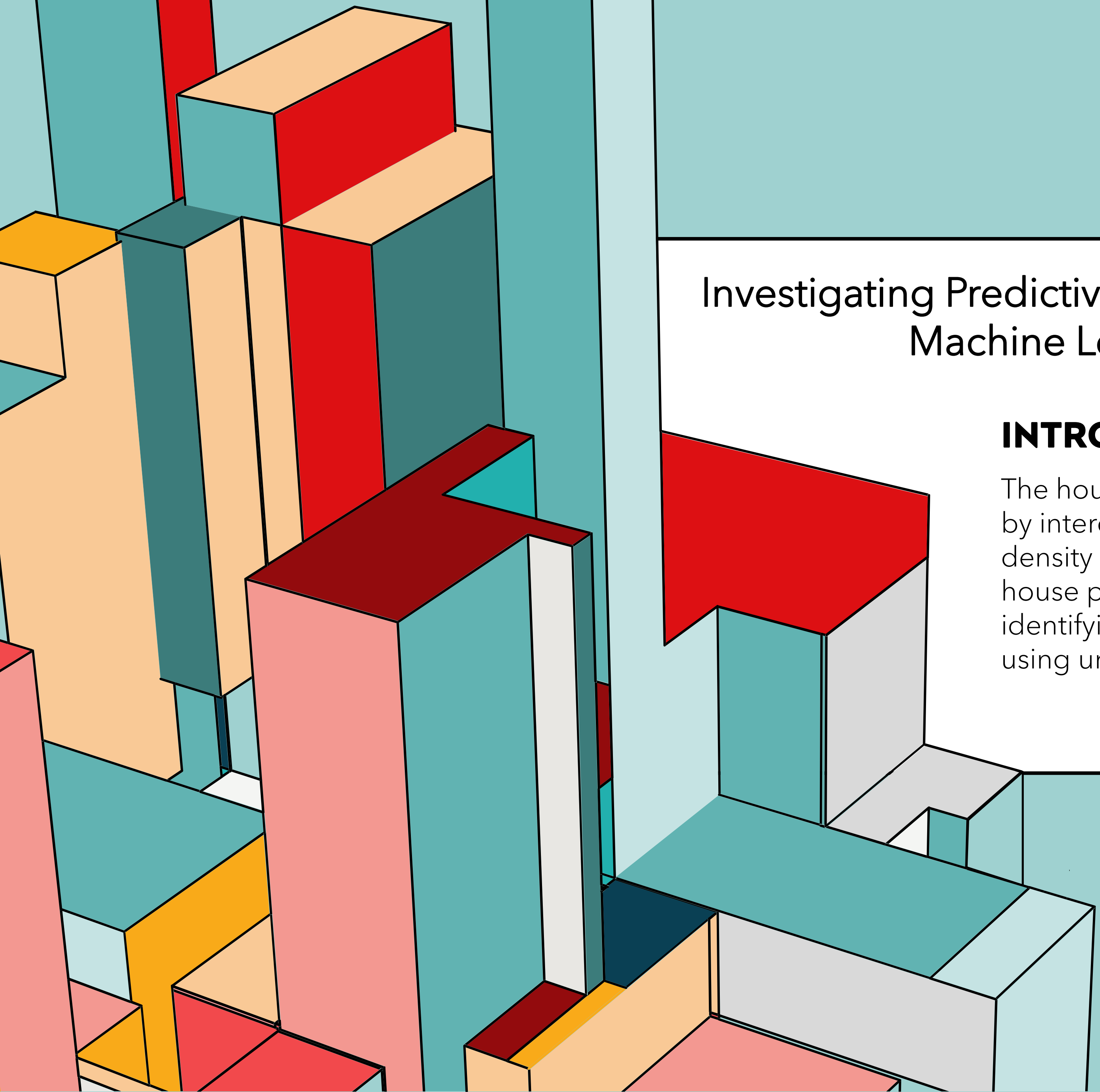


Investigating Predictive Techniques in House Prices using Machine Learning (ML) algorithms

