July 15, 2024

C964: Computer Science Capstone Template

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 3](#_Toc173576530)

[Part B: Project Proposal Plan 5](#_Toc173576531)

[Project Summary 5](#_Toc173576532)

[Data Summary 6](#_Toc173576537)

[Implementation 7](#_Toc173576543)

[Timeline 8](#_Toc173576546)

[Evaluation Plan 9](#_Toc173576547)

[Resources and Costs 10](#_Toc173576550)

[Part C: Application 11](#_Toc173576553)

[Part D: Post-implementation Report 12](#_Toc173576554)

[Solution Summary 12](#_Toc173576555)

[Data Summary 12](#_Toc173576556)

[Machine Learning 13](#_Toc173576560)

[Validation 15](#_Toc173576564)

[Visualizations 16](#_Toc173576569)

[User Guide 16](#_Toc173576570)

[Reference Page 19](#_Toc173576571)

# Part A: Letter of Transmittal

123 Tech Boulevard, Silicon Valley, CA 94000

September 29, 2024

Ms. Priya Sharma

Chief Operating Officer

Arko Holdings

789 Global Avenue, Mumbai, 400001, India

Dear Ms. Sharma,

India's economy has been experiencing unprecedented growth, emerging as one of the fastest-growing major economies globally. This rapid expansion is particularly evident in the real estate sector, driven by increasing urbanization, rising income levels, and favorable government policies promoting infrastructure development. While this growth presents significant opportunities for property investment and management firms like Arko Holdings, it has also introduced complexities in accurately valuing and pricing rental properties.

The volatility and unique characteristics of the Indian real estate market have made it difficult to rely on traditional pricing models. Current strategies, based on international market practices, while effective elsewhere, fail to capture the nuances of India's diverse and rapidly evolving property landscape. This misalignment has led to inefficiencies in pricing, with properties either undervalued or overpriced. The consequences include extended vacancy periods, missed revenue opportunities, and a weakened competitive position in the market.

To address these challenges, our team at Arko Data Solutions proposes a tailored approach that leverages advanced machine learning techniques to align with the specifics of the Indian real estate market. With the power of AI and machine learning, we can accurately forecast rental prices based on key property attributes such as location, size, number of bedrooms, and other critical features. Our solution, developed using data from Indian cities, will ensure that it reflects local market trends and conditions. By analyzing thousands of property records and factoring in unique local variables—such as neighborhood development, proximity to tech hubs, and cultural preferences—our model will deliver precise, data-driven insights for optimal pricing strategies. This will help reduce vacancy rates, boost revenues, and enhance your property management operations.

In addition, the model will be seamlessly integrated into a user-friendly interface, designed to be intuitive for both technical and non-technical users. Your team will be able to input relevant property data and receive rental price forecasts instantly, with real-time adjustments available as market conditions evolve. This flexibility will allow Arko Holdings to make informed decisions, from pricing individual properties to analyzing broader trends across cities, ensuring that your pricing strategies are always aligned with the latest market shifts.

For implementation, we will begin by thoroughly analyzing your dataset, which includes over 4,700 records of rental properties across India. Our team will then develop and train the machine learning model, ensuring both accuracy and scalability for long-term use. Following the model’s completion, we will integrate it into a custom web application that will seamlessly fit into your existing infrastructure. The project is expected to be completed within six weeks. During this time, we will collaborate closely with your team to ensure the final product meets your needs and exceeds expectations. Upon completion, we will conduct rigorous testing and provide comprehensive documentation and training, ensuring ease of use and adaptability over time. The estimated project cost is just under $100,000, covering end-to-end model development, deployment, labor, hardware, and infrastructure costs. A detailed breakdown of costs and the project timeline will be provided with the formal proposal or upon request.

Our team at Arko Data Solutions brings extensive experience in developing and deploying enterprise-level applications and machine learning solutions tailored specifically to the real estate market. As Lead Software Engineer, I have overseen numerous projects of this scope, combining deep expertise in statistical learning with a solid understanding of large-scale application deployment. Our team’s proven track record in delivering high-quality, data-driven solutions positions us to offer a system that will enable Arko Holdings to navigate the complexities of the Indian housing market with confidence.

