III. Phương pháp

3.1. Kỹ thuật nền tảng

3.1.1. Sai số trung bình bình phương (MSE)

Sai số trung bình bình phương (MSE) là thước đo thống kê được sử dụng để đánh giá mức độ chênh lệch giữa giá trị dự đoán và giá trị thực tế trong các mô hình học máy.

Được minh hoạ như sau: $$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

Trong đó: n là số lượng dữ liệu

Yi là giá trị thực tế

Y^i là giá trị dự đoán

Giá trị MSE càng thấp, mô hình càng chính xác hơn. Giá trị MSE càng cao, mô hình càng kém chính xác. MSE được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong nhiều lĩnh vực, bao gồm: Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị liên tục (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại)

MSE được đánh giá dễ hiểu và tính toán đơn giản, có thể so sánh được giữa các mô hình khác nhau. Tuy nhiên mặt hạn chế là nhạy cảm với các giá trị ngoại lai, các giá trị ngoại lai có thể làm tăng giá trị MSE và ảnh hưởng đến đánh giá của mô hình. Trong dự đoán giá bán nhà, MSE được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá bán nhà. Giá trị MSE càng thấp, mô hình càng dự đoán chính xác giá bán nhà.

3.1.2. Căn bậc hai trung bình bình phương sai số (RMSE)

Căn bậc hai trung bình bình phương sai số (RMSE) là thước đo thống kê được sử dụng để đánh giá mức độ chênh lệch giữa giá trị dự đoán và giá trị thực tế trong các mô hình học máy.

RMSE là căn bậc hai của MSE (Sai số Trung bình Bình phương) có công thức như sau: **RMSE = √(MSE)**

RMSE có cùng đơn vị đo lường với giá trị thực tế, giúp dễ dàng đánh giá mức độ sai lệch. Giá trị RMSE càng thấp, mô hình càng chính xác hơn. Giá trị RMSE càng cao, mô hình càng kém chính xác.RMSE được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong nhiều lĩnh vực, bao gồm: Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị liên tục (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại)

RMSE được đánh giá dễ hiểu và tính toán đơn giản. Có đơn vị đo lường, giúp dễ dàng so sánh mức độ sai lệch giữa các mô hình. Tuy nhiên mặt hạn chế là nhạy cảm với các giá trị ngoại lai và RMSE không có ý nghĩa thống kê.Trong dự đoán giá bán nhà, RMSE được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá nhà. Giá trị RMSE càng thấp, mô hình càng dự đoán chính xác giá bán nhà.

3.1.3. R^2: Hệ số tương quan xác định

R^2 (Hệ số tương quan xác định) là thước đo thống kê được dùng để đánh giá mức độ tương quan giữa giá trị dự đoán và giá trị thực tế trong các mô hình học máy. R^2 thể hiện tỷ lệ phần trăm biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập.

Công thức**:** **R^2 = 1 - (SSR / SST)**

Trong đó : R^2: Hệ số tương quan xác định

SSR: Sum of Squared Residuals (Tổng Bình phương Sai số).

SST: Sum of Squared Totals (Tổng Bình phương Toàn phần)

Giá trị R^2 càng cao, mô hình càng có khả năng dự đoán tốt hơn. Giá trị R^2 càng thấp, mô hình càng có khả năng dự đoán kém hơn. Giá trị R^2 không âm và không thể lớn hơn 1. R^2 được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong nhiều lĩnh vực, bao gồm: Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị liên tục (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại)

R^2 được đánh giá dễ hiểu và tính toán đơn giản, có thể so sánh được giữa các mô hình khác nhau. Tuy nhiên mặt hạn chế là không có ý nghĩa thống kê, nhạy cảm với các giá trị ngoại lai.

Trong dự đoán giá bán nhà, R^2 được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá bán nhà. Giá trị R^2 càng cao, mô hình càng dự đoán chính xác giá bán nhà.

3.1.4. Adjusted R^2: Hệ số tương quan xác định điều chỉnh

Adjusted R^2 là phiên bản điều chỉnh của R^2, giúp bù đắp cho việc số lượng biến dự đoán ảnh hưởng đến giá trị R^2. Adjusted R^2 có xu hướng giảm khi thêm nhiều biến dự đoán không quan trọng vào mô hình.