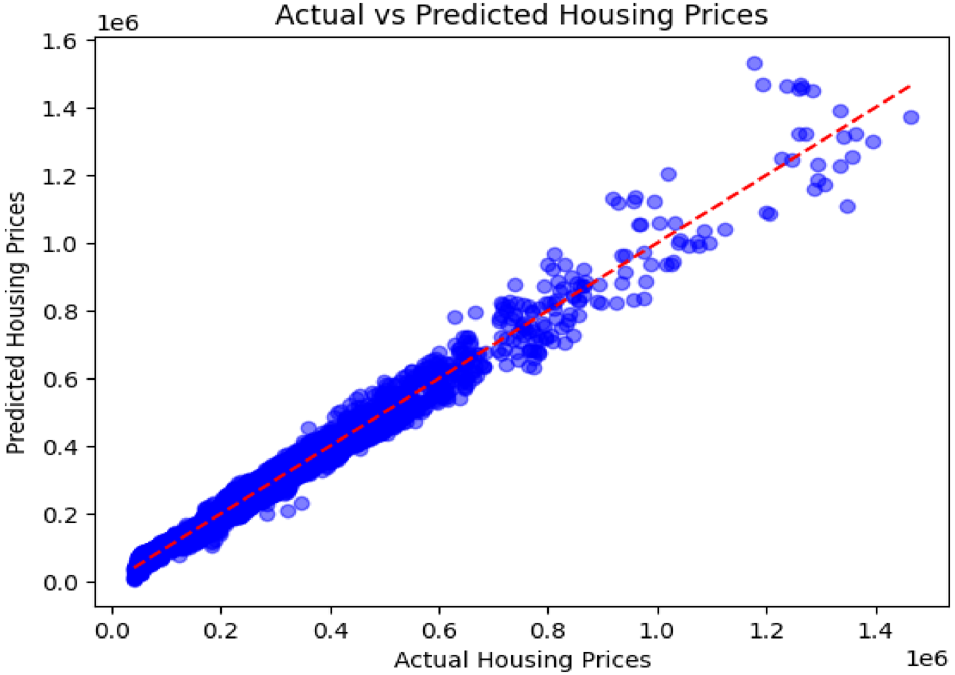
INTRODUCTION

The housing market is a complex ecosystem influenced by interconnected variables such as income, population density ratio, and local economic conditions (ex. Average house price). Predicting house prices, it's all about identifying the best patterns throughout constant testing, using uniqueness and clarity.

POLYNOMIAL REGRESSION

Degree	RMSE	MAE	Model Score(R2)
Degree 1	76017.96	51550.29	0.8432
Degree 2	32298.93	24354.41	0.9717
Degree 3	30543.31	21770.95	0.9747

Linear regression is one of the chosen algorithms to predict housing prices. Since linear regression only identifies linear relationships, to optimise it, polynomial regression was used, which captured non-linear relationships by using different polynomial degrees. By testing the model with different degrees, it was discovered that the polynomial with degree 3 was the best model for the dataset, as it increased accuracy from 0.8432 to 0.9747 with a low root mean squared error (RMSE) compared to linear regression, as shown in the results table.



Mohamed Mohamed [001299125]

RANDOM FOREST

Definition

Random Forest is a type of Machine Learning model that uses an ensemble of decision trees to make its predictions.

Results

Average MSE: 138855339.0289

Average RMSE: 11774.3305

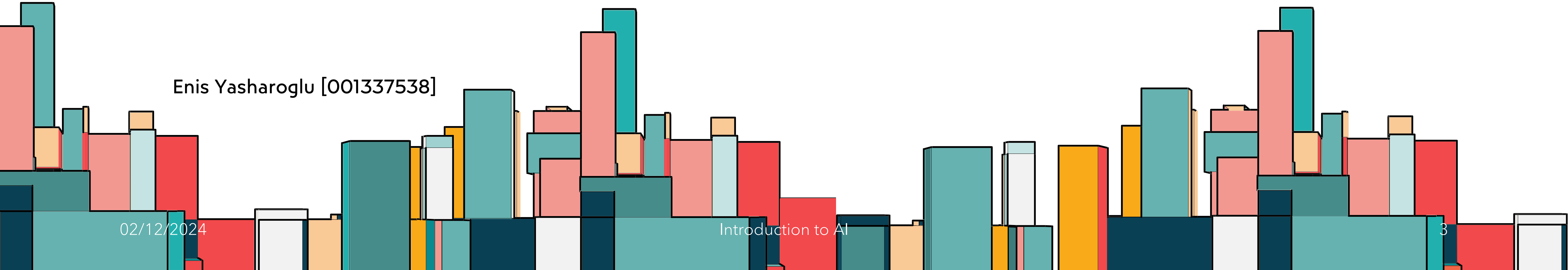
Average MAE: 6242.9464

Average R^2: 0.9959

Advantages of Random Forest

- High Accuracy
- Handles Non-Linear Data
- Robustness to Overfitting
- Estimating Feature Importance
- Handles Both Numerical and Categorical Data