We are confident that our proposed solution will empower Arko Holdings to optimize rental pricing strategies, improve profitability, and maintain a competitive edge in India’s rapidly growing real estate market. I look forward to discussing the next steps with you and am eager to support Arko Holdings in achieving continued success.

Sincerely,

Daniel Akoko

Daniel Akoko  
Lead Software Engineer  
Arko Data Solutions

# Part B: Project Proposal Plan

## Project Summary

India, the world's most populous country, has emerged as one of the fastest-growing economies globally, driven by favorable government policies promoting urbanization and infrastructure development. This rapid expansion has attracted numerous investors, including Arko Holdings, a leading international property investment and management firm. Despite the promising opportunities in India’s real estate market, Arko Holdings faces significant challenges in accurately pricing rental properties across various cities.

Due to fluctuating market conditions, driven by factors such as rapid urbanization, rising income levels, and increasing demand for housing, predicting rental prices has become increasingly difficult. These market dynamics have created uncertainties for developers, investors, property owners, and tenants alike. The diverse nature of India’s housing market, which includes luxury apartments and more modest dwellings, further complicates the task. As a result, Arko Holdings has experienced inefficiencies in their pricing strategies, leading to extended vacancy periods for some properties and undervaluation of others. These challenges have negatively impacted profitability and reduced their competitive advantage in the market.

Arko Holdings, managing a portfolio of over 5,000 rental properties across major Indian cities, requires a solution that allows them to better predict rent prices. This solution should enable them to optimize investment decisions, manage their property portfolio effectively, and offer fair and market-aligned rental prices to tenants.

To meet this need, the engineers at Arko Data Solutions propose developing a robust and intuitive predictive model using advanced machine learning techniques. This model will be capable of accurately forecasting rental prices based on key property attributes, such as location, property size, number of bedrooms, and other correlated features. It will be trained using a comprehensive dataset of over 4,700 rental properties from major Indian cities. By using property data exclusively from India, the model will capture the unique characteristics and trends of the local real estate market. Moreover, it will be capable of adjusting rental price predictions dynamically as market conditions evolve, ensuring that Arko Holdings can respond promptly to market fluctuations.

The key deliverables for this project include a machine learning model designed to predict optimal rental prices across various Indian cities. This model will incorporate property attributes such as the number of bedrooms, property size, and location to produce accurate and generalized predictions. It will allow Arko Holdings to make data-driven decisions for their rental properties, minimizing vacancies and maximizing revenues. In addition, an intuitive user-friendly interface will be developed for stakeholders to interact with the model. Both technical and non-technical users will be able to input key property details and receive instant rental price forecasts, empowering decision-makers with real-time insights for both individual properties and broader market trends. Alongside these deliverables, comprehensive documentation will be provided to ensure transparency and reproducibility, including a user guide with instructions for local installation and usage of the solution.

By implementing this predictive model, Arko Holdings will be able to improve its ability to accurately price rental properties across various cities in India. The solution will enable the company to optimize its pricing strategies, reducing the risk of overpricing or underpricing properties. Accurate pricing will also help reduce vacancy rates, ensuring higher occupancy and maximizing revenue. The model will provide actionable insights that can support strategic decision-making, allowing Arko Holdings to make informed choices about expanding into new markets or adjusting investments in different cities based on predicted rent trends. Overall, this solution will not only enhance profitability but also strengthen stakeholder satisfaction by offering fair and competitive rent prices that align with market conditions.

## Data Summary

**Dataset Link**: [House Rent Prediction Dataset on Kaggle](https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset)

The raw data for this project will be sourced from Kaggle's public dataset, titled the "[House Rent Prediction Dataset](https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset)" This dataset contains over 4,700 records of residential rental properties in India and includes 12 key features such as rent, property size, number of bedrooms, furnishing status, tenant preferences, locality, and other essential property attributes. These features are highly relevant for developing a predictive model for rental prices. The data was pre-collected and made publicly available by third-party contributors, ensuring its reliability and comprehensiveness for our purposes.

