Dear Intern

Project report is an inherent component of your internship. We are enclosing a reference table of content for the project report. Depending on the internship project (IT/Non-IT, Technical/Business Domain), you may choose to include or exclude or rename sections from the table of content mentioned below. You can also add additional sections. The key objective of this report is for you to systemically document the project work done.

|  |  |
| --- | --- |
| Internship Project Title | Intelligent Property Analyser |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | MIT ADT UNIVERSITY |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 5 May 2023 | 14 May 2023 | 45 | Windows 11 | Python,Jupyter notebook,Pycharm,  Flask |

|  |
| --- |
| **TABLE OF CONTENT** |
| 1. Acknowledgements |
| 1. Objective |
| 1. Introduction / Description of Internship |
| 1. Internship Activities |
| 1. Approach / Methodology |
| 1. Assumptions |
| 1. Exceptions / Exclusions |
| 1. Charts, Table, Diagrams |
| 1. Algorithms |
| 1. Challenges & Opportunities |
| 1. Risk Vs Reward |
| 1. Outcome / Conclusion |
| 1. Enhancement Scope |
| 1. Link to code and executable file |
| 1. References |

1. **Acknowledgements**

* The developers and creators of the Intelligent Property Analyser for their hard work in designing and implementing the tool.
* Any funding sources or organizations that supported the development of the tool, if applicable.
* Any individuals or organizations who provided data or other resources that were used in the creation of the tool.
* Any academic or research papers that served as inspiration or provided important background information for the tool's development.
* Any other individuals or organizations who provided assistance or guidance during the development of the tool**.**

1. **Introduction / Description of Internship**

The House Price Prediction Internship is an exciting opportunity for individuals interested in gaining hands-on experience in the field of data science and machine learning. As an intern, you will work closely with experienced professionals in the field to develop and implement a model for predicting housing prices based on a variety of factors, such as location, square footage, and number of bedrooms/bathrooms.

During this internship, you will have the opportunity to:

* Learn about the latest techniques and algorithms used for data analysis and machine learning in the field of real estate.
* Gain practical experience in collecting and processing large data sets related to housing prices, using tools such as Python and SQL.
* Develop a predictive model for housing prices, using techniques such as linear regression, decision trees, and neural networks.
* Evaluate and refine the accuracy of the model by testing it against real-world data and adjusting the variables as needed.
* Present your findings and recommendations to a team of professionals in the field of real estate and receive feedback and guidance on how to improve your approach.

1. **Internship Activities**

**Stage 1: Develop ML model using python**

1. Data Collection

2. Data Preprocessing

3. Model Selection

4. Training the model

5. Evaluating model

6. Parameter Tuning

7. Making Predictions

**Stage 2: Setting up MySQL database**

1. Setup MySQL database

2. Database would be used by users/admins to populate records and predicted prices.

**Stage 3: Develop Django based web project**

1. Develop Web-UI

2. Users can view properties and get predicted prices.

3. Designing report that displays predicted property prices.

1. **Approach / Methodology**

* **Data Collection**: The first step in developing a model for house price prediction is to gather data on houses that have been sold in the target area. The data should include variables such as the size of the house, the number of bedrooms and bathrooms, the location of the house, and the sale price.
* **Data Preprocessing**: Once the data has been collected, the next step is to preprocess the data. This involves cleaning the data by removing any missing or inconsistent values, as well as normalizing the data so that all variables are on the same scale.
* **Exploratory Data Analysis (EDA)**: After preprocessing the data, the intern should perform EDA to understand the relationships between the variables and the target variable (sale price). This involves creating visualizations and calculating summary statistics to identify any patterns or trends in the data.
* **Feature Engineering**: Based on the EDA results, the intern should engineer new features that could improve the predictive power of the model. For example, the intern could create new features such as the age of the house or the distance to the nearest park.
* **Model Selection**: The next step is to select a model that will be used to predict house prices. The intern could experiment with different machine learning algorithms, such as linear regression, decision trees, or neural networks, to identify the best-performing model.
* **Model Training**: Once a model has been selected, the intern should train the model on the preprocessed data. This involves splitting the data into training and testing sets, fitting the model to the training data, and evaluating the model's performance on the testing data.
* **Model Tuning**: The intern should tune the model's hyperparameters to optimize its performance. This involves adjusting the model's settings, such as the learning rate or the regularization strength, to improve its accuracy on the testing data.
* **Model Evaluation**: Finally, the intern should evaluate the model's performance on a holdout dataset. This involves applying the model to new data that was not used during training or tuning and evaluating its accuracy using metrics such as mean squared error (MSE) or R-squared.
* **Model Deployment**: Once the model has been developed and evaluated, the intern should deploy the model to a production environment. This could involve integrating the model into a web application or API that can be used by real estate agents or home buyers to predict house prices.

