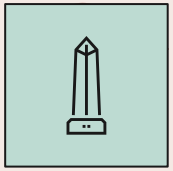
NEW YORK

House Price Estimator

Leticia Genao

PROJECT DEFINITION

Problem Statement:



The goal is to provide accurate price predictions for residential properties in New York City. This tool aids buyers, sellers, and investors by utilizing advanced machine learning techniques to evaluate various property features.

Why Machine Learning?

Traditional methods often involve manual appraisal, which can be biased and inconsistent. Machine learning offers a scalable, objective, and efficient alternative that can handle complex datasets and uncover hidden patterns in the real estate market.

Dataset Overview:

Source: Kaggle

Final Features: bedrooms, bathrooms, square footage, type (one-hot encoded), borough (one-hot encoded)

Target: price

Finale Shape: 3878x12

DATA EXPLORATION AND PREPROCESSING

01

Initial Data Analysis

The data exploration phase involved generating statistical summaries to understand distributions and potential outliers.

02

Data Cleaning Steps

The preprocessing involved handling missing values through imputation, and outlier detection using IQR score methods to ensure model accuracy

03

Feature Engineering

The data was refined by creating a 'borough' feature based on 'locality' mappings, simplifying property 'type' into more general categories, and dropping rows categorized as 'Other' under 'type'.

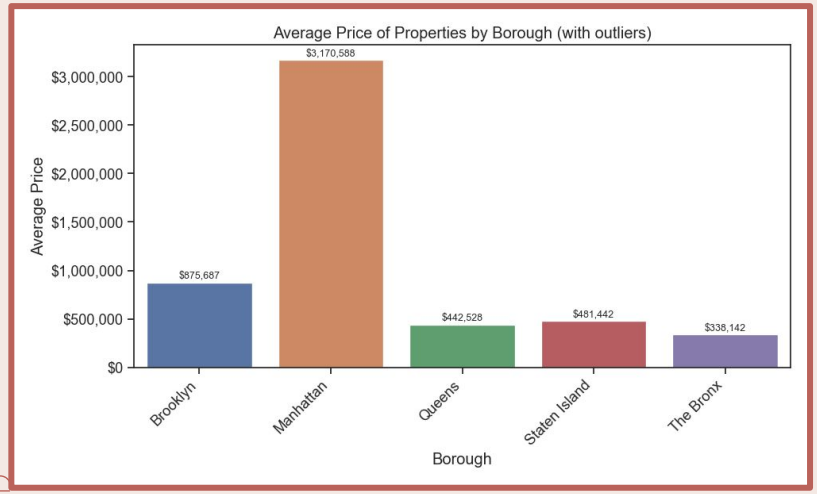
04

Final Model Features

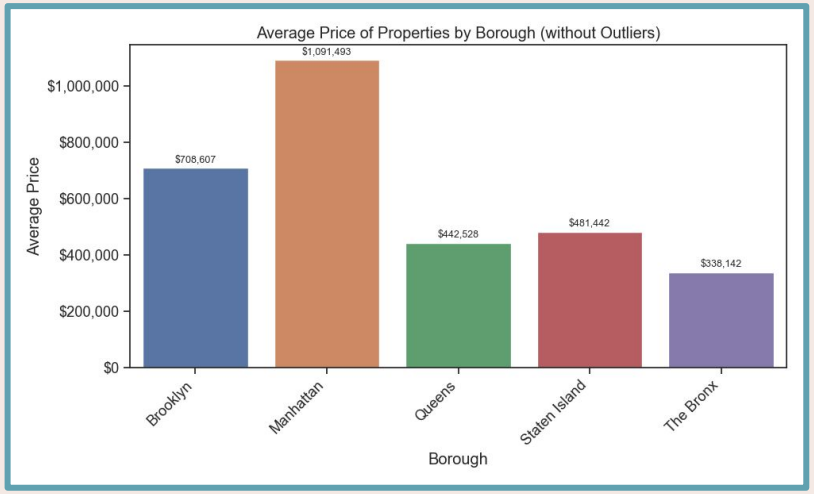
price: Housing price
beds: Number of bedrooms
bath: Number of bathrooms
propertysqft: Property square footage
borough_Manhattan, borough_Queens, borough_Staten Island, borough_The Bronx
type_Condo, type_House, type_Multi-family, type_Townhouse