The data collection process will begin with downloading the dataset in CSV format from Kaggle. Once acquired, an exploratory data analysis (EDA) will be conducted to thoroughly understand the structure of the data and to identify any missing values or inconsistencies. Missing numerical data will be addressed using imputation techniques, such as mean or median imputation, while missing categorical data will be handled using mode imputation. Flags will be created to mark missing values, allowing the model to account for any potential patterns in the absence of data. For skewed distributions, log transformations will be applied to normalize the data, and advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique) will be used to balance any imbalanced categorical variables.

During the design and development stages, the data will undergo feature selection to identify variables that are highly correlated with rental prices. Irrelevant or redundant features with low correlation will be excluded to improve the model's accuracy. Categorical variables will be encoded into numerical representations using techniques such as one-hot encoding, while numerical features will be standardized to ensure consistency across all inputs. The processed data will be used to train the machine learning model, which will be designed to predict rental prices based on the selected features.

As part of the model’s lifecycle, regular updates to the dataset will be incorporated to reflect changes in the Indian rental market. The model will be periodically retrained with fresh data to maintain its accuracy and relevance. Data security and integrity will be ensured by storing the data in secure environments, enabling reproducibility of results and adherence to best practices in data management.

This dataset is ideally suited for the project due to its size, diversity, and recency. With over 4,700 records, it provides a sufficiently large sample size to train a robust predictive model. The variety of features, such as property size, location, and tenant preferences, allows the model to generalize across different market segments and property types, from luxury apartments in major cities to more affordable suburban homes. Additionally, the dataset captures critical attributes that influence rental prices, enabling fine-grained predictions tailored to various market segments. The dataset reflects contemporary trends in India's housing market, making it highly relevant for Arko Holdings' needs.

Ethical and legal considerations have been addressed, as the dataset is publicly available and anonymized. There are no personally identifiable details in the data, ensuring compliance with privacy standards. This mitigates any concerns about data misuse or privacy violations. The dataset is provided for educational and analytical purposes, and proper attribution to its source on Kaggle will be maintained throughout the project, aligning with standard data governance practices.

## Implementation

For the successful development and deployment of the rent price prediction model, we will adopt the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM is a widely recognized, industry-standard framework for data mining and machine learning projects, offering flexibility and alignment with agile development practices. It ensures that the project follows a systematic, iterative process that can adapt to evolving needs and insights throughout its lifecycle. Below is a detailed outline of the project’s implementation plan based on the six phases of CRISP-DM.

tion plan based on the six phases of CRISP-DM.

1. **Business Understanding**  
   In this initial phase, we will work closely with Arko Holdings’ stakeholders to clearly define the business objectives. This will involve identifying key goals, such as optimizing rent pricing strategies, reducing vacancy periods, and improving overall revenue across various Indian cities. Specific use cases, like providing insights to property owners and investors, will be defined. Once the objectives are established, we will translate them into machine learning objectives, ensuring that our model aligns with the business needs.
2. **Data Understanding**  
   The data for this project will be sourced from Kaggle’s publicly available House Rent Prediction dataset, which includes over 4,700 property listings across India. Additional real estate sources may be integrated for comprehensive data coverage. In this phase, we will assess the data quality, completeness, and relevance. Through exploratory data analysis (EDA), we will gain insights into data distributions and key patterns, identify missing or anomalous data, and determine which features (e.g., property size, location, amenities) have the highest correlation with rent prices.
3. **Data Preparation**  
   Data preparation is crucial to ensure that the model is trained on clean, relevant, and well-structured information. We will handle missing data using imputation techniques and perform feature engineering to create new variables based on domain knowledge. For instance, tenant preferences and furnishing status could be used to create derived features that improve model accuracy. Additionally, categorical variables will be encoded (e.g., one-hot encoding), and numerical features will be normalized or standardized. We will also handle outliers and anomalies, ensuring that the dataset is ready for effective model training.
4. **Modeling**  
   In this phase, we will develop and train several machine learning models, starting with baseline models like linear regression. More advanced models such as Random Forest and Gradient Boosting will then be implemented to capture more complex patterns in the data. The dataset will be split into training, validation, and test sets to ensure that the model generalizes well to unseen data. During this phase, we will also fine-tune model hyperparameters using techniques like grid search to optimize performance.
5. **Evaluation**  
   Once the models are trained, they will be evaluated against established performance metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared values. Cross-validation will be used to ensure robustness and avoid overfitting. This phase will help us determine the best-performing model for the prediction task and allow for necessary refinements before moving to deployment.
6. **Deployment**  
   In the final phase, we will deploy the chosen model as part of a scalable, enterprise-grade solution. A cloud-based deployment strategy will be favored to enable easy access and integration with Arko Holdings’ existing systems. The model will be versioned to facilitate updates as new data becomes available, and we will implement a monitoring system to track model performance over time. Additionally, a feedback loop will be established, allowing stakeholders to continuously provide input and ensure that the model remains aligned with business objectives as market conditions evolve.