Enis Yasharoglu [001337538]



CATBOOST

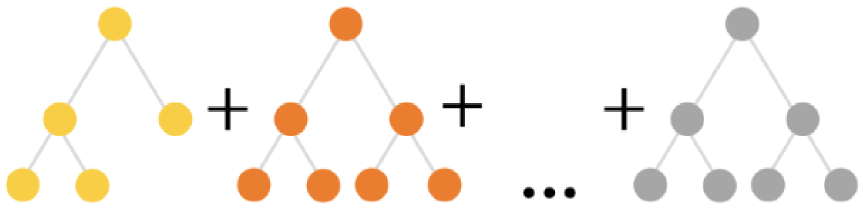
Boosting Algorithm

CatBoost is a boosting algorithm that builds a strong predictive model by combining multiple weak learners, typically decision trees, in sequence. Each tree corrects the errors of the previous ones, steadily improving overall accuracy.

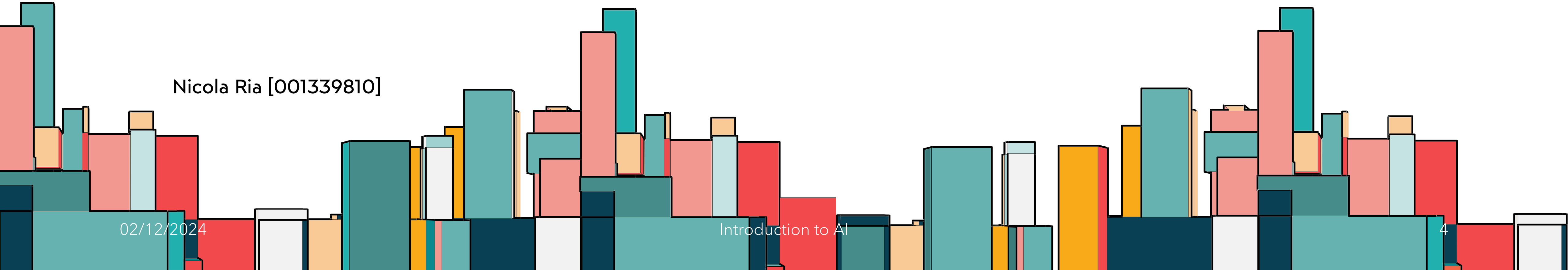
Ordered Boosting

What makes CatBoost unique is its Ordered Boosting technique, which reduces overfitting by carefully structuring how data points are used during training. This ensures the model generalizes well to unseen scenarios.

Boosting algorithm on Decision Trees



Nicola Ria [001339810]



CATBOOST

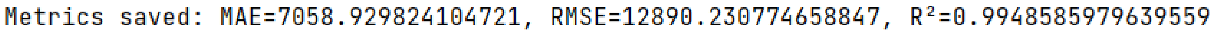
Processing Categorical Features

CatBoost handles categorical features, such as boroughs, natively. This eliminates the need for complex preprocessing and allows the algorithm to leverage these features effectively.

