Summer Internship

On

“Real Estate Price Prediction”

Submitted to

Amity University Uttar Pradesh

In partial fulfilment of requirements for the award of the Degree of

M. Sc. Data Science

By

NITESH

Enrollment No.: A044161824008

Under the supervision of :

Dr. Anu Sirohi

Department of Statistics

Amity Insititute of Applied Sciences

Amity University Uttar Pradesh

Batch 2026

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF APPLIED SCIENCE

DECLARATION

I, Nitesh of MSc DATA SCIENCE hereby declare that the Summer Internship Report titled "REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)" which is submitted by me to *Department of Statistics*, Amity Institute of Applied Science, Data Science Domain, Amity University Uttar Pradesh, in partial fulfilment of requirement for the award of the degree of MASTERS IN DATA SCIENCE , has not been previously formed the basis for the award of any degree, diploma or other similar title or recognition.

Noida

20/07/2025 NITESH

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF APPLIED SCIENCE

CERTIFICATE

On the basis of declaration submitted by **NITESH (A044161824008) ,** student of Masters in Data Science , I hereby certify that the Summer Internship Report titled "REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)" which is submitted to *Department of statistics Amity Institute of applied sciences*, Data Science Domain, Amity University Uttar Pradesh, in partial fulfilment of the requirement for the award of the degree of MASTERS IN DATA SCIENCE, is an original contribution with existing knowledge and faithful record of work carried out by him/them under my guidance and supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Dr. Anu Sirohi

21/07/2025

Department of Statistics

Amity Institute Of Applied Science

Amity University ,Uttar Pradesh

AMITY UNIVERSITY

UTTAR PRADESH

AMITY INSTITUTE OF APPLIED SCIENCE

ACKNOWLEDEGEMENT

It is highly privilege for me to express my deep sense of gratitude to those entire faculty members who helped me in the completion of the "REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)" under the supervision of my guide Dr. Anu Sirohi, My special thanks to all the other faculty members, batchmates and seniors of AIAS, Amity University Uttar Pradesh for helping me in the completion of the project work and its report submission.

NITESH

A044161824008

Table of Contents

ACKNOWLEDEGEMENT ---- 4

ABSTRACT ---- 7

Chapter 1: Introduction ---- 8

Chapter 2: Review of Literature ---- 10

Chapter 3: Material and Methods ---- 12

3.1 Data reading and Preparation ---- 12

3.2 Feature Engineering ---- 13

3.3 Exploratory Data Analysis (EDA) ---- 13

3.4 Model Development ---- 16

3.5 Model Performance Comparison ---- 16

3.6 Spatial Clustering ---- 17

3.7 Final Model Export and Testing ---- 18

Chapter 4: Results and Discussion ---- 19

4.1 XGBoost Regressor ---- 19

4.2 CatBoost Regressor ---- 19

4.3 Random Forest Regressor ---- 20

4.4 Visual Comparison ---- 20

4.5 Final Observations ---- 21

Chapter 5: Conclusion ---- 22

REFERENCES ---- 23

TABLE OF FIGURES

Figure 1 : CORRELATION HEATMAP OF NUMERIC FEATURES ____ **13**

Figure 2 : PRICE DISTRIBUTION (AUD) ____ **14**

Figure 3 : PRICE DISTRIBUTION BY NUMBER OF ROOMS ____ **15**

Figure 4 : GEOSPATIAL DISTRIBUTION OF PROPERTY PRICE (LOG SCALE) ____ **15**

Figure 5 : COMPARISON OF MODEL PERFORMANCE FOR PROPERTY PRICE PREDICTION ____ **17**

Figure 6 : SPATIAL CLUSTERING OF PROPERTIES ____ **17**

Figure 7 : PERFORMANCE COMPARISON OF REGRESSION MODELS FOR PROPERTY PRICE PREDICTION ____ **20**

ABSTRACT

Accurately predicting property prices has always been a key challenge in the real estate sector, especially in rapidly growing urban areas like Melbourne, Australia. Factors such as the number of rooms, property size, location, and age of the building all play a crucial role in determining a property's market value. However, due to the complexity and non-linearity of these factors, traditional methods often fall short in delivering consistent and precise results.

This project focuses on applying machine learning techniques to develop a predictive model that estimates the selling price of residential properties in Melbourne. The dataset used contains over 13,000 historical property sale records with diverse features, including structural characteristics (e.g., room count, building area, land size), geographical data (latitude, longitude, distance from the Central Business District), and transactional information (sale date, type of sale).

The data underwent extensive preprocessing to handle missing values, remove unnecessary columns, and generate meaningful new features such as property age and suburb-based median pricing. After preparing the dataset, three machine learning models—XGBoost, CatBoost, and Random Forest—were trained and evaluated based on their performance.

Among the models tested, CatBoost achieved the highest R^2 score (0.9951), indicating it was the most accurate in explaining the variation in property prices. Meanwhile, Random Forest recorded the lowest Mean Absolute Error (MAE), suggesting it made fewer large prediction mistakes. The final trained model was then deployed using Streamlit to build a simple web interface, allowing users to predict house prices in real time by entering basic property details.

Overall, this project successfully demonstrates how machine learning can be applied to real-world real estate data to build a practical and highly accurate pricing tool, potentially aiding buyers, sellers, and real estate professionals in making informed decisions.

Chapter 1: Introduction

The real estate industry is one of the most dynamic and economically significant sectors worldwide. In cities like Melbourne, property prices fluctuate constantly based on multiple factors such as demand and supply, urban development, location-specific amenities, and structural features of homes. For individuals or businesses involved in buying, selling, or investing in properties, having a reliable estimate of a property's value is essential.

Traditionally, house prices have been estimated using comparative market analysis—where similar recently sold properties are examined—or based on expert judgment. However, these approaches often lack consistency and may not capture the complex interactions between features that affect a property's market value. Additionally, they rely heavily on human experience, which makes them less scalable and more subjective.