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Duration (days)** | **Projected start date** | **Anticipated end date** |
| Project Kickoff | 1 | 2024-10-01 | 2024-10-01 |
| Business Understanding | 3 | 2024-10-02 | 2024-10-04 |
| Data Understanding | 5 | 2024-10-07 | 2024-10-11 |
| Data Preparation | 5 | 2024-10-14 | 2024-10-18 |
| Modeling and Evaluation | 3 | 2024-10-21 | 2024-10-23 |
| User Interface Design | 5 | 2024-10-24 | 2024-11-30 |
| Testing and Deployment | 3 | 2024-10-31 | 2024-11-04 |
| Documentation and User Guide | 2 | 2024-11-05 | 2024-11-06 |
| Final Presentation and Handover | 2 | 2024-11-07 | 2024-11-08 |

Total project duration: 29 working/business days (approximately 6 weeks)

## Evaluation Plan

The evaluation plan for the rent prediction model will incorporate both verification and validation processes to ensure that the machine learning model is accurate, reliable, and aligned with Arko Holdings’ business objectives. This dual approach will facilitate the identification and rectification of any discrepancies or issues throughout the development cycle.

**Verification Process**  
Verification will be implemented at each stage of the model development to ensure that the product meets its specifications and requirements. The process will begin during the **data preprocessing** phase, where data integrity will be rigorously checked for missing values, correct formatting, and appropriate encoding. This initial verification will ensure a clean and consistent dataset, which is vital for effective model training.

In the **feature engineering** phase, we will employ cross-validation techniques to verify that newly created features enhance model performance without leading to overfitting. By splitting the dataset into training and validation subsets, we can assess the contribution of each feature to the model’s predictive capability.

During **model training**, we will implement regularization techniques to penalize overly complex models and maintain generalizability. Additionally, hyperparameter tuning will be performed to optimize the model’s settings for better accuracy and efficiency. Code verification will also be integrated through unit tests and continuous integration practices, ensuring that all components of the model function correctly and efficiently throughout the development lifecycle.

**Validation Process**  
Upon completion of the development phases, we will initiate the validation process to assess the machine learning model's performance on unseen data. The model's predictive accuracy will be evaluated using key performance metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). These metrics will provide a comprehensive overview of how well the model predicts rental prices compared to actual values.

To further substantiate the model's robustness, cross-validation methods, such as k-fold validation, will be employed. This technique involves partitioning the dataset into k subsets and training the model multiple times, ensuring that every data point is used for both training and validation. This process will help identify any potential biases or edge cases that may affect model performance.

Finally, we will validate the model against predefined business objectives to confirm its alignment with market realities and Arko Holdings’ strategic goals. This holistic approach will ensure that any identified biases are addressed and that the final model is not only accurate but also effective in meeting the unique needs of the Indian rental market.

## Resources and Costs

### **Hardware and Software Costs (US-based)**

|  |  |
| --- | --- |
| **Item** | **Cost (USD)** |
| High-performance workstation for model training (2 units) | $6,000 |
| Cloud computing resources (e.g., AWS, Google Cloud) | $5000/month |
| Data visualization software licenses (e.g., Tableau) | $350/month/user |
| Development software licenses (e.g., PyCharm Professional) | $1000/year/developer |
| Version control system (e.g., GitHub Enterprise) | $210/month/user |
| Total Hardware and Software Costs | $6,000 + $5,560/month |