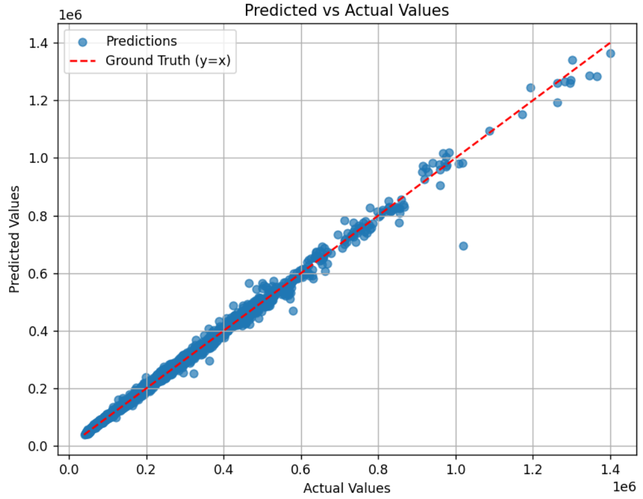
Insights for Policymakers

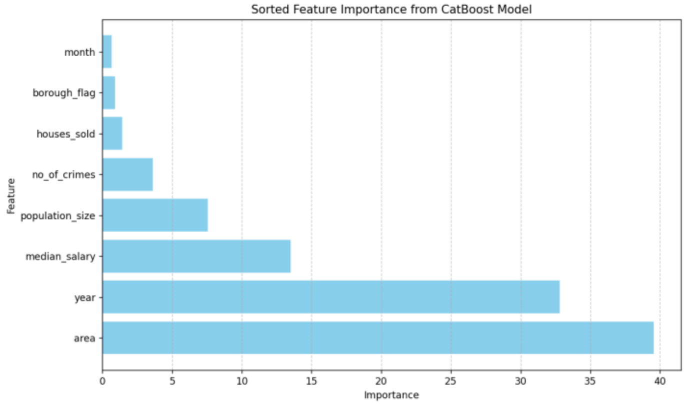
The feature importance analysis highlighted boroughs, population size, median salary, and crime rates as key predictors. This information can help policymakers and real estate investors understand housing market dynamics better.

Model Performance



```
Metrics saved: MAE=7058.929824104721, RMSE=12890.230774658847, R²=0.9948585979639559
```





Nicola Ria [001339810]

XGBOOST

Feature Importance Analysis

It highlighted several influential factor like median salary, population size, number of houses per area, the size of an area, and if it's a borough or not, enabling a better understanding of housing market dynamics.

Gradient Boosting Strength

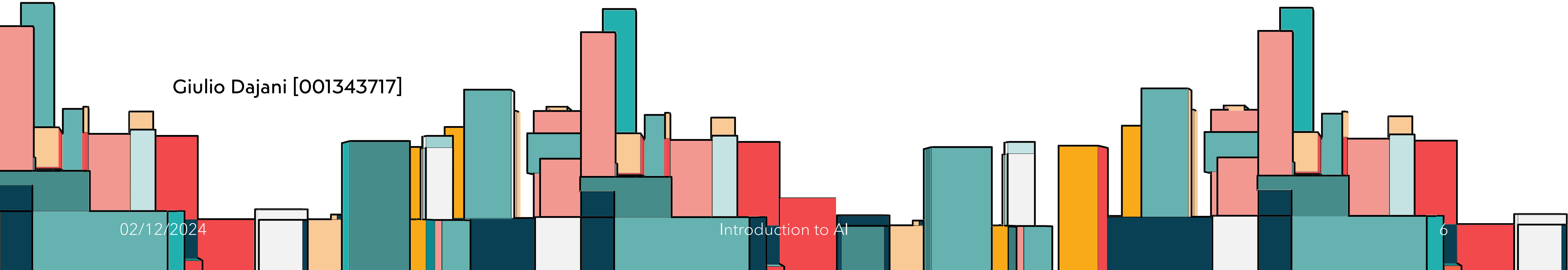
XGBoost produced highly accurate results with a low Root Mean Square Error (RMSE) and a perfect R^2 of 0.99995, meaning the predicted and actual values are almost the same.

Handling Missing Data

The dataset had some gaps, but the algorithm's built-in mechanism helped preprocessing all these information in the best way.

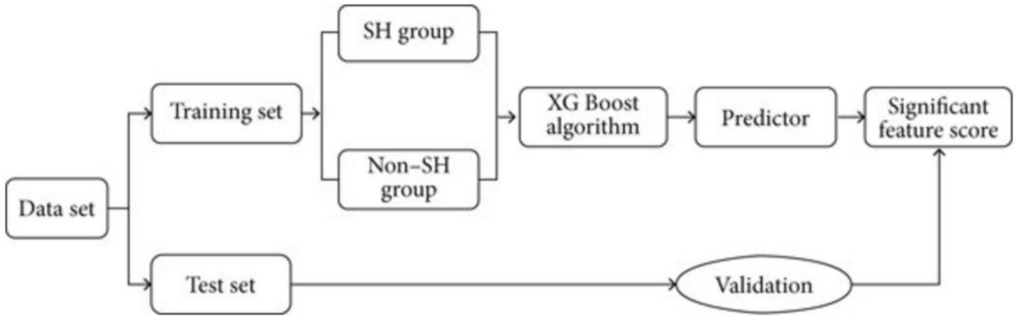
Speed & Scalability

It handled the large dataset efficiently, allowing me to experiment with different hyperparameters while keeping the running time low, improving accuracy.