DATA EXPLORATION AND PREPROCESSING

Price with Outliers



Price without Outliers



MODEL DEVELOPMENT

Algorithm Selection:

Tested: Linear Regression, RandomForest, and GradientBoosting
Selected:GradientBoosting

Training Process:

60% Training split
20% Validation split
20% Testing split

Performance Metrics:

Validation Performance:

R^2 Score:
0.610
Root Mean Squared Error (RMSE):
386,537.91

Test Performance:

R^2 Score:
0.592
Root Mean Squared Error (RMSE):
399,797.70

CONTAINERIZATION WITH DOCKER

Docker Setup

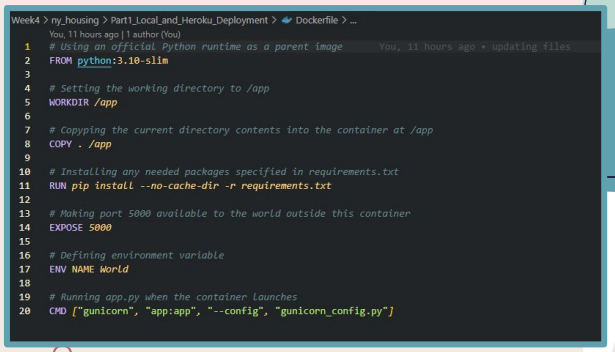
Python environment and dependencies are encapsulated within the Docker container

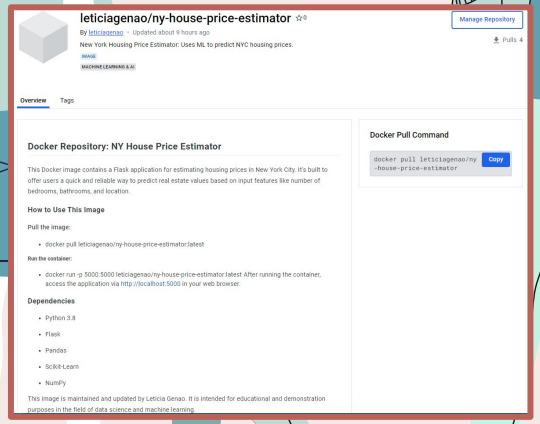
Docker Hub

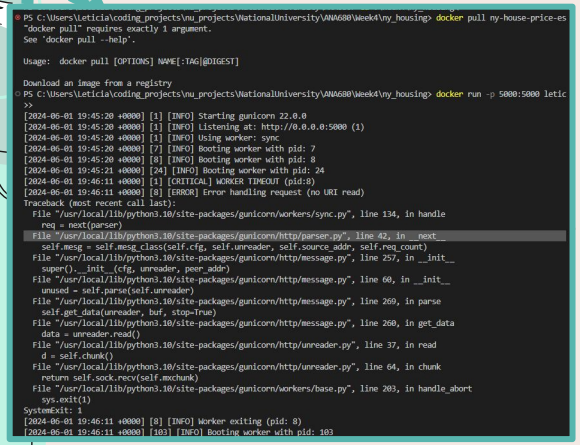
Docker image is made available for public use, for sharing/reproducing the project setup easily

Local Testing

Docker container is tested locally to ensure it functions correctly, mimicking production environment behavior







DEPLOYMENT TO HEROKU

WHY HEROKU DEPLOYMENT?

The Flask-based application was deployed on Heroku, leveraging its robust ecosystem and easy-to-use platform, which simplifies application scaling and management.

LIVE DEMO

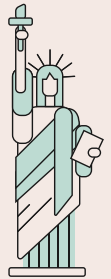
Experience the application firsthand via the live demo on Heroku. Interact with the model to see how it estimates New York housing prices based on various input features!