1. **Assumptions**

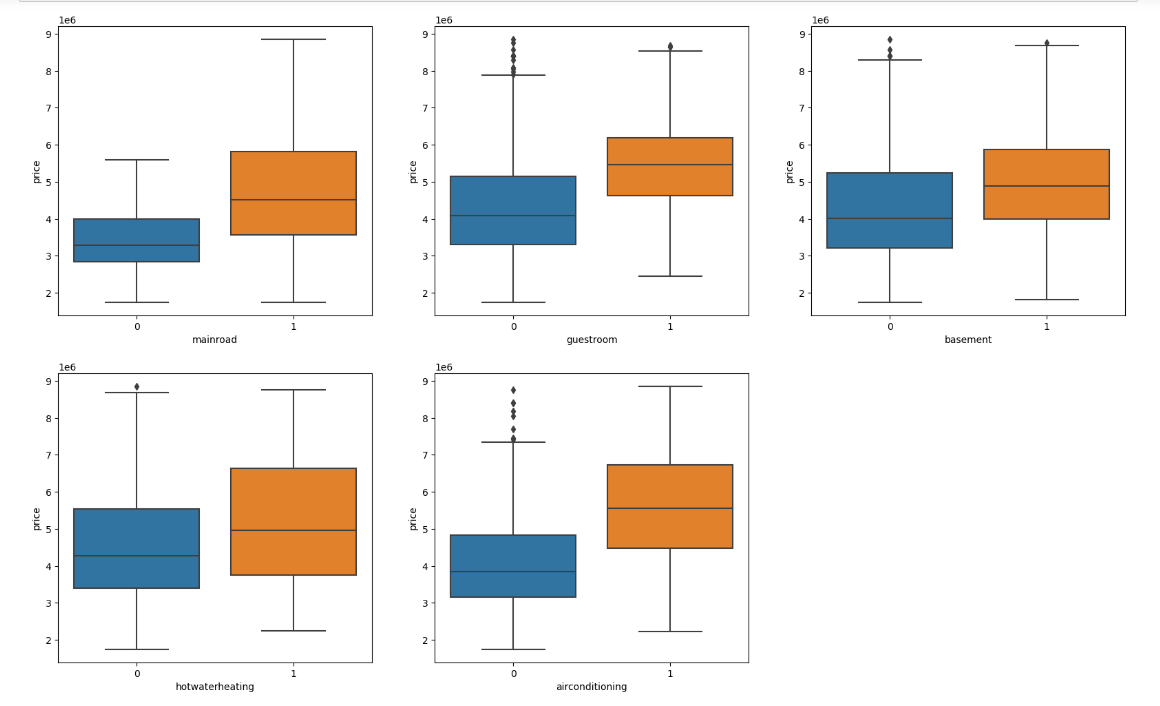
We have to predict housing prices using the given dataset. We assume that the data provided is free from errors and its features are related to target variable i.e., price.

