## Predictive Analytics for Real Estate: Price Prediction of Houses using Machine Learning

PROJECT MANAGEMENT

Ø **Choosing role**: - Project manager has assigned different roles to all the team members that are aligned with their interests and qualifications. These roles include researcher, Solution and design, Development, Testing- QA, Documentation, and Implementation/Deployment.

|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Roles** |
| Husanpreet Kaur | 500195671 | Researcher |
| Nanda Kishore Karicherla | 500197946 | Solution and design |
| Tajdar Unnisa Begum | 500201392 | Project Manager , Development |
| Surinder Singh Katal | 500202036 | Testing and QA |
| Maninder Singh | 500202625 | Documentation |
| Gurjot Singh | 500202820 | Implementation/Deployment |

Ø **Timeline**: For this week data set was provided and with the help of python, we were able to clean the data and filter out all the necessary data required to improve the quality of data. We researched on real estate industry and found that this industry is dependent on multiple factors such as prices, location, size, and locality.

Ø **Platform :** The project makes use of the Python programming language and few data analysis tools, including scikit-learn, pandas, and NumPy. Jupyter Notebook and Google collab is used to set up the development environment, and Git is used to maintain version control. We have used Microsoft Teams and outlook for sharing our researches, outcome, resources, files worked on and discussion.

Ø **Action plan:** The team worked together to specify the project's objectives, goals, and deliverables. The processes necessary to clean and transform the raw data, generate the prediction model, and assess its effectiveness were outlined in a plan of action.

Ø **MEETING DOCUMENTATION AND PROOF OF DISCUSSION**: Most of the discussions were in the class and shared documents on teams and outlook. Below is the screenshot attached.

**SOLUTION FLOW**

Ø **Requirement Finalization:** The team conducted thorough discussions to understand their specific requirements and expectations for the predictive model of Natty City data set.

Ø **Approach:** The team decided to follow a supervised machine learning approach for now, using a regression algorithm to predict house prices based on the provided features. The data will be divided into training and testing sets to evaluate the model's performance.

Ø **Data Cleaning and Transformation:** This week our team has performed data cleaning tasks, handling missing values, outliers, and inconsistent data. Feature engineering techniques will be employed to extract relevant information from the raw data and create meaningful features for the model.

Ø **Model Development:** we will work together to develop and fine-tune the predictive model using appropriate regression algorithms. The model will be trained on the training data and validate using the testing data.

Ø **Evaluation:** The model's performance will be evaluated using various metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared. The team will iteratively refine the model based on the evaluation results.

**PROJECT RESOURCES AND METHODOLOGY**

Ø **Resource Selection:** The team selected Python as the best programming language for data analysis and machine learning activities because of its large library and ecosystem support. The chosen libraries, including pandas and scikit-learn, offered effective tools for model creation and data manipulation.

Ø **Resource Quality Considerations:** The team will be using established documentation, user reviews to assure the credibility and quality of the resources it chose. To maintain the readability and quality of the code, best practices and coding standards will be adhered to.

Ø **Methods of Implementation:** The project will be using a collaborative and iterative process. To encourage a productive workplace, routine code reviews, pair programming, and knowledge exchange sessions will be held. Git version control will make it possible to integrate code quickly and effectively.

Ø **Analytics Techniques:** To obtain insights into the data, spot trends, and spot abnormalities, exploratory data analysis (EDA) methods will be used. The prediction model's significant characteristics will be derived using feature extraction and selection techniques.

**FINDING AND ANALYSIS**

Several features are frequently considered while developing a machine learning model for predicting home prices. These characteristics aid in identifying the crucial elements that affect a home's value. Here are some essential characteristics that are frequently included in models for predicting home price changes:

**DRAWBACKS OR FAILINGS:**

Ø One of the major challenges faced by the team was the presence of missing data in the raw dataset. Imputation techniques were applied to handle missing values, but it may have introduced some level of bias in the model.

**RESOURCES USED:**

[**https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e**](https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e)

[**https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms**](https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms)

Ø **Hours of Labor:** The team utilized AIP class hours, spread across researching on real estate, Exploring data and data cleaning.

Ø **Costs:** The project utilized existing resources on google, and no additional costs were incurred.

In conclusion, this project aims to develop a predictive model for Natty City to estimate house prices based on various features.

[10:09 PM, 2023-06-07] Nanda Canada: Price prediction in the real estate industry requires various techniques to handle different types of data and ensure accurate predictions. Here's an overview of techniques commonly used in real estate price prediction:

1. Handling Missing Values:

- Imputation: Fill missing values using techniques like mean, median, mode, or regression imputation.

- Deletion: Remove rows or columns with missing values if the missingness is low.

- Advanced techniques: Use more sophisticated methods like k-nearest neighbors (KNN) imputation or matrix factorization.

2. Handling Outliers:

- Statistical methods: Use statistical techniques like z-score, interquartile range (IQR), or boxplots to identify and handle outliers.

- Winsorization: Replace extreme values with a predefined percentile value.

- Transformation: Apply data transformations like log transformation to reduce the impact of outliers.