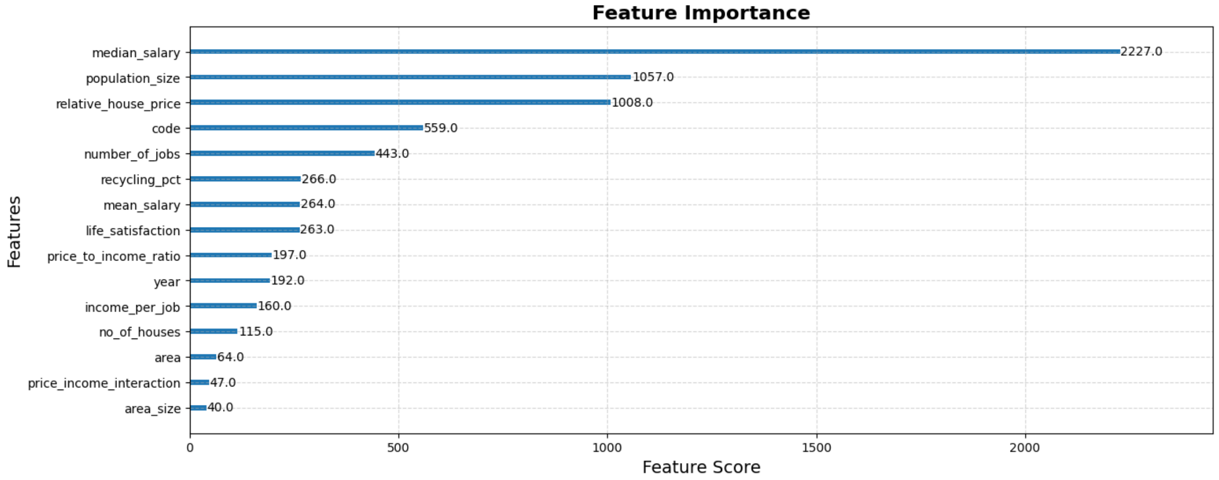


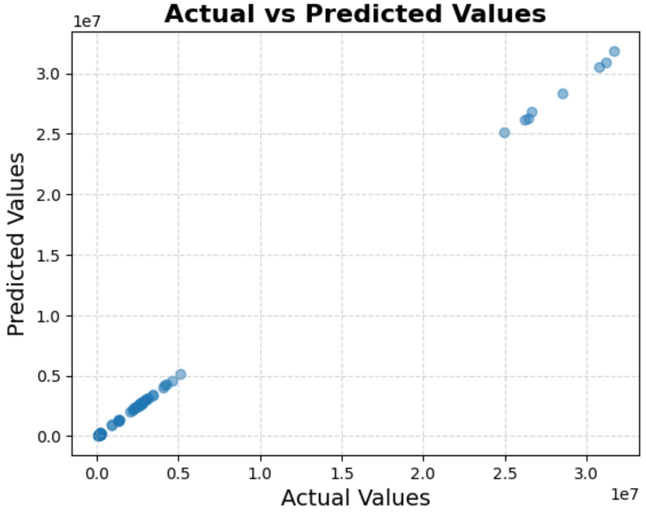
Giulio Dajani [001343717]

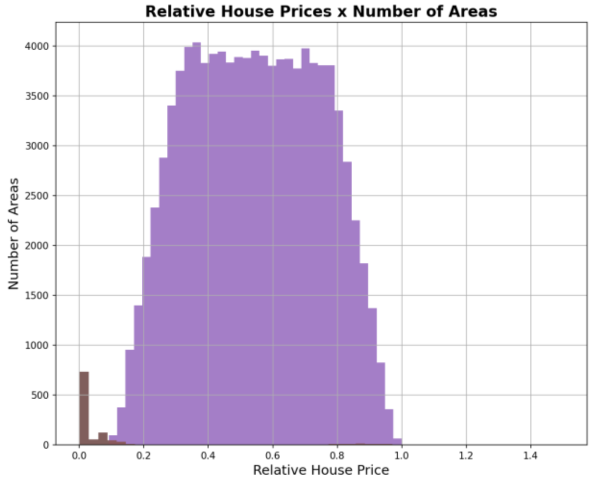
XGBOOST



Flow chart of XGBoost algorithm prediction model training.