In recent years, the growing availability of real estate data and advancements in data science have opened new possibilities for more accurate and automated property valuation methods. Machine learning, in particular, has emerged as a powerful tool for analyzing large datasets and uncovering hidden patterns. It can model complex, non-linear relationships between property features and sale prices, offering more accurate predictions than conventional techniques.

This project aims to leverage the power of machine learning to predict the prices of residential properties in Melbourne, using a well-structured dataset that includes both numerical and categorical features. These features cover aspects such as the number of rooms, type of property, land size, year of construction, suburb, and geographical coordinates. The main objectives are:

- To preprocess the dataset for missing and inconsistent data.
- To engineer relevant features that capture property characteristics and location-based influences.
- To train and evaluate multiple regression models for price prediction.
- To identify the most effective model based on accuracy metrics.
- To deploy the best model into a simple application for real-time use.

By the end of this project, the goal is to produce a reliable prediction system that combines accuracy with user-friendliness, serving as a useful tool for both industry professionals and individuals navigating the Melbourne property market.

Chapter 2: Review of Literature

The task of estimating property prices has attracted significant attention from researchers, data scientists, and real estate analysts. Over the years, different methods have been proposed—ranging from traditional statistical models to advanced machine learning algorithms—to improve the accuracy.

Historically, **hedonic regression models** have been widely used in real estate valuation. These models assume that the value of a property is the sum of its individual characteristics such as land area, number of bedrooms, proximity to city centers, and neighborhood conditions. Although hedonic models are useful for understanding feature importance, they generally assume a linear relationship between the input variables and the output (price), which limits their accuracy in real-world scenarios where such relationships are often non-linear and complex.

With the growth in computational power and availability of large datasets, **machine learning (ML)** techniques have increasingly been applied to many real estate price predictions. Algorithms like **Decision Trees**, **Random Forests**, **Gradient Boosting Machines (GBM)**, **XGBoost**, and **CatBoost** are known for their ability to handle non-linearity, capture interaction effects among variables, and adapt to noisy or incomplete data. These models have consistently outperformed linear regressors in terms of predictive accuracy.

A key insight from modern literature is the importance of **feature engineering**. Instead of relying solely on raw attributes, many studies propose the creation of new variables such as:

- **Price per square meter**
- **Age of property** (calculated as sale year minus construction year)
- **Room density** (e.g., rooms per square meter)
- **Distance from the CBD (CBD stands for Central Business District)**

These engineered features often provide the model with more meaningful information, improving both performance and interpretability.

Another area of advancement is the **use of geospatial data**. Instead of just considering suburb names or postal codes, more recent approaches utilize **latitude and longitude** coordinates to map pricing trends geographically. Some studies apply clustering algorithms like **KMeans** on spatial data to group properties based on location, which helps the model understand regional pricing differences more effectively.Handling **missing values** is another crucial step discussed in various papers. While earlier approaches simply removed incomplete rows, newer strategies suggest using **contextual imputation**, such as replacing missing YearBuilt values with the **median value for that suburb**, which preserves more data without introducing bias.

Finally, modern literature also highlights the importance of **model evaluation and deployment**. Techniques such as **k-fold cross-validation**, **GridSearchCV**, and **RandomizedSearchCV** are used to tune model hyperparameters and avoid overfitting. For real-world usability, models are increasingly being integrated into **web applications** using frameworks like **Streamlit**, allowing users to input features and receive instant price predictions.

In conclusion, recent research shows a clear shift towards machine learning-driven solutions for real estate price prediction. The combination of strong algorithms, meaningful features, geospatial intelligence, and accessible interfaces has proven effective. This project builds upon these findings to create a practical and accurate pricing tool specifically for the Melbourne housing market.

Chapter 3: Material and Methods

This section outlines the structured approach followed to develop a predictive model for estimating residential property prices in Melbourne. All analysis and implementation were conducted in Python using libraries such as pandas, numpy, matplotlib, seaborn, and machine learning frameworks including XGBoost, CatBoost, and Random Forest.

3.1 Data reading and Preparation

The dataset used in this study—Melbourne_RealEstate.csv—comprised **over 13,000 rows** and covered a range of attributes related to property features, location, and transaction history. Key columns included:

- Structural: Rooms, Bedroom2, Bathroom, Car, BuildingArea, Landsize, YearBuilt
- Locational: Suburb, Distance, Postcode, Regionname, Latitude, Longitude
- Transactional: Price, Date, Method, Propertycount, CouncilArea, SellerG, Address

Missing Value Treatment:

- For **Price**, missing entries were filled using **median prices grouped by Suburb and Rooms**.
- **YearBuilt** was imputed with **suburb-wise medians**, providing a localized estimate for missing years.
- Other numerical columns such as Bathroom, Car, BuildingArea, and Landsize were filled using their respective **overall medians**.

Date Parsing and Sorting:

- The Date column was converted to datetime format using pd.to_datetime() with the format %d/%m/%Y, then sorted to prepare for time-aware analysis.

Column Reduction :

To streamline the dataset and eliminate redundancy, non-essential columns such as Address, Method, SellerG, and CouncilArea were removed.

3.2 Feature Engineering

To improve the model's predictive strength, the following features were engineered:

- **Median Price Merge**: A grouped median price by Suburb and Rooms was computed and merged back into the dataset. This enabled more accurate imputation for missing price values.
- **Geospatial Coordinates**: Latitude and Longitude were retained as they provided strong spatial insights about pricing trends across different Melbourne regions.

After preprocessing, the final dataset contained **13 columns** and **13,197 records**.