Công thức : **Adjusted R^2 = 1 - [(1 - R^2) * (n - 1) / (n - k)]**

Trong đó : Adjusted R^2: Hệ số tương quan xác định điều chỉnh

R^2: Hệ số tương quan xác định

n : Số lượng dữ liệu

k : Số lượng biến dự đoán

Giá trị Adjusted R^2 càng cao, mô hình càng có khả năng dự đoán tốt hơn. Giá trị Adjusted R^2 càng thấp, mô hình càng có khả năng dự đoán kém hơn. Giá trị Adjusted R^2 không âm và không thể lớn hơn 1. Adjusted R^2 được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong nhiều lĩnh vực, bao gồm: Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị liên tục (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại)

Adjusted R^2 giúp so sánh hiệu quả giữa các mô hình có số lượng biến dự đoán khác nhau, bù đắp cho việc R^2 có xu hướng tăng khi thêm nhiều biến dự đoán. Tuy nhiên nó không có ý nghĩa thống kê và nhạy cảm với các giá trị ngoại lai. Trong dự đoán giá bán nhà, Adjusted R^2 được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá bán nhà. Giá trị Adjusted R^2 càng cao, mô hình càng dự đoán chính xác giá bán nhà. Cần so sánh giá trị Adjusted R^2 của mô hình Random Forest với các mô hình khác để đánh giá hiệu quả của mô hình.Cả R^2 và Adjusted R^2 đều là thước đo thống kê, do đó không nên sử dụng chúng để đưa ra quyết định cuối cùng về việc lựa chọn mô hình.

3.1.5. Chỉ số MAE (Mean Absolute Error)

Chỉ số MAE (Mean Absolute Error) là thước đo thống kê được sử dụng để đánh giá mức độ lệch trung bình giữa giá trị dự đoán và giá trị thực tế trong các mô hình học máy. MAE là giá trị trung bình của các giá trị tuyệt đối của sai số.

Công thức:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Trong đó : MAE: Mean Absolute Error (Chỉ số Sai số Tuyệt đối Trung bình)

n: Số lượng dữ liệu

Yi: Giá trị thực tế

Xi: Giá trị dự đoán

Giá trị MAE càng thấp, mô hình càng chính xác hơn. Giá trị MAE càng cao, mô hình càng kém chính xác.MAE được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong nhiều lĩnh vực, bao gồm Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị liên tục (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại).

MAE được đánh giá dễ hiểu và tính toán đơn giản, có đơn vị đo lường và it nhạy cảm với các giá trị ngoại lai hơn so với MSE. Tuy nhiên, MAE không có ý nghĩa thống kê và không phạt sai số lớn như MSE. Trong dự đoán giá bán nhà, MAE được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá bán nhà. Giá trị MAE càng thấp, mô hình càng dự đoán chính xác giá bán nhà. Tuy nhiên, cần so sánh giá trị MAE của mô hình Random Forest với các mô hình khác để đánh giá hiệu quả của mô hình. Cả MAE và MSE đều là thước đo thống kê, do đó không nên sử dụng chúng để đưa ra quyết định cuối cùng về việc lựa chọn mô hình. Cần xem xét các yếu tố khác như độ phức tạp của mô hình, khả năng giải thích và khả năng tổng quát hóa khi lựa chọn mô hình.

3.1.6. Chỉ số MAPE (Mean Absolute Percentage Error)

Chỉ số MAPE (Mean Absolute Percentage Error) là thước đo thống kê được sử dụng để đánh giá mức độ lệch trung bình phần trăm giữa giá trị dự đoán và giá trị thực tế trong các mô hình học máy. MAPE thể hiện sai số trung bình theo tỷ lệ phần trăm.

Công thức:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Trong đó : Percentage Error (Chỉ số Sai số Tuyệt đối Trung bình Phần trăm)

n: Số lượng dữ liệu

At: Giá trị thực tế

Ft: Giá trị dự đoán

Giá trị MAPE càng thấp, mô hình càng chính xác hơn. Giá trị MAPE càng cao, mô hình càng kém chính xác. MAPE có thể được so sánh trực tiếp giữa các mô hình dự đoán giá trị phần trăm. MAPE được sử dụng rộng rãi để đánh giá hiệu quả của các mô hình học máy trong các lĩnh vực: Đánh giá mức độ chính xác của mô hình trong việc dự đoán giá trị phần trăm (Hồi quy) và Đánh giá mức độ chính xác của mô hình trong việc phân loại các mẫu dữ liệu (Phân loại).

Mô hình MAPE được đánh giá dễ hiểu và tính toán đơn giản. Có đơn vị đo lường (phần trăm) và cho phép so sánh trực tiếp giữa các mô hình dự đoán giá trị phần trăm. Tuy nhiên, MAPE không có ý nghĩa thống kê, nhạy cảm với các giá trị 0 hoặc gần 0 trong giá trị thực tế và phạt sai số nhỏ nhiều hơn so với sai số lớn. Trong dự đoán giá bán nhà, MAPE được sử dụng để đánh giá mức độ chính xác của mô hình Random Forest trong việc dự đoán giá bán nhà theo tỷ lệ phần trăm. Giá trị MAPE càng thấp, mô hình càng dự đoán chính xác giá bán nhà.

