Project Report: Price Prediction Using Text Embeddings

HousePricePrediction

Introduction: The goal of this project was to develop a machine learning model to predict house prices based on numerical features and text embeddings derived from the area descriptions. The project involved preprocessing the data, generating text embeddings using the OpenAI API, splitting the data into training and testing sets, training a linear regression model, and evaluating its performance.

Data Preprocessing: The dataset consisted of various features such as the number of bedrooms, bathrooms, square footage of living space, square footage of the lot, and the number of floors. Additionally, there was a column named 'area_embedding' containing lists of text embeddings for each area description. The data preprocessing steps included handling missing values, converting the price column from a string format to numeric, and extracting the text embeddings from the list.

Text Embedding Generation: The area descriptions were preprocessed by converting them to lowercase, removing non-alphabetic characters, tokenizing the text, removing stop words, and lemmatizing the tokens. The preprocessed text was then used to generate text embeddings using the OpenAI API. However, there were challenges encountered during this process, such as handling errors related to the API responses and accessing the correct embedding values from the response.

Model Training and Evaluation: The dataset was split into features (X) and the target variable (y). The numerical features and text embeddings were used as input features (X), and the house prices were used as the target variable (y). The data was split into training and testing sets with a test size of 20%. A linear regression model was trained using the training set and evaluated on the testing set using the mean squared error (MSE) metric.

Results and Insights: The trained linear regression model performed reasonably well in predicting house prices based on the given numerical features and text embeddings. However, the performance could be further improved by exploring different models, feature engineering techniques, and hyperparameter tuning. Additionally, analyzing the importance of the text embeddings in the overall prediction process could provide valuable insights into the impact of the area descriptions on house prices.

Challenges and Future Work: One of the main challenges faced during the project was handling errors and extracting the correct embedding values from the OpenAI API responses. Further investigation and error handling mechanisms can be implemented to improve the robustness of the text embedding generation process. Additionally, exploring alternative text embedding models or techniques could be beneficial to capture more nuanced information from the area descriptions.

In conclusion, this project successfully demonstrated the potential of using text embeddings in combination with numerical features for house price prediction. The project provided valuable insights into the preprocessing of text data, integrating text embeddings into a machine learning pipeline, and addressing challenges related to text embedding generation. Future work can focus on refining the models, improving the text embedding process, and incorporating additional features for more accurate price predictions.