**Estimated Labor Time and Costs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Role** | **Team Size** | **Hourly Rate (USD)** | **Estimated Hours** | **Total Cost (USD)** |
| Data Science Team (India) | *4* | $25 x 4 | 128 | $12,800 |
| Software Development Team (India) | *4* | $20 x 4 | 40 | $3,200 |
| Quality Assurance Specialist (India) | *2* | $18 x 2 | 16 | $576 |
| Project Manager (US) | *1* | $110 | 200 | $ 22,000 |
| Total Labor Costs |  |  | 384 | $38,576 |

**Environment Costs (US-based)**

|  |  |
| --- | --- |
| **Item** | **Monthly Cost (USD)** |
| Cloud hosting for web application (e.g., Heroku Professional) | $250 |
| Database hosting (e.g., Amazon RDS) | $100 |
| Continuous Integration/Continuous Deployment (CI/CD) pipeline | $50 |
| Monitoring and alerting system | $75 |
| Ongoing maintenance and support (estimated 20 hours/month at $20/hour) | $400 |
| Total Monthly Application Costs | $2,375 |

### **Total Projected Costs**

* Initial Costs (Hardware, Software, Labor): $91,280
* Monthly Recurring Costs: $7,935

# Part C: Application

The submitted application comprises an interactive user interface built using **ipywidgets** within a Jupyter notebook. This interface allows users to input property attributes and receive predicted rental prices in real time. The machine learning model, which was trained in Jupyter notebooks, utilizes a combined approach with both RandomForest and GradientBoosting algorithms to ensure accurate predictions based on historical rental data. The model was developed using Python libraries, including scikit-learn for model implementation, pandas for data manipulation, and NumPy for numerical computations.

The user interface is designed for ease of use, allowing executives at Arko Holdings to input key property variables such as the number of bedrooms, property size, area type, city, furnishing status, and more. Once the user submits the form, the application processes the input and outputs a rental price prediction using the trained model. The prediction is displayed directly within the notebook, providing real-time insights.

Following the ‘User Guide’ provided in Part D, evaluators can successfully run the application in a Jupyter notebook environment, such as Google Colab or a local machine. The provided instructions guide users through uploading the notebook and required files, running the model, and interacting with the embedded UI.

All necessary files, including the Jupyter notebooks used for training the model, the CSV dataset, and any supporting code, are submitted to ensure that evaluators can access and review the project comprehensively.

# Part D: Post-implementation Report

## Solution Summary

Arko Holdings, our sister company under Arko Inc., faced a significant challenge in predicting rent prices across their extensive property portfolio in India. The rapidly evolving real estate market, influenced by urbanization, rising incomes, and fluctuating housing demand, rendered their existing pricing strategies ineffective, leading to prolonged vacancies and property undervaluation. This directly impacted their profitability and market competitiveness.

To address this issue, the engineers at Arko Data Solutions developed a sophisticated machine learning model, leveraging data from over 4,700 Indian residential properties. Our solution combines Gradient Boosting and Random Forest Regressors to accurately predict optimal rent prices based on key features such as the number of bedrooms and bathrooms, property size, location, and furnishing status. The model's high accuracy, demonstrated by low RMSE and high R-squared values, enables Arko Holdings to swiftly adapt to market changes.

We further enhanced the solution with an intuitive user-friendly interface, allowing users to input property details and receive real-time predictions. This comprehensive tool empowers property owners and investors to make informed decisions, thereby reducing vacancy rates and maximizing revenue. The seamless integration of our model into Arko's existing systems ensures scalability and long-term usability, facilitating strategic expansion into other markets like Mumbai and Delhi. By providing data-driven insights and streamlining the pricing process, our solution not only resolves Arko Holdings' immediate challenges but also positions them for sustainable growth and enhanced competitiveness in the dynamic Indian real estate market.

## Data Summary

## The dataset used in this project was sourced from [Kaggle](https://www.kaggle.com/datasets/iamsouravbanerjee/house-rent-prediction-dataset), comprising over 4,700 records of rental properties located across major Indian cities. Each record in the dataset contained detailed information about various property attributes critical to predicting rental prices. These attributes included the number of bedrooms, hall, and kitchen (BHK), the total size of the property in square footage, and the type of area classified as either Carpet Area, Super Area, or Built Area. Additionally, the dataset recorded the city in which the property was located, the furnishing status (categorized as Furnished, Semi-Furnished, or Unfurnished), tenant preferences (such as Bachelors, Family, or Bachelors/Family), and the number of bathrooms. Other important features included the point of contact (whether the owner, agent, or builder was the primary contact for the property), the floor level, and the total number of floors in the building.