3. Handling Categorical Data:

- Label Encoding: Convert categorical variables into numerical labels.

- One-Hot Encoding: Create binary columns for each category of a categorical variable.

- Target Encoding: Replace categorical values with the mean or median of the target variable for each category.

- CatBoost Encoding: Encode categorical variables using gradient boosting techniques.

4. Handling Date Data:

- Feature Extraction: Extract relevant features like year, month, day, or season from the date.

- Time Series Techniques: Use time series models such as ARMA, ARIMA, VARIMA, Holt-Winters, or Prophet to capture temporal patterns.

5. Handling Address Data:

- Geocoding: Convert address data into latitude and longitude coordinates using geocoding services.

- Distance Calculation: Calculate distances between properties or to important locations using Haversine or Euclidean distance.

6. Handling Name Data:

- Feature Extraction: Extract meaningful information from names, such as title or length of the name.

- NLP Techniques: Apply natural language processing (NLP) techniques like word embeddings or sentiment analysis if names contain textual information.

7. Feature Selection Techniques:

- Correlation Analysis: Identify features with high correlation to the target variable.

- Recursive Feature Elimination (RFE): Iteratively select the most important features using a machine learning model.

- L1 Regularization: Use techniques like Lasso regression to shrink coefficients and select relevant features.

8. Feature Ranking Techniques:

- Information Gain: Measure the amount of information a feature provides about the target variable.

- Random Forest Importance: Utilize the feature importances provided by random forest models.

- Mutual Information: Assess the dependency between a feature and the target variable.

9. Feature Binning Techniques:

- Equal Width Binning: Divide continuous features into bins of equal width.

- Equal Frequency Binning: Divide continuous features into bins containing an equal number of observations.

- Decision Tree Binning: Use decision tree algorithms to automatically determine optimal bin boundaries.

10. Feature Generation and Combination Techniques:

- Polynomial Features: Generate polynomial combinations of existing features.

- Interaction Terms: Create new features by multiplying or combining existing features.

- Domain-specific Knowledge: Incorporate domain expertise to engineer relevant features.

11. Kernel Techniques:

- Support Vector Machines (SVM): Use SVM with kernel functions like linear, polynomial, or radial basis function (RBF) kernels for regression.

- Gaussian Processes: Employ Gaussian processes with appropriate kernel functions to model non-linear relationships.

12. Evaluation Techniques for Time Series and Regression Models:

- Mean Absolute Error (MAE): Calculate the average absolute difference between predicted and actual values.

- Root Mean Squared Error (RMSE): Compute the square root of the average squared difference between predicted and actual values.

- R-squared: Measure the proportion of the variance in the target

Time series techniques are widely used in real estate price prediction to capture temporal patterns and forecast future prices. Here are some commonly used time series models:

1. ARMA (Autoregressive Moving Average):

- ARMA(p, q) models combine autoregressive (AR) and moving average (MA) components.

- AR component captures the linear relationship between the current value and previous values.

- MA component models the dependency between the current value and past forecast errors.

- Suitable for stationary time series with no trend or seasonality.

2. ARIMA (Autoregressive Integrated Moving Average):

- ARIMA(p, d, q) models extend ARMA by including differencing (d) to make the time series stationary.

- Differencing eliminates trends and seasonality by subtracting previous values from current values.

- Effective for non-stationary time series with trends and/or seasonality.

3. VARIMA (Vector Autoregressive Integrated Moving Average):

- VARIMA(p, d, q) models are an extension of ARIMA for multivariate time series.

- VARIMA models consider the interdependencies among multiple variables to make forecasts.

- Suitable when real estate price prediction involves multiple correlated variables.

4. Holt-Winters:

- Holt-Winters models capture trends and seasonality in time series data.

- Simple Exponential Smoothing (SES) is used to model level, Holt's Linear Exponential Smoothing (Holt's) for trend, and Seasonal Exponential Smoothing (SES) for seasonality.

- Particularly useful for time series with trend and seasonality.

5. Other Sequence Models:

- Long Short-Term Memory (LSTM): A type of recurrent neural network (RNN) capable of capturing long-term dependencies in sequential data.

- Gated Recurrent Unit (GRU): Another variant of RNN similar to LSTM but with a simplified architecture.

- Prophet: An open-source library developed by Facebook for time series forecasting, incorporating multiple components such as trends, seasonality, and holiday effects.

When applying these time series techniques, it's essential to properly preprocess the data, including handling missing values, removing outliers, and ensuring stationarity if necessary. Additionally, model evaluation techniques like cross-validation and metrics such as mean absolute error (MAE) or root mean squared error (RMSE) are commonly used to assess the performance of the models.

• <https://www.kaggle.com/code/yasserh/housing-price-prediction-best-ml-algorithms>

• <https://www.kaggle.com/code/chanakyavivekkapoor/house-price-prediction>

<https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/>

<https://stackoverflow.com/questions/58756515/onehotencoder-object-has-no-attribute-get-feature-names>

<https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60>

<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>

• <https://www.kaggle.com/code/gaurav896/nashville-housing-initial-analysis>

• <https://www.kaggle.com/code/gauthampughazh/house-sales-price-prediction-svr>