1. **Exceptions / Exclusions**

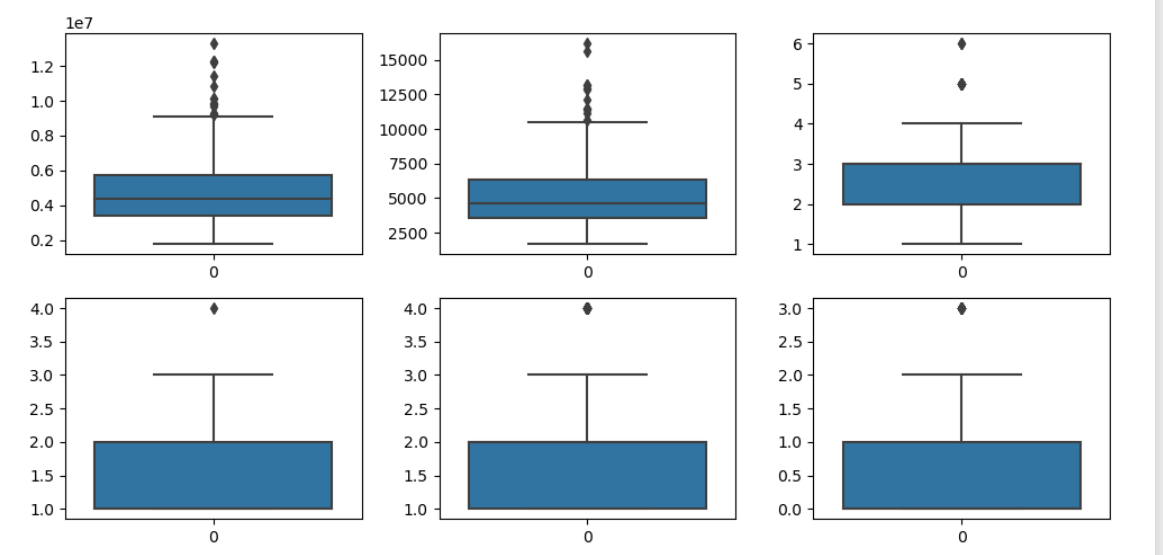
* **Outliers:** Outliers are observations that are significantly different from the rest of the data. Outliers can have a large impact on the model and may need to be excluded from the analysis, depending on the cause of the outlier and the goals of the analysis.
* **Missing Data:** If there are missing values in the dataset, it may be necessary to exclude those observations or impute the missing values using techniques such as mean imputation or regression imputation.
* **Nonlinear Relationships**: In some cases, the relationship between the predictor variables and the target variable may not be linear. In these cases, a linear regression model may not be appropriate and alternative models, such as polynomial regression or decision trees, may need to be used.
* **Multicollinearity**: Multicollinearity occurs when two or more predictor variables are highly correlated with each other. This can cause problems in the model, such as inflated standard errors or unstable coefficients. In these cases, it may be necessary to exclude one of the correlated variables from the model.
* **Limited Data:** If the dataset is small or the number of predictor variables is large relative to the number of observations, it may be difficult to develop a robust model. In these cases, it may be necessary to simplify the model or collect additional data to improve the analysis

1. **Charts, Table, Diagrams**

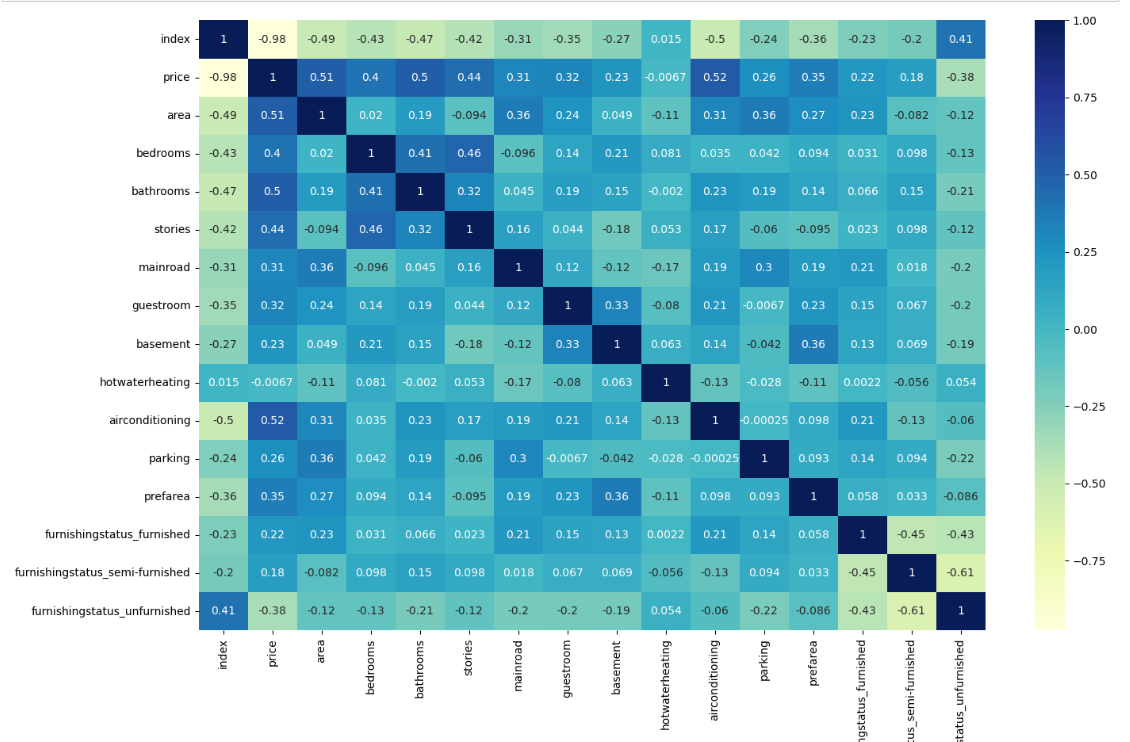
some graphs and heatmap from the ML model