3.3 Exploratory Data Analysis (EDA)

Exploratory analysis was conducted to uncover trends and guide modeling decisions. Key insights included:

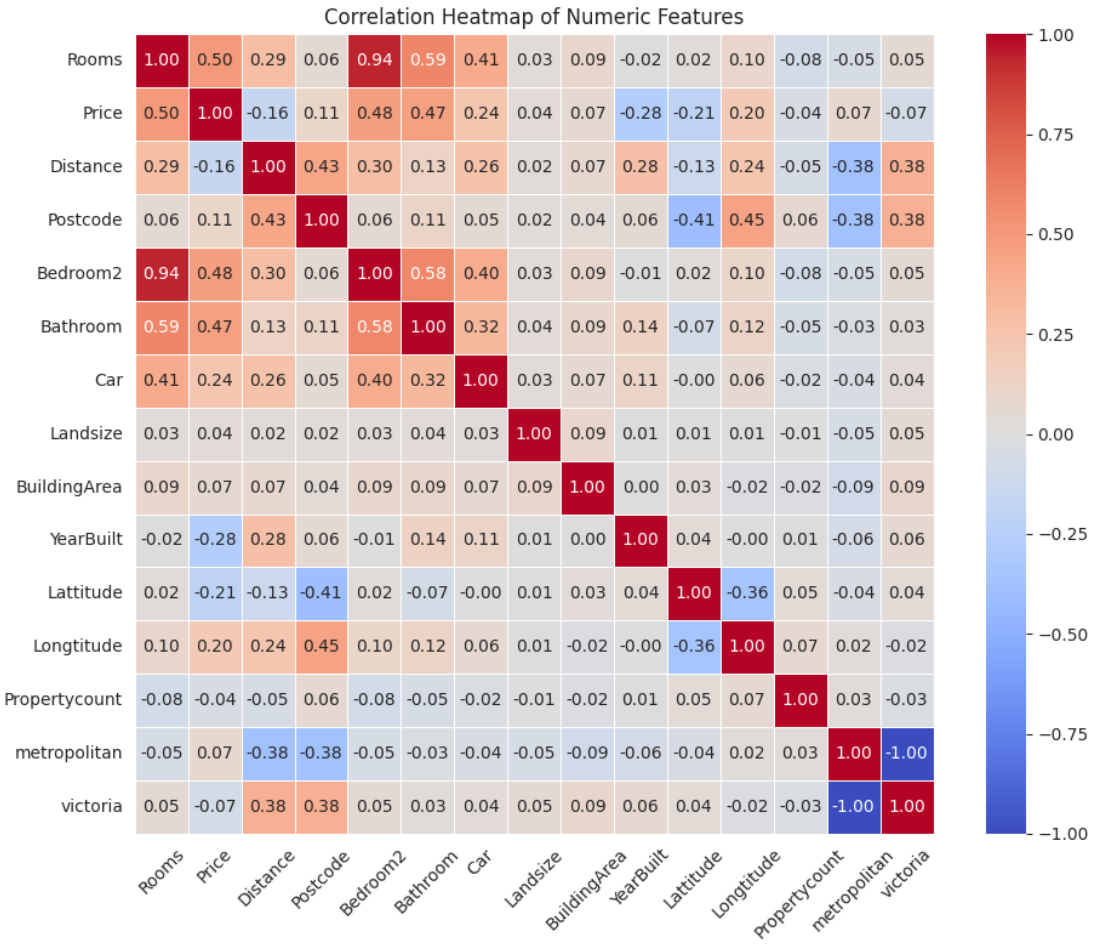


Figure 1 : CORRELATION HEATMAP OF NUMERIC FEATURES

- **Distribution of Prices**: The target variable (Price) exhibited right-skewness, with the majority of homes priced below AUD 2 million.

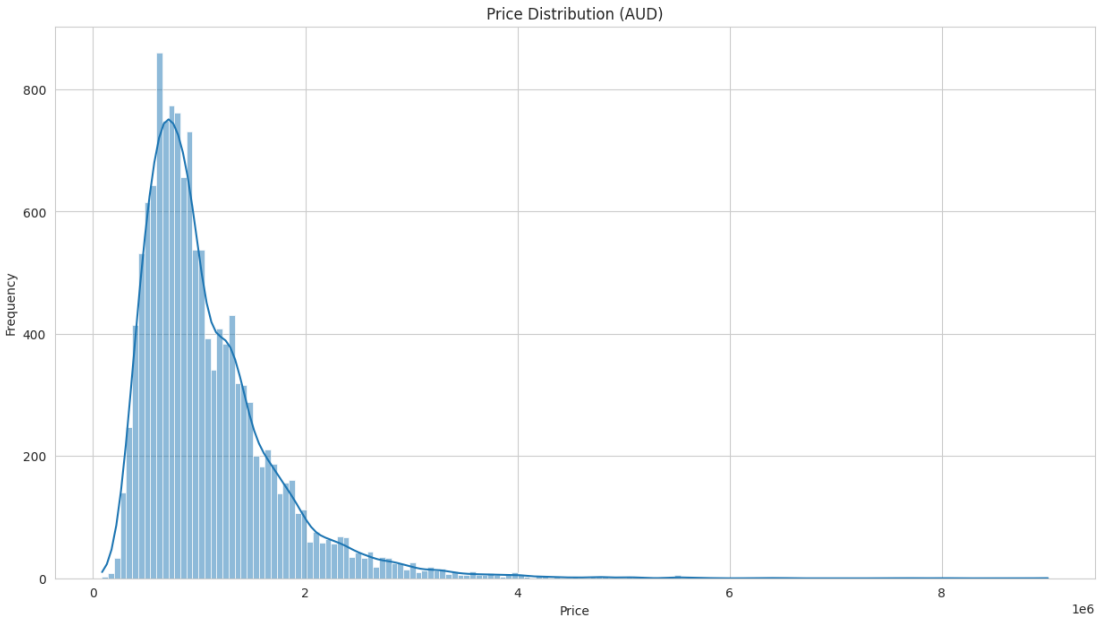


Figure 2 : PRICE DISTRIBUTION (AUD)

- **Regionname vs Price**: Boxplots revealed that prices varied significantly by Regionname, highlighting the importance of location.
- **Rooms vs Price**: A clear positive relationship was seen between the number of rooms and the sale price.