Giulio Dajani [001343717]

ALGORITHM COMPARISON

Linear Regression

It's easier to understand and is computationally efficient for smaller datasets but assumes a linear relationship between variables and struggles to handle missing data.

CatBoost

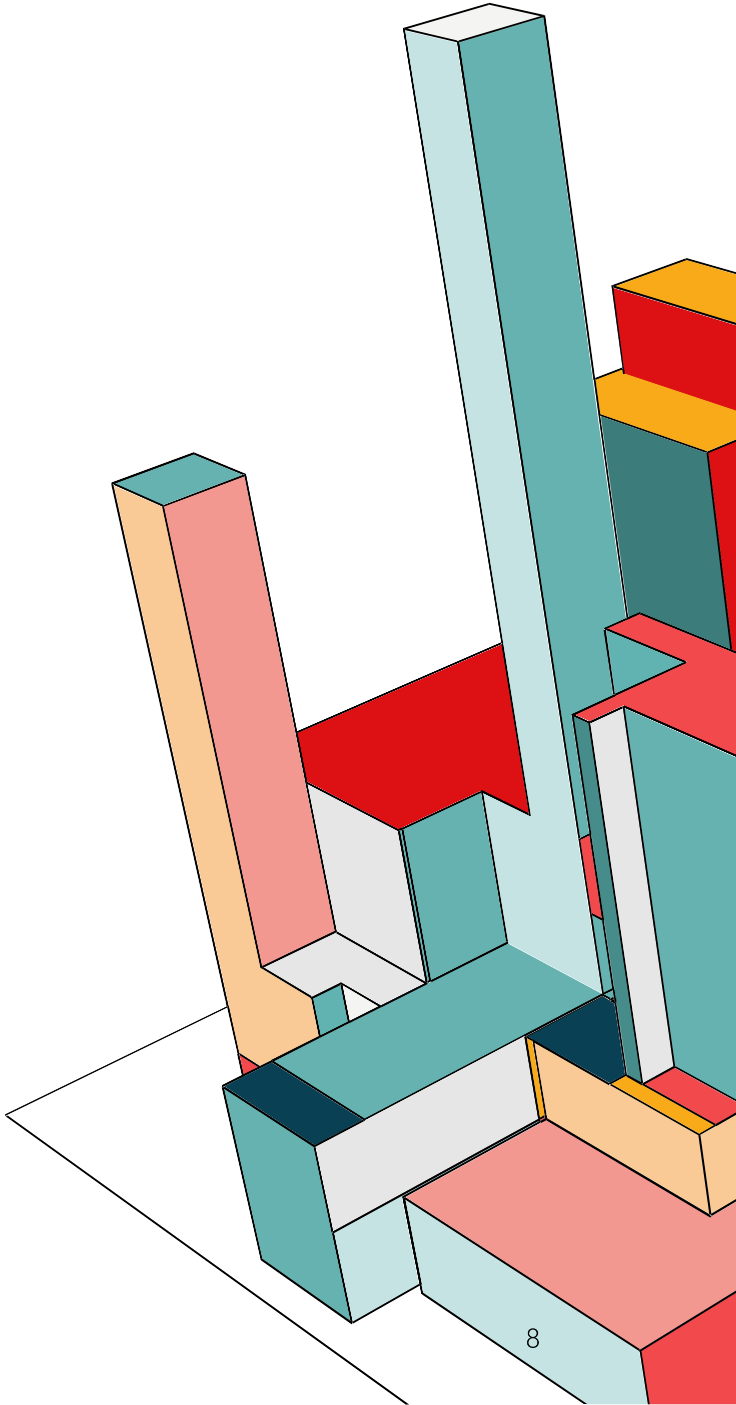
Faster training on some datasets, making preprocessing simpler but computationally heavier on particular configurations and slightly less "mature" in documentation.