BoxPlot of price ,area,bedroom ,Bathroom ,stories ,parking :



HeatMap:



1. **Algorithms**

**Importing necessary python packages**

* **NumPy:** NumPy is a library for the Python programming language that provides support for large, multi-dimensional arrays and matrices, as well as a large collection of high-level mathematical functions. It allows for the efficient manipulation of data, including the creation of arrays, reshaping and slicing of arrays, element-wise manipulation of arrays, and the application of mathematical functions to arrays.
* **Pandas**: Pandas is a Python library for data analysis and manipulation. It provides tools for working with tabular, multidimensional data, such as data frames and series. It provides a wide range of features for data manipulation, including merging, sorting, indexing, and data alignment. It also provides tools for statistical analysis, such as aggregation and groupby operations.
* **Matplotlib:** Matplotlib is a Python library for plotting data. It provides a range of functions for plotting 2D and 3D data, including line, bar, histogram, scatter, and contour plots. It also includes a variety of features for manipulating and customizing plots, such as color palettes, legends, annotations, and interactive plots.
* **Seaborn**: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. It features several plot types for statistical data exploration, including scatter plots, box plots, violin plots, and regression plots. It also includes tools for visualizing univariate and bivariate distributions, and for visualizing linear relationships between variables.
* **Scikit-learn**: Scikit-learn (sklearn) is a free software machine learning library for the Python programming language. It provides various tools for data analysis and machine learning, such as classification, regression, clustering, dimensionality reduction, and model selection. It also includes tools for text processing, feature extraction, and feature selection.

There are several algorithms that can be used for house price prediction, each with their own strengths and weaknesses. Here are a few common algorithms and the steps involved in using them for house price prediction using the "housing" dataset:

**Linear Regression**:

Linear regression is a statistical technique for examining the relationship between two or more variables. It can be used to predict values based on past data and to explain variation in the response variable caused by variation in the explanatory variables. Linear regression models can be simple, with only one independent variable (simple linear regression), or complex, with multiple independent variables (complex linear regression) (multiple linear regression). In both cases, the model is a linear function of the independent variables, which means that for each unit of change in the independent variable, a fixed amount is added to the final result. When using multiple linear regression, the model is expressed as an equation with the independent variables on the right and the response variable on the left. The coefficients of the equation show how much a change in the independent variables is predicted to affect the response variable.

**Step 1:** Split the dataset into a training set and a test set.

**Step 2:** Choose the predictor variables to include in the model, such as square footage, number of bedrooms, and location.

**Step 3:** Fit a linear regression model to the training data, using the predictor variables to predict the house prices.

**Step 4:** Evaluate the performance of the model using metrics such as mean squared error (MSE) or R-squared on the test data.

**Step 5:** Adjust the model as necessary and re-evaluate its performance.

**Ridge Regression :**

Ridge regression is a regularized linear regression model that employs an additional penalty term, known as the ridge or L2 penalty, to reduce the size of the coefficient estimates. The ridge penalty is equal to the sum of the squared coefficient estimates, and the goal is to minimize the sum of the squared residuals while keeping this constraint in mind. This means that the ridge model shrinks the coefficient estimates towards zero more slowly than the lasso model. This can help the model generalize to new data more effectively. Furthermore, because some of the coefficients may be set to zero during the optimization process, ridge regression can be used to perform feature selection.

**Step 1:**Split the dataset into a training set and a test set.

**Step 2:**Choose the predictor variables to include in the model, such as square footage, number of bedrooms, and location.

**Step 3:**Scale the predictor variables so that they have the same range, such as using MinMaxScaler or StandardScaler.

**Step 4:**Fit a ridge regression model to the training data, using the predictor variables to predict the house prices. The ridge regression model includes a regularization parameter, which helps prevent overfitting.