## In the data processing stage, several steps were taken to prepare the data for effective modeling. Initially, the dataset underwent cleaning to handle missing values, remove duplicates, and correct any inconsistencies in formatting, such as adjusting spaces in column names. After the initial cleaning, feature engineering was employed to extract additional useful attributes. For example, the floor column was split into two distinct features: floor\_level and total\_floor, which provided more granular information about the property. Additionally, the posted\_on column, which indicated when the property was listed, was transformed into temporal features such as month, day, and weekday, allowing the model to capture potential seasonal trends in rental prices.

## The distribution of the rent feature, which was initially right-skewed, was normalized by applying a log transformation. This transformation reduced the impact of extreme values and made the data more suitable for machine learning models. To prepare categorical variables for modeling, label encoding was applied. This converted non-numeric attributes, such as city names, area types, and furnishing statuses, into numerical representations that the models could interpret. Finally, the dataset was split into training and test sets, with 80% of the data used for training the model and 20% reserved for testing and evaluating model performance. This approach ensured that the model could be trained effectively while retaining a separate dataset for unbiased validation.

## Machine Learning

The goal of this project was to develop a machine learning model capable of predicting rental prices based on various property attributes. This model was designed to help Arko Holdings create more accurate and dynamic pricing strategies, ultimately minimizing vacancies and maximizing revenue. To achieve this, we leveraged several machine learning techniques and tools to ensure that the model could handle the complexity of the housing market in India.

In the development of the machine learning solution, Python libraries such as pandas and NumPy were used for data preprocessing. These libraries facilitated cleaning and preparing the dataset, which included handling missing values, transforming categorical features, and splitting the data into training and test sets. The scikit-learn library was central to building the machine learning models, with Matplotlib and Seaborn used for visualizing data trends and feature relationships.

Two machine learning algorithms were selected for this task: the **RandomForestRegressor** and the **GradientBoostingRegressor**. RandomForest is a powerful ensemble method that constructs multiple decision trees, each trained on different subsets of the data, and averages their predictions. This approach ensures stable and accurate results, particularly in regression tasks with complex feature relationships, like those in rental property pricing. GradientBoosting, another ensemble method, iteratively refines its predictions by focusing on the errors made by previous trees, further improving the model's accuracy. This method excels at fine-tuning predictions, especially in datasets where noise and outliers might otherwise skew the results.

The machine learning development process involved several stages. First, data preprocessing was performed, during which the dataset was cleaned and relevant features, such as bhk, size, area\_type, city, and furnishing\_status, were transformed and encoded. The data was then split into training and test sets using an 80/20 ratio to evaluate model performance effectively. After preprocessing, the models were trained. Both the RandomForest and GradientBoosting algorithms were trained on the full feature set, with hyperparameters tuned through cross-validation to optimize model performance.

The final step involved combining the predictions from both models to improve accuracy. The ensemble approach was employed by averaging the predictions from RandomForest and GradientBoosting, leveraging the strengths of each model. RandomForest's generalization capability and GradientBoosting's iterative error correction provided a balanced, robust prediction mechanism. This combination ensured that the model not only made accurate predictions but also handled the variability in the data effectively.

The selection of these algorithms was driven by their suitability for handling large datasets and complex, non-linear relationships between features and target variables. RandomForest was chosen for its ability to generalize well to unseen data and prevent overfitting through the use of bagging techniques. GradientBoosting, on the other hand, was selected for its ability to iteratively improve upon predictions, making it a good complement to RandomForest. Other models, such as Linear Regression, were initially tested, but they did not perform as well, given the complexity of the feature space and the nature of the rental price prediction task.

By combining these two models, the solution capitalized on the strengths of both, ensuring a high level of predictive accuracy and robustness. This approach allowed the model to generalize well to new, unseen data while maintaining low error rates, making it an effective tool for Arko Holdings in their pricing strategy efforts.