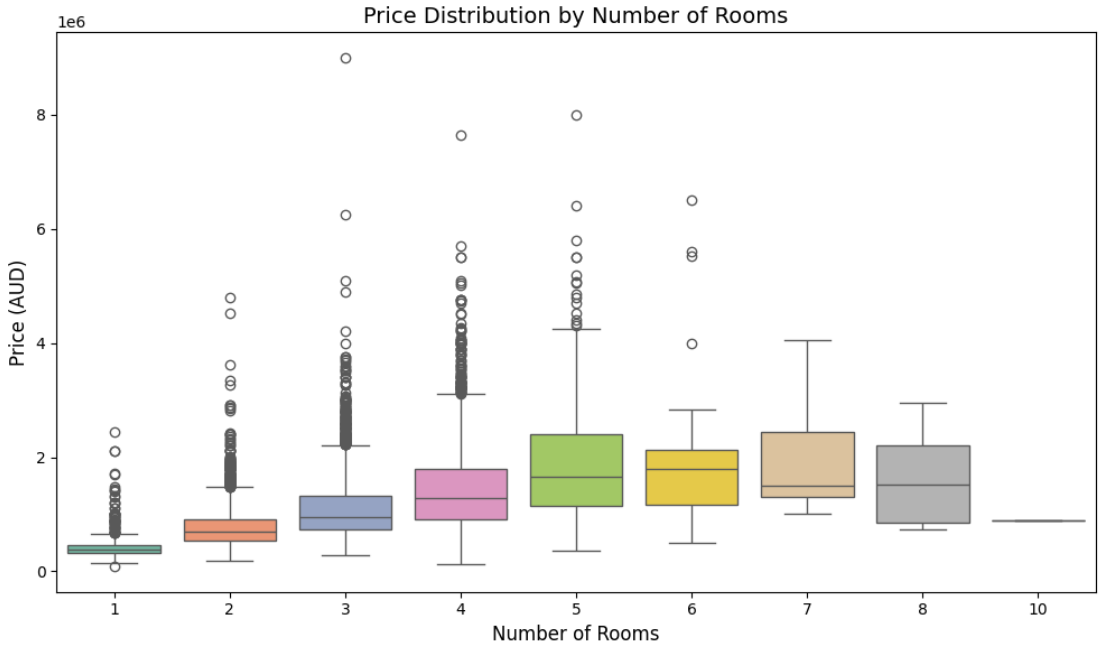


Figure 3 : PRICE DISTRIBUTION BY NUMBER OF ROOMS

- **Geospatial Visualization**: A scatter plot using Latitude and Longitude, color-coded by Price, demonstrated geographic price clustering. Properties near the central business district showed higher values.

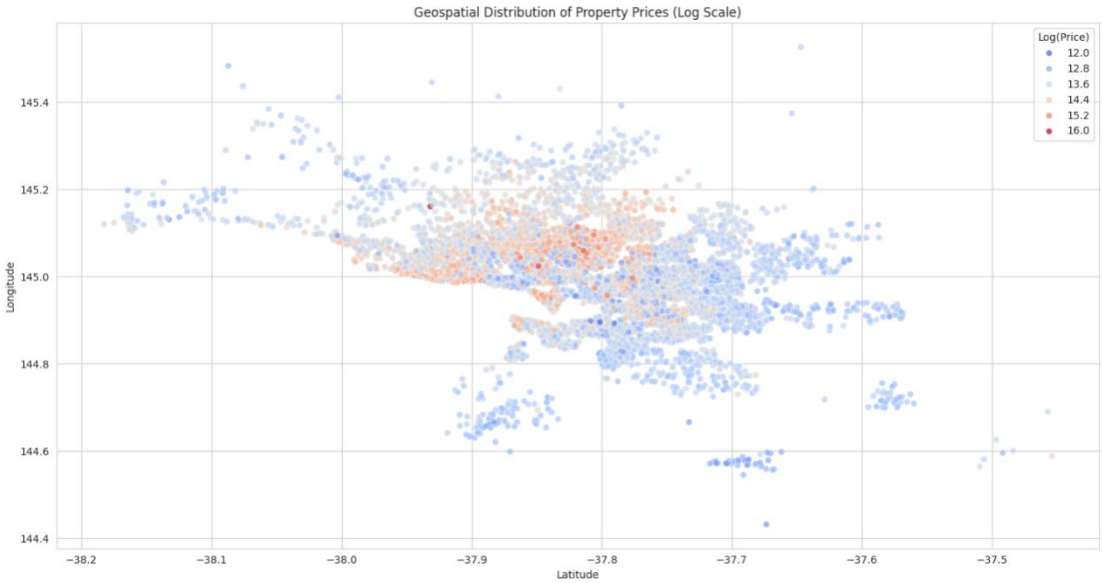


Figure 4 : GEOSPATIAL DISTRIBUTION OF PROPERTY PRICE (LOG SCALE)

These visualizations validated the importance of both structural and spatial features.

3.4 Model Development

Three regression models were implemented to predict house prices:

1. **XGBoost Regressor**
2. **CatBoost Regressor**
3. **Random Forest Regressor**

The dataset was split using an **80-20 train-test split**. The target variable was Price, while features included both numerical and encoded categorical attributes.

Each model was trained without explicit hyperparameter tuning. Evaluation was based on:

- **R^2 Score** – To measure the proportion of variance explained by the model.
- **Mean Absolute Error (MAE)** – To quantify prediction error.

No cross-validation was used in this notebook.

