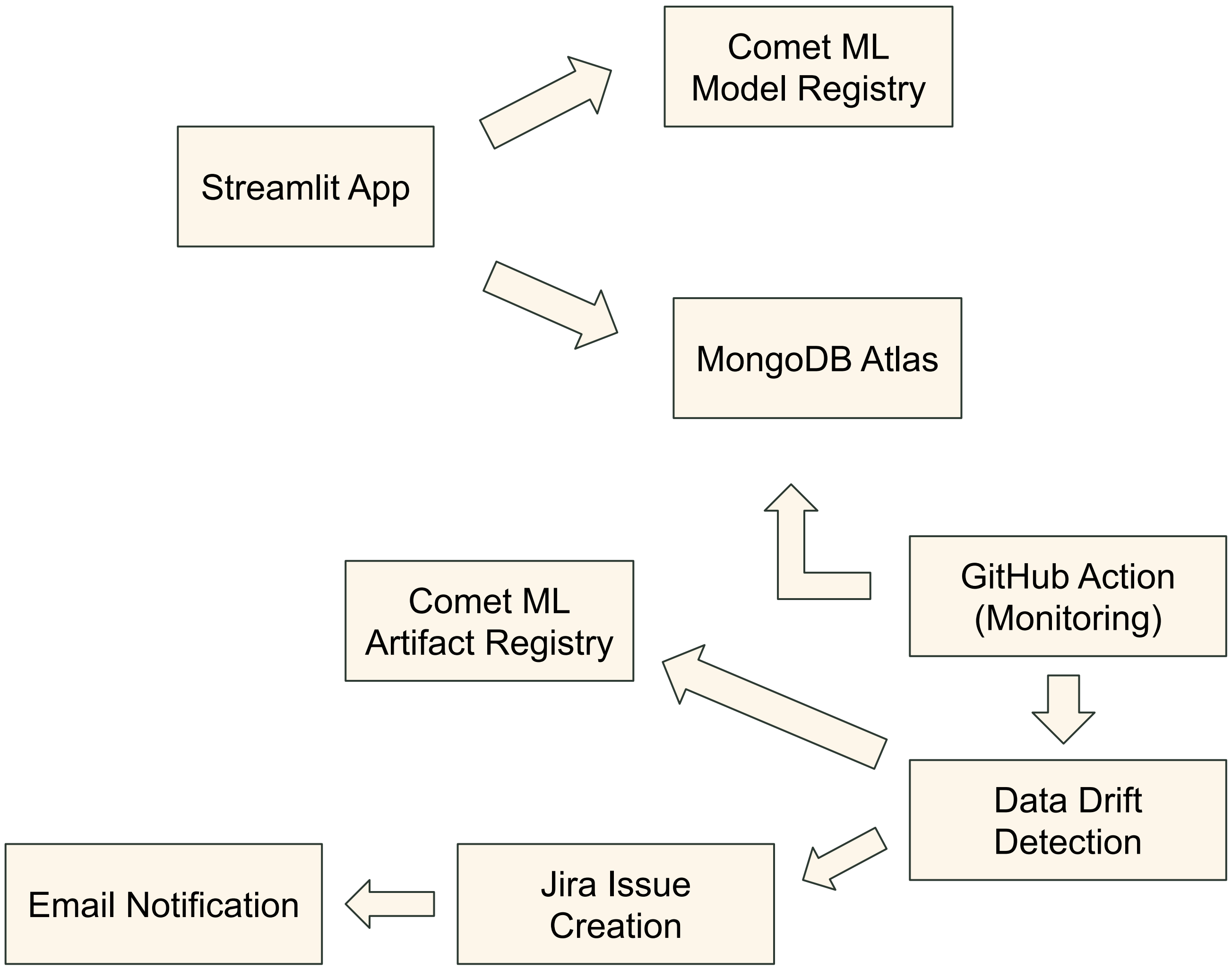
Best Hyperparameter

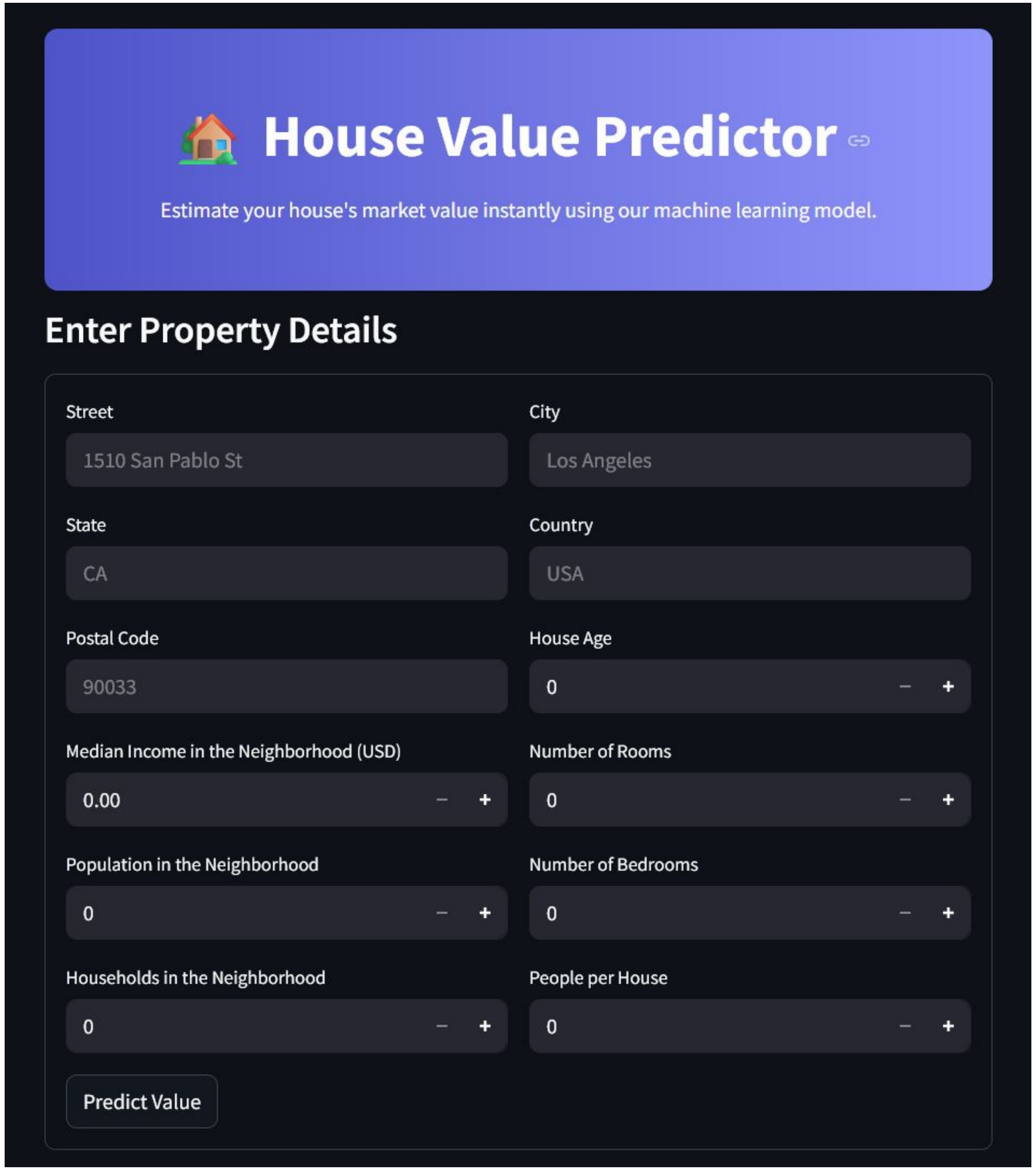
Hyperparameter	Value
Number of Estimators	135
Criterion	“poisson”
Min Samples per Leaf	2

Metrics Value

Metric	Value
RMSE	46395.6
MAPE	16.4%
MAE	29509.7
R2	83.6%

9. Deploy, Monitoring, Alert





Final Thoughts

Limitations:

1. lacks an automated retraining mechanism
    - ❏ delayed availability of ground truth labels (i.e. actual property sale prices)
    - ❏ must perform manual retraining once new labeled data becomes available
2. marginally performance improvement w.r.t. the manual estimations
    - ❏ most likely due to the low number of tuning trials

Conclusion:

We developed and deployed data-driven machine learning system to predict houses prices in California. The model achieved acceptable performance, allowing the real estate agency to shift resources away from the labor-intensive task of manually estimating house prices.

Thank You

for your attention