3.2 Đề xuất phương pháp

TensorFlow Decision Forests (TF-DF) là một thư viện mã nguồn mở được phát triển bởi Google, cung cấp các công cụ mạnh mẽ để xây dựng và triển khai mô hình rừng quyết định (decision forest) trên nền tảng TensorFlow. TF-DF mang đến nhiều lợi ích cho việc phát triển mô hình học máy, bao gồm:

- Hiệu suất cao: TF-DF được tối ưu hóa cho hiệu suất cao, giúp tăng tốc độ đào tạo và dự đoán mô hình.

- Khả năng mở rộng: TF-DF có thể xử lý lượng dữ liệu lớn một cách hiệu quả, phù hợp cho các ứng dụng học máy quy mô lớn.
- Dễ sử dụng: TF-DF cung cấp API đơn giản và dễ sử dụng, giúp bạn dễ dàng xây dựng và triển khai mô hình rừng quyết định.
- Tính linh hoạt: TF-DF hỗ trợ nhiều loại mô hình rừng quyết định khác nhau, cho phép bạn lựa chọn mô hình phù hợp nhất với nhu cầu của mình.

TF-DF bao gồm các thành phần chính sau:

- Bộ xây dựng mô hình: Cung cấp các công cụ để xây dựng các loại mô hình rừng quyết định khác nhau, bao gồm Random Forest, Gradient Boosting Trees, và XGBoost.
- Bộ xử lý dữ liệu: Cung cấp các công cụ để chuẩn bị dữ liệu cho việc đào tạo mô hình, bao gồm xử lý thiếu dữ liệu, mã hóa dữ liệu, và chia tập dữ liệu.
- Bộ đánh giá mô hình: Cung cấp các công cụ để đánh giá hiệu suất mô hình, bao gồm tính toán các chỉ số đánh giá như độ chính xác, độ F1, và AUC.
- Bộ triển khai mô hình: Cung cấp các công cụ để triển khai mô hình vào sản xuất, bao gồm lưu mô hình, xuất mô hình sang TensorFlow Lite, và sử dụng mô hình trong các ứng dụng web và di động.

TF-DF là một thư viện mạnh mẽ và dễ sử dụng cho việc xây dựng và triển khai mô hình rừng quyết định trên nền tảng TensorFlow. TF-DF cung cấp nhiều tính năng và lợi ích giúp phát triển mô hình dự đoán giá nhà.

Hồi quy tuyến tính là một mô hình thống kê được sử dụng để dự đoán giá trị của một biến phụ thuộc dựa trên một hoặc nhiều biến độc lập. Mô hình sử dụng phương trình tuyến tính để mô tả mối quan hệ giữa các biến.

- Đơn giản và dễ hiểu: Hồi quy tuyến tính là một mô hình học máy đơn giản và dễ hiểu. Mô hình này sử dụng phương trình tuyến tính để dự đoán giá nhà dựa trên các

biến độc lập. Nhờ vậy, việc triển khai và giải thích mô hình trở nên dễ dàng hơn so với các mô hình học máy phức tạp khác.

- Hiệu quả: Hồi quy tuyến tính là một mô hình học máy hiệu quả. Mô hình này có thể học nhanh và dự đoán chính xác với lượng dữ liệu nhỏ. Điều này giúp tiết kiệm thời gian và chi phí khi triển khai mô hình.
- Khả năng giải thích: Hồi quy tuyến tính cung cấp khả năng giải thích cao. Mô hình này cho phép xác định được mức độ ảnh hưởng của từng biến độc lập đến giá nhà. Nhờ vậy, người dùng có thể đưa ra quyết định đầu tư sáng suốt hơn.
- Khả năng mở rộng: Hồi quy tuyến tính có thể dễ dàng mở rộng để xử lý lượng dữ liệu lớn. Điều này giúp mô hình phù hợp với việc dự đoán giá nhà trong các thị trường bất động sản lớn.
- Tốc độ xử lý nhanh: Hồi quy tuyến tính có tốc độ xử lý nhanh hơn so với các mô hình học máy phức tạp khác. Điều này giúp mô hình phù hợp với việc dự đoán giá nhà trong thời gian thực.

Sử dụng Hồi quy tuyến tính để dự đoán giá nhà có nhiều ưu điểm, bao gồm đơn giản, hiệu quả, khả năng giải thích, khả năng mở rộng, tốc độ xử lý nhanh. Do đó, Hồi quy tuyến tính là một lựa chọn tốt cho việc dự đoán giá nhà trong thực tế, đặc biệt là khi dữ liệu có sẵn là tuyến tính.