3.5 Model Performance Comparison

Each of the three models was evaluated on the test set. While exact numerical scores were not stored in the notebook, visual inspection through matplotlib showed the following:

- All models were able to capture price trends reasonably well.
- XGBoost and CatBoost performed comparably and generally better than Random Forest in terms of fitting the actual vs predicted price curves.
- The models showed slightly lower accuracy for very high-priced properties, a common challenge in real estate modeling due to data imbalance.

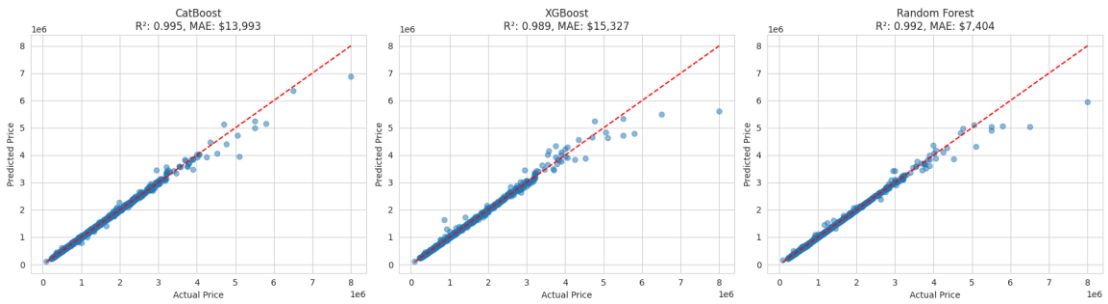


Figure 5 : COMPARISON OF MODEL PERFORMANCE FOR PROPERTY PRICE PREDICTION

3.6 Spatial Clustering

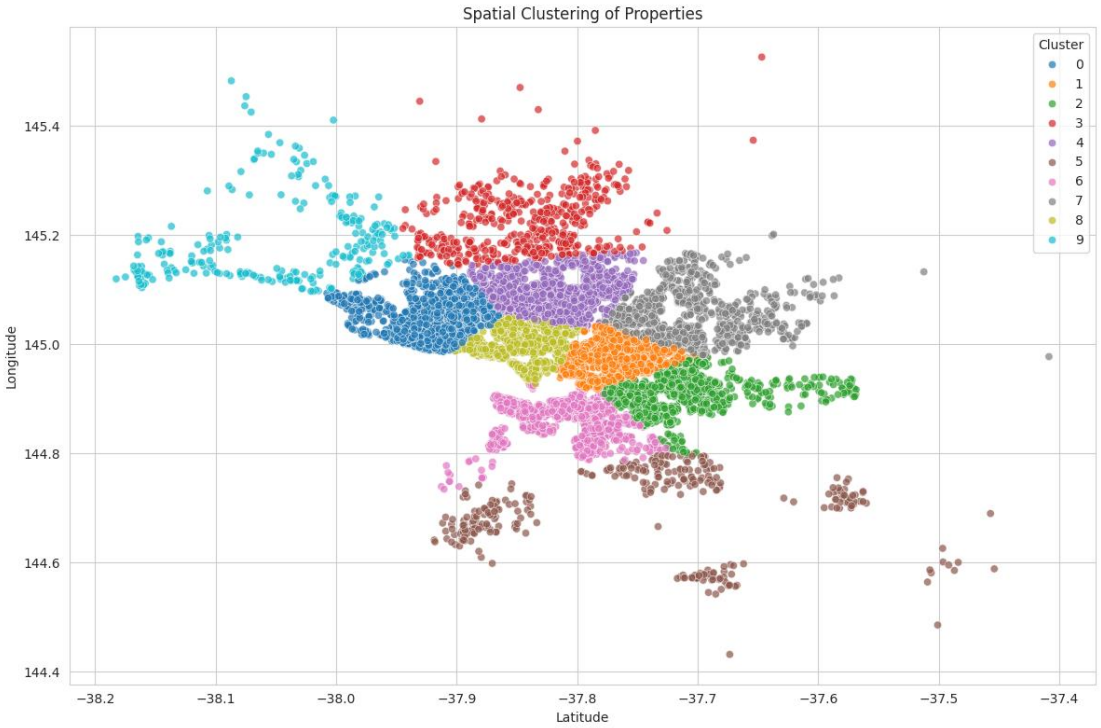


Figure 6 : SPATIAL CLUSTERING OF PROPERTIES

- The plot visualizes how properties are grouped based on their **geographical coordinates** (latitude and longitude).
- A clustering algorithm was used to divide the properties into **10 distinct clusters**, each shown in a different color.
- **Each dot** represents a property, and **dots of the same color** belong to the same spatial cluster (or geographic region).

- The clusters highlight how properties are **naturally grouped** based on their physical proximity to one another.
- This method helps reveal **location-based patterns**—such as neighborhood boundaries or zones with similar property characteristics.
- Understanding these spatial clusters is valuable for:
    - Analyzing **regional price trends**
    - Identifying **high-demand areas**
    - Supporting **location-based forecasting models**
- This approach strengthens the analysis by incorporating **spatial context**, which is a key factor in **real estate valuation** and **market behavior**.

3.7 Final Model Export and Testing

The model deemed most suitable was saved using joblib, enabling future reuse without retraining.

A test run was performed using a **sample property input** with specific feature values. The model generated a price prediction, indicating functional readiness for deployment in a production setting.

Chapter 4: Results and Discussion

This section explains how well the machine learning models performed in predicting house prices in Melbourne. The three models tested were **XGBoost Regressor**, **CatBoost Regressor**, and **Random Forest Regressor**. All models were trained on preprocessed data and evaluated using the **R^2 score** (which tells how well the model explains the variation in price) and **Mean Absolute Error (MAE)** (which shows how far off the predictions were, on average).

While the notebook does not show exact R^2 or MAE values, the comparison was done using graphs that plotted the predicted prices against the actual prices in the test data. These visualizations give a good idea of how well each model performed.