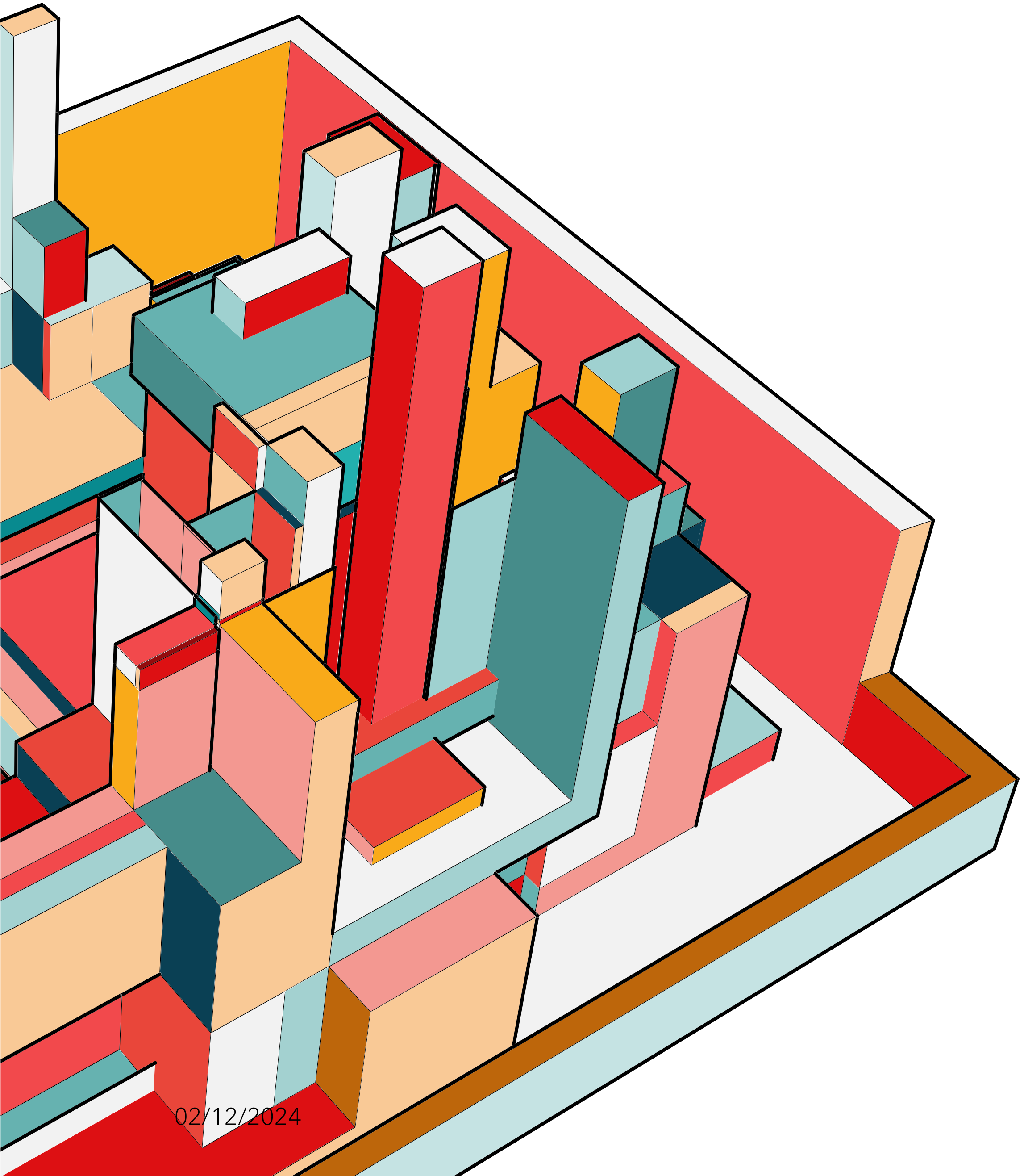
Random Forest Model

Best for general-purpose tasks, Random Forest excels at handling non-linear data and feature interactions, is robust to overfitting, and works well without extensive tuning.

XGBoost

Faster training on medium-big size datasets, making preprocessing faster and computationally less heavy with detailed description documentation.





CONCLUSION & FUTURE WORK

Beyond the deployment of the algorithms for providing real-time housing price predictions, there are a lot of improvements that can be done such as incorporating data like transportation accessibility, and crime rates per educational levels ratio to enhance prediction accuracy; combining XGBoost with another model like CatBoost or Random Forest for a better performance; reducing error rates by implementing advanced techniques; developing a mechanism for the model to retrain periodically and creating insights into housing trends and predictions; expanding the deployment of the web application to other cities and countries. There's always space for improvements and upgrades but of course this project provides a robust starting point for building smarter and more accurate than ever.

THANK YOU

Giulio Dajani [001343717]

Mohamed Mohamed [001299125]

Nicola Ria [001339810]

Enis Yasharoglu [001337538]