## Validation

The performance of the machine learning models was validated using multiple metrics to ensure both accuracy and robustness in predicting rental prices. The primary metric used was Root Mean Squared Error (RMSE), which measures the average magnitude of prediction errors. The combined model, which averaged the predictions from the RandomForest and GradientBoosting models, achieved an RMSE of 0.396. This value indicates that, on average, the model’s predictions deviated from the actual rent prices by approximately 39.6%. The relatively low RMSE suggests that the model performed well in minimizing large prediction errors.

Another key metric used was the R-Squared (R²) score, which explains how much of the variance in the rental prices is captured by the model. The combined model achieved an R² score of 0.822, meaning that it could explain over 82% of the variation in rental prices. This high R² score reflects the model's ability to account for the relationship between property attributes and rent, indicating that the model generalizes well to different properties within the dataset.

Additionally, the Mean Absolute Error (MAE) was calculated to assess the average difference between the predicted and actual rental prices. The model's MAE value was 0.295, providing a direct interpretation of the average prediction error in terms of absolute rent values. This metric, combined with RMSE and R², gave a comprehensive view of the model's performance across different types of errors.

These validation results demonstrated that the combined model consistently outperformed individual models, providing higher accuracy and generalization to unseen data. The ensemble approach, which combined RandomForest’s generalization strength and GradientBoosting’s fine-tuning capability, effectively captured the complexities of the Indian rental market. By leveraging these two models, the solution delivered predictions that were not only accurate but also reliable in varied market conditions.

## Visualizations

The application generates various visualizations to analyze data and model performance:

A graph with a blue line

Description automatically generated

**[Figure 1:** *Rent Distribution using displot, qqplot and boxplot***]**

**A screenshot of a graph

Description automatically generated**

**[Figure 2:** *Rent overview – before & after log transformation using stripplot***]**

**A screenshot of a graph

Description automatically generated  
[Figure 3:** *Categorical features’ univariate analysis using countplot***]**

**A screenshot of a graph

Description automatically generated**

**[Figure 4:** *Univariate analysis of Numerical features using subplots* **]**

**A screenshot of a graph

Description automatically generated  
[Figure 5:** *Bivariate analysis of Rent and Numerical features using scatterplot***]**

**A graph of different colored lines

Description automatically generated**

**[Figure 6:** *Bivariate analysis of Rent and Categorical features using kdeplot***]**

**A colorful squares with white text

Description automatically generated**

**[Figure 7:** *Correlation Matrix***]***.*

A screenshot of a computer

Description automatically generated  
  
**[Figure 8:** *Non descriptive visualization of the UI interface after prediction***]**

## User Guide

This guide was made to ensure a smooth process setting up the **Arko Data Solution’s Rent Price Prediction Model** in google collab and interacting with the UI directly within the notebook. Follow these steps below:

1. **Access Google Colab**: Open your browser and go to [Google Colab](https://colab.research.google.com/).

*Alternatively, you can access the notebook hosted on Google colab by clicking* [*here*](https://colab.research.google.com/drive/1iCKG9b6JfjrZQHY4IPKOADU0hoc5ckPL?usp=sharing) *and lauching the notebook in Colab with the ‘Open with Google Colab’ button at the top of the page*

1. **Upload Files**: In the Colab interface, click on the **Files** tab on the left side. Upload the Jupyter Notebook (.ipynb file) and the CSV file required for the model.
2. **Run the Notebook**: At the top of the Colab interface, click on **Runtime** and select **Run All**. This will execute all the code cells in the notebook, including loading the data, training the models, and preparing the UI.
3. **Wait for Model Training**: The program may take a few minutes to run, especially during the hyperparameter tuning phase for the Gradient Boosting Regressor. Please be patient while the notebook processes the data.
4. **Locate the Embedded UI**: Once the notebook has finished running, scroll down through the notebook to find the embedded user interface (UI). The form will be displayed in one of the output cells, ready for interaction.
5. **Input Property Details**: In the embedded UI, fill out each field with the appropriate data for the property you wish to evaluate (e.g., BHK, size, area type, city, etc.).
6. **Predict Rent**: After entering all the required details, click the **Predict Rent** button. The model will process your input and display the predicted rent price directly in the UI.
7. **View Prediction**: The predicted rent price will be displayed using the **Arko Data Solution’s rent price model**. The prediction is formatted to reflect Indian Rupees (₹).