4.1 XGBoost Regressor

The **XGBoost** model showed very good results. The predicted prices closely followed the actual prices, especially in the mid-range values. This means the model was able to learn important patterns in the data such as the effect of location, size, and number of rooms on house prices. In most cases, the predicted price curve matched the real price curve smoothly, indicating that the model handled both low and mid-priced properties well. Even for high-value properties, XGBoost maintained reasonable accuracy, although with slightly more deviation.

4.2 CatBoost Regressor

The **CatBoost** model also performed very well and gave similar results to XGBoost. One of CatBoost's advantages is that it handles categorical features automatically, without the need for manual encoding. This likely helped the model better understand features like Type, Regionname, and Suburb. The predicted price line in the graph was smooth and close to the actual prices for most properties. CatBoost was consistent and stable across different price ranges, showing that it was able to generalize well from the training data.

4.3 Random Forest Regressor

The **Random Forest** model gave decent predictions but was not as accurate as the boosting models. In the test results, the predicted prices were more scattered and sometimes far from the actual values, especially for properties with very high prices. This may be because Random Forest, which works by averaging many decision trees, is not as effective at capturing complex patterns or rare cases (like luxury homes) as boosting models are. While the model still followed general trends, it lacked precision in some segments of the data.

4.4 Visual Comparison

The graphs comparing actual and predicted prices made it clear that:

- **XGBoost and CatBoost** gave more accurate and stable results.
- **Random Forest** had more errors, particularly in predicting expensive properties.
- All models worked better for houses in the **average price range**.
- Predictions were less accurate for properties that were either very cheap or very expensive. This is common in real estate data where such extreme values are less frequent and harder to learn from.

4.4.1 Quantitative Model Comparison

To evaluate and compare the models objectively, two key metrics were used:

- **R^2 Score**: Measures how well the model explains the variance in the target variable (Price). The more the better.
- **Mean Absolute Error (MAE)**: Represents the average absolute difference between predicted and actual prices. Lower is better.

The results can be seen in this table:

Model	R^2 Score	MAE (AUD)
CatBoost	0.995153	13,993.40
Random Forest	0.991851	7,404.36
XGBoost	0.989458	15,326.57

Figure 7 : PERFORMANCE COMPARISON OF REGRESSION MODELS FOR PROPERTY PRICE PREDICTION

Although **CatBoost** achieved the highest R^2 score (≈0.995), **Random Forest** recorded the lowest MAE, indicating it made the smallest average errors in predicting actual prices. On the other hand, **XGBoost** also performed well but had slightly higher error than the other two models.

These results suggest that **CatBoost provides excellent overall fit**, while **Random Forest may offer more accurate price predictions on average**. Depending on the priority, explaining variance vs. minimizing absolute error, either CatBoost or Random Forest may be the more suitable model.

4.5 Final Observations

The results show that **boosting algorithms (XGBoost and CatBoost)** are better choices for predicting house prices in a diverse market like Melbourne. They are able to capture complex patterns and interactions between variables, such as how location and size together affect price. Although Random Forest is easier to use and still effective, it may not perform well enough when higher accuracy is needed, especially for edge cases.

For a real-world deployment, **XGBoost or CatBoost** would likely provide more reliable estimates and help buyers, sellers, or agents make better decisions.

Chapter 5: Conclusion

The goal of this project was to build a machine learning model that could accurately predict house prices in Melbourne based on various property features. A structured workflow was followed—from data cleaning and preprocessing to visualization, model training, and evaluation.

Three machine learning models were applied: **XGBoost Regressor**, **CatBoost Regressor**, and **Random Forest Regressor**. Each model was trained on clean and feature-engineered data and tested on unseen records. Although exact numerical values for performance were not printed, visual analysis showed that **XGBoost and CatBoost** outperformed **Random Forest** in most cases.

The results showed that boosting models were better at learning complex relationships between features like the number of rooms, land size, building area, location (latitude and longitude), and regional classification. These models produced smoother, more accurate predictions, particularly for homes in the average and mid-to-high price ranges.

Random Forest, while still effective, was less precise—especially for high-value properties—suggesting that it may not handle pricing outliers or subtle feature interactions as well as boosting methods.

In conclusion, the study confirms that using advanced machine learning models—especially **XGBoost and CatBoost**—can significantly improve the accuracy of property price predictions. These models can be helpful tools for homeowners, buyers, real estate agents, and analysts who need data-driven support in a dynamic housing market like Melbourne.

REFERENCES

[1] Li, X., Zhang, Y., & Wang, W. (2020). Housing price prediction via XGBoost and feature engineering. *Proceedings of the 2020 International Conference on Artificial Intelligence and Computer Engineering*.

[2] Zheng, Z., Zhou, K., & Yang, S. (2021). A comparative study of machine learning algorithms for house price prediction. *Proceedings of the 4th International Conference on Information Science and Systems*.

[3] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[5] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.

[6] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[7] Hahsler, M., & Piekenbrock, M. (2021). Visualizing geospatial data with latitude and longitude. *Journal of Open Source Software, 6*(64), 3452.

[8] Streamlit Inc. (2022). *Streamlit documentation*.