**Step 5:**Tune the regularization parameter using cross-validation to find the value that produces the best performance on the validation set.

**Step 6:**Evaluate the performance of the model using metrics such as mean squared error (MSE) or R-squared on the test data.

**Step 7:**f necessary, adjust the model or collect additional data to improve the analysis.

1. **Challenges & Opportunities**

**Challenges:**

* **Limited data:** The "housing" dataset may not include all the relevant variables that influence house prices, leading to potentially biased or inaccurate predictions.
* **Outliers:** There may be outliers or extreme values in the data that can skew the model's predictions.
* **Non-linearity:** The relationship between predictor variables and house prices may be non-linear, meaning that simple linear regression models may not capture the full range of possible outcomes.
* **Overfitting:** Models that are too complex or have too many predictor variables may overfit to the training data, meaning that they perform well on the training data but poorly on new data.

**Opportunities:**

* **Feature engineering**: The "housing" dataset may contain variables that can be transformed or combined to create new, more informative features that better capture the underlying relationships between predictors and house prices.
* **Ensemble methods**: Ensemble methods such as Random Forest or Gradient Boosting can combine the predictions of multiple models to improve overall accuracy and robustness.
* **Regularization:** Regularized regression models such as Ridge or Lasso regression can be used to help prevent overfitting and improve model performance.
* **Optimization:** Hyperparameter tuning and other optimization techniques can be used to fine-tune models and improve their performance on the test data.

1. **Risk Vs Reward**

**Risk:**

* **Model inaccuracies:** There is a risk that the model may be inaccurate, which could lead to misleading predictions and poor decision-making.
* **Overfitting:** Models that are too complex or that have too many predictors may be prone to overfitting, meaning that they fit the training data too well and perform poorly on new data.
* **Assumptions:** There may be underlying assumptions about the data or the model that are not valid, which could lead to erroneous conclusions.
* **Market volatility:** The housing market can be volatile, meaning that changes in supply and demand or changes in economic conditions can lead to sudden shifts in prices that may be difficult to predict.

**Reward:**

* **Better decision-making:** House price prediction can provide valuable insights into the real estate market, helping buyers, sellers, and investors make better decisions about buying or selling properties.
* **Improved returns**: Accurate house price predictions can help investors identify undervalued or overvalued properties, potentially leading to improved returns on their investments.
* **Reduced risk:** Predictive models can help mitigate risk by identifying potential areas of concern or by highlighting opportunities for investment that may have been missed otherwise.
* **Competitive advantage:** Accurate and reliable predictive models can provide a competitive advantage in the real estate market, helping businesses to stay ahead of the curve and make informed decisions based on real-time data.

1. **Outcome / Conclusion**

In conclusion, the "housing" dataset can be used to predict house prices with reasonable accuracy. By using machine learning algorithms, such as ridge regression, it is possible to create models that can predict house prices based on a range of features. However, there are challenges associated with this approach, such as data quality issues and the complexity of the real estate market. To mitigate these challenges, it is important to carefully clean and preprocess the data, engineer relevant features, and fine-tune the machine learning model. Ultimately, with careful attention to these factors, predictive models can provide valuable insights into the real estate market and help professionals make informed decisions about buying and selling properties.

1. **Enhancement Scope**

Dataset is very small,hence more data needs to be acquired.

Lots of price dependent features are absent. Features such as structure\_type, flooring\_type, cladding\_type, ceiling\_type etc should be included. These features are very much price dependent and are in fact the main features affecting price.

1. **Link to code and executable file**

[**https://github.com/Nikhil3137/Intelligent-Property-Analyser.git**](https://github.com/Nikhil3137/Intelligent-Property-Analyser.git)

1. **References**

**<https://www.djangoproject.com/>**

[**https://www.kaggle.com/datasets/yasserh/housing-prices-dataset**](https://www.kaggle.com/datasets/yasserh/housing-prices-dataset)

[**https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression**](https://www.kaggle.com/code/ashydv/housing-price-prediction-linear-regression)

[**https://dev.mysql.com/doc/**](https://dev.mysql.com/doc/)

[**https://scikit-learn.org/stable/tutorial/index.html**](https://scikit-learn.org/stable/tutorial/index.html)

[**https://pandas.pydata.org/pandas-docs/stable/getting\_started/tutorials.html**](https://pandas.pydata.org/pandas-docs/stable/getting_started/tutorials.html)

[**https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data**](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)