Task Prompt: Predicting California Property Close Price (Final Sales

Accessing the datasets:

1. Download FileZilla Client from https://filezilla-project.org
2. Establish FTP connection using Host: ftp.boxgrad.com, Username: data@idxexchange.com, Password - Real_estate123$, Port: 21
(Type in the password instead of copying and pasting).
3. Datasets are in raw/California and have the prefix 'CRMLSSold'
4. Please make a copy of the 'CRMLSSold' files and download to your local machine, and do not modify any of the files in the raw folder. Download the latest 6 months of data.
5. Document containing the meta data for the datasets is 'Trestle Property MetaData.pdf' and is in the 'resources' folder
6. Please use only observations with PropertyType="Residential" and PropertySubType="SingleFamilyResidence" when constructing the model.

Background:

You have been provided with datasets containing historical close price (final sales price) data on real estate properties sourced from CRMLS (California Regional Multiple Listing Service), including various features such as living area, number of bedrooms, bathrooms, lot size, and close prices. Your task is to develop a machine learning model to predict the close price of a particular property based on its characteristics.

Task Description:

1. **Data Exploration:** Begin by exploring the dataset to understand its structure, features, and any patterns present in the data. Identify relevant features that could influence the close price of a property.
2. **Data Preprocessing:** Preprocess the dataset as needed, including handling missing values, encoding categorical variables, and scaling numerical features if necessary. Split the data into training and testing sets for model evaluation.
3. **Model Selection:** Choose an appropriate machine learning model for predicting property close prices. You may start with a simple model like linear regression or explore more complex models such as decision trees, random forests, or gradient boosting.
4. **Model Training:** Train the selected model using the training data. Tune hyperparameters if applicable to optimize model performance.
5. **Model Evaluation:** Evaluate the trained model using appropriate metrics such as R-squared on the testing data. Assess the model's performance and identify areas for improvement.

6. **Prediction:** Once the model is trained and evaluated, use it to predict the close price of a particular property based on its characteristics. Provide the predicted close price as the output of the model.
7. **Documentation:** Document the entire process, including data exploration findings, preprocessing steps, model selection rationale, training methodology, evaluation results, and the prediction process.

Deliverables:

- Python script containing the code for data preprocessing, model training, evaluation, and prediction.
- Documentation detailing the steps taken during the analysis and reasoning behind decisions made at each stage.
- Live presentation via Zoom summarizing key findings, model performance, and the prediction process for stakeholders.

Additional Notes:

- Ensure that the code is well-organized, commented, and follows best practices for reproducibility and readability.
- Experiment with different features, models, and hyperparameters to improve prediction accuracy.
- Seek feedback from team members or domain experts to validate the model's effectiveness and refine the approach if necessary.