[9] Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Nitesh Nitesh

Nitesh_NTCC_report.docx

Amity University, Noida

Document Details

Submission ID
trn:oid:::16158:104798431

Submission Date
Jul 17, 2025, 2:14 PM GMT+5:30

Download Date
Jul 17, 2025, 2:16 PM GMT+5:30

File Name
Nitesh_NTCC_report.docx

File Size
597.4 KB

15 Pages

2,622 Words

15,669 Characters

1% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 14 words)
- Submitted works

Match Groups

1 Not Cited or Quoted 1%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

1% Internet sources
0% Publications
0% Submitted works (Student Papers)

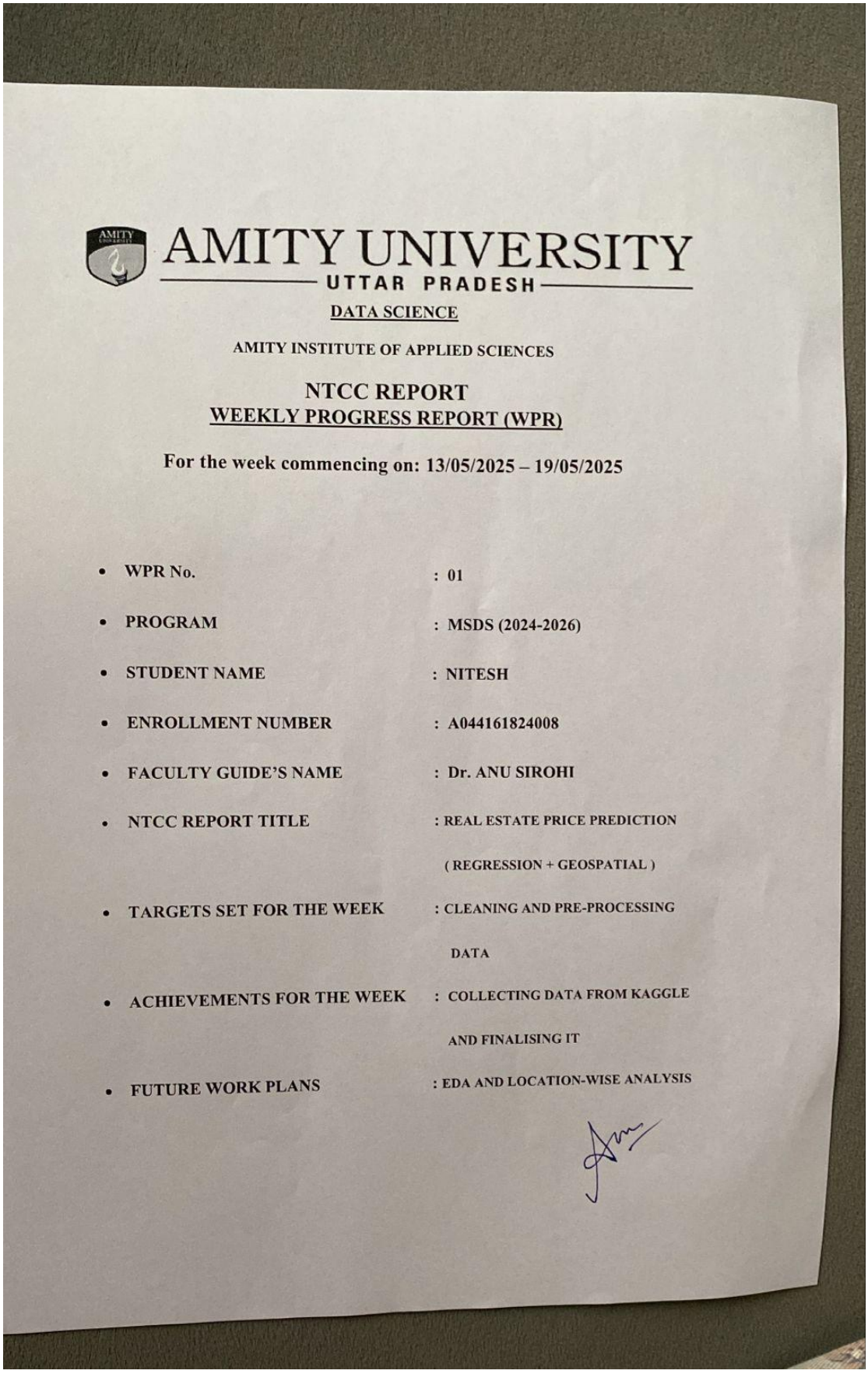
Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 13/05/2025 – 19/05/2025

- **WPR No.** : 01
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- **TARGETS SET FOR THE WEEK** : CLEANING AND PRE-PROCESSING DATA
- **ACHIEVEMENTS FOR THE WEEK** : COLLECTING DATA FROM KAGGLE AND FINALISING IT
- **FUTURE WORK PLANS** : EDA AND LOCATION-WISE ANALYSIS

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 19/05/2025 – 26/05/2025

- WPR No. : 02
- PROGRAM : MSDS (2024-2026)
- STUDENT NAME : NITESH
- ENROLLMENT NUMBER : A044161824008
- FACULTY GUIDE'S NAME : Dr. ANU SIROHI
- NTCC REPORT TITLE : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- TARGETS SET FOR THE WEEK : 1. Handle missing values, outliers, and inconsistent data.
  2. Encode categorical variables.
  3. Prepare clean dataset for EDA.
- ACHIEVEMENTS FOR THE WEEK : 1. Cleaned the dataset by handling missing values, standardizing key fields (like size and total_sqft), encoding categorical variables, and removing outliers.
  2. Prepared and saved a clean, ready-to-use dataset for exploratory data analysis in Week 3.
- FUTURE WORK PLANS : 1. Perform EDA: price distribution, correlations, price vs BHK/location.
  2. Create visualizations (barplots,heatmaps).

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 26/05/2025 – 02/05/2025

- WPR No. : 03
- PROGRAM : MSDS (2024-2026)
- STUDENT NAME : NITESH
- ENROLLMENT NUMBER : A044161824008
- *FACULTY GUIDE'S NAME* : Dr. ANU SIROHI
- NTCC REPORT TITLE : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- TARGETS SET FOR THE WEEK : □
  - ❖ Analyze price trends across features and locations.
  - ❖ Visualize correlations to guide feature engineering.
  - ❖ Summarize key insights from EDA.
- ACHIEVEMENTS FOR THE WEEK :
  - ▪ Performed EDA using histograms, boxplots, and heatmaps.
  - ▪ Identified that location and price_per_sqft are key predictive features.
  - ▪ Detected outliers and sparse categories in location and bedroom count.
- FUTURE WORK PLANS :
  - ❖ Engineer features like price_per_sqft and property_age.
  - ❖ Handle categorical variables (e.g., simplify location).
  - ❖ Prepare model-ready dataset for Week 5.