CHALLENGES:

Deploying on Heroku involved overcoming challenges related to dyno management and Github Actions

Efforts were made to ensure that the application remains responsive and efficient by managing how dynos sleep and wake up to handle user requests without delay.

https://nyhouseprice-c155a038476a.herokuapp.com/



AWS INTEGRATION AND CHALLENGES

AWS Services Used

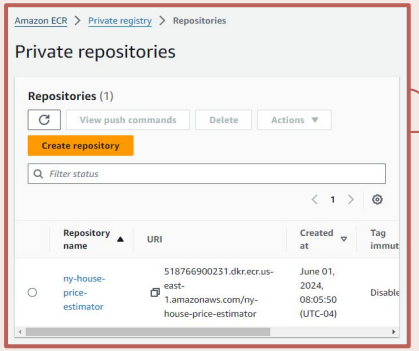
Utilized AWS Elastic Container Registry (ECR) for Docker image management and S3 for secure model data storage

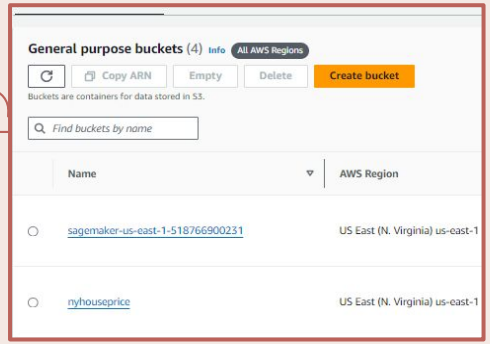
SageMaker Notebook

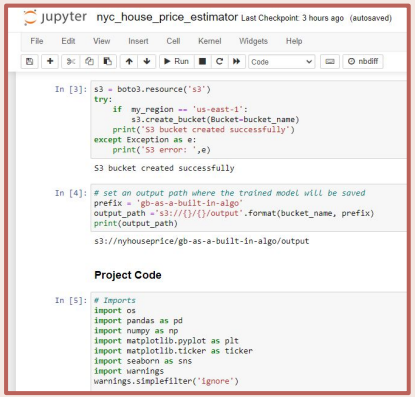
Leveraged AWS SageMaker's cloud-based Jupyter notebooks for model development

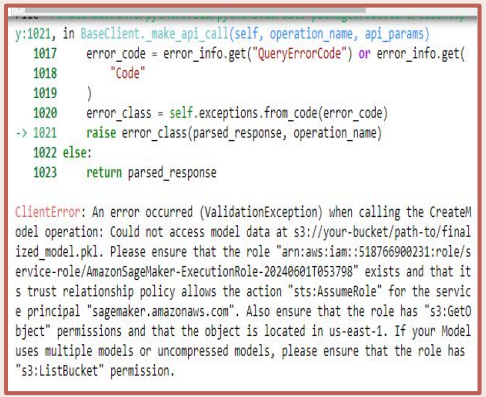
Deployment Issues

Encountered significant challenges with SageMaker deployments, particularly issues related to IAM permissions and endpoint configurations









Lessons Learned and Best Practices

Challenges Overcome:

This project taught the complexity of deploying machine learning models at scale, including handling various AWS services and ensuring proper access permissions

Best Practices

Established best practices including strict version control, comprehensive testing across environments, and detailed logging for issue resolution.

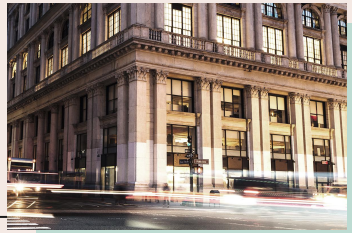
Future Improvements

Plans to enhance the model by incorporating interaction features, and integrating improved ML techniques.

Conclusion and Q&A

Summary

The NY Housing Price Estimator is a testament to the power of machine learning in transforming real estate analysis and prediction.



Future Directions

Aiming to incorporate more predictive factors to enhance its accuracy and utility

Q&A

Comments & Questions?

RESOURCES

- Kaggle. (n.d.). *New York housing market*. Retrieved from https://www.kaggle.com/datasets/nelgiriyewithana/new-york-housing-market/data
- NYC Office of the Comptroller. (2024). *Spotlight: New York City's homeowner housing market*. Retrieved from https://comptroller.nyc.gov/reports/spotlight-new-york-citys-homeowner-housing-market/
- Investopedia. (2024). *Top U.S. housing market indicators*. Retrieved from https://www.investopedia.com/articles/personal-finance/033015/top-us-housing-market-indicators.asp
- Krish Naik. (n.d.). *Tutorial 7-Build,Train,Deploy Machine Learning Model AWS SageMaker- Predicting Test Data Endpoints*
- YouTube. Retrieved from https://www.youtube.com/watch?v=XSsnPtHRZ6A&ab_channel=KrishNaik