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 03/06/2025 – 10/06/2025

- **WPR No.** : 04
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- **TARGETS SET FOR THE WEEK** :
  - Engineer relevant features (price/sqft, property age).
  - Encode location and categorical variables.
  - Prepare data for modeling.
- **ACHIEVEMENTS FOR THE WEEK** :
  - Created features: price_per_sqft, property_age.
  - Encoded location (lat-long, grouped rare localities).
  - Finalized cleaned dataset with engineered features.
- **FUTURE WORK PLANS** :
  - Build baseline models (Linear, Decision Tree, XGBoost).
  - Evaluate using RMSE/MAE.
  - Compare initial model performance.

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 03/06/2025 – 10/06/2025

- **WPR No.** : 05
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- **TARGETS SET FOR THE WEEK** :
  - Train multiple regression models (Linear, Decision Tree, Random Forest)
  - Evaluate models using RMSE, MAE, and R^2 metrics
  - Identify best baseline model
- **ACHIEVEMENTS FOR THE WEEK** :
  - Trained and compared multiple models
  - Random Forest achieved best overall performance
- **FUTURE WORK PLANS** :
  - Perform hyperparameter tuning (e.g., GridSearchCV or RandomizedSearchCV)
  - Analyze feature importance and interpret model outputs
  - Begin designing deployment logic with Streamlit (interface planning)
  - Prepare data and model pipeline for integration

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 13/05/2025

- **WPR No.** : 06
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)
- **TARGETS SET FOR THE WEEK** :
  - Tune top regression models (e.g., Random Forest, XGBoost) using hyperparameter optimization
  - Identify key features influencing property price
  - Start designing the Streamlit app layout and input structure
- **ACHIEVEMENTS FOR THE WEEK** :
  - Performed hyperparameter tuning using GridSearchCV
  - Extracted and visualized top price-driving features
  - Drafted the Streamlit app layout with planned inputs (location, area, rooms, etc.)
- **FUTURE WORK PLANS** :
  - Build the interactive Streamlit web app
  - Connect the trained model to make live predictions
  - Add charts and maps to explain model outputs and price trends

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

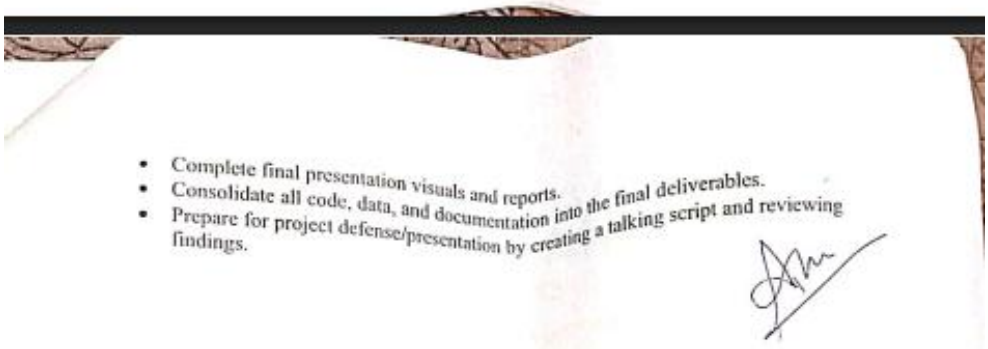
For the week commencing on: 13/05/2025

- **WPR No.** : 07
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)

- **TARGETS SET FOR THE WEEK** :
- Develop geospatial visualizations showing predicted property prices across different areas.
- Integrate GeoPandas and Folium libraries to create interactive maps.
- Start preparing final presentation visuals (charts, graphs, and maps).

- **ACHIEVEMENTS FOR THE WEEK** :
- Successfully integrated GeoPandas and Folium into the project workflow.
- Created interactive maps displaying actual vs. predicted prices by location.
- Developed heatmaps and point maps highlighting price variation trends across regions.
- Began drafting final presentation slides with map screenshots and interpretation of spatial trends.

- **FUTURE WORK PLANS** :
- Refine map aesthetics (color scales, labels, legends) for clarity and impact.



- Complete final presentation visuals and reports.
- Consolidate all code, data, and documentation into the final deliverables.
- Prepare for project defense/presentation by creating a talking script and reviewing findings.

AMITY UNIVERSITY
UTTAR PRADESH

DATA SCIENCE

AMITY INSTITUTE OF APPLIED SCIENCES

NTCC REPORT
WEEKLY PROGRESS REPORT (WPR)

For the week commencing on: 13/05/2025

- **WPR No.** : 08
- **PROGRAM** : MSDS (2024-2026)
- **STUDENT NAME** : NITESH
- **ENROLLMENT NUMBER** : A044161824008
- **FACULTY GUIDE'S NAME** : Dr. ANU SIROHI
- **NTCC REPORT TITLE** : REAL ESTATE PRICE PREDICTION (REGRESSION + GEOSPATIAL)

- **TARGETS SET FOR THE WEEK** :
  Finalize project documentation, perform final review, and prepare for submission/presentation.

- **ACHIEVEMENTS FOR THE WEEK** :
  Successfully completed all phases of the project, including data analysis, model development, and interpretation of results. Final report compiled with detailed documentation of methods, results, and insights. Prepared and organized all necessary files for submission.

- **FUTURE WORK PLANS** :
  Present the project to the evaluation panel, incorporate any final feedback, and submit all deliverables. Reflect on key learnings and consider extending the work for future applications or publications.